UNIVERSIDADE D
COIMBRA

Catarina Sofia Henriques Maçãs

Time-Series Visualization: Highlighting Patterns and Deviations

Doctoral thesis submitted in partial fulfilment of the Doctoral Program in Information Science and Technology supervised by Professor Fernando Jorge Penousal Martins Machado and presented to the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

June 2021

UNIVERSIDADE D
COIMBRA

Catarina Sofia Henriques Maçãs

Time-Series Visualization: Highlighting Patterns and Deviations

Tese no âmbito do Programa de Doutoramento em Ciências e Tecnologias da Informação orientada pelo Professor Doutor Fernando Jorge Penousal Martins Machado e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Junho de 2021

Time-Series Visualization: Highlighting Patterns and Deviations

Supervised by Fernando Jorge Penousal Martins Machado as Associate Professor of the Department of Informatics Engineering of the Faculty of Sciences and Technology of the University of Coimbra

This work was supported by Fundação para a Ciência e a Tecnologia under the grant SFRH/BD/129481/2017

FCT Fundação para a Ciência e a Tecnologia

Co-financed by:

POCH CENTRO 2020 PORTUGAL 2020 UNIÃO EUROPEIA Fundo Social Europeu

The perception of meaningful beauty is a lubricant
for the mind's gears and a boost for memory
— Alberto Cairo

Abstract

Information Visualization is usually perceived as an analytical tool with its roots in scientific reasoning. However, with the creation of programming languages directed to the design communities, along with the democratisation of data, Information Visualization expanded its conceptual boundaries to explorative and user-oriented areas. Nowadays, Information Visualization is used in different scientific and social domains, and works with different types of data, being time-based datasets common. The analysis of time-series implies understanding the evolution of data attributes over time. To achieve this, approaches from data mining and statistical areas, or even from visual design, are adopted to represent trends and patterns. The visual analysis of these temporal patterns and their disruptions is important in several fields of knowledge and are often an enticing and revealing output for the interested user.

Given the relevance of the representation of trends and patterns in time-varying data, we focus our research on the development of visual mechanisms to map and synthesise complex data and enable the user to gather more information with less effort. Our visual explorations are an integral part of the design practice and will be embraced as a necessary aid to improve the understanding and accessibility of information. This thesis aims to explore the use of time-series visualization tools to allow the user to explore the data and answer specific problems in business domains, more specifically, in the retail, banking, and online shopping domains. Overall, our investigation includes: the application of existing principles of time-series visualization in business; the development of visualization models capable of highlighting temporal patterns intrinsic to the datasets; and, the development of visual models able to adapt to the user's aesthetic preferences.

Different visualization tools were developed to enable the visualization of time-series data from the business domain. Firstly, we explored the representation of deviations to facilitate the analysis of the Portuguese consumption over time. This research project was developed in collaboration with SONAE, one of the most important Portuguese retail companies. With their data, we developed a set of visualization models to optimise their operations by improving the understanding of how the consumption values are distributed along time within their product hierarchy. Secondly, we explored the repre-

sentation of fraudulent actions in finance. The visualization models were developed in collaboration with Feedzai, an important company in the fraud prevention domain. We developed two visualization tools to enable Feedzai's analysts to study the evolution over time of a set of transactions and detect possible cases of fraud more efficiently. Finally, we developed two visualization models to investigate the aesthetic dimension of the Portuguese consumption. These works aim to highlight the temporal patterns and rhythms in a simple and pleasing way. Also, these works aim to be a step towards captivating the big public through aesthetic experiences, luring them to further explore the data in a more analytical form.

Overall, our visualization tools and models enabled us to (i) represent the deviations in consumption over time; (ii) represent suspicious and fraudulent activities in the finance domain; (iii) investigate the aesthetic domain in time-oriented data and represent visually the rhythms of consumption. Through our work, we demonstrated the feasibility of representing temporal patterns and their disruptions to enhance the understanding of the data and increase the knowledge available for business intelligence.

Resumo

A Visualização de Informação é normalmente vista como uma ferramenta analítica com raízes no raciocínio científico. No entanto, com a criação de linguagens de programação direccionadas para as comunidades de design, bem como a democratização dos dados, a Visualização de Informação expandiu os seus limites conceptuais para áreas mais exploratórias e orientadas para o utilizador. Hoje em dia, a Visualização de Informação é usada em diferentes domínios científicos e sociais e trabalha com diferentes tipos de dados, sendo os dados temporais comuns. A análise de dados temporais, requer conhecimento sobre a evolução de um conjunto de dados ao longo do tempo e adquirir este conhecimento implica adoptar abordagens tais como da análise exploratória de dados, análise estatística, ou até design, de forma a promover a representação de tendências e padrões. A análise visual destes padrões temporais e respectivas perturbações pode ser, simultaneamente, cativante e reveladora para o utilizador interessado e, por isso, deve ser aplicada em várias áreas do conhecimento.

Dada a relevância da representação de tendências e padrões em dados temporais, focamos a nossa investigação no desenvolvimento de mecanismos visuais para mapear e sintetizar dados complexos que permitam ao utilizador adquirir mais informação com menos esforço. As explorações visuais resultantes da nossa investigação são uma parte integral da prática do design e serão adoptadas como uma ajuda necessária para melhorar a compreensão e acessibilidade dos dados. Assim, pretendemos usar ferramentas de visualização de dados temporais de forma a permitir ao utilizador explorar os mesmos e obter respostas a problemas específicos no domínio empresarial, mais especificamente, nos sub-domínios da venda ao retalho, da banca e das compras *online*. Em termos gerais, a presente investigação inclui: a aplicação de conhecimento já existente sobre visualização de dados temporais ao domínio empresarial; o desenvolvimento de modelos capazes de enfatizar padrões temporais intrínsecos aos dados; e o desenvolvimento de modelos visuais capazes de se adaptar às preferências do utilizador.

No contexto desta tese, foram desenvolvidas diferentes ferramentas de visualização de forma a permitir a visualização de dados temporais. Inicialmente, foi explorada a representação de desvios por forma a facilitar a análise dos consumos dos Portugueses ao longo

do tempo. Este primeiro projecto de investigação foi financiado pela SONAE, uma das mais importantes empresas Portuguesas de venda a retalho. Com estes dados, desenvolvemos um conjunto de modelos de visualização para optimizar as operações da empresa através da melhoria do conhecimento interno sobre como os valores de consumo estão distribuídos no tempo. Numa segunda fase, foi explorada a representação de actividades financeiras fraudulentas. Este estudo foi financiado pela Feedzai, uma das mais importantes empresas no âmbito da prevenção de fraude. Foram desenvolvidas duas ferramentas de visualização de forma a permitir aos analistas da Feedzai o estudo da evolução temporal de um conjunto de transacções financeiras e a detecção mais eficiente de possíveis casos de fraude. Finalmente, foram desenvolvidos dois modelos de visualização de forma a investigar a dimensão estética dos consumos dos Portugueses. Com estes modelos, pretendemos enfatizar os padrões temporais e ritmos de forma simples e apelativa. Pretendemos dar um passo no sentido de capturar o interesse do grande público através de experiências estéticas, incentivando-os a explorarem, posteriormente, os dados de forma analítica.

De forma geral, através das nossas ferramentas de visualização e modelos de visualização conseguimos: (i) representar os desvios no consumo ao longo do tempo; (ii) representar actividades financeiras suspeitas ou fraudulentas; e (iii) investigar o lado estético dos dados temporais, representando visualmente o ritmo dos consumos. Através do nosso trabalho demonstramos a viabilidade da representação de padrões temporais e respectivas perturbações de forma a melhorar o conhecimento dos dados e a aumentar o conhecimento disponível em inteligência de negócio.

Acknowledgments

First of all, I would like to thank my supervisor, Professor Penousal Machado, for supporting my ideas, pushing me to go beyond, and, especially, for his patience and guidance in this long journey.

To the Department of Informatics Engineering of the University of Coimbra, particularly the Cognitive and Media Systems Group of the Centre for Informatics and Systems, a big thanks for providing the necessary conditions for the development of this thesis.

I would like to thank all my colleagues—and friends—at the Computational Design and Visualization Lab. (CDV). This thesis would not be possible without their support and companionship. My biggest thanks to Ana Rodrigues, Filipe Assunção, João Correia, Nuno Lourenço, and Pedro Martins, for sharing this path with me, and for participating in different pieces that make up this thesis. To Bruna Sousa, João Cunha, and Mariana Xavier, thank you all just for being there with me. A special thanks to my work buddy Evgheni Polisciuc, for sharing and discussing projects and ideas.

To my sister, a heartfelt thanks for her love, support, and confidence in my choices and my work. To Adriana, a big thanks for the relaxing conversations over a cup of coffee and for her genuine friendship over all these years.

A special and heartfelt thanks to Tiago Martins for walking by my side, and for supporting me in my ups and downs, no matter what. Without you this journey would not be the same. For all your trust, care, and love, thank you.

À minha mãe, um obrigada especial, que por mais anos que passem não deixa de cuidar de mim, de ser o meu porto de abrigo, de me apoiar incondicionalmente e de me inspirar. Obrigada do fundo do coração por todo o teu amor e compreensão.

Thank you all!
Obrigada a todos!

Ao meu pai.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence. 120, 121, 122, 215, 242, 256, 257

ATO Account Takeover. 9, 119, 124, 153, 154, 157, 158, 159, 170, 173, 175, 178, 179, 184, 185, 249

BA Bot Attack. 124, 154, 155, 157, 163, 178, 249

BMU Best Matching Unit. 126, 139

EA Evolutionary Algorithm. 68, 216, 226, 227, 228, 229, 235, 237, 238, 242, 250, 257

EEG Electroencephalography. 56

FMSOM Frequency Neuron Mixed Self-Organising Map. 129, 130

HCI Human-Computer Interaction. 5, 191, 256

IEC Interactive Evolutionary Computation. 12, 14, 68, 189, 191, 226, 228, 229, 243, 250, 252, 256

ISP Internet Service Provider. 149

MAS Multi-Agent Systems. 11, 68, 191, 215, 216, 226, 243, 256

ML Machine Learning. 120, 121, 123, 125, 153, 154, 155, 156, 158, 166, 168, 169, 170, 173, 185, 256, 257

SD Standard Deviation. 175, 176, 177

SOM Self-Organising Maps. 10, 125, 126, 127, 128, 129, 130, 132, 134, 135, 137, 138, 139, 141, 143, 144, 147, 148, 150, 183, 249

U-Matrix Unified Distance Matrix. 127

Part I

THE BEGINNING

1

Introduction

"Visualization offers a method for seeing the unseen. It enriches the process of scientific discovery and fosters profound and unexpected insights"
— McCormick et al. [195]

The visual representation of abstract concepts is one of the most basic sign-systems developed to store, interpret, and communicate information. Moreover, with the aid of the intrinsic properties of our visual perception, we can connect and understand the relationships between graphics, data, and the real world. Thus, the application of Information Visualization in our daily lives, and, more specifically, in different business domains, can enhance the processes of decision making, analysis, and comprehension of real-world patterns. For these reasons, Information Visualization is a research field of utmost importance and relevance to study [22].

Information Visualization has its origins in the process of observation and record keeping. In the early days, these recordings took the formats of tables, timelines, and calendars. With the development of copperplate engravings, and the expansion of publishing industries in the sixteenth and seventh centuries, visual imagery evolved into more complex representations that can present and produce knowledge. Due to the influence of important scientific figures, such as Johannes Kelps, Galileo Galilei, or Isaac Newton, in the application of visual forms for the presentation of intellectual research, especially in statistics, the interest in graphical means of expression increased substantially [84, Ch.3]. This also caused Information Visualization to be commonly seen only as an analytical tool and has its roots in scientific reasoning [105, 197, 278]. However, the creation of programming languages directed to the design communities, along with the democratisation of data, expanded Information Visualization conceptual boundaries, practitioners, and audiences to a more explorative and user-oriented field. Nowadays, Information Visualization is used in different scientific and social domains, and works with different types of data. Time-based data is among the most common and relevant data types. Although this type of data is ubiquitous, it can

be complex to represent, especially when more than one dimension is involved [13].

Unlike other data types, time-oriented data has an inherent semantic and can be represented in different granularities of time, such as days, weeks, months, or years. When analysing it, usually, the main goal is to understand the evolution of data attributes over time [14]. To achieve this, the visual mapping of the different attributes must be comparable. However, for a broader understanding of the information, approaches from data mining and statistical areas, or even from visual design, must be adopted to ease the detection of trends and patterns [3]. The visual analysis of these temporal patterns and their disruptions are of vast importance in several fields of knowledge, and are often enticing and revealing for the interested user [14].

Given the relevance of the representation of trends and patterns in time-varying data [44], we focus our research on their visual mapping to synthesise complex data, and to enable the user to gather more information with less effort. Also, we study how patterns can be found during the analysis and representation phases. These visual explorations are an integral part of the design practice and will be embraced as a necessary aid to improve the accessibility of information, improving its acquisition and understanding by the users [25, 52]. More specifically, the present thesis aims to explore the relevance of time-series visualization tools in the business context, more precisely, in the retail and finance domains. Although the visualization of time-series has already been explored, and several adequate models have been presented, there is still a need for investigating visualization models capable of working with specific domain tasks.

Overall, our research focus on the application of existing principles of time-series visualization in the business domain with the intent to facilitate the acquisition of information in *Business Intelligence*; the development of visualization models to better highlight the temporal patterns intrinsic to the datasets; and the development of visual models capable of adapting to the user's aesthetic preferences. Additionally, our contributions include: (i) a literature review of the state of the art in time-series visualization; (ii) the development of two visualization tools to represent the deviations in the Portuguese consumption; (iii) two visualization tools for the analysis of financial fraud; (iv) two adaptive timelines to facilitate the analysis of consecutive events; (v) a user-profiling system developed to characterise a set of financial transactions; (vi) two visualization tools that emphasise the aesthetic and rhythmic nature of the Portuguese consumption; and (viii) a system to interactively evolve visual artefacts according to the user preferences.

1.1 Motivation

We consider that design, colour theory, and Human-Computer Interaction (HCI) can, and should, be combined with programming and data analysis skills in order to promote the creation of Information Visualization models that are visually related to the context of the data, meet specific goals, and, therefore, are tailored to the users' needs and goals. Additionally, by exploring different visualization models for time-oriented data, we aim to expand the knowledge on the methods and applications of different representations.

The motivation to focus on time-oriented data lies in the fact that a significant part of data is time-based [4]. The need to present time-series is common in many fields, and the challenges to do so are always varying between datasets, tasks, and domains. When dealing with time-series, it is important to answer the question of how it can be represented visually so it is possible to make visible the trends, correlations, and patterns that may be hidden. Thus, it is important to reflect such events correctly, and aid the reader in perceiving how things change over time.

Ultimately, the challenge is to create novel visualization models, or adapt known ones to emphasise and respond to the specificities of each visualization problem. We do not aim to create a model that fits all. We argue that each visualization should be created to respond to the specific questions of a particular problem. For this reason, this thesis aims at studying domain-specific time-series visualizations, rather than creating complex general-purpose visualization systems. With our visualization models, we aim to augment the users' understanding of the data while engaging them during the analysis of the visualization.

1.2 Research Questions

Our research hypothesis is that the visual highlight of temporal patterns and trends can be used to create valuable visualization tools, and promote a better analysis and comprehension of the characteristics of time-oriented data. The main goal of this thesis is to explore alternative visual approaches for the representation of time-series patterns in different real case scenarios. In this way, we aim to highlight the need for custom made visualization models, which give answers to specific tasks. As such, this thesis is centred on the development of domain-specific visualization tools. Additionally, we widen our research area to the creation of visualization tools for broad audiences,

exploring the role of aesthetics in visualization, and its effect on the understanding of the data.

Summarising, the main objective is to contribute to the business domain with visualization tools that can be incorporated into the analysis workflow, allowing the company's analysts and their customers to explore, analyse, and present time-oriented data. With this in mind, the following research questions arise:

- Which visualization models exist that can be used to structure time?—To overview how time-oriented data has been visualised, an analysis of the state of the art must be made. The overview of all visualization approaches would be overwhelming and is not the main focus of this thesis. A deeper analysis of visualizations for time-oriented data, in general, can be seen in [3, 4, 209, 301]. We provide an overview of how time-oriented visualization has been applied throughout time, and describe the ways to visually represent the independent variable time.

- Can the representation of time-oriented patterns and their disruptions be useful and valuable for the analysis of real-world data?—In real-world scenarios, time-oriented datasets tend to be complex, multivariate, large, and cyclical. Hence, the visualization of every datum may lead to overplotting, clutter, and occlusion of important elements, hindering the exploration and acquisition of insights. Focusing the visualization model on the representation of specific patterns may overcome these issues by structuring the data, and easing the acquisition of valuable insights. However, several questions arise, including: which mechanisms should be applied to structure, summarise, and emphasise intrinsic data patterns; and, whether such mechanisms should be applied in the pre-analysis of the data or during the construction of the visualization model.

- Can the positioning of the variable time influence the reading of time-oriented patterns and enhance the detection of anomalies, producing valuable insights?—The choice of a visualization model in time-oriented data may depend not only on the specific requirements of each dataset, context, and users, but also on the structure and characteristics of the data. As such, the following questions may arise: how should time variables be mapped into the visual space; and, which are the most appropriate variable arrangements for different tasks.

- Can temporal patterns from business datasets be beautified and simplified for wider audiences?—To be able to explore different

representations of time-oriented data, the exploration of aesthetics can lead to visualization models that are enticing and able to lure the user's attention. For this reason, the exploration of the aesthetic side of the visualization models is another objective of this thesis. This research question is intertwined with the previous ones, however, it focuses on explorative approaches. Hence, questions such as, can aesthetics be applied without reducing functionality, and can the users be able to adapt the visual models to their aesthetic taste, become relevant in this context.

Addressing the above research questions implies the following sub-objectives:

- Study of the state of the art in time-oriented visualization;
- Identify the audience, tasks, and requirements for each visualization model;
- Design and develop visualization tools to allow the analysis of real-world time-oriented data for a specific domain;
- Explore the arrangement of time variables, and study the most appropriate model;
- Explore visualization solutions capable of representing rhythmic trends, patterns and their ruptures in time-oriented data;
- Explore visual representations of data for wider audiences and with emphasis on aesthetics;
- Assess the impact, efficiency, and efficacy of the developed visualization models.

In summary, we intend to explore visually the representation of time-oriented data and improve its analysis in real use cases. More specifically, our models are intended to facilitate the analysis of time-oriented data about consumption in retail and financial fraud. From a methodological perspective, the research and development of each visualization model are driven by the principles of human-centred design and development. We will follow an experimental methodology based on the Three Cycle View of Design Science Research [124] that enables quick iterations between the visualization development phases, its evaluation, and subsequent guidelines for its refinement. This methodology has three components. Firstly, in the Relevance Cycle, we define the research context. We identify the opportunities,

problems, and acceptance criteria for the ultimate evaluation of the results. Secondly, in the Design Cycle, we follow a loop of research to develop and evaluate the visualization. In this cycle, we will follow the methodologies proposed by Wilkins [300] and Fry [106]. Both methodologies are iterative and promote the analysis and improvement of previous steps and the experimentation and refinement of different approaches. Finally, the Rigor Cycle connects with the central Design Cycle through an iterative exchange of knowledge, both from scientific foundations and from the visualization validation. We choose this type of methodology to promote the analysis and improvement of previous steps and the experimentation and refinement of different visual strategies.

Several approaches can be used to validate the contributions in the Information Visualization domain (e.g., algorithm performance analysis, user study, qualitative discussion). Munzner [211] organised a set of validation methods according to the contribution type. For instance, for contributions related to the creation of a visualization system or tool, one should concentrate on providing arguments relative to the perceptual principles and information visualization theory, and the description of usage scenarios. Also, to validate an information system, a user study—which is conducted in laboratory settings with sets of tasks to measure quantitative performances in terms of time and accuracy—should be used. Hence, to validate our visualization tools, we will focus, mainly, on the validation of qualitative and quantitative metrics (e.g., efficiency, efficacy, and user experience) through user studies [300], usage scenarios, and critical analysis (i.e., discussion of the results). Due to the heterogeneity of the works present in this thesis, their validation encompasses different subsets of the validation approaches referred above.

Finally, it is important to note that we direct our research to represent abstract time-oriented data in two-dimensional representations. Therefore, spatio-temporal and three-dimensional models of time-oriented data are out of the scope of this thesis.

1.3 Contributions

The present Section describes our contributions to the representation and analysis of time-oriented patterns. Overall, we contributed with a set of visualization tools that extract and highlight time-series patterns to tackle specific problems. In the following paragraphs, we highlight and aggregate the main contributions that result from this thesis' research into five groups:

- LITERATURE REVIEW. We reviewed the state of the art of time-oriented visualization, from its origins to more contemporary visualization models. We overviewed the main topics in Information Visualization (e.g., visual perception, semiology, validation) and described the time-oriented data domain (visualization models, taxonomies, main tasks, and interaction techniques). Additionally, we analysed the work related to the topics of our research: radial and calendar representations, visualization of financial fraud, and the role of aesthetics in visualization.

- DEVIATIONS IN CONSUMPTION. We developed two visualization tools that aim to represent the deviations in the Portuguese consumption values over time. Both tools concern the application of calendar-based visualizations. One focuses on a regular calendar structure, and the other focuses on a radial one. We performed a user study to validate and evaluate the effectiveness of our models, also focusing on the differences between regular and radial calendars. This work resulted in two publications:

  [178] C. Maçãs et al. "Time-series Application on Big Data - Visualization of Consumption in Supermarkets". In: *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015*. Ed. by J. Braz, A. Kerren, and L. Linsen. SciTePress, 2015, pp. 239–246. DOI: 10.5220/0005307702390246. URL: https://doi.org/10.5220/0005307702390246

  [185] C. Maçãs and P. Machado. "Radial Calendar of Consumption". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 96–102. DOI: 10.1109/iV.2018.00027. URL: https://doi.org/10.1109/iV.2018.00027

- TIME DEPENDENT PATTERNS OF FRAUD. We developed two visualization tools for the analysis of financial fraud. In the first, we focus on bank data, and our goal is to facilitate the analysis of different types of transactions and detect suspicious behaviours. In the second, we focus on the highlight of fraudulent patterns in online shopping, namely Account Takeover (ATO). We perform two user studies with fraud analysts to perceive the efficacy and efficiency of both tools. This work resulted in one publication:

[188] C. Maçãs, E. Polisciuc, and P. Machado. "VaBank: Visual Analytics for Banking Transactions". In: *24th International Conference Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*. 2020, pp. 336–343. DOI: 10.1109/IV51561.2020.00062. URL: https://doi.org/10.1109/IV51561.2020.00062

- ADAPTIVE TIMELINE. For the tools mentioned above, we apply a dynamic timeline that aims to: enable the interaction with the data, and overview the behaviours along time. Both timelines adapt their granularities according to the time span of the data and the canvas space. This work was published in the 24th International Conference on Information Visualization:

[188] C. Maçãs, E. Polisciuc, and P. Machado. "VaBank: Visual Analytics for Banking Transactions". In: *24th International Conference Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*. 2020, pp. 336–343. DOI: 10.1109/IV51561.2020.00062. URL: https://doi.org/10.1109/IV51561.2020.00062

- USER PROFILING. We developed two profiling systems based on Self-Organising Maps (SOM) to enable the characterisation of the consumption in retail, and bank transactions. To distinguish different behaviours, we also focused on the design of custom glyphs to characterise the different elements. This work was disseminated through two publications:

[187] C. Maçãs, E. Polisciuc, and P. Machado. "GlyphSOMe: Using SOM with Data Glyphs for Customer Profiling". In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 3: IVAPP, Valletta, Malta, February 27-29, 2020*. Ed. by A. Kerren, C. Hurter, and J. Braz. SCITEPRESS, 2020, pp. 301–308. DOI: 10.5220/0009178803010308. URL: https://doi.org/10.5220/0009178803010308

[188] C. Maçãs, E. Polisciuc, and P. Machado. "VaBank: Visual Analytics for Banking Transactions". In: *24th International Conference Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*. 2020, pp. 336–343. DOI: 10.1109/IV51561.2020.00062. URL: https://doi.org/10.1109/IV51561.2020.00062

- AESTHETICS OF TIME-ORIENTED DATA. We developed two visualization tools to represent the "rhythms" of the retail consumption values. In the first, we rely on the visual animation of the consumption rhythms. In the second, we add sonification to create a multimodal experience. Additionally, we implemented a Multi-Agent Systems (MAS) to represent the retail consumption over time. We performed user testing to study the aesthetics of the resulting visual artefacts. This work resulted in several publications:

[179] C. Maçãs, P. Cruz, P. Martins, and P. Machado. "Swarm Systems in the Visualization of Consumption Patterns". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Q. Yang and M. J. Wooldridge. AAAI Press, 2015, pp. 2466–2472. URL: http://ijcai.org/Abstract/15/349

[183] C. Maçãs and P. Machado. "The Rhythm of Consumption". In: *5th Joint Symposium on Computational Aesthetics, Sketch-Based Interfaces and Modeling, and Non-Photorealistic Animation and Rendering, Expressive 2016 - Posters, Artworks, and Bridging Papers, Lisbon, Portugal, May 7-9, 2016, Proceedings*. Ed. by E. Akleman, L. Bartram, A. Çamci, A. G. Forbes, and P. Machado. Eurographics Association, 2016, pp. 11–12. DOI: 10.2312/exp.20161259. URL: https://doi.org/10.2312/exp.20161259

[184] C. Maçãs and P. Machado. "The Rhythm of consumptions". In: *IEEE VIS Arts Program*. Baltimore, EUA, 2016

[186] C. Maçãs, P. Martins, and P. Machado. "Consumption as a Rhythm: A Multimodal Experiment on the Representation of Time-Series". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 504–509. DOI: 10.1109/iV.2018.00093. URL: https://doi.org/10.1109/iV.2018.00093

[179] C. Maçãs, P. Cruz, P. Martins, and P. Machado. "Swarm Systems in the Visualization of Consumption Patterns". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Q. Yang and M. J. Wooldridge. AAAI Press, 2015, pp. 2466–2472. URL: http://ijcai.org/Abstract/15/349

[181] C. Maçãs, N. Lourenço, and P. Machado. "Interactive Evolution of Swarms for the Visualisation of consumptions". In: *Interactivity, Game Creation, Design, Learning, and Innovation - 7th EAI*

International Conference, ArtsIT 2018, and 3rd EAI International Conference, DLI 2018, ICTCC 2018, Braga, Portugal, October 24-26, 2018, Proceedings. Ed. by A. L. Brooks, E. Brooks, and C. Sylla. Vol. 265. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2018, pp. 101–110. DOI: 10.1007/978-3-030-06134-0_11. URL: https://doi.org/10.1007/978-3-030-06134-0_11

[182] C. Maçãs, N. Lourenço, and P. Machado. "Evolving visual artefacts based on consumption patterns". In: *Int. J. Arts Technol.* 12.1 (2020), pp. 60–83. DOI: 10.1504/IJART.2020.107693. URL: https://doi.org/10.1504/IJART.2020.107693

- INTERACTIVE EVOLUTION OF ARTEFACTS. We developed an Interactive Evolutionary Computation (IEC) framework that enables the automatic generation of visual representations of the Portuguese consumption, in accordance to the user preferences. Our framework, although being guided by the user, also produces visual novelty, enabling the user to choose from a diversified set of visual artefacts. This framework was also validated through user testing, resulting in two publications:

[181] C. Maçãs, N. Lourenço, and P. Machado. "Interactive Evolution of Swarms for the Visualisation of consumptions". In: *Interactivity, Game Creation, Design, Learning, and Innovation - 7th EAI International Conference, ArtsIT 2018, and 3rd EAI International Conference, DLI 2018, ICTCC 2018, Braga, Portugal, October 24-26, 2018, Proceedings*. Ed. by A. L. Brooks, E. Brooks, and C. Sylla. Vol. 265. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2018, pp. 101–110. DOI: 10.1007/978-3-030-06134-0_11. URL: https://doi.org/10.1007/978-3-030-06134-0_11

[182] C. Maçãs, N. Lourenço, and P. Machado. "Evolving visual artefacts based on consumption patterns". In: *Int. J. Arts Technol.* 12.1 (2020), pp. 60–83. DOI: 10.1504/IJART.2020.107693. URL: https://doi.org/10.1504/IJART.2020.107693

1.4 Outline

The present thesis is divided into four main parts: Part I, which introduces the main theme of this thesis—the representation of patterns

in time-oriented data; Part II, which presents our investigation in the visualization of the deviations of the Portuguese consumption; Part III, which describes our research in the visualization of fraud patterns in finance; Part IV, which presents our exploratory approaches in the beautification of the Portuguese consumption; and Part V, which summarises the work done in the scope of the thesis. Part II (Deviations) and Part III (Patterns) describe the works developed in the context of funded research projects with two top-tier Portuguese companies: SONAE and Feedzai. Hence, these two parts present visualization tools that are our response to real-world problems. Part IV (Rhythms) follows a more explorative approach to visualization, having a greater degree of freedom in terms of functionality. In the following paragraphs, we further describe each part.

In Part I, we introduce our thesis, present our motivation, research hypothesis, and contributions (Chapter 1). Then, we provide the background concepts of Information Visualization (Chapter 2) and the main characteristics of time-oriented data (Chapter 3). We close this Chapter with a review of the state of the art in time-oriented visualization (Chapter 4).

In Part II, we present our work on the representation of the Portuguese consumption. We start by introducing the domain, the SONAE company, the data, and the tasks and requirements (Chapter 5). Then, we discuss our first approach for the representation of the deviations over time in a reduced and simplified calendar structure (Chapter 6). We proceed, in Chapter 7, by reimplementing the calendar in a radial layout and testing its efficiency through a user study. Finally, in Chapter 8, we overview our work and draw the final conclusions.

In Part III, we present our work on the representation of fraud patterns in finance. Similarly to Part II, we first introduce the domain and the Feedzai company (Chapter 9). Then, we divide this part into two distinct visualization models, developed for two distinct purposes. In Chapter 10, we discuss the visualization of bank transactions and the highlight of atypical transactions. In Chapter 11, we present a visualization tool for the detection of fraud in online transactions. In both Chapters, we validate the visualization tools through a user study, performed with the company's analysts. In Chapter 12, we summarise both visualization models.

In Part IV, we present our experiments in the beautification of consumption data. In Chapter 13, we introduce our motivation for aesthetic explorations and contextualise our works. In Chapter 14, we present a visualization model to characterise the changes in consumer preferences and enhance the representation of this data through a multimodal approach. In Chapter 15, we make use of a swarming sys-

tem to paint the consumption values in a canvas and test the viability of a IEC system to create visual artefacts which are in accordance with the user aesthetic preferences. In Chapter 16, we overview our work and make final conclusions.

Finally, Part V synthesises the work done in the scope of this thesis, presenting the overall conclusions and answers to our research questions, highlighting the main contributions and achievements, and indicating future work.

2

Visualization as a scientific domain

The visual representation of data, concepts, or ideas, is one of the main contributions of Information Visualization. Through visualization, it is possible to make sense of abstract data, organise information, and produce knowledge [276]. Although Information Visualization artefacts appeared many years ago in an era of limited observation and storage tools (see Chapter 3), nowadays, Information Visualization refers to the field of computer science [150]. After its widespread use and theorisation [22, 278, 279] it was introduced to scientific computing and, in 1987, finally defined as an independent self-contained research field by McCormick et al. [195]. The computer brought to visualization the analytic power, graphic output, and interactive manipulation needed to ease the comprehension of complex datasets that otherwise would be laborious and time-consuming to analyse [276]. In the present days, Information Visualization is commonly defined as a computing method capable of transforming large-scale collections of abstract data into simple, structured, visual representations that enable the analysis, discovery, and communication of the known and unknown [44, 105].

Information Visualization brings together the capabilities of human visual perception and the processing power of computers to support the discovery and understanding of structures and unknown connections in data. Hence, by exploiting the human perceptual system that is very efficient in processing visual inputs, Information Visualization is able to reduce information overload and enable the exploration of complex patterns and relationships [2, Ch.1]. Currently, the Data Visualization field has evolved into three branches, i.e., Scientific Visualization, Information Visualization, and Visual Analytics. In general terms, Scientific Visualization deals with three-dimensional phenomena, such as medical and biological data; Information Visualization is concerned with non-numeric, non-spatial, and high-dimensional data, such as textural data and networks [56];

and Visual Analytics is concerned with supporting users to perform analytical reasoning through interactive visual interfaces [272].

Information Visualization may be applied in a wide range of fields such as scientific research, data mining, financial data analysis, market studies, medicine, and social studies [260]. Different fields can have different goals of analysis, which may lead to different ways of presentation. These goals of analysis can be summarised in three main groups: (i) explorative analysis, in which no a priori hypotheses about the data are given and the goal is to get new insights or to define the hypotheses; (ii) confirmative analysis, in which a priori hypotheses about the data are already formed, and the goal is to verify the hypotheses; and, (iii) presentation, in which facts about the data were already investigated and defined, and the goal is to communicate and disseminate the analysis results [4, 292].

Overall, through visualization one can transform the invisible into the visible world and represent complex information in a simple and structured way so the user can read the displayed information more rapidly [26, Ch.14]. However, as visualization is based on visual mappings—which are models of the data and not the data itself, as it has no intrinsic form—visualization is always subject to interpretations of the meanings of line, colour, shape, space, form, or arrangement [26, Ch.14]. For this reason, it is important to understand how we interpret internally such representations and how the visual mappings can represent different types of data [22, 279]. In the following Sections, we present a summary of some factors that guide the visualization field, from Gestalt principles to validation techniques.

2.1 Visual Perception

Information Visualization can explore our perceptual and cognitive capabilities to ease the interpretability of abstract data [4, Ch.1] [26, Ch.3]. For this reason, visualizations should be designed based on the basic characteristics and constraints of the human visual perception and cognitive abilities, such as preattentive processing [294], Gestalt principles [298], or sense-making theory [226].

Visual perception can be defined as a human internal process that is used to collect data and interpret what is being seen [149, Ch.2] [26, Ch.27]. Hence, the act of seeing is an active, and almost unconscious, exercise in which we search and select only the pertinent visual information to complete a certain task [149, Ch.2]. The knowledge of how we interpret images can help define how to represent data.

For example, our perceptual system can compare two images which are side by side more easily by flicking our eyes between the two, than if those images appear in a slide show, one after the other. In this last case, the images would have to be stored in our short-term visual memory and we would have to memorise how they appeared, which reveals to be less accurate than the first case [301, Ch.1]. Also, through a more thorough study of the process of interpreting visual objects, Ware [293] found that only one to three visual objects can pass through our cognition processes and be stored into our visual working memory. This may also contribute to lessening the accuracy of comparing images that are being presented in a slide show.

We easily recognise patterns and combine them into meaningful structures. We can distinguish different objects from the environment and, through our experiences, thoughts, and values, interpret and create meaning from them [26, Ch.27]. For this reason, perception is subjective, and it can vary depending, for example, on age, gender, culture, and economic, historical, political, religious, and social factors. Also, we are constantly updating our knowledge of the world through past and present experiences. This makes our internalised knowledge influence the interpretation of objects, which may lead to the interpretation of the same stimulus in different ways at different times [26, Ch.27]. One of our most known perceptual abilities is our recurrent attempt to establish order to make sense of patterns. This process is made in accordance with certain laws, or principles, which can be studied and used to improve the interpretation of visualization models [26, Ch.27].

In 1912, a group of German psychologists, namely Max Westheimer, Kurt Koffka, and Wolfgang Köhler [294, Ch.6], founded what is known as the *Gestalt School of Psychology*. In Gestalt Psychology, the whole is more than the sum of its parts. Visual perception cannot be understood simply by analysing the elements of a scene separately, since our interpretations depend on the relations among these elements. To study how we interpret images, and the elements within the images, Westheimer, Koffka, and Köhler developed a set of *Gestalt Laws* of pattern perception to describe how we see patterns in visual displays. These laws can be translated into a set of design principles for Information Visualization [294, Ch.6]. We briefly discuss some of the Gestalt laws that we found more relevant for the field of Information Visualization[1]:

[1] For more details on gestalt principles refer to [26, 226, 294, 298]

Figure/Ground. A figure is something that is perceived as being in the foreground. The ground is whatever lies behind the figure [294, Ch.6]. According to this principle, we select some

elements in a picture as the figure, the object of interest, and the remaining parts constitute the ground. In Information Visualization, one should make sure that data entities are represented graphically to be perceived as figures, and not the ground. Overall, the content and message must stand out as clearly as possible from the ground [26, Ch.27] [294, Ch.6].

Similarity. According to this principle, we tend to perceive and group objects based on the similarity of their properties, such as brightness, colour, orientation, texture, shape, or size [26, Ch.27]. In Information Visualization, one should use similar properties to represent similar data attributes. On the contrary, to distinguish data attributes, one should use distinct properties. For example, colour and texture are separate channels and, for this reason, can be used together as users can easily attend to either one pattern or the other [294, Ch.6].

Contrast. Contrast is of major importance for our perception of images. We tend to distinguish and group elements that form natural opposites. Also, as there can be no large without small, and no small without large, the differences in similar characteristics will create emphasis [26, Ch.27]. For example, we can highlight elements in images by contrasting their colour in relation to the others, enhancing the perception of the intended message [26, Ch.27].

Continuity. This principle refers to continuation and simplicity. We are more likely to build visual entities from smooth, continuous visual elements, rather than elements that contain sudden changes in direction [294, Ch.6]. For this reason, the continuity principle can be used to indicate relatedness between elements in Information Visualization [26, Ch.27].

Proximity. Things that are close together tend to be perceptually grouped together [294, Ch.6]. According to this principle, individual elements will be perceptually grouped based on their proximity in space or in time [26, Ch.27]. In Information Visualization, visual elements representing related information should be put close together so we can analyse the visualization model in less time and with less effort [294, Ch.6].

Connectedness. Objects which are connected visually by lines are visually perceived as being related to one another. This concept is applied, for example, in node-link diagrams and networks, which represent relationships between concepts [294, Ch.6].

Closure and Common Region. A region enclosed by a contour makes the elements within that region to be considered as related. This principle is related to the proximity principle, however, Palmer [221] showed that the common region is a stronger organising principle than simple proximity [294, Ch.6]. In Information Visualization, visual elements which are related to each other should be placed in an enclosed region.

Lastly, it is important to refer that two or more principles can be used simultaneously. However, it is also to note that if the principles agree, the effect will be stronger, whereas if the principles disagree, the effect will be weaker, and one principle will take over the other [26, Ch.27].

2.2 Visual Variables

In Information Visualization, most datasets contain information that has no obvious visual manifestation. This makes the visual representation of information a non-trivial process, as visual entities must be used so we can interpret them correctly [48]. In *Semiology of Graphics* [22], Jacques Bertin presents one of the most important approaches to structure the visual representation of data. Bertin was inspired by the field of semiotics—the study of sign systems—that studies how one object can represent another object and helps to identify in which way a map symbol (the signifier) representing a data variable (the signified) comes to mean something to the user (the interpretant) [245]. Bertin divided the visual mapping into two parts: *mark* and *visual variable*. The mark is a basic visual element that represents a data element. Bertin initially defined three marks: point, line, and area. Later, with the advance of technology, Carpendale [48] added two more marks: surface and volume. Below is a short description of each mark:

Point. Has no size, it only represents a position in space.

Line. Has length but no width. It can represent a boundary, connection, separation, or edge. The line's meaning can change if its position changes. Changing its visual characteristics (e.g., thickness, colour) does not change its meaning.

Area. Has length and width. It can represent anything that has a measurable size. The area's meaning can change if its size, shape, or orientation change.

Surface. Similar to the area, but in a three-dimensional space. It can represent connections, volume separations or volume edges. The surface's meaning can change if its position, size, shape, or orientation change.

Volume. Has length, width, and depth in a three-dimensional space. It can represent anything that has a measurable size. The volume's meaning can change if its size, shape, or orientation changes.

In addition to marks, information can be expressed in the visual properties of the mark, referred to as *visual variables*. Hence, a visual variable describes the basic encodings of data and defines how a mark can vary [48, 245]. Bertin defined seven visual variables: position, size, value, texture, colour, orientation, and shape. However, several expansions to this list of visual variables have been made in subsequent publications. For example, Mackinlay [191] added angle, slope, area, volume, density, colour saturation, connection, and containment. MacEachren [176] identified transparency, crispness, and resolution to depict uncertainty, regarding data quality or reliability [299]. More recently, Ward et al. [292] included motion as a dynamic visual variable. These new visual variables were added to Bertin's list mainly due to the advent of the computer and the digital world. Whereas, Bertin's means of communication was the paper, more recent visualizations focus on computer displays, which can display more complex visual structures [161, Ch.3].

To predict how each visual variable is processed by our perceptual system, Bertin studied the principles of perceptual psychology and defined which visual variable is the most appropriate for different tasks [245]. To classify visual variables according to whether their changes enable the performance of the different tasks, Bertin defined four levels of organisation:

Selective. A visual variable is defined as selective if its variation allows us to select it from a group [48]. It should be relatively easy to isolate visually the distribution of a particular symbol across the visualization using a selective visual variable. Bertin believed shape to be the only visual variable that is not selective [245].

Associative. A visual variable is defined as associative if its variation allows us to perceive different marks as a group. Bertin believed location, shape, orientation, colour hue, and texture to be associative visual variables [245]. With a dissociative visual variable, one variation dominates the visual perception, and

we cannot make groups. Bertin believed that size and colour value are dissociative visual variables [245].

Ordered. A visual variable is defined as ordered if its variation supports ordered readings [48]. Variations in ordered visual variables are perceived as rankings, as we interpret one variation to be more or less than another. Bertin believed location, size, colour value, and texture to be ordered visual variables. Later, MacEachren [176] also referred to colour saturation, crispness, resolution, and transparency as ordered visual variables and texture as marginally ordered.

Quantitative. A visual variable is defined as quantitative if its variation allows us a numerical reading. Note that the difference between two visual variables can be seen as numerical, but not necessarily with precise numbers (e.g., one line can be seen as the double of another) [48]. Bertin believed quantitative perception to be restricted to location and size [245].

Creating different visual representations comes from choosing which visual variable is the most appropriate to represent each aspect of information. The ability to make these choices can be greatly enhanced by understanding how a change in a particular visual variable is likely to affect the performance of a particular task [48]. Next, we present the visual variables and their characteristics concerning the four levels of organisation, derived from previous research [22, 177, 245]. Our descriptions are a summary of several other studies [48, 161, 176, 245, 292, 299]:

Position. Describes the position of the mark relative to a coordinate frame. It is considered to be the most important visual variable and indispensable for the representation of data. The spatial arrangement of graphics is the first step in reading a visualization, and for this reason, it takes visual primacy over the others [292]. However, in three-dimensional spaces, its use can be problematic, due to occlusion [48, 245].

Size. Describes the amount of space occupied by the mark. Size can be seen as the length of a line, the area of a shape, or the volume of an object. It is effective at representing quantitative information, particularly ratio-level data, as marks can be scaled to the data values they represent [299]. Size easily maps to interval and continuous data variables. Line lengths are easier to estimate than areas and volumes [48, 245, 292].

Shape. Describes the external form of the mark: points, lines, areas, volumes, and their compositions. Shape is effective at representing nominal data attributes but may be unsuitable for quantitative or ordered data attributes [299]. Shape can be associated with a meaning and become a sign of that, however, the link between meaning and shape must be explicit [48, 245]. When using shape, it is important to be careful with small shapes, as they can be indistinct, and to consider how well different shapes can be differentiated. Also, in the same visualization, different shapes must have a similar area and complexity, to avoid visually emphasising one or more unintentionally [292].

Brightness. Describes the relative amount of light of a mark. Variation in brightness results in different shades of grey, where typically, light shades correspond to low data values, and dark shades to high data values [299]. Quantitative readings of this visual variable can be difficult but it can be used to define order, as it is possible to say one is lighter or darker than the other [48]. It is important to note that human perception cannot distinguish between all brightness values. For this reason, for accurate values' distinction, brightness should be used with a reduced brightness scale [292].

Hue. Describes the dominant wave-length of the mark. Hue is the only colour variable appropriate for representing nominal information. For numerical values, there is some discussion about their adequacy. Some refer to hue as not quantitative, as two hues will not be read numerically and have no order (e.g., changes from red to green cannot be seen as one having more colour than another) [48]. Others refer that hue can represent ordinal and numerical data so long as the hues are limited in number and logically sequenced [161, 243, 299]. For this reason, the mapping of hue is usually done by defining hue maps that can represent the relationship between brightness ranges and hue values [292].

Saturation. Describes the spectral peakedness of the mark across the visible spectrum [245]. It is associated with ordered data variables or uncertainty in data. Saturation is also used to enhance hue differences [299].

Orientation. Describes the direction or rotation of the mark compared to an initial state [245]. Orientation differences are readily discriminated and thus, effective at representing nominal

information. Orientation is less suitable for quantitative information, as it generally lacks an inherent order [299]. However, changes in angles are commonly used to represent numbers in pie charts.

Texture. Describes the fill pattern within the mark [245]. It is a composite variable often defined as an intersection between other visual variables [176]. For example, dashed and dotted lines are some of the textures applied to the line mark. Texture can be rapidly differentiated, as long as a small number of distinct types exist [292]. It is suited for nominal information, and may also be used to suggest quantitative information [299].

Grain. Describes the variability of granularity [48]. It is usually associated with texture.

Arrangement. Describes the graphic layout of a mark. The visual variable arrangement varies from regular to irregular and is best suited for representations of nominal information [245, 299].

Crispness. Describes the sharpness of the boundary of the mark. Crispness is also referred to as "fuzziness" [245]. MacEachren et al. [177] found that crispness was the most effective visual variable for representing uncertainty in the context of point symbolisation.

Resolution. Describes the spatial precision at which the mark is displayed [245].

Transparency. Describes the opaqueness of a mark [245].

In summary, this set of visual variables can be distributed among different types of information: Nominal (concepts of differentiation or classes), Ordinal (concepts that can be ordered), and quantitative (an ordinate scale that can be manipulated mathematically). For example, visual variables, such as hue, orientation, and shape are appropriate for encoding nominal information. Visual variables that are ordered but not quantitative, such as brightness, saturation, crispness, resolution, and transparency are appropriate for encoding ordinal information. For ordinal variables, colour coding can be an alternative, in which each point or interval can be visualised using a unique colour from a colour scale. However, this should be applied with care [4, Ch.4]. Finally, visual variables that are quantitative, such as position and size, are appropriate for encoding numerical information. These variables, especially position, can also be applied for ordinal and nominal information [245]. Position and size are also more efficient

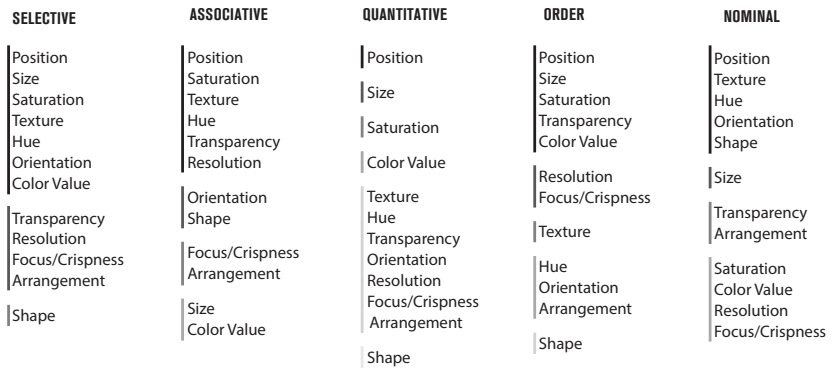

FIGURE 2.1: Visual Variables sorted by data type according to the knowledge retrieved from the state of the art. The higher the position of the visual variable the more appropriate it is to represent a data value with a certain characteristic. Grouped variables are approximate in terms of adequacy.

than using colour or other visual variables such as texture, shape, or orientation to represent quantitative information [191]. However, more recent research has shown that brightness is quite suitable for quantitative values, and has a higher chroma produces the visual effect of higher importance for humans [161, Ch.3].

Visualization can make use of conjunctions—the use of two or more visual variables in a mark [299]. Conjunctions can be applied for redundant symbolisation, strengthening the graphic encoding of one attribute, or for representing multiple attributes. In the latter case, a conjunction can be homogeneous, applying the same visual variable in different ways, or heterogeneous, applying different visual variables to represent the pair of attributes [245, 299]. Conjunctions can also be separable or integral. In separable conjunctions, the variation of each attribute can be "seen" without one restricting the other. For example, when combining size and hue on the same mark, both visual variables are separable as they are considered to be dissociative. These conjunctions can be used to map independent variables that have different scales, as there is no assumed correlation and the user can analyse each attribute individually [90]. In integral conjunction, it is difficult to separate each attribute. This type of conjunction can be applied when the correlation between dependent attributes is more important than the attributes themselves [90]. For example, one can use saturation and hue on the same mark to emphasise the correlation between two attributes.

To answer to the question of which visual variables to use, Cleveland and McGill [68] and Mackinlay [191] suggested a ranking of visual variables depicting the suitability of each visual variable to encode data according to the data types (quantitative, ordinal, or nominal data). Similarly to previous authors, in Figure 2.1 we use the knowledge retrieved from [22, 245, 299] and summarise it in a ranked system according to the authors' findings in relation to their suitability.

2.3 Tasks

Most Information Visualization models can be divided into two components: the representation of data and the interaction with the model. The first deals with the mapping of the data elements to the visual domain, and the second deals with the communication between the user and model [310]. As interaction enables users to manipulate and interpret visualizations, it also aids users to uncover insights about the data. Commonly, the search for insights can be grouped according to the goals of the user. The activities that need to be carried out to accomplish such goals are seen as visualization tasks [253]. The investigation of visualization tasks aims to find and group recurring tasks and use the knowledge derived from them to improve the design and evaluation of visualization models [253]. As a wide range of tasks can be defined to accomplish different purposes, researchers have been working on different or cumulative approaches to systematising them.

One of the earliest approaches to the systematisation of tasks was proposed by Bertin [22]. In his approach, a task is characterised by question type and reading level. The question types are derived from the data elements present in the data, and for this reason, there can be as many types of questions as there are data elements in the dataset [22]. Bertin explains this with an example of a dataset that contains two elements: date and price. From these two elements, two types of questions are possible: "On a given date, what is the price of stock X?" and "For a given price, on what date(s) was it attained?" [8]. For each question type there are three reading levels: elementary (single data element); intermediate (group of elements); and overall (all data). As Bertin's approach derives tasks directly from the structure of the data, it is not biased by any data analysis methods and tools. Adrienko and Adrienko [8] also explored this concept but reduced the reading levels to elementary tasks, which address single elements, and synoptic tasks, which address groups of elements and/or the whole dataset.

In literature, other approaches can be found which relate the tasks with the user intents and can be characterised as user-centred. Such approaches are commonly referred to as high-level interactions (those between the user and the information space) [5, 50, 296, 317]. For example, Amar et al. [5], define a set of low-level analysis tasks that are typically performed with visualizations. These primitive tasks—retrieve value, filter, compute derived values, find extremum, sort, determine ranges, characterise distribution, find anomalies, cluster,

and correlate—accommodate specific questions that might be asked in each task or be used together for more complex questions.

Other authors focus on interaction techniques and can be considered system-centric. Such approaches are considered to characterise low-level interactions (those between the user and the software interface) [62, 143, 259, 281, 291]. One of the most known low-level task's taxonomy is referred to as the *Information Seeking Mantra*, by Shneiderman [259]—"overview first, zoom/filter, details on demand". Later, Keim et al. [145] took this mantra to the Visual Analytics field and defined the *Visual Analytics Mantra*—"Analyse First; Show the Important; Zoom, Filter and Analyse Further; Details on Demand". Yi et al. [310] also defined a taxonomy of interaction intents—select, explore, reconfigure, encode, abstract/elaborate, filter, and connect—that can aid in the process of knowledge discovery or hypothesis confirmation.

In recent years, several studies have been carried out to create a more general approach to tasks. For example, Pike et. al [225] relate high- and low-level interactions with high- and low-level user tasks and goals. These relations enable mutual feedback between users, goals, and tasks and the interactive capabilities of information [225]. Brehmer et al. [33] defined a multi-level topology which aims to fill the gap between low-level tasks and high-level tasks. This task topology, which is not domain-specific, aims to create flexible and concise descriptions of tasks of varying complexity and scope. It encapsulates the tasks in three questions: (i) why the task is performed; (ii) how the task is performed; and, (iii) what does the task pertain to [33]. Finally, Schulz et al. [253] tries to consolidate the task taxonomies existent in literature, using six dimensions of the design space: Why, What, Where, Who, When, and How. To describe the tasks, they ask for answers to the following questions:

- Why is a task pursued?—specifies the task's goal;
- How is a task carried out?—specifies the task's means;
- What does a task seek?—specifies the data characteristics;
- Where in the data does a task operate?—specifies the target, as well as the cardinality of data entities within that target.
- When is a task performed?—specifies the order of tasks.
- Who is executing a task?—specifies the type of user.

2.4 Interaction

Reasoning is commonly dependent on dynamic and exploratory analysis which enable the research of problems from different points of view. From this exploration, the flexibility of our thought processes allows us to gather deeper understanding and novel insights which, in most cases, could not be detected with automated methods. As visual representations can provide an initial direction to the data and its meaning, by adding appropriate interaction mechanisms, users can gain a deeper understanding of the data [2].

In Information Visualization, a wide range of interpretations and taxonomies of interactivity exist. At a higher level, Information Visualization models and interfaces can be categorised in two groups of interactivity: linear and nonlinear [26, Ch.14]. In linear interactivity, the user can move forward or backwards in predefined sequences [262]. Examples of such interactivity are present in slide shows, progress bars, and timelines. Although with a low level of interactivity, this type of interaction facilitates the guidance of the user through the visualization and enables the visualization designers to tell a story as they please, creating a top-down design. When creating this type of visualization the aim is usually to apply storytelling or to communicate rapidly and efficiently [26, Ch.14]. A nonlinear approach refers to a bottom-up design [26, Ch.14]. This type of interaction is characterised by the freedom given to the user to analyse and interact with what is being presented. There is no guided story in this type of visualizations and the users can explore the data and manipulate the graphics by filtering, selecting, and searching the data. An Information Visualization model with this nonlinear structure is often classified as an explorative visualization [26, Ch.14].

In Information Visualization, the interaction methods should be designed according to the user's demands. The users will not only interpret the outcome of the visual and automated methods, they will also explore and drive the whole exploration process [2, Ch.1]. For this reason, interaction methods should allow: (i) the visual overview of all data; and, (ii) the ability to drill down into areas of interest, preserving the context of the information space. Shneiderman refers to this process in his *Visual Information Seeking Mantra* [259]. Shneiderman created a taxonomy in which he defines seven types of interaction techniques: overview, zoom, filter, details-on-demand, relate, history, and extract. Some works exist that propose other techniques or rename the ones defined by Shneirderman [145, 310]. For example, Yi et al. [310], identified the following categories of

interaction: (i) select, to mark data items of interest; (ii) explore, to show some other data; (iii) reconfigure, to rearrange the data spatially; (iv) encode, to change the visual appearance of a data value; (v) abstract/elaborate, to show more or less detail; (vi) filter, to select or show data that match a certain condition; and, (vii) connect, to highlight related data items. Other works do not aim at providing a framework for categorising techniques, instead, they describe a set of methods and their utility. An early example of such works is present in the book of Cleveland and McGill, *Dynamic Graphics* [65], that provides guidance, describes sets of techniques for interaction and displays the results of the interaction.

2.5 Evaluation

Information Visualization researchers have no common ground on what methodological approaches should be used for evaluation. Some researchers argue that the measurement of time and error is sufficient to evaluate a visualization model, whereas other researchers refer that, as visualization is typically exploratory in nature, the interaction and gathered insights should also be evaluated [149]. Visualization models that are created from scratch, are often defined and created upon the designer assumptions. However, analysts or the final user may not think the same way as designers. For this reason, visualization models should be evaluated on whether they fulfil their aims and meet the expectations of the users [149]. To understand if a visualization accomplishes its goals, Mackinlay [191] defined two criteria: (i) expressiveness, a visualization should encode all and only the facts in the data; and, (ii) effectiveness, a visualization should be easy to read and to interpret, being effective at exploring the capabilities of both the output medium and the human visual system [161, Ch.4]. Later, in 2000, Schumann and Müller [254] defined a third criteria to be satisfied: appropriateness—a visualization should be employed only when its value and cost can benefit the achievement of a given task. To ensure that visualization models achieve these three criteria, evaluation should be performed before, during, and after a visualization project [26, Ch.29].

Overall, different types of projects can use different evaluation methods, which can be categorised into two types: summative and formative. Summative evaluation is characterised by the use of quantitative data (e.g., time metrics or performance) to validate the visualization model, emphasising measurable outcomes. It can be useful to validate the performance of a model, however, it does not pro-

vide the reasons for the performance results or how a model can be improved [26, Ch.29]. To address this issue formative evaluations, which are the foundation of iterative/participatory design, can be used. These evaluations aims to find how models are analysed and interpreted, emphasising an interpretative analysis of the observations. These studies can reveal unexpected use patterns or other potential issues that are interfering with the readability of the visualization. However, as these studies are time-consuming, due to the evaluation process and analysis of its results, usually they are tested with a small number of people, which may lead to generalisations and inaccurate conclusions [26, Ch.29].

When performing visualization evaluations a few key aspects should be taken into account. First, it is important to select carefully the participants of the study. One of the best options is to randomly select the participants from a target population, which is mainly the intended audience of the visualization model [26, Ch.29]. Then, the data to collect from the design study should be carefully thought about as it depends on the information being visualised and on the main goals of the visualization model. Although other metrics can be used, most studies use four key metrics: time, accuracy, preference, and insight [26, Ch.29]. To collect time, most studies define tasks to be performed by users. These tasks should be defined to enable the determination of the effectiveness of different design approaches. Time will be measured and compared across studies. Regarding accuracy, usually, participants are questioned about their perception of the visualization model and the insights they could retrieve from it. In relation to preferences, they can be measured when there is a comparison to be made. Participants may have to view more than one design and select the preferred one. However, this metric can be inaccurate or lead to wrong conclusions on effectiveness. For example, participants could prefer a visualization model for other reasons than its efficiency. An alternative to testing preference may be to test how confident people feel when using particular designs. Finally, insights can also be measured by counting the number of insights gained from looking at a design. This can be useful to determine the impact of visualization in terms of knowledge transmission and discovery. Such insights can be counted or recorded qualitatively [26, Ch.29].

There are also different methods to evaluate the validity of a visualization model. In usability tests, participants work with a visualization system to solve a set of tasks in a laboratory environment. During the tests, precise measurements (e.g., speed and correctness) of their performance are taken, their interactions, insights, and difficulties are observed, and post-test questionnaires or interviews can

also be used. In heuristic evaluations and inspections (often referred to as expert reviews), a set of evaluators examine the visualization tool with the goal to find inconsistencies or other issues. The observed problems with the investigated tool are often collected using a set of heuristics to focus on important aspects of the tool [2, Ch.1]. In case studies, a set of domain users describe how the visualization system is used with real work tasks. These studies are conducted in cooperation between researchers and domain experts and are often run over longer periods of time, allowing the experts to become familiar with the system. These studies can give a realistic understanding of the system's strengths [2, Ch.1].

Several studies have been made to identify the advantages and disadvantages of different evaluation methods. From these studies, McGrath [175] identified important factors that should be achieved in evaluation studies: generalisability, precision, and realism. Following the line of thought of McGrath, Carpendale [49] discussed the various approaches in quantitative and qualitative evaluations and emphasised the importance of qualitative approaches in the evaluation of visualization models. Plaisant [227] also discussed the challenges of evaluation, whereas Zhu [318] focused more on the definition and measurement of effectiveness in visualization through accuracy, utility, and efficiency. North [217] focused on how to measure insight and Munzner [212] proposed a nested model for visualization design and evaluation. In this model, Munzner subdivides the process of creating visualizations into four nested levels: domain problem characterisation, data/operation abstraction design, encoding/interaction technique design, and algorithm design. Then, for each of these levels, Munzner defends the use of different evaluation methodologies due to the different goals and validation constraints of each level [149].

2.6 Final Remarks

In the present thesis, and to facilitate the understanding of data characteristics, we explore the perceptual and cognitive capabilities of humans to recognise and detect visual patterns [4, 26]. To create visualization models that can be easily understood, we apply the state of the art knowledge about the characteristics and properties of the different visual variables [4, 22, 191]. For example, as position is considered to be the most important visual variable and the first to be read in a visualization [292], its use to represent time-dependent variables is privileged. Hence, by using this visual variable, the user can easily perceive the temporal features of the data. Also, as hue is

considered to be an appropriate visual variable for encoding nominal information, it is applied to differentiate and highlight different categories within the different datasets.

As we are interested in developing visualization tools, the application of interaction techniques is required. However, it is not the main focus of this thesis and, for this reason, simple techniques are applied to enable the users to freely explore and guide the whole analysis process [2]. We follow Shneiderman's *Visual Information Seeking Mantra* [259], to allow: (i) the visual overview of all data; and, (ii) the ability to drill down into areas of interest. Additionally, in the majority of the tools presented in this thesis, other techniques are applied such as select, reconfigure, abstract/elaborate, filter, and connect, proposed by Yi et al. [310].

Finally, to assess the tools on whether they fulfil their aims and meet the expectations of the users, different evaluations are conducted during and after each visualization project. More specifically, summative evaluations are performed to gather quantitative data about the time and accuracy of the participants when performing a certain task. With this, it is possible to assess the visualization model, emphasising measurable outcomes. Additionally, formative evaluations are performed to find how models are analysed and interpreted, emphasising an interpretative analysis of the observations. We focus on two main validation methods: user studies and usage scenarios. In the first, the final users of the tool work with it to solve a set of tasks in a laboratory environment. Also, the participants are invited to think out loud, so it is possible to better understand their analysis process. In the second, it is described how each tool can be used to analyse the data and perform a set of tasks. With both evaluations, we aim to give a realistic understanding of the system's strengths.

3

Visualising Time

Time-oriented datasets have at least one dimension associated to time and are datasets in which the temporal aspects play a key role in the analysis [4, Ch.1] [161, Ch.2]. This type of data is commonly used in a wide range of domains, such as business, finance, science, and politics, as it enables the study of past events, the comprehension of the present, and the prediction of future outcomes [276].

Temporal variables have been stored, analysed, and used to structure our lives since many centuries ago. For example, one of the first recordings of time-oriented data is a bone engraving that resembles the cycles of the moon [4, Ch.3]. Over time, more complex recordings such as calendars evolved with the emergence of agricultural communities and the seven-day cycle of the week spread to social, biological, and geological contexts [161, Ch.2]. Today, more advanced calendars, such as Gregorian or Chinese calendars, can be easily found and understood. The same applies to other segmentations of time, such as academic semesters, or financial quarters. All of these examples characterise time as being unidirectional and as a way to structure events [4, Ch.3].

Time can be modelled in different ways depending on the particular problem of its application [101]. To do it properly, it is important to understand time and its characteristics. According to Aigner et al. [4], time can be characterised according to (i) scale, time can be ordinal, discrete, or continuous; (ii) scope, time can be point-based or interval-based; (iii) arrangement, time can be linear or cyclic; and, (iv) viewpoint, time can be ordered, branched, or with multiple perspectives.

In terms of organisation, time can be aggregated in different granularities, be associated with different time primitives, and be of the determinate or indeterminate type.

Granularity is an abstraction of time that eases and structures our lives. Time can be aggregated in different granularities, such as

years, months, and days of the week, and multiple granularities can be used in the same system, as we are used to seeing in calendars [4, Ch.3]. Several aspects should be taken into account when defining the most appropriate granularity for the visualization at hand. For example, the number of data elements per granule should be taken into account so it is possible to achieve a better compromise between (i) the oversimplification of the data, caused by the exaggerated level of aggregation—which can also lead to the occlusion of important facts; and, (ii) the creation of visual clutter, caused by an excess of data points per granule. Also, the definition of the granularity level should take into account the context of the use of the visualization, or give to the user the possibility to choose the best granularity level.

With time primitives, time variables can be defined in three different ways: (i) instant, which is a single point in time; (ii) interval, which is a section of time constituted by two instants—the beginning and end of the interval; and, (iii) span, which is a duration of time, without a defined position in time (e.g., 2 days) [4, Ch.3].

To deal with time-oriented data, and in a certain way, with any type of data, is to deal with uncertainty. In time-oriented data, one must consider whether there are or are not absolute knowledge on the time specifications or time primitives. An indeterminate time variable is an uncertain reference of time (e.g., “when the earth was formed” or “the activity started around two o’clock”) and a determinate time variable is any temporal value to which we have complete knowledge of the temporal aspects [4, Ch.3].

During the creation of a time-oriented visualization model, it is important to analyse the time variables concerning (i) scale: whether a time variable is quantitative, (i.e., discrete or continuous) or qualitative (nominal or ordinal data); (ii) time structure: whether the time variables can be placed in a linear (with no overlapping occurrences), cyclic (if there are cyclic behaviours), or branching way (if there are multiple parallel occurrences); (iii) frame of reference: whether there is an inherent spatial layout (geospatial data) or not (abstract data); and, (iv) number of variables: whether there is only one data variable associated to time (univariate), or more (multivariate) [4, Ch.3].

Finally, to enhance the analysis of the visualization, time interaction techniques, such as time scrolling and zooming, can also be applied. By applying such techniques one aims to support the detection of patterns and to ease the understanding of the behaviours within the data [161, Ch.4].

In the following Sections, we will focus on how to represent time-oriented data (Section 3.1), what to analyse in this type of data (Section 3.2), different types of of interactions (Section 3.3) and

existent taxonomies (Section 3.4) of time-oriented data.

3.1 Representation

To map the dimension of time into visual variables, Information Visualization models can use our inherent understanding of time and its progression [4, Ch.4]. In the following paragraphs, we refer to some visual mappings of time and how visual variables are applied to represent it.

Time can be mapped directly into the display space (e.g., position, colour, saturation), resulting in static visualizations, or into the physical time, resulting in dynamic visualizations. In the former, time-oriented data is mapped in a single visual representation, which typically does not change over time. In the latter, time-oriented data is mapped to sequential frames, resulting in visualizations that change over time [4, Ch.4].

Most static visualizations represent all information simultaneously and map time into one display dimension [22], usually the horizontal axis. Another technique of mapping time is to apply small multiples, where each element of the small multiples represents the data at a particular point in time [276]. Static visualizations are advantageous as they allow the overview of all data in a single-frame, enabling the user to concentrate on the dependency of time and data and easing the visual comparison of different sections of time [4, Ch.4]. For this reason, static visualizations are useful to detect trends or find temporal patterns, which typically involve the visual comparison of several time-points [276].

In dynamic visualizations, several frames of the visualization model are rendered successively to show the changes in the data values [2] [4, Ch.4]. In this type of visualization, interpolation techniques are commonly used in small datasets to augment the number of time steps, and aggregation techniques are applied to decrease the length of animation when the datasets are too large [303]. Dynamic visualizations are commonly applied when most of the display space is required to represent characteristics and relationships of other data elements, as is the case with geographical data, multivariate data, and graph structures. In general, dynamic visualizations are well suited to communicate the overall development and major trends of the data being analysed [4, Ch.4] [2]. However, when using this type of visualization, it is important to take human perception into account, so the changes between frames are perceptible (see Section 2.1). This is especially important with more complex visualizations or multi-

variate time series, as users may be unable to memorise and follow changes, making the information imperceptible. Nonetheless, this can be overcome by enabling the user to change between frames interactively [4, Ch.4] [2].

Most univariate visualization approaches map the time dimension into the x-axis and the other data elements in the y-axis (e.g., line and bar charts) [2, Ch.1]. As in these charts, time is seen as a linear structure, time-dependency is immediately perceived and is recognised easily, facilitating the interpretation of the temporal characteristics of the data [4, Ch.4]. However, time can also be cyclical (e.g., seasons of the year, days of the week) and to emphasise such characteristics, polar coordinates are commonly applied. In fact, the application of a polar coordinate scheme can be seen in ancient depictions of time, such as the first sundials in which a polar system was implemented due to the sun rotations around the earth [277]. Unlike linear representations, which use only one dimension to represent the passage of time, in radial representations two dimensions are used (i.e., angle and radius). These representations allow the creation of more elaborate models and can emphasise the cyclic patterns in data [276].

As previously stated, time is commonly represented through position and rarely mapped to other visual variables. However, other mappings can be used as long as the characteristics of the visual variable match the characteristics of the time variable. For example, when using colour hue, the colour scale must be capable of communicating order so the users can easily interpret the visualization and relate data items to their temporal context [4, Ch.4]. In this case, strategies such as the heatmap metaphor can be used. For example, a double-ended colour scale can be applied in cases in which a zero-point divides the scale into two sections. This concept can be used, for example, when the zero point defines the present time, and the other endpoints the past and future [301].

When time is interpreted in a discrete way and referred to as a duration between two occurrences, other visual variables, such as transparency, colour, or saturation, can be applied. Size and length are good visual variables to represent duration as they facilitate the comparison between events. When considering time as a categorical variable—for example, when time is categorised into days of the week—it is possible to use visual variables which are categorical in nature, as long as they can represent the time order. One example of such visual variables is shape, which can map time to symbols, glyphs, or dashlines [301].

When visualising time-oriented data, one dimension of the display space is usually occupied for the representation of the time

variable, diminishing the possibilities of encoding the dependent variables [4, Ch.4]. With multivariate data, the visualization can get more complex to enable the representation of relationships among data elements [276]. Visual variables, such as shape, size, texture, or colour, can be used to represent all data variables, but they may lead to overlapping elements and visual clutter. One solution is to use three-dimensional visualizations and make use of a third dimension, the z-axis, to allow the visualization of more complex and volumetric structures [4, Ch.4]. This kind of representation raises some discussion in the Information Visualization community, with researchers arguing both for its advantages and disadvantages. Some argue that two dimensions are sufficient for effective data analysis and that adding a third dimension may introduce unnecessary difficulties (e.g., hidden information on back faces, information occlusion, or information distortion). Others argue that having just two dimensions for the visual mapping might not be enough for large and complex datasets. Nonetheless, it depends on the problem at hand. For example, to represent flows and scientific data, three-dimensional visualizations are well accepted and often used [4, Ch.4].

3.2 Analysis

When dealing with time-oriented data, users are commonly interested in understanding the evolution of data over time. It is important to enable users to compare the different data values in the different time positions, so they can detect trends, patterns, and have a better understanding of the behaviours within the data. Detecting trends, correlations, and patterns in a visual representation are of utmost importance to gain knowledge from data and its intrinsic relations [4, Ch.5]. However, the direct mapping of data into a visual model may fail to represent the relationships within the data, making them difficult to understand [4, Ch.5] [2, Ch.3]. Additionally, to integrate all-time primitives and their dependent values in a single visualization may lead to a cluttered representation difficult to interpret. For such reasons, with more complex datasets, it is important to make an a-priori analysis of the data. For example, through automated data analysis methods one can extract meaningful features that will then be visualised to enhance the understanding of the data, giving it context and interpretability. Also, through interaction techniques that support the navigation in time, large datasets can be better analysed in detail [2].

The pre-analysis of time-oriented data aims to facilitate and re-

duce the computation workload of visualization models and to reduce the user perceptual effort in interpreting the visualization. A wide range of methods can be applied to ease temporal analysis tasks. For example, one can employ statistical aggregation operators (e.g., sum, average, minimum, maximum), methods from time-series analysis, or data mining techniques [34, 64, 165, 202]. Aigner et al. [4, Ch.6] enumerates a set of analytical methods that can be applied in time-oriented data: (i) classification, the characterisation of a dataset, sequence, or subsequence as one of a predefined set of classes; (ii) clustering, concerned with grouping data into clusters based on similarity (e.g., hierarchical, partitional, and sequential clustering); (iii) search and retrieval techniques, which involves searching for a-priori specified queries and represent only the corresponding subsets of large volumes of data, enabling the user to locate exact or approximate matches for a given date; (iv) pattern discovery, concerned with automatically discovering patterns in data (e.g., sequential pattern, periodic pattern, or temporal association rules); and, (v) prediction, which is related to the prediction of likely future behaviours.

3.3 Interaction

Through our ability to recognise visual patterns and through the combination of visualization and interaction mechanisms, it is possible to create visualization tools that ease the analysis of data and enhance the discovery of new knowledge or the verification of a-priori knowledge [4, Ch.6]. Interaction aids users to understand the visualization structure, boosting their confidence in the reading of the visualization and in the actions that they might have to take. Additionally, interaction can also increase curiosity which may enhance the detection of hidden patterns and be particularly helpful when exploring unknown data. With time-oriented data, the aforementioned factors may enhance the understanding of temporal patterns, the relations between data values, facilitate the understanding of past events, and, consequently, facilitate the prediction and determination of future ones [2, Ch.3].

Generic approaches of visualization usually fail to provide methods that enable and promote the direct exploration of time-oriented data, which is essential for a successful visual analysis. To address this problem, interactive approaches, such as the navigation in time and the possibility to switch between different levels of temporal granularity, can be applied [2, Ch.3]. To navigate in time and conduct the visual analysis of time-oriented data, direct interactions with the

visual representation or more advanced "brushing" techniques can be considered [80, 120, 259]. To use different temporal granularities it is necessary to present the data at different levels of graphical and semantic detail. Strategies like *Overview+detail* can address this issue by presenting an overview of the data and details separately. Another approach is to enable the user to interactively zoom into details or return to the overview. In this approach, overview and details can be presented together. Also, there are two types of zooming: (i) semantic, in which the zoom is performed in the data space; and, (ii) graphical, which operates in the presentation space [4, Ch.5].

Multiple view techniques can also be used to enhance the analysis of time-oriented data. These techniques aim to allow different views on the data at the same time, enabling the analysis of different perspectives and even of different levels of temporal aggregation and supporting the visual analysis and decision making. In addition to multiple views, coordination methods—the propagation of interaction originated from one view to all other views—can also be applied to facilitate the reasoning about time-oriented data [2, Ch.3] [4, Ch.5].

To fully support the knowledge discovery process, visualization models of time-oriented data can take Keim's *Visual Analytics Mantra* into account: "Analyse first; Show the important; Zoom, filter and analyse further; Details on demand" [148]. With this method, the analysis of a visualization model is an iterative process, in which the user focuses on different parts of the visualization, analyses it from different perspectives, and seeks for answers to questions that were already defined or may arise in this process. By starting with an overview of the data, the user can identify and select time periods to analyse in more detail. From there, the user might want to analyse related or similar time periods or return to the overview to analyse the data from a different point of view. In summary, with this process, the user can form a mental model of the data and develop a deeper understanding and insight [4, Ch.5].

In 2011, Aigner et al. [4] adapted the taxonomy defined by Yi et al. [310] and applied it to describe a set of interaction mechanisms for the analysis of time-oriented data:

Select, mark something as interesting. In time-oriented data, users may want to mark temporal events of interest, so they can visually highlight something intriguing, or memorise important results;

Explore, show something else. Users may have to interactively visit different parts of the time domain. Also, users may want to use

alternative visual variables to arrive at an overall view of the data;

Reconfigure, show a different arrangement. Different spatial arrangements of time and respective data values can show different perspectives of the data. This is evidenced by the distinction between linear and cyclic representations of time;

Encode, show a different representation. Similarly to reconfigure, different visual encodings of data values can have an impact on how the data values are perceived along time;

Abstract/Elaborate, show more or less detail. During visual analysis, it is important to show more detail of the data (e.g., the specific values of certain time points) but also to have simpler schematic representations (e.g., the representation of the average values over time);

Filter, show something conditionally. Users may want to search for particular information or evaluate a certain hypothesis about the data in specific time periods. For this reason, it is important to restrict the visualization to show only those data items that fulfil the user criteria;

Connect, show related items. Users may want to highlight data that is similar to previously selected data values. In time-oriented data, this can be emphasised by highlighting certain variables over time.

Also, Aigner et al. [4] refer to tasks that are not specific to time-oriented data, such as **Undo/Redo**, in which the user may be able to go forward or backwards in his/her interactions, and **Change Configuration**, in which the user may be able to adjust the configuration of the interface, especially in cases of multiple views visualization models.

3.4 Taxonomies

Since its earliest days, the Information Visualization field has grown and the number of different visualization models expanded, especially with the evolution of computational methods. For this reason, several taxonomies have been created so it is possible to organise the different visualization models and study their efficiency and usability in each specific case. These taxonomies can be useful to (i) guide users which need to find the visualization models that can fit their datasets; (ii) guide research and enable researchers to know where their works

fit in the visualization field; and, (iii) to understand the field as a whole, enabling its rapid progression [277]. Chengzhi et al. [59] analysed existent taxonomies and grouped them according to: (i) the design model used to define the visualization [45, 60, 61, 277]; (ii) the data type that is the basis of the visualization [44, 259]; (iii) the visual mappings used to represent data [144]; (iv) the interaction techniques applied to enable the analysis of the visualization [35, 62]; and, (v) the user intents, tasks, and skills [59, 91, 241, 281, 296, 312]. Also, other taxonomies exist that try to be more comprehensive, being based in multiple factors [143, 212, 223]

Of the aforementioned taxonomies, none is focused on the structure of time and a small number refers to this type of data. The majority of the taxonomies describe one or more levels of the visualization process but do not describe the process itself. Aigner et al. [3] organises visualizations for time-oriented data in three different levels: data, time, and visualization model. In the first, visualizations can be grouped according to their frame of reference (abstract or spatial), and their variables (univariate or multivariate). Concerning time, visualizations can arrange time in linear or cyclic manners, and the time primitives can be defined as instant or intervals. Finally, concerning the visualization model, they can be static or dynamic, and be projected into two dimensions, or three dimensions. Goralwalla et al. [110] focus on the data models and define an object-oriented framework for time-oriented data. In this framework, time is firstly characterised by its structure (primitive type, domain, determinacy), then its representation, mainly focused on calendars; the order of time (linear or branching); and the history (the values associated with time) [161, Ch.8].

3.5 Final Remarks

Different approaches can be used to model time and, to do it properly, it is important to understand time and its characteristics. We model time according to the main tasks of the visualization model, and for this reason, explore different possibilities. For example, concerning scale, in the majority of our visualization models, time is considered as being ordinal and discrete. Additionally, time is explored as being point-based or interval-based. Finally, in terms of arrangement, time is defined as being linear or cyclic.

In terms of mapping, the majority of the developed tools explore static visualizations as time is mapped directly into the display space (e.g., position). This approach is useful to detect trends or find

temporal patterns, which typically involve the visual comparison of several time-points. Nonetheless, dynamic visualizations are also explored in which time is mapped into the physical time. With this mapping, the evolution of time-dependent variables and the major trends of the data being analysed, can be emphasised

In this research, we aim to enable the user to detect trends, patterns, and have a better understanding of the behaviours within the data. To facilitate this, and during the development of the tools, an a-priori analysis of the data is made to understand how the patterns should be presented. Also, statistical aggregation operators (e.g., sum, average, minimum, maximum) and data mining techniques (e.g., clustering) are applied to facilitate and reduce the computation workload of the visualization models and to reduce the user perceptual effort in interpreting the visualizations.

Additionally, to promote the direct exploration of time-oriented data, auxiliary timeline visualizations are applied. These timelines are intended to enable the user to navigate in time and conduct the visual analysis of time-oriented data. This can also be seen as a multiple view technique as we allow different views on the data at the same time, enabling the analysis of different levels of temporal aggregation and supporting decision making. With this method, the analysis of a visualization model is an iterative process, in which the user focuses on different parts of the visualization and analyses it from different perspectives. Finally, to facilitate the reasoning about time-oriented data, coordination methods are used between the timeline and the main visualization model.

4

Historical Context

The concept of time is constantly present in our daily lives. We check our watches several times a day in an almost compulsive way just to position ourselves in the temporal space. We look at calendars and mark our events so we are able to structure and plan our lives [286]. This need to study and analyse our position in time—enabling us to learn from the past, understand the present, and predict the future—made time a common variable in many application domains, such as medicine, business, science, history, politics and humanities [4, Ch.1] [57]. Data which comprises time as an independent variable is often defined as time-oriented data [209]. The representation of time-oriented data is the main focus of this thesis and in this Chapter, we review its history and applications.

Unlike to other data variables, time in itself contains many distinct characteristics. It has a semantic structure, which enables us to characterise and segment time in different granularities, such as minutes, hours, days, weeks, months, and years. For this reason, time can be represented through different calendar structures with different granularities and can be aggregated in different time intervals rather than be represented only with time points [301, Ch.1]. Also, although time can be seen as a continuous line with a past, present, and future, time can contain cyclic occurrences in which events repeat themselves in regular cycles (e.g., seasons) or irregular cycles (e.g., school breaks and holidays) [4, Ch.1]. Consequently, to properly analyse data that is inherently related to time, first, it should be treated in accordance to the time characteristics and visualization goals and second, the analytical methods should be defined accordingly to the tasks we aim to accomplish [4, Ch.1].

Time-oriented data is in the basis of statistical thinking [57, Ch.1]. Among the earliest techniques to aggregate and archive temporal data is the list, in which events are named, dated, and organised in sequential order, and the table, a more complex system that can

store multiple related elements per time event [26, Ch.1]. An early example of such a system is the Eusebius's chronicle of c. 300 BCE, that organised Christian, Pagan, and Jewish history in a series of parallel columns [95]. A drawback of these systems is that they fail to give information about the time intervals between events and to provide an overall sense of scale. To solve this, time started to be mapped arithmetically, highlighting trends, connections, and relations [26, Ch.1].

From the earliest examples of visual representation of time to the most recent representations, the line is the most used element [244, Ch.1]. This association between line and time can be explained by the similarities between the characteristics of the visual line and the concept of time. In this regard, W. Mitchell mentioned "We speak of 'long' and 'short' times, of 'intervals' (literally, 'spaces between'), of 'before' and 'after'—all implicit metaphors which depend upon a mental picture of time as a linear continuum ... Continuity and sequentiality are spatial images based in the schema of the unbroken line or surface" [201]. This linear metaphor can also be found in genealogy and evolutionary trees used to represent the relationships among entities throughout time.

The first modern timelines appeared in the eighteenth century to replace the previous models: lists, tables, and symbolic shapes. These new representations of time-oriented data implied a novel use of space, in formats that we continue to recognise in modern Information Visualization [26, Ch.1]. This transition from lists and tables to geometric representations of time—timelines—was made to simplify and create visual schemes that could communicate the uniformity and directionality of time, enabling the patterns in the data to emerge more easily [26, Ch.1]. From chronology, the visualization of time-oriented data evolved to statistical graphics and is, in the current days, applied in many domains, such as economics [53], biology [198], and social media [314]. In all, from the first timelines to the new and modern representations of time-oriented data, the main goal of time-oriented visualization is to enable the correct understanding of sequence, the interval between events, and causality, resulting in time structures that enhance both new and previous knowledge [26, Ch.1].

In the following Sections, we refer to the main happenings in the time-oriented visualization history, from the early days to the twenty-first century. Most Sections refer to a single century and important works of the visualization domain will be highlighted. With the advance of computation and the appearance of tools that facilitate the creation of time-oriented visualizations, new techniques

and visualization models started to appear exponentially. For this reason, from the twenty century onwards the visualization projects are grouped according to the visual display of the independent variable, time.

4.1 Pre-eighteenth Century

The first records of time-oriented data appeared in ancient Rome and Greece in the form of lists of priests and magistrates carved in stone. The first tabular timeline was attributed to the Christian theologian Eusebius of Caesarea who designed the *Chronicle* in the fourth-century (Figure 4.1) [244, Ch.2]. Other lists and tabular forms were created throughout the centuries mainly to record significant events and chronological information. In the eleventh century, appeared what is considered to be the first time-series representation—an anonymous illustration of the positions of seven planetary bodies over time and space (Figure 4.2). In this visualization, the horizontal axis represents time, divided into 30 intervals, and the vertical axis represents the inclination of the planetary orbits [3, Ch.2] [278].

In the sixteenth century, with the development of instruments and techniques for precise observation and measurement of physical quantities and geographic positions, the idea of representing and recording knowledge started to emerge in the form of, for example, recordings of mathematical functions in tables (e.g., the trigonometric tables of Georg Rheticus in 1550). The first modern cartographic atlas, the *Theatrum Orbis Terrarum* of Abraham Ortelius, was created in 1570 [57, Ch.1.2].

In 1612, Christopher Scheiner (1573–1650) introduced what is currently known as *small-multiples* [279] to represent the changing configurations of sunspots over time (Figure 4.3) [57, Ch.1.2]. In this

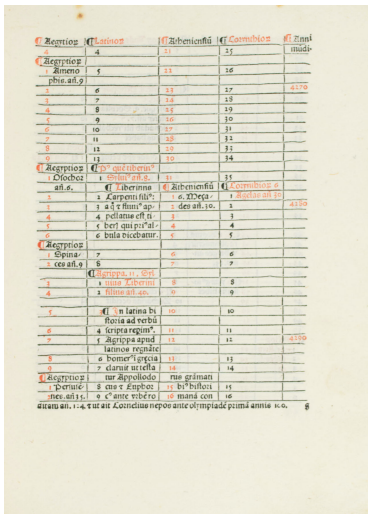

FIGURE 4.1: *Chronicon*. Eusebius of Caesarea, 1483. This *Chronicon* contains several records of Greek and Roman events until the fourth century.

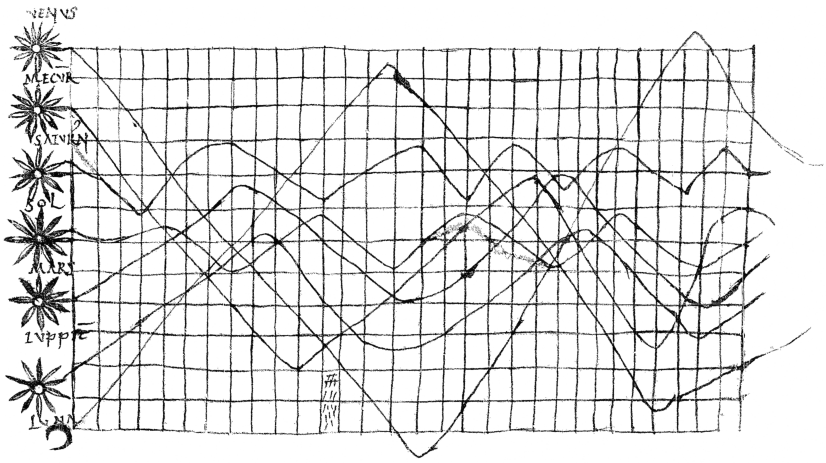

FIGURE 4.2: Planetary movements, 11th century. This representation by an unknown astronomer depicts the movements of seven planetary bodies over time.

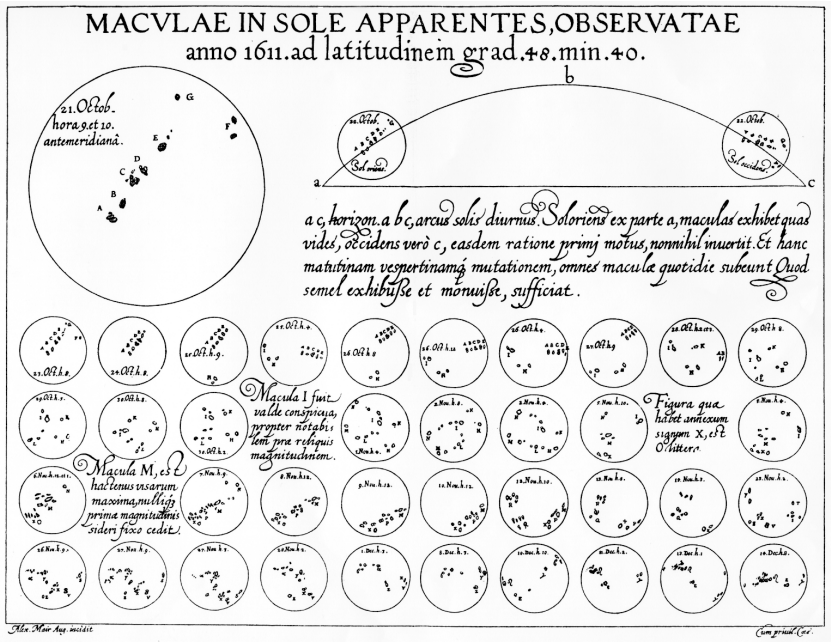

FIGURE 4.3: Sunspot Observations. Christopher Scheiner, 1612. A small-multiples representation from Scheiner's *Tres Epistola*, showing the changes of sunspots over time.

visualization, a set of circles representing the sun, between 23 October 1611 and 19 December 1611, are displayed in a grid, arranged from left to right, top to bottom. The labels in each spot show seven groups of sunspots changing their positions along time [105]. With this visualization, Scheiner aimed to analyse and reflect upon the complexities of observing a rotating sun from a rotating and orbiting earth [279].

In the seventeenth century, two fields of study emerged, the demographic statistics and the 'political arithmetic', that aimed to augment the state knowledge about matters related to wealth, population, agricultural land, taxes, and commercial matters, such as insurance and annuities based on life tables [57, Ch.1.2]. These two new fields and the need to study them promoted the development of new ways of representing, and share their results. For this reason, this century initiated and expanded visual thinking, as illustrated by the example created by Scheiner [57, Ch.1.2].

4.2 Eighteenth Century

Until the eighteenth century, the field of time-oriented visualizations did not have major visual developments when compared with the development in, for example, cartography [244, Ch.4]. However, with the advances in statistical theory and the systematic collection of empirical data, the eighteenth century is characterised by the widespread development of abstract graphs. As the recording of economic and political affairs augmented, new visual representations were invented

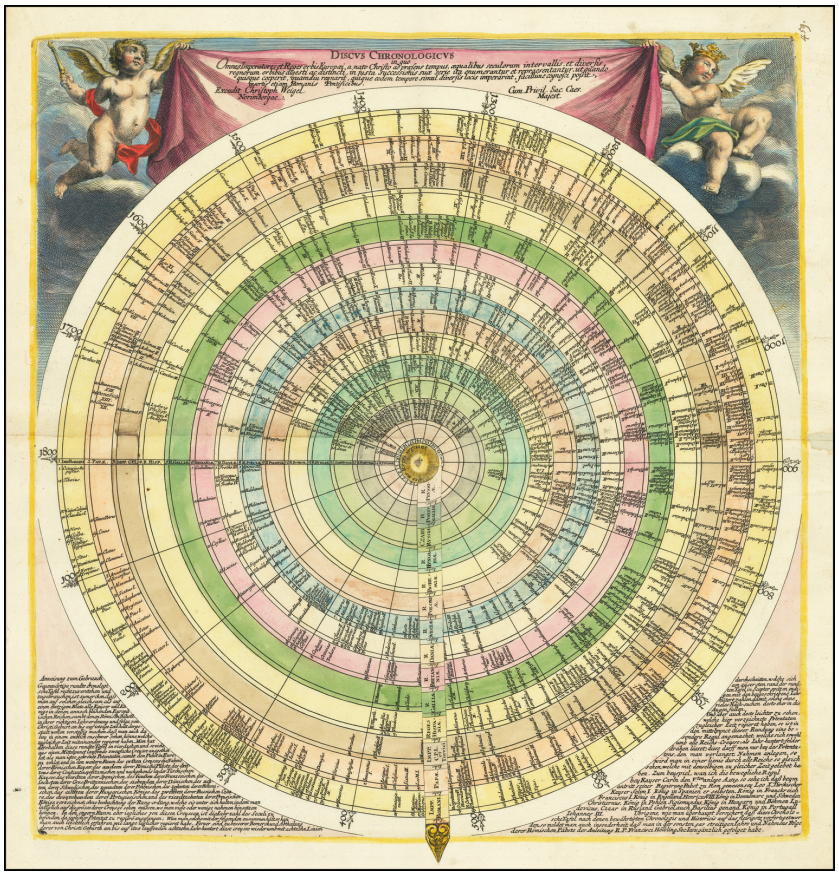

FIGURE 4.4: *Discus chronologicus*. Christoph Weigel, 1720s. *Discus* is a volvelle, a paper chart with a pivoting central arm. Weigel applies a circular layout with rings representing kingdoms and radial wedges representing centuries. The names of kingdoms are printed on the moveable arm. It provides a chronology of the Holy Roman Emperors, Popes and kings of Europe from AD to 1721.

to portray them and facilitate their understanding and analysis [57, Ch.1.2]. At the same time, several technological innovations provided the necessary tools for the production and dissemination of graphical works. Some of these innovations, which facilitated the reproduction of data representations, were three-colour printing, invented by Jacob le Blon (1667–1741) in 1710, and lithography, invented by Aloys Senefelder (1771–1834) in 1798 [57, Ch.1.2].

The use of tables continued to have an impact on the representation of chronological data. However, new arrangements and symbolic representations started to be used with tables. In 1720 Christoph Weigel of Nuremberg (1654–1725) published a *Discus chronologicus*, a chart resembling a clock face with a pivoting central arm—a volvelle—, that organises history from the birth of Christ until 1721 (Figure 4.4). The sequence of years is represented in the radii and the concentric rings represent each one of the fourteen European kingdoms. Although there was little space left for handwriting, Weigel of Nuremberg wanted to represent both sequence and synchrony of historical events in a single sheet, enabling this data to be seen and analysed in one view [26, Ch.1] [244, Ch.4].

In 1718, Girolamo Andrea Martignoni, published a large radial chart about the history and changes of the Roman empire through

FIGURE 4.5: Chart of the Roman Empire. Martignoni, 1718. In this chart of history, Martignoni took inspiration from geographic maps and focused on the representation of the Roman Empire. His chart is a visual summary of history that can be read by following the events along the centuries or the histories of major families.

FIGURE 4.6: *Atlas Historicus*. Johann Georg Hagelgans, 1718. Hagelgans' atlas condenses the work of Roman Historian Eusebius of Caesarea into a single canvas with a grand allegorical view across the top, extending from Mount Sinai on the left to the area around Eden on the right, with Jonah's whale in the center of the image.

time (Figure 4.5) [26, Ch.1]. Although Martignoni applies a circular approach similar to that of Weigel of Nuremberg, instead of using a minimal clock face, he uses a metaphorical approach to represent geographic space and historical time [26, Ch.1] [244, Ch.4]. With his method, 'rivers' and 'landmasses' represent the evolution and changes in the territory over time: the streams at the top of the chart represent the nations conquered by the Roman Empire; the streams at the bottom, represent the nations that emerged from the empire; and the lake at the centre, represents the empire itself. Also, Martignoni avoided text labels so the reader could be driven by the visual experience of information [244, Ch.4]. With this approach, he tried to emphasise the value of the data by condensing it into a single view and enabling three ways of analysis: by events, by century, and by great families [26, Ch.1].

Also in 1718, Johann Georg Hagelgans published a political and

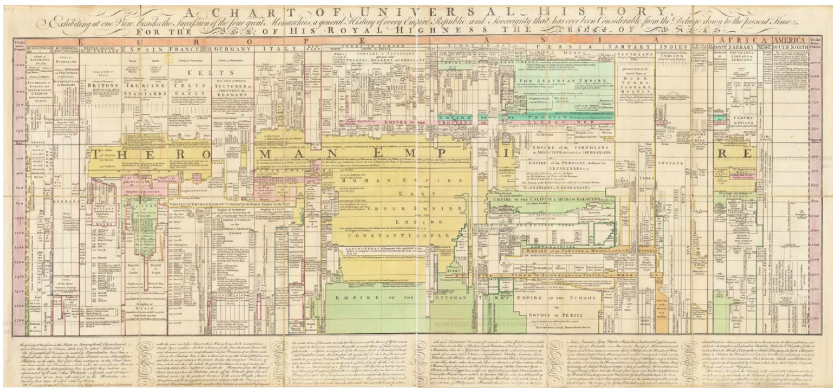

FIGURE 4.7: *A Chart of Universal History*. Thomas Jefferys, 1750s. Jefferys introduced a new era of graphic representation of time. Instead of dividing the data into discrete, indexed cells, he made the space of the chart a continuous field.

military historical Atlas—*Atlas Historicus*—in which he applied the Eusebian's method of representation but, instead of using text to describe the events, he drew a vast amount of small images of soldiers, statesmen, and political figures (Figure 4.6). Similarly to Martignoni and Weigel of Nuremberg, he aimed to represent all information in a single image. His attempt was well succeeded, however, it ended to be a very large work. To enable the reader to understand the meaning of each figure, his Atlas had a dictionary of eighty symbols and their correspondence to the ways how kings died and how the crowns were acquired [244, Ch.4].

Whereas at the beginning of the eighteenth century, symbolic representations were common, in the 1750s, more abstract representations of time-oriented data started to appear. In the 1750s, the cartographer Thomas Jefferys (1719–1771) created *A Chart of Universal History* (Figure 4.7). Like previous works, Jefferys displays all of its data in a single image, making it all visible at once. However, and in contrast to the previous tables, Jefferys' atlas aims to represent the changes over time of geographic areas and, for this reason, his atlas should be scanned like a geographic map. In *A Chart of Universal History*, the columns represent the nations and the years are positioned from top to bottom. Then, in contrast to Eusebius' tables, Jefferys's work changes the width of the columns according to its geographic area. With this representation, empires which were geographically vast but short-lived, have a short and wide representation, and empires which were geographically compact but long-lived, are tall and narrow. To emphasise empires that were both large and long-lived, they are coloured. Each empire's colour is used in other columns to indicate regions belonging to the same empire [244, Ch.4].

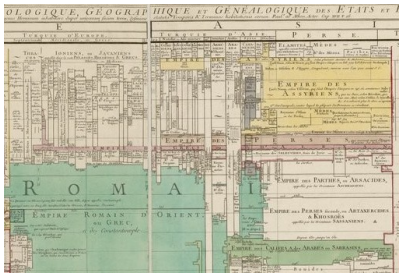

FIGURE 4.8: *Mappemonde Historique*. Jean-Louis Barbeau de la Bruyère, 1750. This chart is an Historical World Map, a chronological and genealogical map of the states and empires of the world.

To allow patterns to emerge, and similarly to Jefferys' chart, Jean-Louis Barbeau de la Bruyère (1710–1781) created a chart in which space is organised horizontally and time vertically (Figure 4.8). By using a single axis for space, the two dimensions of a common map

FIGURE 4.9: *Carte Chronographique*. Jacques Barbeu-Dubourg, 1753. A 16.5 metre long scrollable timeline. It is a moving chronology which reviews all ages of the world, famous figures, and events.

had to be composed into one, which made each country to have only two neighbours [26, Ch.1]. The width of a column represents the geographical extent of each kingdom and the importance of a kingdom is highlighted by a variation in the lettering. Also, he applied different types of lines to represent different relations between adjoining countries. Barbeau de la Bruyère and Jefferys charts remove pictorial aspects and adopt a series of graphical codes that anticipate the quantitative visualizations of the following centuries. In these representations, every point has a meaning, being positioned according to place and time [26, Ch.1].

JACQUES BARBEU-DUBOURG

Jacques Barbeu-Dubourg (1709–1779) created the earliest known modern interactive timeline. His *Carte Chronographique*, of 1753, is assembled on rollers in a wood case which can be scrolled back and forth (Figure 4.9). In this chart, Barbeu-Dubourg plots all events since the "beginning of time" until his time [4, Ch.2]. All years are plotted from left to right and are equally spaced. On the bottom of the chart, there are the names of queens, kings, murderers, and short phrases describing some events. Those events are visually represented with a line which length represents the duration, and symbols are drawn to represent different types of events. All these elements are grouped by territory. The differentiating point in his work is the different scales applied to the temporal axis which decrease as we move backwards in time (and up the sheet). This puts in perspective the time variable, where the past time occupies less space than recent

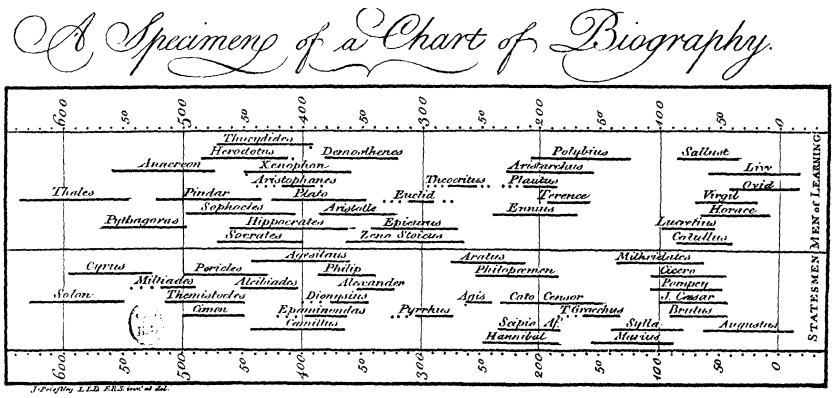

FIGURE 4.10: *Chart of Biography*. Joseph Priestley, 1765. A detailed chart of history that visualises the lifespans of approximately two thousand individuals.

times [26, Ch.1]. With his mechanism, Barbeu-Dubourg aimed to ease learning, enabling the events to be organised in the reader's memory with less effort. Barbeu-Dubourg's work represents a turn in how time-oriented data could be represented, transforming it in "a science that is entertaining, and so to speak mechanised, which speaks to the eyes and the mind" [26, Ch.1].

JOSEPH PRIESTLEY

Although Eusebius's tables could facilitate the understanding of intersecting trajectories in history, the timeline emphasised wider patterns and could give more rapidly the overall picture [244, Ch.1]. Joseph Priestley (1733–1804) understood this and, with an approach similar to that of Barbeu-Dubourg, developed a biographical timeline to represent the lifespans of 2000 famous historical statesmen and men of learning, from 1200 B.C. to A.D. 1750 (Figure 4.10) [57, Ch.1.2] [4, Ch.2]. In *Chart of Biography* the horizontal lines depict the lifespan of each historical figure, which provide an intuitive visual encoding for concepts as historical progress, also seen in Barbeu-Dubourg's work [244, Ch.1]. However, Priestley used a new encoding in this work: the visual representation of different ranges of temporal uncertainties. He used one, two, or three dots to encode the degree of uncertainty of the dates of birth or death [26, Ch.1]. Barbeu-Dubourg and Priestley understood the effectiveness of a simple style of "writing", however, Priestly also recognised the importance of the empty space: "the thin and void places in the chart are, in fact, not less instructive than the most crowded" [234]. Such empty spaces could highlight interruptions and emphasise more dense moments in time [26, Ch.1].

In 1769, Priestley's *New Chart of History* (Figure 4.11) can be seen as an appropriation of Jefferys's visual concepts, but it adds the definition of strict graphic encodings for the translation of historical data into a visual medium [244, Ch.4]. In the horizontal axis, the

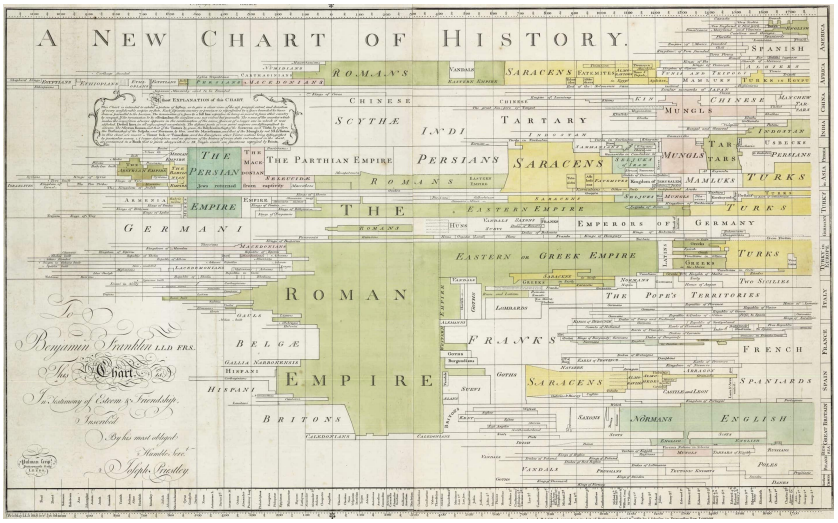

FIGURE 4.11: *New Chart of History*. Joseph Priestley, 1769. This chart allows readers to search for dependencies in events and to distribute them throughout time in a structured way.

chart is marked at intervals of one hundred years. Between these marks, dots indicate decades. Labels are present in the top and bottom of the graph and the dates are connected by vertical lines to facilitate reading. Each Empire is grouped by country and positioned along the vertical axis. Bigger empires are coloured in distinct colours. For Priestley, the efficiency of visualization was one of its most appealing characteristics. For him, the visual analysis of a chart could bring insights more promptly than the reading of the same amount of information [26, Ch.1]. Priestley charts were designed for the general reader and for scholars. Due to their simplicity and efficiency, he believed that his charts could accomplish the goals of both, reducing the amount of time needed to understand history [26, Ch.1] [244, Ch.4]

► WILLIAM PLAYFAIR

"A man who has carefully investigated a printed table, finds, when done, that he has only a very faint and partial idea of what he has read"
—William Playfair [229]

William Playfair (1759–1823) is one of the most important personalities in statistical graphics. Playfair recognised the ability of charts and graphs in exploiting human perceptual and cognitive capacities, and, although there were already a significant collection of economic statistical graphs, he could represent such data in different and modern ways [26, Ch.3]. Playfair believed that "making an appeal to the eye when proportion and magnitude are concerned, is the best and easiest method of conveying a distinct idea" [26, Ch.3]. Being considered as the founding father of modern statistical graphs [4, Ch.2], Playfair is considered to be the inventor of the time-series line, bar, and pie charts; and radial and silhouette graphs[26, Ch.3] [57, Ch.1.2]. Playfair understood the importance of our cognitive and perceptual capacities. He believed that graphs could boost our memory, as he referred to visual memory as more robust than words or numbers. In

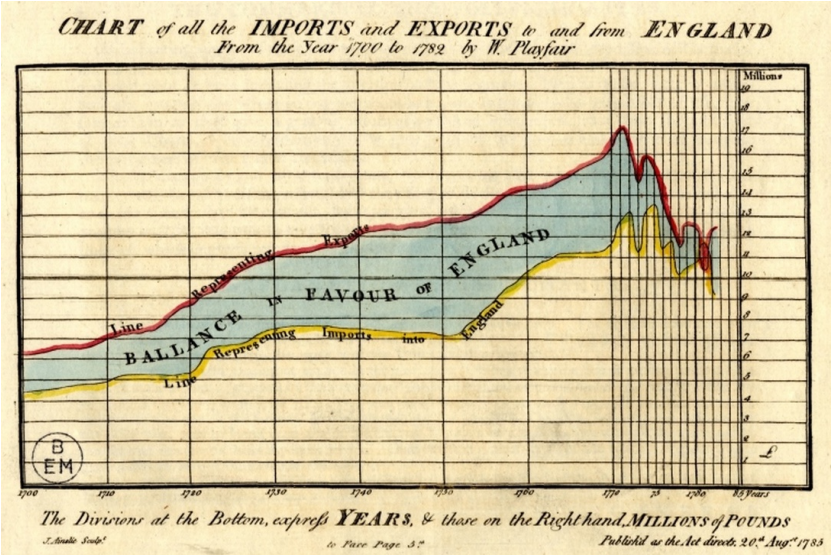

FIGURE 4.12: *Imports and Exports to and from England*. William Playfair, 1786. Chart of all the imports and exports to and from England from the year 1700 to 1782.

terms of representation, Playfair also referred that visual comparisons of objects of different sizes were made more rapidly, intuitively, and almost as accurately, than the mental calculations of numbers [230].

In 1786, Playfair published what is considered to be the first known time-series chart depicting economic data in the *Commercial and Political Atlas* [229]. In this book, Playfair applied for the first time what now are considered to be standard representations: the area, bar, and line charts [4, Ch.2]. In the eighteenth century, visually displaying statistical data was seen as lacking accuracy, which made statistical graphs uncommon [26, Ch.3]. However, Playfair understood their versatility and usefulness for displaying data, their ability to ease the comparison of information—which through tables would make that process time-consuming and complex—and their ability to make instantly visible trends, differences, and associations [26, Ch.3]. In his book, Playfair applied a basic structure of time×quantity which become one of the most recognised forms to depict time-oriented data in the modern days [301, Ch.1] [244, Ch.4]. At the beginning of the book, Playfair overviews all data and only then shows the details of individual countries' trade in the subsequent Chapters. Nowadays, This technique can be seen in the form of a known Mantra: "overview first, zoom and filter, then details-on-demand" [259]. An example of such overview charts can be seen in Figure 4.12. This chart summarises the imports and exports of England from 1700 to 1782. The yellow line represents the imports and the red line the exports. The shade of blue represents positive balances, and the shade of red, near 1781, represents negative balances. The use of colour served to emphasise how accumulated amounts have varied [26, Ch.3]. Another interesting point to note in this graph is the visual focus and,

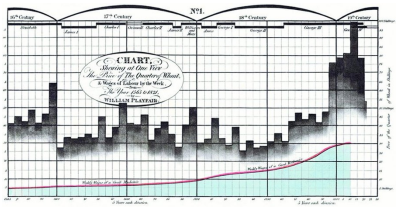

FIGURE 4.13: *Wheat and Labour*. William Playfair, 1821. Chart "Showing at One View the Price of the Quarter of Wheat, and Wages of Labour by the Week, from 1565 to 1821". In this chart, William Playfair enabled the visual comparison of trends.

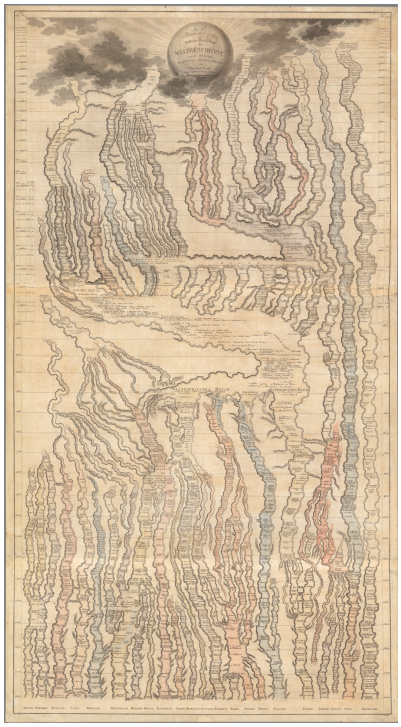

FIGURE 4.14: *Stream of Time*. Friedrich Strass, 1804. This map shows a universal history of mankind, beginning in the depth of unrecorded history and running down to the author's time.

in consequence, more detail on the years from 1760 onward [301, Ch.1]. The same rationale was also implemented in other charts as, for example, one representing the imports and exports to and from Denmark and Norway, in which the shades emphasise the differences in balances [4, Ch.2].

A later graph by Playfair (Figure 4.13), demonstrates how Playfair was able to represent visually the data and ease its analysis [244, Ch.1.2]. In this single chart, Playfair applied three parallel time series to depict the weekly wages of a mechanic (line plot at the bottom), the price of a quarter of wheat (bar graph in the centre), as well as historical context (timeline at the top) over more than 250 years [4, Ch.2]. With this graph, he could demonstrate that workers had their salaries improved in most recent years [244, Ch.1.2].

Playfair's charts are a major mark in the history of data visualization and can be regarded as the first quantitative charts of time [301, Ch.1]. Playfair was able to represent: (i) several data variables in the same graphic; (ii) actual, missing, and hypothetical data through different line characteristics (e.g, solid or broken); and, (iii) accumulated or total amounts, through filled areas between curves [301, Ch.1]. From his innovative visualization models some can be highlighted: the time-series line and bar chart; the divided surface area chart, ideal for showing trends in the variation of two or more time-series; the good design practices in terms of titles, textual descriptions, framing, labelling of axes, and gridlines; the contextualisation through time period indicators and event markers; and the proper use of colour-coding, usually to emphasise qualitative differences between time series or the quality and quantity of the varying accumulated amounts [26, Ch.3]. Obviously, his works also have shortcomings and can be improved (see, e.g., Tufte's analysis [278]). Nevertheless, these limitations are largely surpassed by his innovative contributions.

4.3 Nineteenth Century

During the first half of the nineteenth century, there was a considerable growth in statistical graphics, and graphical analysis of natural and physical phenomena began to appear regularly in scientific publications. Most of the modern forms of data representation were developed, such as the bar, line, and pie charts, histograms, time-series, and scatter plots [57, Ch.1.2]. Also, with the development of photography, film, and other imaging technologies, the recording of time-sequenced phenomena was facilitated, widening the application areas of time-oriented visualization [244, Ch.1].

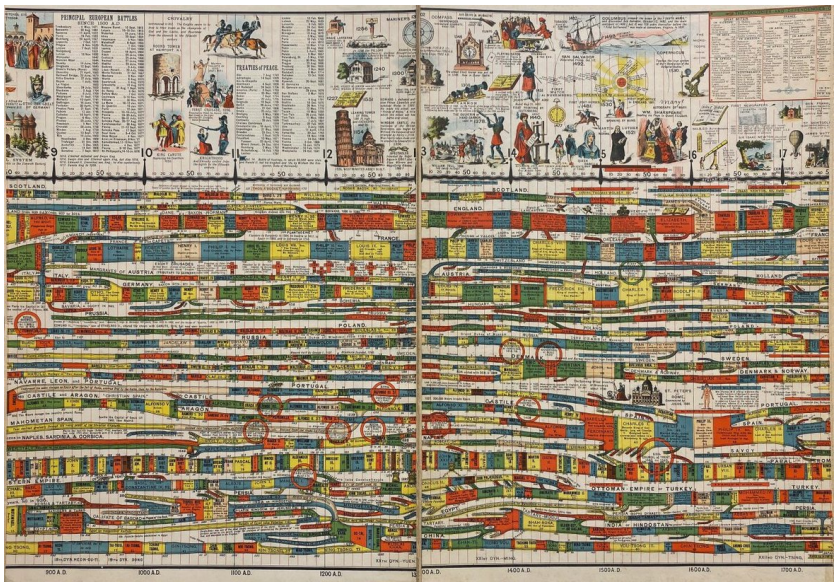

FIGURE 4.15: *Deacon's Synchronological Chart of Universal History*. Edmund Hull, 1890. It is a pictorial and descriptive chart, which maps the World's Great Empires. It is a complete geological diagram of the Earth drawn by Professor Edward Hull.

In the nineteenth century, other timelines appeared in great numbers, as, the *Compend of General History*, by Samuel Whelpley, in 1806, and the *Elements of History, Ancient and Modern* by Joseph Emerson Worcester, in 1833, both revealing the inspiration in the work of Priestley [244, Ch.5]. In 1804, Friedrich Strass published an influential chronological chart entitled *Stream of Time* (Figure 4.14). This chart was similar in terms of metaphoric approach to some charts of the previous century and, like Priestley, Strass believed that the graphic representation of history had more advantages than the textual one. Through the visual medium, it was possible to reveal order, scale, and synchronism, without the complex perceptual task of memorisation and calculation, required in tables. However, Strass applied a different approach to the geometrically regular organisation of Priestley's charts. For Strass that regular organisation implied a uniformity in history that could be misleading [244, Ch.4]. Another example of a graphical representation of historical information through timelines is the *Deacon's Synchronological Chart of Universal History*, originally published in 1890 and drawn by Edmund Hull (Figure 4.15), in which, contrasting with the work of Strass, the temporal axis flows from left to right, and not from top to bottom [4, Ch.5].

As in the previous centuries, the use of graphics to show statistical data in the socioeconomic domain is still applied. The charts representing the United States census of 1870 by Francis A. Walker, are an example of such applications (Figure 4.16). In his *Fiscal Chart of the United States*, Walker gives to the reader the ability to analyse the debt and compare the growth in expenses and revenue among departments along time. He used three charts to represent: (i) the public debt evolution between 1789 and 1870, positioned in the mid-

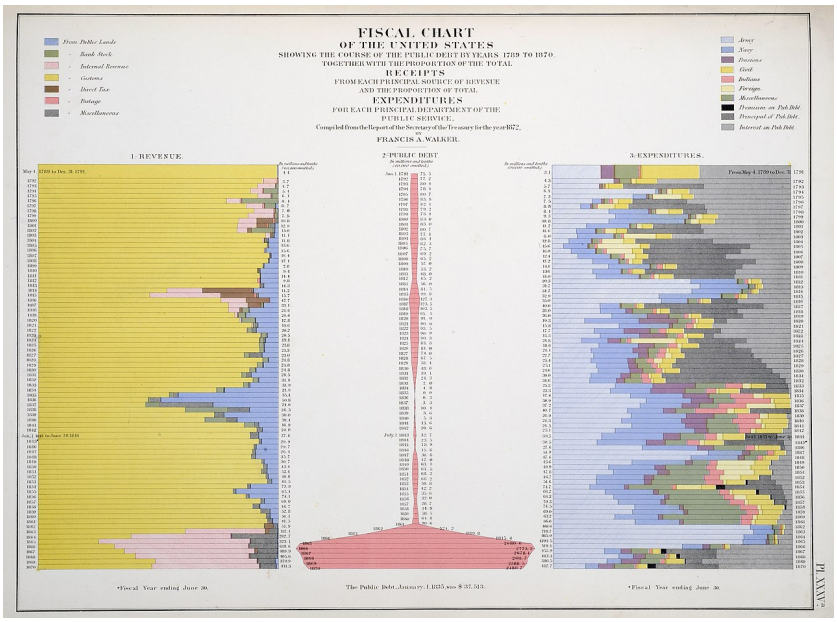

FIGURE 4.16: *Fiscal Chart*. Francis A. Walker, 1874. The *Fiscal Chart* of the United States represents the course of the public debt from 1789 to 1870. In this char it is also represented the proportion of the total receipts and the proportion of total expenditures.

dle; (ii) the proportion of total receipts from the principal sources of revenue, in the left; and, (iii) the proportion of total expenditure for each principal department of the public service, on the right.

With the establishment of statistical offices all around Europe and the recognition of the importance of numerical information for social planning, commerce, politics, and transportation, numerous new time-oriented formats began to appear [57, Ch.1.2]. For example, in medicine, large amounts of information started to be generated and stored, leading to the application of graphical representations to aid the analysis of complex datasets, such as fever curves and Electroencephalography (EEG) [4, Ch.5]. In transportation, an important approach for the representation of time-oriented data was created by Etienne-Jules Marey in the 1880s (Figure 4.17). It shows graphically the train schedule for the track Paris–Lyon. In the vertical axis, the stations are ordered according to their distances and sequence. Time is then represented horizontally, and the individual trains are represented by diagonal lines from top-left to bottom-right (Paris–Lyon) and bottom-left to top-right (Lyon–Paris) [4, Ch.5]. In April 1912, the Marconi chart, visually similar to the one presented by Etienne-Jules Marey, became iconic. After the sinking of the Titanic, the chart showed that at the time of the Titanic collision at 11:40pm, ten ships carrying Marconi operators were in wireless range. However, some had signed off the wireless network at 11:30pm and others were fifty-eight miles out, delaying their arrival to the scene by two hours [244, Ch.6].

Some developments extended graphics to display more than two variables simultaneously, in a three-dimensional space. For example,

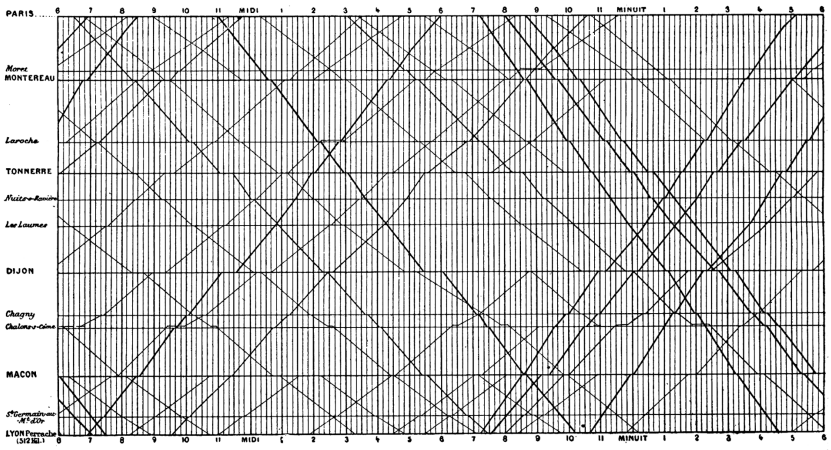

Fig. 7. Graphique de la marche des trains sur un chemin de fer, d'après la méthode de Ibry.

FIGURE 4.17: *Train Schedule*. Etienne-Jules Marey, 1875. A train schedule from Paris to Lyon, designed by Ibry and published by Étienne-Jules Marey.

Luigi Perozzo, in Italy, constructed a three-dimensional surface to plot the population distribution along time (Figure 4.18) [57, Ch.1.2]. In his *Stereogram*, he represents the number of the male population (in the z-axis) per year (in the x-axis) and per age group (in the y-axis). Another approach to the representation of time-oriented data in three-dimensions is the work of Emma Willard. Unlike Perozzo, Williard opted to contradict the lean, undecorated, anti-metaphorical charts of time-series, and project the stream of history in the space of a Renaissance memory theatre [244, Ch.6]. In her *Temple of time* of 1846, a temple is shown in perspective, with timelines on the ceiling and floor, and columns representing key historical figures. She argued that to understand chronology by simply memorising dates, was as difficult and inefficient as it is to learn latitudes and longitudes without the study of maps. She also referred that, as in geography, in which the relation between places is what it is important to know, in chronology, those relations between past and present events is what constitutes the useful knowledge [26, Ch.1].

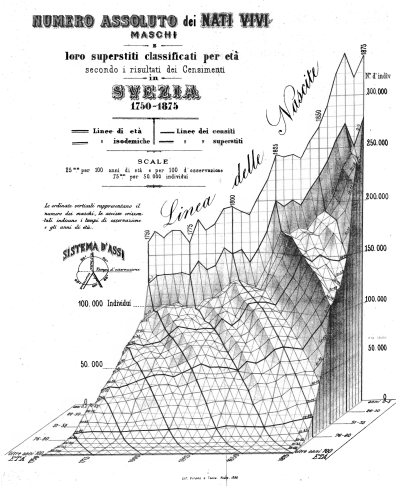

FIGURE 4.18: *Stereogram*. Luigi Perozzo, 1879. Luigi Perozzo's area chart of census data shows the male population of Sweden between 1750 and 1875 by age group.

Relatively to the dissemination of official government statistics on population, trade, commerce, social, moral and political issues, several reports containing data graphics were published all around Europe. At the same time, some standards for the development of graphical representations were proposed by the International Statistical Congresses, which had begun in 1853 in Belgium. In this area, the most notorious works were the Albums *Statistique Graphique* published annually by the French Ministry of Public Works from 1879 to 1897 under the direction of Émile Cheysson [57, Ch.1.2].

▸ CHARLES MINARD

Charles Joseph Minard is an important figure in the visualization domain, especially for his representations of flows and numerical data

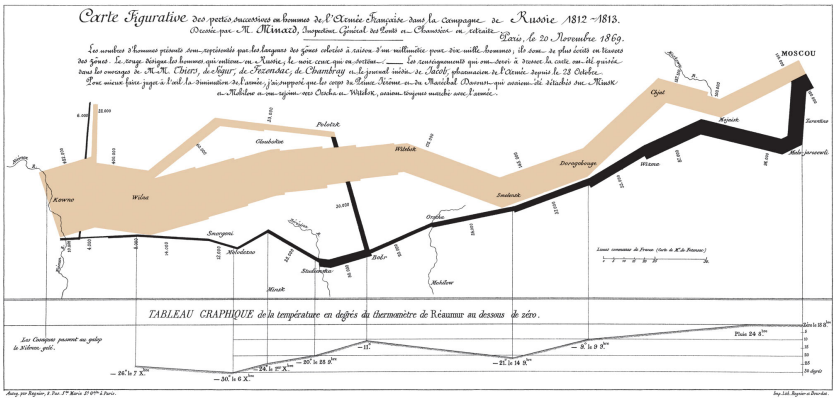

FIGURE 4.19: *Napoleon's 1812 Russian campaign*. Charles Minard, 1869. This chart shows the number of men in Napoleon's 1812 Russian campaign army, their movements, and the temperature they encountered on the return path.

on geographic maps. Concerning time-oriented data, in 1869, Minard developed a flow map which is considered to be a masterpiece of the visualization of historical information (Figure 4.19) [4, Ch.5]. Edward Tufte [278] even called it the "best graphic ever produced" [57, Ch.1.2]. In *Carte Figurative des Pertes Successives en Hommes de l'Armée Française Dans la Campagne de Russie* 1812–1813, Minard created a thematic map that depicts the successive casualties of the French army in the Russian campaign. This flowchart, which is simultaneously a geographic representation of the path made by the army and a time-series of the declining number of army elements, produced beautiful results but, most importantly, questioned the premise of the straight line to depict time-oriented data [244, Ch.1].

Minard was capable of representing in two dimensions six different variables that provide context, enable the reader to make relations between the data and promote exploration. In this sense, the geographic information (latitude and longitude) guides the reader in the display and provides context. The horizontal axis is both time and longitude and the vertical axis shows position. The width of the path shows the size of the total numbers of survivors along time, which demonstrated to be an easily interpretable and highly effective mapping. Colour distinguishes the direction of the soldiers' path. Finally, to facilitate the linking between time, position, and temperature, Minard uses vertical lines. This enabled Minard to relate two charts with different quantities and different appearances, and linking them together [301, Ch.1].

▸ FLORENCE NIGHTINGALE

The social and political uses of graphics are also evidenced in the polar area charts (called 'rose diagrams') invented by Florence Nightingale to prove the bad sanitary conditions in battlefield treatment of soldiers (Figure 4.20) [57, Ch.1.2]. Florence Nightingale recorded the mortality data for two years and created a novel diagram to communi-

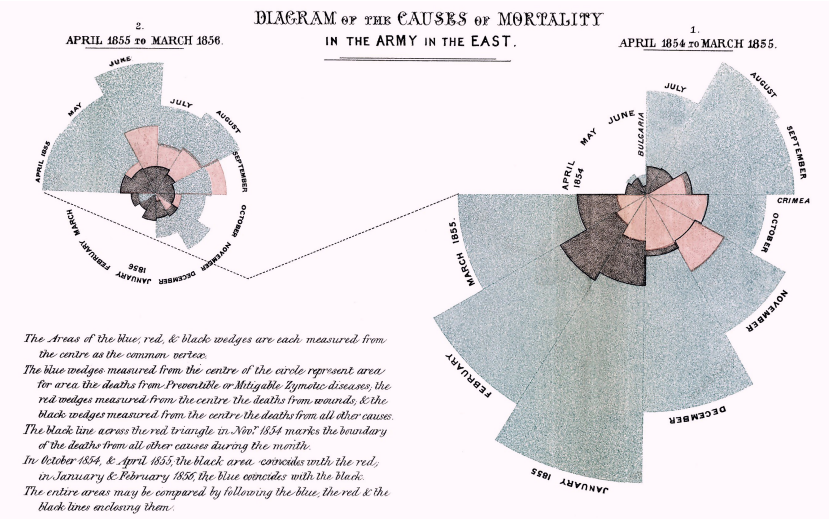

FIGURE 4.20: *Causes of Mortality in the Army*. Florence Nightingale, 1858. This chart shows the number of deaths that occurred from preventable diseases (in blue), those that result from wounds (in red), and those due to other causes (in black).

cate her findings [4, Ch.5]. This representation consists of circularly sequenced wedges that represent time points and convey quantitative data through size. All wedges of the rose diagram have the same angle. This chart revealed clearly the fact that more soldiers were dying due to preventable diseases caught in the hospital, and therefore caused by its bad conditions, than from wounds made in battle [4, Ch.5].

4.4 Twentieth Century

By the mid-1930s the enthusiasm from the eighteenth century for visualization was surpassed by the rise of quantification and formal, often statistical, models in the social sciences. For many statisticians, numbers, parameter estimates and standard errors were precise, whereas graphic representations were considered to be "just images incapable of stating a 'fact' to three or more decimals" [57, Ch.1]. However, some examples of statistical graphs were developed. For example, in 1901, Arthur Bowley dedicated two Chapters in his *Elements of Statistics* to charts and discussed: (i) frequency and cumulative frequency curves through graphs; (ii) the effects of different scales and baselines on visual estimation of differences and ratios; (iii) how to smooth time-series graphs; and, (iv) how 'historical diagrams' with two or more time-series could be shown on a single chart to enable comparison [57, Ch.1]. Also, in the USA, John W. Tukey recognised visual data analysis as a legitimate branch of statistics [280]. Tukey developed a wide variety of new, simple, and effective graphic displays, under the concept of 'exploratory data analysis', such as stem-leaf plots and boxplots [57, Ch.1.2].

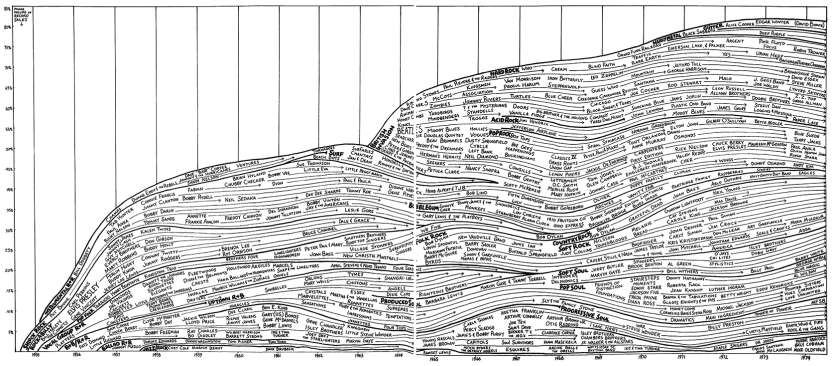

FIGURE 4.21: *Rock'N'Roll is here to Pay*. Chapple and Garofalo, 1977. This chart the history and politics of the music Industry, from 1955 to 1974.

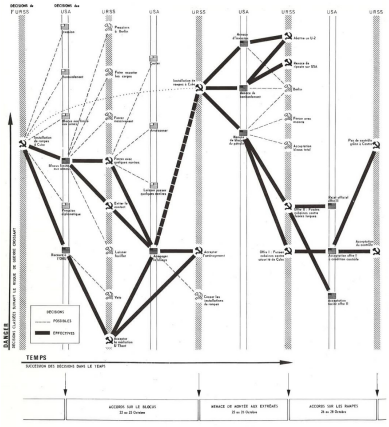

FIGURE 4.22: *La Crise Cubaine de 1962*. Jacque Bertin, 1983. Time-series of the decisions that led to the crisis in Cuba in 1962.

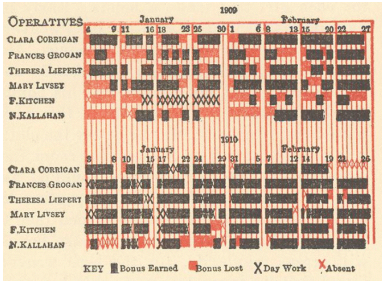

FIGURE 4.23: *La Crise Cubaine de 1962*. This chart represents the distribution of operatives along time.

In this century, studies and comparisons of the efficacy of various graphic forms began to appear and a set of standards and rules for the graphic presentation were made [11, 279] [57, Ch.1]. In France, Jacques Bertin published the *Sémiologie graphique* in 1967 [22]. In this book, Bertin organised the visual and perceptual elements of graphics according to their features and relationships in data [57, Ch.1.2]. Bertin also developed a set of different approaches to the representation of time-oriented data. For example, in 1983, he represented the decisions, possible decisions, and outcomes over time, in a structure similar to the current parallel coordinates, enabling the comparison between decisions through time (Figure 4.22).

In 1977, Chapple and Garofalo illustrated the Rock'n'Roll history (Figure 4.21) that depicts the protagonists and developments in the area as curved lines that are stacked according to the artists' percentage of annual record sales. Years later, the *Theme River* technique [121] can be seen as a further, and more formal, development of this idea [4, Ch.2].

With the developments in industrialisation in the late nineteenth century, the need to optimise resources and prepare time schedules were central to the improvement of productivity. This led to the study of optimisation of work processes and to the development of a timeline-based chart which could be intuitive, and could represent time-oriented processes. This chart is nowadays known as *Gantt chart* and was developed by Henry Laurence Gantt (Figure 4.23). Another visualization model employed in the domain of business is the radial chart of Willard Cope Brinton, in 1934. Brinton aimed to analyse the working time and leisure time in 1932 and chose the radial layout to reduce the space needed to present all information. In his chart, Brinton uses the outer rings to represent days without work and the inner rings to represent the hours worked during the day. Also, he makes use of colour (green) to indicate night hours [4, Ch.2].

The application of visualization in social areas also started to grow. For example, in the New York Times, visualization of time-series

was used to represent the weather statistics of 1980 (Figure 4.24). In this visualization, monthly and yearly aggregates are displayed along with more detailed information on temperature, humidity, and precipitation. This graph represents the temperature, precipitation, and relative humidity along one year in a single visualization [4, Ch.2]. Finally, in this century, new paradigms of direct manipulation for visual data analysis started to appear, as it is the example of linking, brushing, selection, and focusing [57, Ch.1.2].

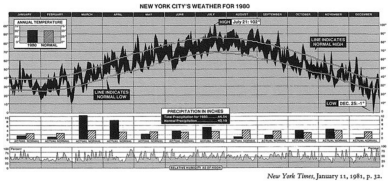

FIGURE 4.24: *New York's City Weather*. New York Times, 1981. This chart shows 2200 entries which summarise the trends and patterns in weather in New York City in 1980.

4.5 Recent Times

The development of hardware and software in the twentieth century facilitated the production of graphics, increased the visual methods available, and enabled the creation of visualizations in every desktop computer [57, Ch.1]. The computer display also facilitated the visualization of multiple temporal events in the same place, enabling the representation of more information in a compact space, and interactive exploration, such as zooming and filtering [4]. All these improvements make it hard to present chronologically and succinctly the most recent developments in time-oriented visualization. As discussed in Section 3.4, most taxonomies focused on time-oriented visualization do not sort the visualization models by representation types. However, we argue that this can aid in the understanding of how time has been mapped in Information Visualization. For this reason, in this Section, we group recent projects of time-oriented visualization according to how they project the independent time variable in linear or radial projections or with multiple granularities. Also, we aggregate other time-oriented visualizations that do not fit in any of the previously mentioned points. It is also important to note that three-dimensional and geospatial visualizations are out of scope, as in this thesis such projections are not explored. Other surveys in time-oriented visualization can be consulted [4, 171, 209, 261, 301].

▸ LINEAR REPRESENTATIONS

The most popular method to visualise time-oriented data is the line chart and its variants. The line chart, introduced by Playfair in 1786, is currently widely used in areas such as finance [21], clinical reports [313], and weather forecasting [37]. Multiple techniques have been proposed to improve the line chart usability and expressiveness [257]. For example, Berry and Munzner [21] developed a set of interaction techniques to enhance the dynamic manipulation and visualization of large time-oriented data through line charts. Zhao

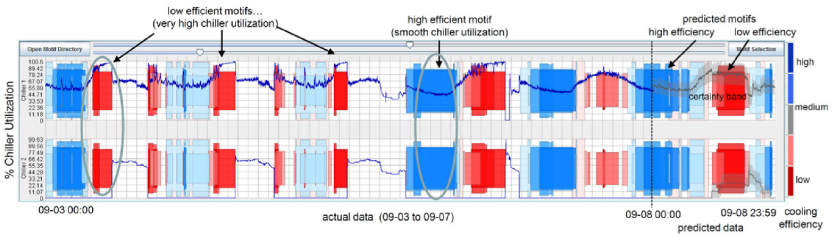

FIGURE 4.25: Frequent Patterns. Hao et al. [119], 2012. Visual representation of frequent patterns in chiller use. This is a multivariate time-series, in which time is represented along the x-axis, the y-axis represents the use of two chillers, and the red and blue rectangles represent low and high cooling efficiency, respectively.

et al. [315], with *ChronoLenses*, aimed at improving the exploration of time-oriented data, by allowing on-the-fly transformation of data points in a focus area. In terms of representing multivariate datasets, and in addition to the line chart, one can use other visual variables along the time axis as seen in the work of Hao et al. [119] (Figure 4.25), or change the line thickness along time [10, 37]. Also, line charts have been used to represent the evolution of time-oriented data in multiple view applications [53, 198, 313], or as an additional view to enable the analysis of data in more detail in grid layouts [196].

In contrast to the aforementioned line charts, which are typically used to represent the evolution of a certain variable at different time points, visualization can also be used to represent the evolution and occurrence of time events, which are usually related to continuous data and defined as time intervals. To represent such data, a common visualization technique is the timeline, which facilitates the understanding of event sequences and their relations. In most timelines, the events are represented through lines. However, unlike line charts, in which the dependent variable changes the line position, in timelines, different events are placed along the same timeline—overlapping if occurring at the same time—or along different straight lines (i.e., timelines)—positioned in parallel to the y-axis [72] or x-axis [74, 207, 228, 288]. To represent relationships among timelines, a common technique is to rearrange their position along the time axis, improving the representation of similarity and/or relationship [74, 173]. In addition to lines, bars are also used to represent events and respective durations or the time between events (Figure 4.26)[71, 115, 207, 228, 304, 305].

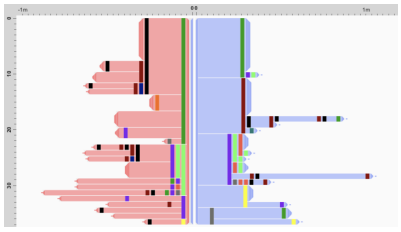

FIGURE 4.26: Temporal Event Sequence. Monroe et al. [207], 2013. This chart shows the UNC offense (blue) vs. the Mary-land offense (red) in a basketball game.

The linear representation of time-oriented data can also be achieved with more complex approaches, in which, for example, the area is used to represent the evolution of a certain value over time [73, 81, 122, 159, 174]. For example, in Horizon [122], the authors were able to increase the data density of a time-series and, at the same time, decrease the space needed to represent it. In this technique, a filled line chart is segmented in equal parts along the vertical axis, and all segments are layered on top of each other with a colour scale

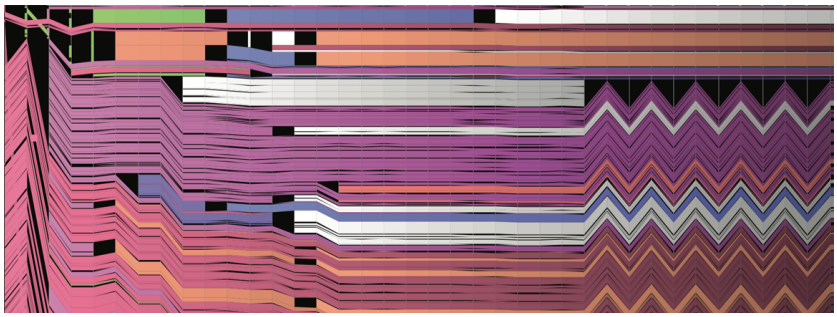

FIGURE 4.27: History Flow. Viégas et al. [288], 2004. Close-up of the Visual representation wikipedia content changes.

representing different value ranges. The use of colour is also common in area charts to represent additional information as, for example, qualitative values [1, 94]. Due to the compactness and ability to summarise the overall and individual temporal trends, the concept of stacked bars was also applied in area charts [257]. In this technique, multiple time-series are represented as stacked layers, whose area varies to represent the value changes. One of the most known 'stacked areas' is the *Theme River* [121]. To show information beyond the overall and individual trend changes, several extensions to the Theme River have been proposed. For example, Cui et al. [73] added visual information about splitting/merging branches between layers to show the inter-layer relationships during their evolution through time. Also, in the work of Shi et al. [257], in addition to the Theme River technique, a stacked bar is applied to represent the distribution of content changes at different time points.

To represent the evolution of different categories along with different time points, parallel coordinates and similar approaches can also be applied. For example, Viégas et al. [288] introduced *History Flow*, a visualization model that aims to reveal patterns within the Wikipedia content (Figure 4.27). In this exploratory tool, time points are positioned along the x-axis, and multiple bars, which represent a certain text added to a wiki document, are drawn, changing in size along time, as the text is modified. Another approach was developed by Gruendl et al. [114], in which a semi three-dimensional visualization is created by representing a parallel-coordinate visualization model, parallel to the x- and y-axis, and by representing time in depth, in the z-axis.

Other visual elements can be used to represent value changes over time. The most common is the use of bar graphs. For its simplicity, such technique is commonly applied in multiple views [27, 152, 290] to provide more details over the data. However, it can also be used as a complement to other techniques such as treemaps [117] (Figure 4.28). Nonetheless, other visual variables we also applied, such as line slope [32], colour [153, 222], and size [139, 314].

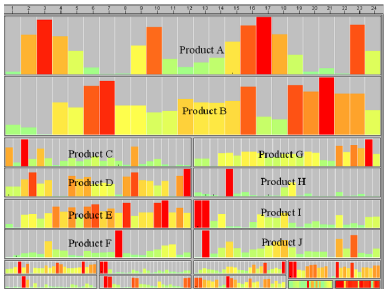

FIGURE 4.28: ID-Map. Hao et al. [117], 2005. ID-Map of product sales time series. The most important product (A) is positioned on top, and the most unimportant, at the bottom of the treemap.

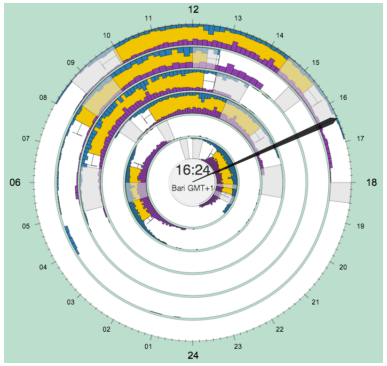

FIGURE 4.29: CBP tool. Buono et al. [39], 2014. Circular visualization of the activities of iOS developers in a team of six, starting from 16pm up to the next 24 hours.

RADIAL REPRESENTATIONS

William Playfair and Florence Nightingale are among the pioneers of radial visualizations [78], a term introduced only later by Hoffman et al. [125]. This type of representation has been widely applied and several methods arose [83]. In the work of Diehl et al. [78], a series of guidelines have been proposed concerning the type of projections (i.e., cartesian or polar) that should be used according to the type of data. This study resulted in a series of observations such as: (i) in terms of effectiveness, cartesian coordinates improved answer correctness and answer time; (ii) in terms of memorability, memorising single cell was easier in radial projections; and, (iii) in terms of encoding, the most important dimension should be encoded in sectors (angle) and not in rings.

In time-series representation, most visualization models apply a linear/cartesian representation. Nevertheless, when the goal is to represent cyclic patterns, especially recurring events along the 24 hours of the day, radial representations can be used. This type of projection has been proven to be advantageous in the analysis of time sequences and the identification of temporal trends [39] (Figure 4.29). Radial representations can be divided into two main groups: the representation of data variables in concentric circles, or along a spiral.

In concentric circles, time can be mapped to angle, and dependent variables are mapped, for example, into the bar's height that changes along the circle [39, 319]. In this type of representation, multivariate data are normally represented in individual concentric circles. For example, in the work of Prieto et al. [96], 12 line charts are placed around a circle to represent the distribution of events per day of the week in each one of the 12 months. With this technique, the authors aim to represent temporal data in a compact view with multiple temporal granularities. Other examples exist in which time is mapped into each concentric circle [40, 147, 269, 309, 316]. In most of these representations, time evolves from the inner circles to the outer circles, and multiple categories are placed in different wedges of the circle [147, 309]. Additionally, time can be mapped into both angle and concentric circles [15, 194], which result in multiple granularities. For example, in the work of Mariano et al. [194], each circle represents a year, which is divided in 12 months (12 angles), and the dependent variable (quantitative) is represented through colour. In the state of the art of radial time-oriented visualizations, an early example also exists in which time is mapped in a zigzagging way through the different wedges of a circle, which represent different data dimensions. In this work, colour is used to represent the evolution of the dependent

variable through time [9].

Spirals are another method to represent time-oriented data. In these cases, time evolves through the spiral, and usually different laps represent different intervals of a certain time granularity (e.g., different days, weeks, months). In most cases, such technique is used with univariate data, the duration of events is represented through the length of circular rectangles [82], and quantitative values are represented through colour [275, 295] (Figure 4.30), circle area, or bar length [47].

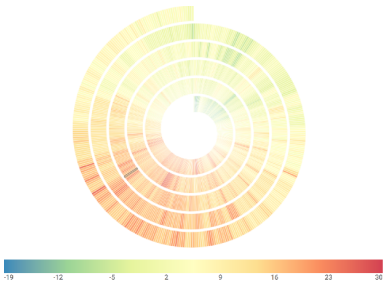

FIGURE 4.30: Spiral Visualization. Tominski et al. [275], 2008. The spiral spiral visualises a subset of 1.825 temperature values throughout 365 segments per spiral cycle that emphasise the yearly temperature fluctuation.

Although the majority of radial representations of time-series are defined by concentric circles or spirals, the state of the art includes another approach that aims to join the radial projection with parallel coordinates. For example, Tominski et al. [274] presents two radial approaches—*Time Wheel* and *Multicomb*—to relate different variable values in the same time point, enhancing multidimensional data browsing and analysis.

MULTIPLE GRANULARITIES

A common structure of time in our everyday life is the calendar. Despite being a well-known model, such a structure is less common to represent time-oriented data. This is a technique in which, usually, data is aggregated by day, and each day is aligned by days of the week and put visually together by month. Nonetheless, a calendar can have different granularities. For example, in the work of Sokol et al. [266], to enable the detection of patterns at different time levels, the authors create calendars in which the data can be aggregated in different granularities (i.e., by month, day, or hour). In the majority of calendar visualizations, each cell is filled with a colour to represent the range of values of a certain quantitative variable, such as temperature [69], work overload [55], or the number of network attacks [266]. Colour can also be used to represent an abstraction of the data as, for example, the distribution of certain clusters along time [172, 268, 282]. In the work of Van Wijk et al. [282], clusters are calculated to identify patterns and trends in multiple time-series (Figure 4.31). Each cluster is represented by a colour, which is then distributed along with the calendar cells. There are other projects in which more complex representations are used to represent the data on each specific date. For example, in the work of Huang et al. [130], a circular treemap is positioned in each day cell to represent the relationships between team members, messages, and conversational threads. In the work of Viégas et al. [286], colour and size are used to represent qualitative and quantitative data variables, respectively.

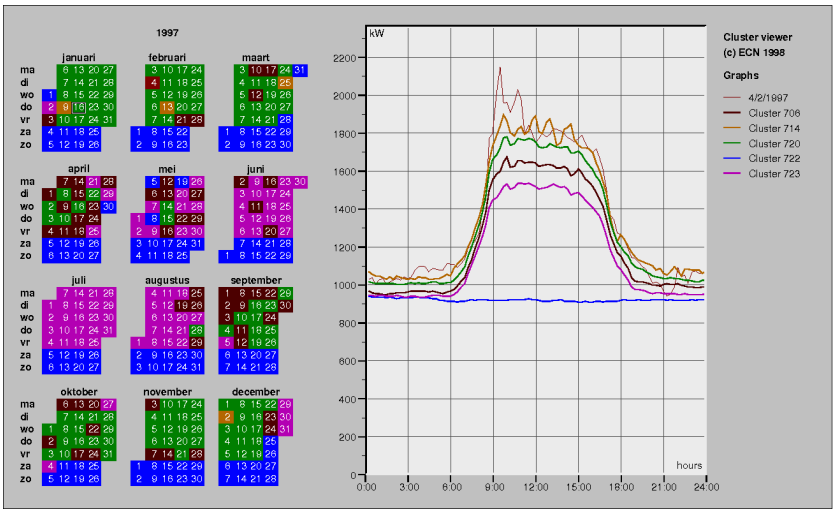

FIGURE 4.31: Cluster and Calendar. Van Wijk et al. [282], 1999. Cluster analysis of power demand by ECN. On the left, the calendar view, and on the right the cluster representation.

OTHER TECHNIQUES

In the previous examples, time is commonly projected sequentially in linear or radial projections. However, other visual representations can be used to represent time-oriented data. To represent duration, the work of Keim et al. [146] uses a matrix to distribute horizontally and vertically two independent variables: starting time, and ending time, respectively. Then the quantitative values which are comprised between each pair are represented through colour. With this representation, time is indirectly represented by a cell which represents the pairs starting and ending times (i.e., duration). In the work of Hao et al. [118], a grid is also used to simplify the representation of time-oriented data. In this project, time is aggregated by a predefined resolution level, and the corresponding values are defined with colour, enabling the overview of the data and the emphasis of existing patterns. In the work of Kumar et al. [160], time is not represented through position but through the analysis of a curved line that connects different events along time. In space, the events are positioned through an algorithm that places similar events close together. Finally, time variables can also be mapped into time itself. In the work of Moere [203], a flocking system is implemented to represent the changes in value over time. In this representation, time is mapped into physical time, and as time passes the boids change their position according to their similarity with other boids.

4.6 Final Remarks

The most common goal in time-oriented visualizations is to represent the evolution of dependent variables and detect temporal patterns visually [257]. To do so, statistical and data mining techniques, such

as clustering and temporal aggregation, can be implemented. These techniques can reduce visual clutter and help users understand data at different granularities. The definition of a visualization model to represent time-oriented data should be done in accordance to its ability to reveal trends, periodicity, disruptions, and temporal patterns [171, 282, 295]. Such visualizations can then be applied in coordinated views [139, 196, 313, 314] or single views [14, 117, 203]. For example, Wijk et al. [282] and Viégas et al. [286] aggregate and represent time-oriented data in a calendar model to reveal patterns and trends on multiple granularities. LiveRAC [196] use a multiple view visualization tool to represent time-oriented data in a grid-based layout, facilitating the comparison of data values side-by-side at multiple levels of detail. Interaction is also an important aspect, as it can enhance the on-demand analysis of temporal data [315, 316], allowing users to select, filter, zoom, and transform what they are seeing and, thus, enhancing data exploration. For example, ChronoLenses [315], provides a set of interaction techniques to support exploratory tasks, such as the exploration in detail of different parts and different granularities of the original data.

The most classic examples of time-series visualization (i.e., bar and line charts) focus on presenting univariate datasets with a small number of time-steps [15]. However, time-oriented data often features multiple, heterogeneous, dependent variables measured for long periods of time, imposing to the visualization of real-world time-oriented data significant challenges [315]. As the datasets get more complex, problems such as clutter, occlusion (due to overlapping visual elements), and an emerging need for scrolling interaction start to appear [117, 118]. To solve this, visual representations of time-series can take advantage of people's perceptual abilities to process information and detect structure in visual patterns [209], making it significantly easier for users to discover trends and new insights at different scales, but also to identify anomalies in the data [38, 196]. Examples of visualization approaches to multivariate time-oriented data can be seen in [107, 139, 194, 222, 269].

There has been comparatively little research on how to support more elaborate tasks typically associated with exploratory visual analysis of time-series, e.g., visualising derived values, identifying correlations, or identifying anomalies beyond outliers. Such tasks typically require deriving new time-series from the data, visualising them, and relating them to the original data [315].

In the present thesis, the aim is to support the fulfilment of domain-specific tasks through the creation of visualization tools that enable the analysis and extraction of trends and patterns through their rep-

resentation in a more appealing, intuitive, functional, and aesthetic manner. With these tools, the aim is to empower the users by improving their relationship with the data and allowing them to explore the data and answer specific problems in the business domain, more specifically, in retail and finance.

The research from this thesis contributes to the state of the art on Time-Oriented Visualization with a set of visualization tools that adapt and combine different techniques from Information Visualization (e.g., histograms, scatterplots, timelines) to tackle different problems in time-oriented visualization: (i) improve the representation of atypical values hidden in cyclic time-series; (ii) overview and highlight consecutive events that should not be missed; and (iii) represent cyclic time-series and capture important trends. Additionally, we contributed to the state of the art on Information Visualization with the development of: (i) two adapted calendar structures that emphasise weekly deviations; (ii) a complex glyph with different levels of readability; (iii) two adaptive timelines; (iv) a visualization model that uses MAS to represent time-oriented data; and (v) the application of Evolutionary Algorithm (EA) and IEC to evolve time-based visual artefacts according to the users' preferences.

Part II

DEVIATIONS

5

Representing Deviations in Retail

"Not only does time structure our lives, it instills our daily activities with meaningful rhythms"
— Viégas et al. [286]

This part of the thesis concerns our research on the representation of trends in time-series through the use of calendar structures. Our goal is to take advantage of the familiarity with calendars to enable a quick overview of all data and, at the same time, enable a detailed analysis of each day and the comparison between weeks and months. In this research, we explore two calendar structures: linear and radial. In the first, and unlike common calendar structures, the days of the week are aligned horizontally. In the second, each month is placed radially and the days of the week are aligned in concentric circles. More specifically, in this part of the thesis we focus on (i) the representation and characterisation of the Portuguese's consumption patterns over time; (ii) the development of two visualization calendars to identify the consumption patterns and deviations; (iii) the analysis and interaction mechanisms used through usage scenarios; and, (iv) the presentation of the results of a user case study that compares linear and radial calendar structures.

In the following Section, we present the context of the projects herein presented. Then, we describe the visualization models in Chapters 6 and 7, and discuss our findings in Chapter 8.

5.1 Context

The age of Big Data emerged from the technological advances that enabled new and innovative ways to collect and store unprecedented high amounts of data. As a consequence, Big Data changed how people work within organisations by allowing access to unparallel amounts of data in new contexts [20, 135]. Hence, with more information about their workflow and results, organisations were able to intensify their ability to make decisions based on data [237].

To improve business intelligence, organisations in the retail do-

main focus, for example, on the understanding of their customer's consumption habits (i.e., what, where, and when they shop). Such analysis can lead to the discovery of new sales opportunities, and to a better understanding of the profitability across products and customers [237]. This type of information is commonly related to time, and therefore, the analysis of time-oriented data— which can be characterised as a sequence of data points indexed over equal or unequal periods [4, 14]—is one of the main tasks for the organisations' analysts. In a business environment, analysts need tools to understand large amounts of time-oriented data in a single place. For example, such tools should be able to highlight moments of high and low consumption values within their product hierarchy so they can adjust their stocks or even prepare new discount events. More specifically, analysts need to understand the impact of their promotions and external events, and identify and analyse the different consumption patterns over time.

By mapping data attributes into visual properties (e.g., position, size, shape, and colour) and by making use of our intrinsic ability to detect visual patterns [26, Ch.27], Information Visualization can be a powerful tool to make sense of data, help analysts discern and interpret patterns in data and aid in the completion of the analysts' tasks [99]. Through visualization models, analysts who work on business intelligence can synthesise, interpret, and present complex amounts of information [142]. More specifically, they can explore the data and detect recurring patterns and disruptions that may influence the business performance [269].

In this part, we apply Information Visualization techniques for the analysis of customer consumption in the retail domain. We worked with SONAE—a Portuguese retail company[1]—that provided us with a dataset containing the consumption values of their customers in 729 Portuguese hyper- and supermarkets, from May 2012 to April 2014. SONAE challenged us to explore this data and develop visualization models capable of representing the consumption variation throughout time and assist the SONAE's analysts in the identification of periodic consumption patterns and atypical values. The SONAE's analysts (i.e., the stakeholders) are not experts in the area of Information Visualization but need to understand how the consumption patterns change over the weeks and what is the impact that daily life events (e.g., Christmas, summer vacations, school beginnings) and previous promotions had on their sales. Through this knowledge, they aim to improve their promotions and stock management. Hence, our visualization models must comply with one main requirement: represent all data simply and efficiently, enabling the rapid identifi-

[1] SONAE (https://www.sonae.pt/en/) is a multinational corporation that manages a wide portfolio of businesses and smaller companies. Amongst various successful businesses, the retail sector plays a significant role in the company.

cation of patterns and deviations, while requiring less exploration to understand the data. To fulfil this requirement, we aim to follow Tufte's principle of data-ink maximisation and use the display space efficiently, minimising the use of ink/pixels and maximising data density [279]. In more detail, the visualization models were developed to enable the analysts to: (i) visually explore the consumption evolution over time; (ii) detect patterns and periodic behaviours in the different product categories of the SONAE's retail chain; (iii) emphasise the atypical consumption values caused by different temporal events; and, (iv) compare the consumption values on different days of the week.

Our approach to solve the analysts' requirement and avoid cluttered visualizations was to focus on the analysis of preprocessed data, more specifically on the previously calculated deviations to the typical consumption values across several product categories. To position time-dependent variables, we explored the calendar structure. The familiarity of the analysts, without prior knowledge in Information Visualization, to the calendar's structure, was ideal since the analysts could easily understand the distribution of the consumption values in the calendar intuitively. Hence, through the calendar visualization we aim to provide an effective and efficient qualitative overview of the Portuguese's annual behaviours, and at the same time, provide a more detailed analysis of the different days. We explored two different approaches to the calendar. In the first, and unlike the common calendar structure, the days of the week are all aligned horizontally (i.e., the same days of the week of all months are in the same row). In the second, we use a radial calendar, structured similarly, but in a more compact layout, which enables us to place selectable data within easy reach for the analyst [83] and increase the readability of the data.

With the calendar visualizations, we were able to show the consumption behaviours and highlight the deviations from typical consumption days. Also, we could demonstrate that the pre-analysis and formulation of statistical values, such as deviations, can contribute to a better summary of the time-series.

5.2 Related Work

Analysing quantitative data involves examining one or more relationships between values that change over time. With time-oriented data, in which time is the independent variable of a set of dependent-independent variables [66], the dependent values can be represented

with different shapes such as lines, dots, bars, or areas that vary over time in size, colour, or position. In the following subsections, we briefly refer to the main techniques used, with special emphasis on calendar structures and radial representations.

5.2.1 *Linear Time-series*

There are many examples of linear representations of time-series, such as the Horizon-Graphs [122], and History Flow [288], and many others can be found (see Chapter 4). The first known time-series based on economic data was published in 1786 in William Playfair's book, *The Commercial Political Atlas*. *Exports and imports to and from Denmark & Norway from 1700 to 1780* is one example of such graphs in which Playfair compares the balance of trade highlighting through colour the difference between import and export data points [278].

Another notable and more contemporary example of a time-series representation is the Streamgraph [42] that stacks areas to represent changes over time within different categories while conveying, at the same time, total volumes. Its layout emphasises the legibility of individual layers, arranging them in a distinctively organic form. However, this type of graph has some problems that we intend to avoid. First, since there is no space between the stacked areas, the changes in one area influence the shape of the surrounding areas, leading to possible misinterpretations of the variations. Also, when the number of areas to represent increases, the readability of the heights of each area and the discernibility among others tend to be compromised. Finally, this visualization model should only be applied when the task is to analyse different categories.

In the previous examples, time is represented linearly and the different categories are represented in the same graph. Another visualization technique that can be used to visualise multiple categories is the small-multiples. Small-multiples are small visualizations of reduced size, indexed by category, and that can be ordered in different ways, according to the context of its use [279]. This technique enforces visually the reader to immediately, and in parallel, compare the differences among objects, relying on an active eye to select and make contrasts [279]. One example of small-multiples is the Tufte's *Consumer Reports, 47 (April 1982)* that shows the frequency-of-repair for automobiles during 6 years [278]. In this visualization, each table represents a car, each column represents a year, and each row represents the evaluation of the typical trouble spots in a car. Each circle is representative of an evaluation that goes from "Much better than average" (white circle) to "Much worse than average" (black

circle). With this visualization, it is possible to compare and distinguish visually which car had more problems and how these problems evolved. More examples of this type of visualization can be found, such as the Flowstrates [31].

Another technique to represent time-oriented data is calendar-based visualizations. For example, in the work of Wijk and Selow [282] a clustering technique is used to identify patterns and trends on multiple time scales: days, weeks, months, and years. To detect monthly patterns, Wijk and Selow aggregate similar days and visually represent those similarities by colouring the different calendar slots. In the work of Viégas et al. [286], a calendar is also used to represent the exchanged emails of a specific user and enable the users to explore their email patterns. This visualization represents the busiest days through size, and how personal a day of emails is through a coloured heat map.

All the above visualization models are efficient in representing time-series. Nonetheless, we intend to overcome some limitations of their approaches. In regards to the linear time-series visualizations, such as line or area charts, they do not allow the analysts to easily depict different consumption values on specific days, which is an important task in this project. To perform such tasks, a calendar structure is a better option as it has more granularities and the days are represented individually. Also, it is not our main intent to represent all categories in the same visualization, so the Streamgraph [42] does not fit our requirements. Regarding the calendar visualizations, they rely practically only on colour to represent quantitative values. In this project, we aim to use other methods to better represent the information, as colour may not be the best visual variable for quantitative data (as seen in Chapter 2). Finally, most of the visualization models do not align all weekdays. This makes it difficult to compare the consumption values on the same day of the week between different months.

5.2.2 *Radial Time-series*

To understand the evolution of data attributes over time, one can structure the time-dependent variables in several ways, including: linearly and radially. The previous examples fall in the first category and the following examples fall into the second one. Radial layouts are well known for their ability to represent periodical patterns and to use the visualization space efficiently [47, 83, 210, 295]. Draper et al. [83] stated that radial models can provide more valuable insights from the data, due to their efficient use of space which eases the

comprehension and interaction of the user. Additionally, Diehl et al. [78] stated that radial visualizations can be more appropriate for focusing on a particular dimension, which is aligned with our intent to represent univariate data (the growth or decline of consumption values).

Radial visualizations position visual data points along a circle, ellipse, or spiral [295]. Some of the most commonly used radial models are pie charts [169, 264], radial bar charts [39, 206, 316], and radar plots [83, 247]. William Playfair's *Statistical Breviary of 1801* [267] and the rose diagram of Florence Nightingale [214] are early examples of radial visualizations. Among the first radial time-series visualizations is the early work of William Farr's *Temperature and mortality of London* [92], which shows London's evolution of temperature and mortality in every week, from 1840 to 1850. In this model, each year is represented by a circle, equally divided in the 52 weeks of the year, and through colour, one can identify the number of deaths and temperature values [275].

Radial visualizations were applied mainly to improve the visualization of periodic behaviours [295]. In 2004, Keim et al. [147] developed *Circle View* to compare continuous data over time. In 2012, Clever Franke divided equally a circle in twelve months to represent the relationship between real weather data and the social media reactions to it [102]. In the work of Paolo Buono et al. [39], the daily routine of work team members is represented in a radial model where each ring represents a twenty-four-hour workday of one team member. Finally, in *KronoMiner*, Jian Zhao et al. [316] focus on non-intrinsically-periodic linear sequences of ordered time points.

The use of spirals to represent time-oriented data is also a well-known technique. For example, in 1998, John V. Carlis and Joseph A. Konstan [47] introduced a spiral visualization to highlight serial data attributes along the spiral axis and periodic attributes along the radii. In 2001, Weber et al. [295] applied a spiral to compare data elements, both in a neighbourhood and between cycles, and to identify patterns and periodic behaviours. In 2008, Tominski et al. [275] implemented a two-tone colouring in a spiral visualization to allow users to read the data values more precisely. In 2013, Xiaoji Chen [58] used a spiral model to represent the evolution of air quality in major Chinese cities.

In the linear examples, the use of a calendar in a grid disposition makes it necessary to use more space, when comparing to the same amount of data in a radial manner. For this reason, we also aim to explore the distribution of the data values in a radial calendar. In our research of radial time-series visualizations, we did not find a radial

model with a structure similar to linear calendars. Our radial calendar differs from the previous radial visualizations as follows. In the case of the Circle View, they use small-multiples to represent different groups, and in our radial exploration, we aim to represent the two years of data in a single visualization, enabling a faster comparison between values. In the work of Paolo Buono et al. [39], as the team grows, the visualization model has also to grow in size, making it difficult to compare all team members in one visualization, and not complying with the aforementioned requirement. In KronoMiner [316], the option to use different charts also requires more screen space and might divide the user's visual attention. While the use of spirals to represent time-series can be advantageous to identify cyclical patterns, they do not fit our goal of discerning the weekly patterns. For this task, it is advantageous to align each week of the month, facilitating their comparison. By using a spiral, this would only be obtained if we defined its period (one lap) as a week, which would culminate in a cluttered visualization. We argue that in addition to spirals, other radial visualizations can be equally efficient to represent cyclical patterns, and, at the same time, use less space. Additionally, we intend to explore these constraints and create a visualization that enables: (i) the detailed comparison between different weeks of consumption; and, (ii) the identification of yearly periodic behaviours in the same space.

5.3 Data

The dataset of this project has a total of 278 GB containing information about 2.86 billion transactions made in 729 hypermarkets and supermarkets, from May 2012 to April 2014. A transaction represents a product bought in a certain retail store and has the following key attributes: date, time, price, quantity, and product, store, and customer identifications. When shopping in retail stores, customers tend to use their client cards to accumulate discounts and other benefits, enabling the association between customer and purchase. Therefore, the transactions are directly associated with a customer through a unique customer card. The cards can be shared among family members, not implying one customer per card. The dataset refers to a total of 6.6 million unique customer cards (typically, a single card is shared by the members of a household) which allows us to define this dataset as a representative sample of the Portuguese consumption patterns—in 2007, Portugal had around 10,553,000 inhabitants [132]. The distribution of purchases per card can be seen in Figure 5.1. We can

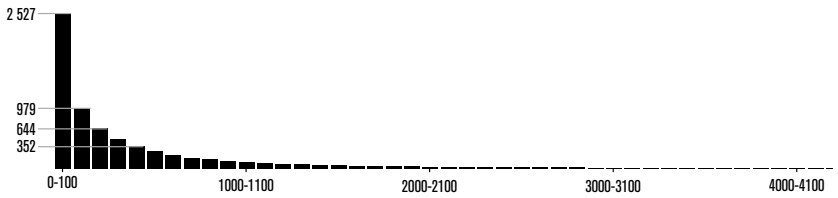

FIGURE 5.1: Histogram of the number of customer cards (y-axis) and the total of transactions (x-axis). Approximately 2.5 millions of customer cards had made from 0 to 100 transactions in the two years range. The histogram ends with a single card with approximately 19150 transactions.

FIGURE 5.2: Scheme of the six levels of the SONAE's product hierarchy.

perceive that the majority of the clients made less than 500 transactions during the two years. Additionally, the highest point of the histogram refers to 2.5 million customers that made less than 100 transactions.

A particularity of this dataset is that all products are identified through a well-defined product hierarchy with six levels, starting in the Department—the highest level—to the product itself, as illustrated in Figure 5.2. For this project, we focus on each product category separately, having a total of 7 Departments, 37 Business Units, 178 Categories, 1031 Sub-Categories, and 5498 Unit Bases. For the majority of our analysis, we discarded the product level as it enables only a narrow view of the consumption values, and we aim for a broader analysis.

5.4 Tasks and Design Requirements

To be able to define the tasks and requirements, we held several meetings with SONAE's analysts. In these meetings, they discussed their main requirements and goals for the analysis of the consumptions data. The SONAE's analysts were not interested in a detailed analysis of the consumption values for each product. Instead, their main requirement was to have an overview of all purchases in specific product categories over the two years, without getting unnecessary clutter. They required simple visualization models that enabled the rapid detection of past atypical consumption values, and could be easily communicated within the team members. For this reason, our goal was to explore a set of alternatives to augment the perception of periodic behaviours and the understanding of how the consumption values vary in time within specific product categories.

To fulfil the goals stated above, the SONAE's analysts aided us to define a set of tasks to which our visualization models should comply:

T1 Analyse the yearly consumption patterns. The analysts need to overview the annual consumption so they can understand the consumption fluctuations along the months;

T2 Compare weeks of a specific month or set of months. To have a better understanding of the consumption values along the month, the analysts need to compare the consumption values over the different weeks;

T3 Analyse in detail the days of the week. The analysts also need to perform a more detailed analysis of the consumption values and their variation within weeks. Also, this detailed analysis should allow them to better understand the impact of specific promotion events and other festivities (e.g., Christmas, popular festivities);

T4 Detect and characterise atypical consumption behaviours. The analysts need to have a mechanism to facilitate the detection of atypical consumption values and study their periodicity, so they can enhance their stock management.

To promote the perception of periodic behaviours, we explored a set of visual alternatives. In the following Sections, we describe the preprocessing of the data and present the results of our initial visual explorations.

5.5 Data Analysis and Preprocessing

FIGURE 5.3: Colour scheme for the identification of each Department.

The first step in the creation of a visualization model is often to preprocess and transform the data to extract meaningful units [145]. As the SONAE's analysts were not interested in the analysis of individual consumption values—their main requirement for this project was to have an overview of all purchases in the different months without getting unnecessary clutter—we started by preprocessing our data. We aggregated each transaction per Department, the highest level in the product hierarchy, and per hour so it was possible to get an overview of the data and detect recurrent patterns.

There are seven different Departments in the dataset: Grocery (biscuits, cereals, frozen foods, hygiene, and cleaning products); Fresh Food (fresh meat, fish, vegetables, and fruits); Food & Bakery (bread, cakes, and coffee); Home (household essentials); Leisure (books, office supplies, pet care, and bricolage); Textile (clothing); and Health (with products from nutrition to beauty). To extract initial clues of how the consumption values evolved, we plotted the raw data of each department into an area chart, where the y-axis represents the consumption value and the x-axis the time. We opted to use an area chart, coloured according to Figure 5.3, instead of a more simple line graph, as we aimed to put the seven results side by side. With

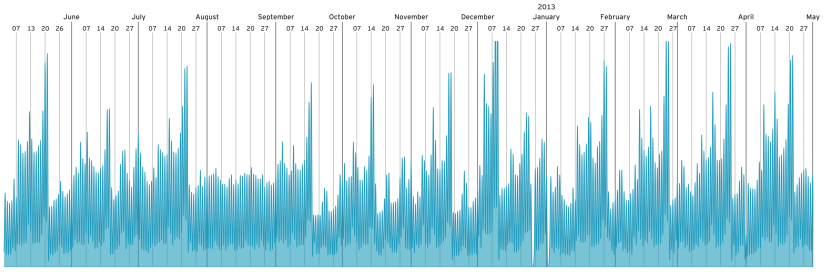

FIGURE 5.4: Visualization of the Health Department, from May of 2012 to October of 2014 with an aggregation of transactions at every one hour.

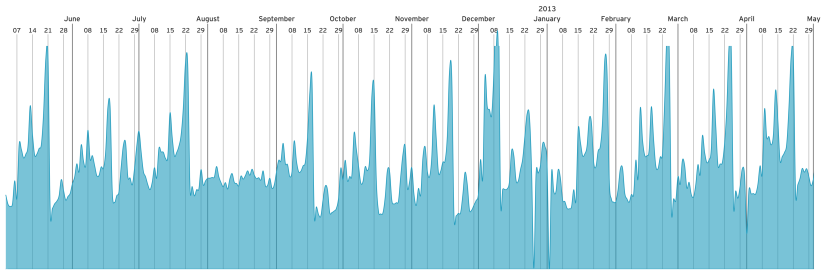

FIGURE 5.5: Visualization of the Health Department, from May of 2012 to October of 2014 with an aggregation by every 3 hours.

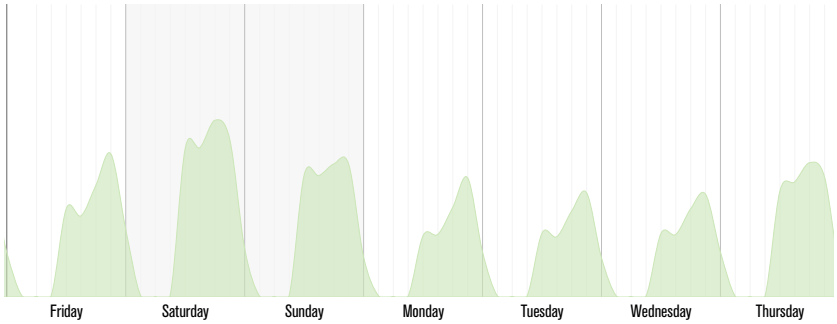

FIGURE 5.6: First week of June of 2012 of the Grocery Department, with an aggregation of transactions at every 3 hours. We can easily see that the customers tend to buy more at lunch time and in the evening.

the coloured areas, it was easier to identify each department at a first glance. We used Catmull-Rom splines[2] to represent the continuity of time in the data and to represent values across time intervals, circumventing the discrete nature of bar charts.

To efficiently explore the data, we implemented the following interactive features for the initial graphs: the navigation over time and the possibility to expand or compress the visualization time window. To smooth the high density of spikes (Figure 5.4), we defined a set of possible time aggregations—1, 3, 6, 12, and 24 hours—to be chosen from. By doing so, we diminished the graphical noise and enhanced the representation of general patterns (Figure 5.5).

When looking into the different days of the week (Figure 5.6), we saw that, in general, customers tend to shop more at the end of the evening, from 16:00 to 20:00. It was perceptible that a common periodic daily behaviour repeated through all departments—at the beginning of the day the consumption values are low, they grow during lunchtime, and have their peak before dinner time. With the daily aggregation, this information is lost, however, it was already well

[2]Catmull-Rom splines are smooth parametric curves that interpolate between a set of points. This method does not require the definition of additional control points for the curves [51].

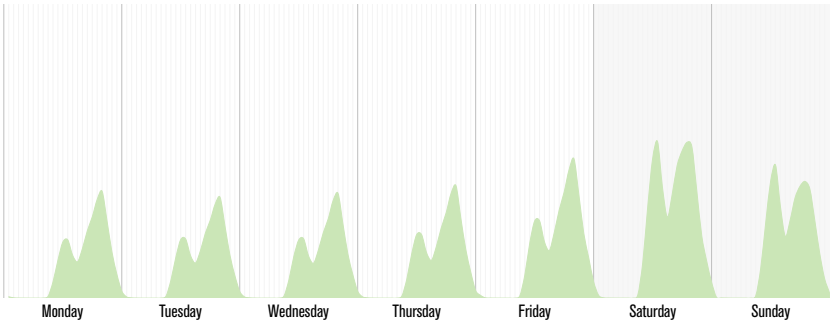

FIGURE 5.7: Week-based baseline of the Frozen Food Business Unit of Grocery Department. This was created through the mean of all values by day of the week.

known by the analysts and such detailed analysis was not their main focus. Hence, for our final model, we focused only on a more general view, where all data is aggregated by day, enabling the overview of all purchases in the different years without getting unnecessary clutter.

We could also perceive a recurrent weekly behaviour during most of the weeks: customers tend to shop less from Monday to Thursday; and on Friday and Saturday, the consumption reaches the weekly maximum, which starts dropping on Sunday. We found that this weekly behaviour is generalised across all Departments. Taking this into account, it was necessary to create a mechanism to emphasise atypical days that stand out from this typical behaviour. To do so, the typical day of consumption should be defined and the deviations from it should be highlighted.

5.5.1 *Week-based Baselines*

To explore the deviations from the typical consumption values, we implemented another visualization model. First, we defined a week-based baseline, to which the deviations are computed. With this week-based baseline, we were able to represent the difference in consumption values (i.e., deviation) between a certain hour and day of the week from the dataset and the same hour and day of the week of the week-based baseline. The week-based baselines represent hourly data aggregations and hence the week-based baseline has 168 points. Finally, the week-based baselines were defined through two methods: the mean computation and a clustering technique.

The week-based baseline computed through the mean is extracted by summing the consumption values for each hour of each day of the week and dividing it by the 168 hours (Figure 5.7). Different product categories showed different week-based baselines, which represent different consumption behaviours. For example, in the Frozen Food Department, during the week the sales are higher in the evening, from 17:00 to 20:00, but during the weekends, the sales are higher at the beginning of the day. The Coffee Shop is the Business Unit that

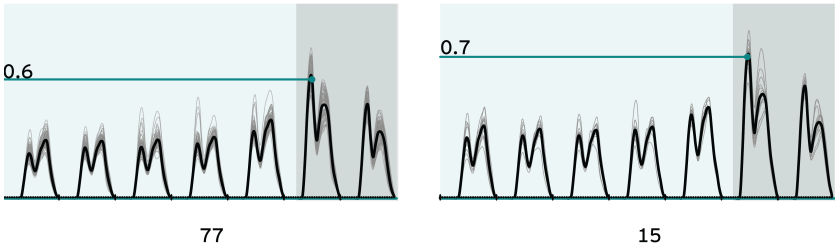

FIGURE 5.8: Small-multiples visualization model applied to the week-based baseline. The weekends are marked with a darker color. Here we present two clusters of Fruits and Vegetables of the Fresh Food Department. The values are normalised by the highest consumption value of the Business Unit in all dataset.

differs the most from the others. In this Business Unit, we can see that, unlike the others, we have three main consumption peaks: (i) in the morning, from 9:00 to 11:00; (ii) in the middle of the day, from 13:00 to 15:00; and, (iii) in the evening, from 17:00 to 19:00. As the product categories have different ranges, every week-based baseline was normalised between the corresponding minimum and maximum values in each hour, so we could better analyse their behaviours.

Considering that each week has abrupt differences in consumption values throughout the timespan, it would be naive to rely only on the mean to extract accurate baselines. For this reason, we implemented another approach in which the baselines are computed by the clustering of the most frequent weekly consumption pattern. Considering the previous time aggregation (i.e., hourly), each week is represented by a sequence of 168 values. The values are normalised as mentioned before. The clustering algorithm is then reduced to comparing the weekly sequences of consumption values among themselves, and grouping similar sequences to determine clusters of weekly consumption behaviours. Having two sets A and B, our measure of similarity s=1−d, where d is the Euclidean distance.

Two sets are considered similar if d is less than a certain threshold. Our clustering approach is a centroid-based algorithm that assigns points to a cluster accordingly with their distances to the cluster's centroid. $\mathcal{S}$ is the set of every day or every week of consumption values in the dataset. A sequence $S_i \in \mathcal{S}$ is then a sequence of n=24 values, for a day, or a sequence of n=168 values, for a week. If O_j is the set of all the *j-th* values of the sequences in $\mathcal{S}$, then the centroid of $\mathcal{S}$ is the sequence $(\bar{O}_j)_{j=1}^{n}$, where $\bar{O}_j$ is the arithmetic mean of the values in a set O_j. Given a set $\mathcal{S}$ of p points and a threshold *eps*, our algorithm computes a list of clusters as described in Algorithm 1. When running the algorithm for every week of each product category, the week-based baseline is defined by the centroid of the cluster with more elements, meaning, the representation of the most frequent consumption behaviour.

In Figure 5.8, we have a small-multiple visualization of the first two clusters of the week-based baselines of the Fruits and Vegetables

Algorithm 1 Clusters definition

```
1  function CLUSTER(S, eps)
2      create list C
3      for all p ∈ S do
4          lastDist ← 1
5          ct ← null
6          for all c ∈ C do
7              d ← dist(p, cluster.centroid)
8              if d < eps and d < lastDist then
9                  lastDist ← d
10                 ct ← c
11         if ct != null then
12             add p to ct
13             compute centroid fot ct
14         else
15             create cluster nc with p
16             add nc to C
17     return C
```

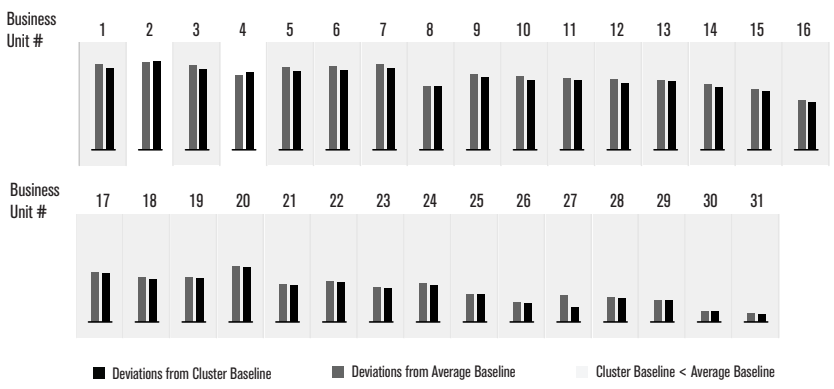

FIGURE 5.9: This graph represents the average deviation from the week-based baselines created through the simple average and the clustering methods with all the dataset. Each column represents one of the 31 Business Units. When the cluster baseline's deviation is lower than the average baseline the corresponding area is filled with light grey.

Business Unit. As we can see, the typical week represents 73% of the 105 weeks. The clusters are sorted by the number of individuals, going from the cluster with more individuals, to the cluster with fewer individuals. In this example, we can see that the highest consumption moments tend to occur during the weekends.

Having two algorithms to detect the week-based baseline, we compared the results obtained by both. To do so, we calculated the deviation between each baseline to all weeks of the dataset (Figure 5.9). When comparing the results, we can see that the cluster baseline has the lowest deviations. By grouping weeks with common behaviours, we can determine the clusters whose centroids are more balanced according to the dataset than when computing the mean. Also, the baselines based on the mean tend to have higher values than

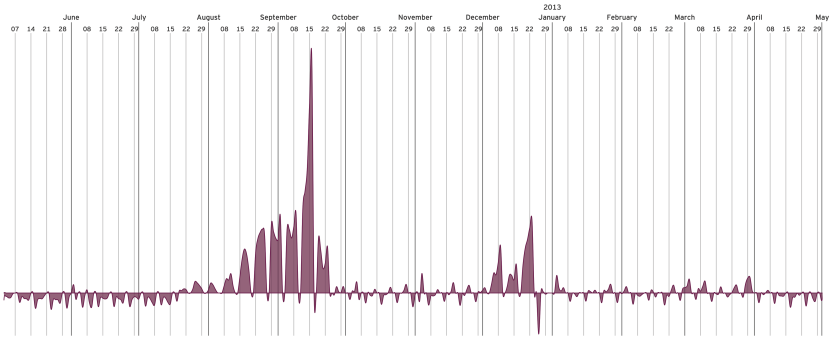

FIGURE 5.10: Visualization of the Business Unit Culture of Leisure, from June of 2012 to April of 2013, with an aggregation of transactions at every 24 hours. It is visible that people tend to buy more on this Business Unit in September and December, coinciding with the beginning of school and Christmas.

the ones extracted through clustering, as the mean is more influenced by atypical days than the clustering technique. For these reasons, we opted to use only the week-based baseline defined by the clustering technique.

With the week-based baselines defined for every product category, we created a variation of the previous visual approach to analyse the deviations. By subtracting the values of a certain hour and day of the week to the baseline's values in the same period, we get the deviation length from the baseline in that period of time. Having the baselines represented as a straight line on the graph, we placed each deviation value above or below that line, depending on whether the value is bigger or smaller than the baseline value (Figure 5.11). This way we represent which time period has values above or below the baseline and how much it distances itself from the typical ones.

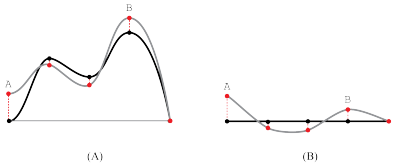

FIGURE 5.11: On the left schematic we can see the baseline, in black, and the consumption line, in grey. Here a set of points are marked and the distance between them are the deviations of the consumption values to the baseline. On the right schematic, these deviations are translated to the new visualization approach.

We tested this visualization model with every Business Unit using the week-based baselines and analysed its deviations. For example, the deviations from the typical consumption for the Culture Business Unit are represented in Figure 5.10. The consumption values were very high from August to October and also in December, probably caused by the beginning of the school period and the Christmas holidays, respectively. We can also see a drop in consumption between days 24 and 29, which matches with the Christmas period. With this visualization model, we managed to eliminate the periodic repetition, and emphasise moments of greater or lesser importance. Although the deviations are clear and we can easily understand what is above or below the baseline, it is difficult to search for specific days and compare their values at a first glance without interaction.

To overcome the aforementioned issues, our final visualization applies a small-multiples technique for the representation of the deviations in each day. We plot each day mark in a structure with more time granularities, the calendar, to better compare the deviations on each day, week, and month, promoting the detection of weekly and monthly patterns.

6

Calendar I—A Linear Approach

To get a general overview of the deviations from the baseline for the whole dataset, we developed a calendar visualization to improve the comparison among daily deviations and emphasise the temporal moments when a certain deviation pattern occurs. To structure the days on the calendar, we position each month from left to right and the days of the week are positioned from top to bottom, from Monday to Sunday, respectively. Each day of the month is placed on the corresponding row and column, making the same days of the week to be horizontally aligned in the visualization. With this structure, the data can be easily interpreted as it corresponds to our experience of looking at calendars. Also, the calendar structure enables the users to identify: (i) long-term trends by analysing the calendar as a whole; (ii) individual trends for each weekday by analysing the calendar rows; and, (iii) weekly patterns by analysing the calendar columns [3].

For the representation of the daily deviations, we defined two different approaches. In the first, we use a linear scale to map the consumption values, highlighting the deviation from the week-based baseline through colour. This approach was well received by the analysts but proved to make it difficult to read small deviations. In the second approach, we overcome this issue and apply a range scale with 5 different levels, emphasising both small and high deviations. Both approaches were implemented in Java and using Processing, an open-source graphical library, to render the visualization[1]. In the following subsections, we describe each of them.

[1] https://processing.org

6.1 First Approach

We took advantage of the small-multiples concept and represented each day as a small visualization of the consumption and deviation

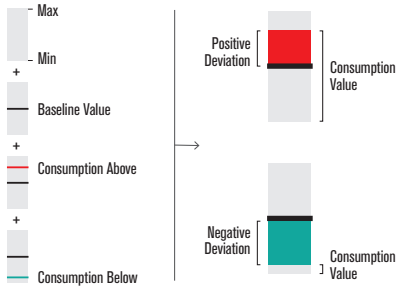

FIGURE 6.1: Scheme of the representation of a day in the Calendar visualization.

values. Each day is represented by a grey rectangle (Figure 6.1). We choose to colour the rectangle with a light colour so each day could be easily distinguished from the others while the consumption values could be emphasised. The top and bottom edges of the rectangles represent, respectively, the maximum and minimum consumption values of the represented product category. Hence, this rectangle's height maps the range of possible values within a certain product category, working as a small scale in which the consumption values of each day are mapped into.

The week-based baseline is drawn with a black horizontal line positioned over the rectangle. For each row of the calendar (from Monday to Sunday) the line will be placed at different heights, according to the baseline's value for the corresponding day of the week. To represent the deviations in each day, we draw a rectangle between the baseline and the consumption value in that respective day. For this reason, the rectangle can be above or below the baseline, if the deviation is positive or negative, respectively. As such, the height of the rectangle is directly proportional to the deviation value on that day. Also, the rectangle is coloured in red, if it has a positive deviation, or persian green if it has a negative deviation.

With this method, we can represent in a single calendar the consumption values and at the same time, the deviations, emphasising the temporal patterns. Also, we can have two levels of information: (i) a general overview of all days where it is possible to see the highest deviations among the different days; and, (ii) a more local view to compare how much the consumption of one day has deviated from the baseline.

6.1.1 *Usage Scenario*

To describe the insights that can be retrieved from the visualization model we present three usage scenarios with two different product categories and one product in particular. Although this last type of analysis is not a priority for SONAE, we want to have this level of detail for the sake of completeness. We aim to test the adequacy of our visualization model for the detection of different consumption patterns.

The calendar for the *Culture* Business Unit can be seen in Figure 6.2. In this calendar, our eyes are drawn to the red columns visible both in September of 2012 and 2013, fulfilling **Task 1** [**T1**] (see Section 5.4). These months have particularly high consumption values and the maximum values occur in the third weekend of both months [**T2**]. These high values can be explained by the fact that:

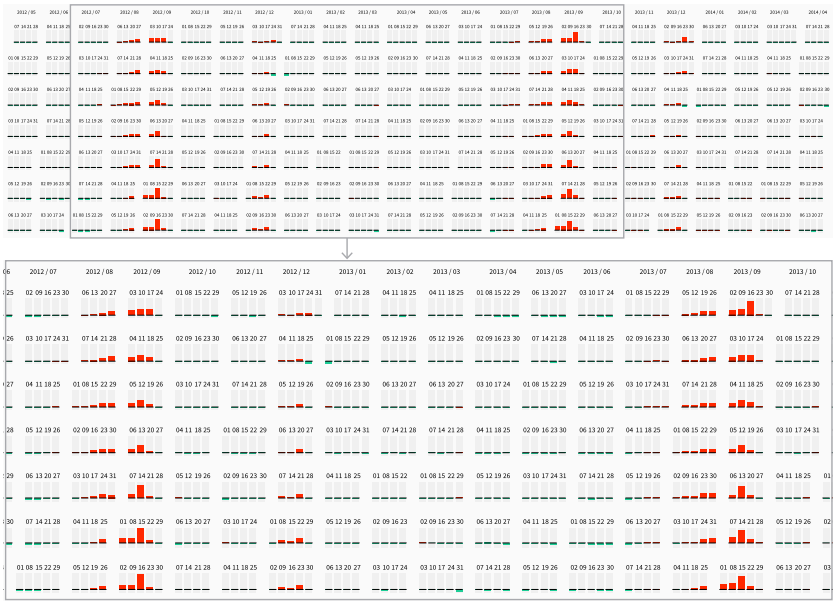

FIGURE 6.2: Visualization of the Business Unit Culture of the Leisure Department, from May of 2012 to April of 2014. With all 730 days being visualised we can perceive that, for example, in September the consumption tends to rise, especially at the end of the month.

(i) this Business Unit includes products such as school supplies and books; (ii) the majority of Portuguese students return to school in September; and, (iii) SONAE usually launches promotional campaigns on books and school materials in September. Also, we can see high consumption values in the final of August—due to its proximity to the beginning of school—and in December, although with lower positive deviations [**T4**]. When comparing December in both years, we can see that 2013 was slightly better in terms of sales than 2012 [**T1**]. As the typical consumption values in this Business Unit are small—making the week-based baseline to be close to the minimum value—the negative deviations are difficult to detect. However, we can perceive that from January 2013 to July 2013, the deviations tend to be negative, especially during the Saturdays between April and June, where the negative deviation is more accentuated [**T3**].

The calendar of the *Drinks* Business Unit can be seen in Figure 6.3. We can perceive that the consumption values in this Business Unit do not have many atypical days [**T1, T4**]. In the two years, we can see the same behaviour: from June to September the sales tend to be slightly higher than the usual, and in December the sales have the highest positive deviations [**T4**]. In December, this high consumption can be associated with the Christmas festivities and the preparations for New Year's Eve. In fact, we can see that the highest consumption tends to occur near the 24 and 31 of December [**T2**]. Also, it is interesting to see a high deviation on July 23 of 2012, which coincides with a Portuguese festivity which occurs on 24—the *São João* which is celebrated especially in the north of Portugal. The fact that it occurred on a Saturday may explain the high consumption values and explain

FIGURE 6.3: Visualization of the Business Unit Drinks of the Food Department, from May of 2012 to April of 2014.

the lower values of 2013 since the 24 of July occurred on a Monday [**T3**]. Also, we can see that the months from January to April have lower consumption values. The exceptions occur in March 2013 from 28 to 31 and April 2014 from 17 to 20, where the consumption values are slightly above the normal. These exceptions can be explained by their proximity to the Easter celebrations. With these results, we can understand the impact that celebrations and festivities have on the purchase of drinks, and consequently, we can understand how they may influence the consumption values of this Business Unit.

When analysing the consumption of a specific product—Codfish—we only detected high positive deviations (Figure 6.4). We can perceive that the highest consumption values occur in December and end of July and beginning of August, in both years [**T1**]. However, it is interesting how the consumption values of December changed in those years. In 2012, we can see that the second week is the week with the highest positive deviations, whereas in 2013, the high consumption values are more evenly distributed by the three first weeks, being the second week also the one with the highest consumption values [**T2, T4**]. It is also possible to detect a high consumption of codfish in May 2012, in the middle of the second week, that does not occur again in 2013, at least with the same intensity [**T1, T2**].

6.1.2 *Discussion*

In this first exploration of the calendar visualization, we applied the concept of small-multiples to visualise the days and respective consumption deviations in the calendar grid. This calendar visualization highlights the deviations from the baselines along time [**T4**],

FIGURE 6.4: Visualization of the Cod Fish Product of the Food Department, from May of 2012 to April of 2014.

eliminating periodic repetition and emphasising moments of greater importance, while still enabling the comparison between different consumption days [**T3**]. With the calendar visualization, we can have a qualitative and quantitative analysis of the consumption values over time and understand behaviours that repeat over the months [**T2**] and years [**T1**]. Also, through the analysis of the usage scenarios, we could access that with this model it is easy to understand when the consumption value is high or low and how the deviations tend to evolve.

After showing the results to the company's analysts, some important aspects were highlighted. Although the calendar visualization was able to give a good overview of the consumption over time when comparing in detail different consumption days, it was difficult to perceive small differences. Due to the reduced size of the daily deviation representation—henceforth referred to as day mark—, the use of a linear scale to represent the consumption values difficult the detection of differences in small ranges. As the main aim was to be able to overview all consumption data but, at the same time, have a more detailed view on the deviations on each week or day, it was important to improve the visualization model concerning this distinction between different value ranges. For this reason, we proposed another version for the representation of the day mark of consumption values.

6.2 Second Approach

For the second version of the day marks, the grey rectangle no longer represents the scale between the minimum and maximum values,

but only marks the day on the calendar. In this second approach, we applied a different scale for the representation of the deviation values. We maintained the distinction between positive and negative deviations but, instead of drawing a rectangle whose size increases proportionally to the consumption value, we defined five distinct scale ranges between which the deviation value may fall. Each of these ranges has a different representation and for this reason, they are easier to distinguish. With this method, small differences between deviation values may still not be noticeable, as they may fall into the same range, but the analysts will be able to have a more detailed analysis of the values, as they will be grouped according to their statistical similarities. Additionally, the representation of the deviations is no longer dependent on the space left between baseline and maximum and minimum values.

To define the ranges, we divided the deviations into positive and negative classes, and for each one computed the first, second, third, and fourth quartiles. The fourth quartile represents 97% of the positive or negative values, and the remaining 3% represent the outliers. The range between the week-based baseline value and the first quartile represents the first range of values. The range between the first quartile and the second represents the second range of values. And so on, until the last range of values that contains the values between the fourth quartile and the maximum value. This range is called the outliers range.

In this new approach, the baseline is always drawn in the middle of the rectangle. This decision was taken since, in the previous approach, the baselines could get too close to the rectangle's upper or lower edges. This made the deviation values to be undetectable due to the lack of space to draw them perceptibly. By putting the baseline in the middle, the positive and negative deviations have the same space to be drawn, having a similar visual impact. In this approach, the analysts can still analyse the visualization model in terms of how much one day deviates from the normal consumption value, but cannot analyse the exact differences in consumption values among the different days of the week. Nonetheless, such close analysis—the understanding of the exact consumption value in each day—was not the main goal of the analysts.

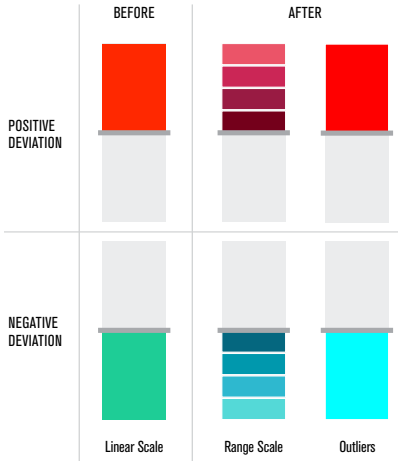

FIGURE 6.5: Differences between the first and second approach. Whereas in the first, we applied a linear scale, in the second, we applied a range scale and emphasised the outliers.

As stated before, we use five different representations for the 5 possible ranges of values. In Figure 6.5, we can understand their visual distinction. The more the consumption value deviates from the baseline, the more complex the day mark gets in terms of composition. Also, we aimed to highlight visually the consumption values that deviate the most from the rest—the outliers. To do so, we opted

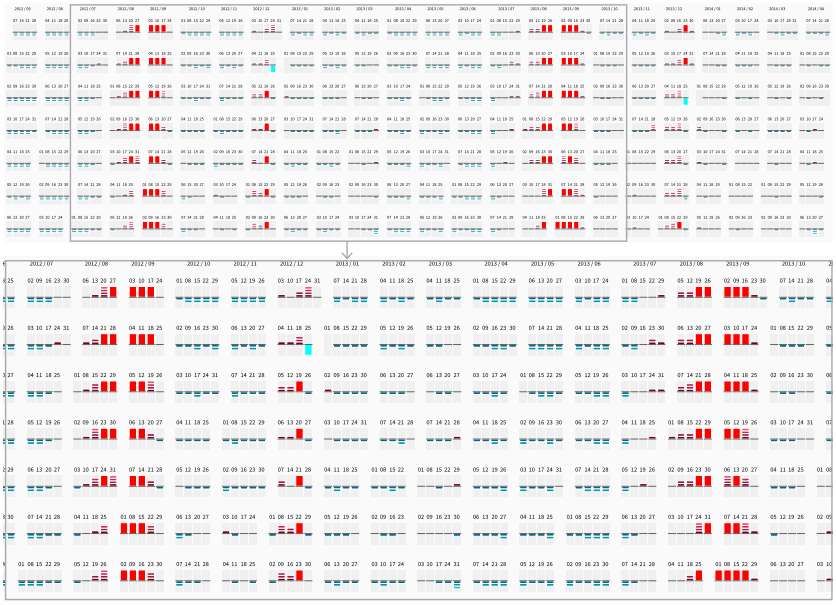

FIGURE 6.6: Visualization of the Culture Business Unit of the Leisure Department, from May of 2012 to April of 2014.

to use a full rectangle whose size occupies the whole section of the deviation.

In addition to shape complexity, we used saturation to better distinguish the different levels of deviation. The higher the deviation the more saturated the colour. Also, we changed the colours from red and Persian green to red and light blue, for positive and negative deviations, respectively. This was done as, in the previous model, the visual distinction between Persian green and red was not possible for colour blind people and the use of Persian green for negative deviations could be misinterpreted. Therefore, we opted to use warm colours for positive deviations and cold colours for negative deviations. Also, we maintained the red colour for the positive deviations as we are unconsciously redirected towards red colours and we aimed to emphasise the positive deviations. This could aid the analysts to more easily be aware of such high deviations and prevent future stock shortage situations, caused by the high consumption of certain products.

6.2.1 *Usage Scenario*

For the analysis of the new approach, we decided to use the same usage scenarios as in the previous one. We aim to perceive how the changes in the day mark changed the reading of the visualizations.

When analysing the *Culture* Business Unit (Figure 6.6), we can easily detect the high consumption in September and December of both years as in the previous approach [**T1**]. In this model, it is easier to see the consumption values on this Business Unit drop from April to the middle of July [**T2**]. We can see that, in 2014, the weekends in

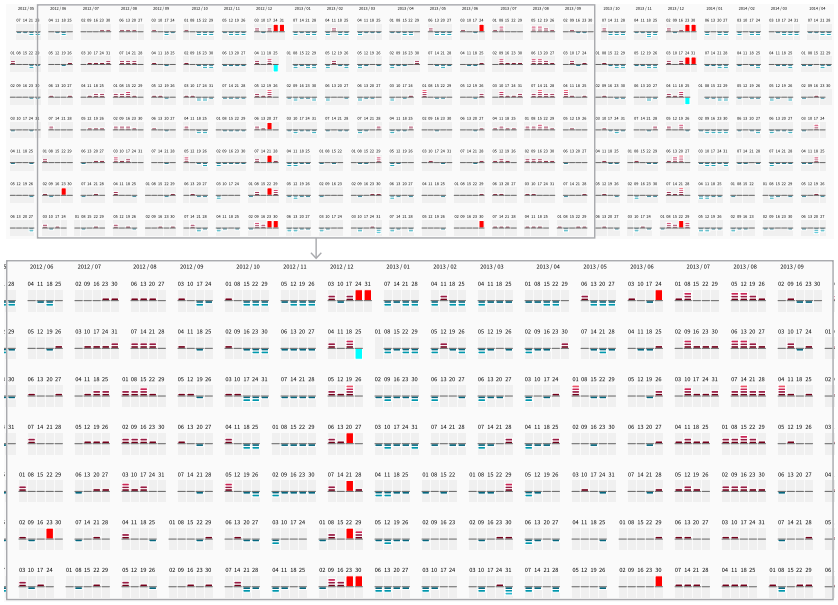

FIGURE 6.7: Visualization of the Drinks Business Unit of the Food Department, from May of 2012 to April of 2014.

the first three months were slightly better than the days of the week [**T3**]. Also, those weekends had higher consumption values than the ones in the same period in 2013, which have in fact higher negative deviations [**T4**]. Also, with this level of detail, we can see positive deviations during the weekdays in March 2013, from 27 to 29. These dates are near the *Good Friday* celebration, a Christian holiday. In 2014, there was also a growth in consumption near April 18, also the date of the *Good Friday* holiday [**T4**]. From this observation, we can see that even holidays which may not be specifically related to the products sold in the Business Unit, may also impact its sales.

In the *Drinks* Business Unit (Figure 6.7), we can instantly perceive the high consumption in December and we can easily distinguish the days with outlier consumption values—about two to three days before December 25 and January 1st [**T1**]. We can better analyse the highest consumption from middle June to September [**T2**]. For example, in August, we can see that the deviations are higher during the week than during the weekends, which means that the consumption in this Business Unit during the week grew more than during the weekend [**T3**]. In the previous approach, this was barely recognisable. Concerning the negative deviations, it is possible to see that in the first three months of the year, people tend to shop less. This behaviour can be seen both in 2013 and 2014 [**T1, T4**]. However, there are some exceptions, as, near February 14 of 2013, in which there are positive deviations, probably due to the celebrations of Valentine's day.

With the previous approach, we hardly could perceive any high deviations in the sales of codfish, besides the one in December. With

FIGURE 6.8: Visualization of the Cod Fish Product of the Food Department, from May of 2012 to April of 2014.

this one (Figure 6.8), the positive deviations can be easily spotted in December, but also in July and August both in 2013 and 2014 [**T1, T4**]. Also, we can perceive a common drop in sales during January in both years. With this approach, we can see more positive consumption values in November of 2012 than in 2013 [**T1**], something that was not perceptible with the previous model. Finally, it is possible to see that the deviations of the consumption values are somewhat irregular, as no visible pattern is preserved throughout the different weeks. Nevertheless, the deviations, in general, tend to be small and negative, with the exceptions of the periods referred to previously [**T1, T4**].

6.2.2 *Discussion*

With the second approach, it was possible to emphasise the highest deviations without neglecting the smaller ones. This was due to the use of different ranges in a range scale of 5 levels. With this approach, smaller deviation values can still be seen. Although they may be less noticeable than higher deviation values, they are still visible to the analyst. In the previous approach, this was not the case, as the linear scale did not emphasise smaller deviations. In fact, since the data had outliers, the majority of the deviations were imperceptible.

The analysis of the negative deviations also improved significantly. As the positive outliers deviated greatly from the normal consumption values, the majority of the baselines were too close to the minimum values. This reduced the space to represent negative deviations and made those deviations almost imperceptible. With this second approach, it was possible to give similar emphasis to the negative and

positive deviations, which improved the reading of the negative deviations. With this approach, we can also better understand the growth of consumption during the different weeks and have a more detailed analysis of the deviations.

A major difference between both approaches is that the specific consumption value is not represented in the second one. Instead, we represent to which range of values a certain deviation value belongs. This makes it impossible to detect the highest consumption value of the dataset. Nonetheless, we can now separate visually the days with the most atypical consumption, highlighted by the outlier representations. To overcome this issue and enable the analyst to visualise the specific consumption value of a certain day, we implemented some interaction functionalities that are described next.

6.3 Graphical Interface

For the analysts to be able to interact with the visualization model and be able to analyse properly the data, we implemented a tool in Java—rendered with Processing—in which the analyst can: (i) choose the product category within the 6 levels of the product hierarchy; (ii) visualise the corresponding calendar of consumption; (iii) mouseover each day and analyse the corresponding consumption and deviation values; (iv) inspect the week-based baseline of the category product; and, (v) visualise the maximum and minimum values of the selected product category.

In the first panel of the tool, the analyst can visualise a list of the first seven Departments of the product hierarchy. By clicking on one, the analyst will see a list of the Business Units contained in that Department. The analyst can continue to drill down on the product hierarchy by clicking on a Business Unit or visualise the calendar of the selected Department. In the latter case, the analyst must click on a button positioned on the bottom right corner of the application ("Ver"). By doing so, this first panel will slide up and the corresponding calendar will be visible for the analyst to explore.

In the calendar panel (Figure 6.9), the visualization model occupies the majority of the upper part of the canvas. In the bottom part of the canvas, there is additional information about the consumption values. On the left side, the analyst can see the week-based baseline of consumption, represented through a line chart, and the maximum and minimum values of the selected product category. Then, by hovering one day mark in the calendar, the analyst has access, in the bottom right corner, to the corresponding consumption and deviation

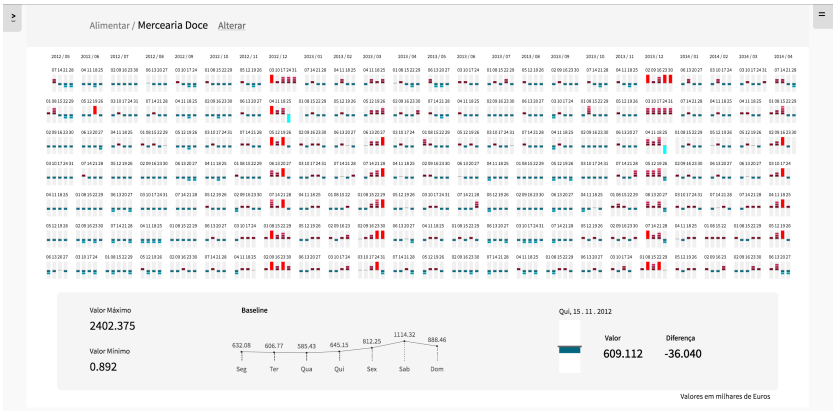

FIGURE 6.9: Graphical interface of the application. The calendar visualization is in the upper part, and additional information (e.g., average values of the week) is placed in the lower part.

values and date. Finally, the analyst can, at any time, select another product category by clicking on the button on the upper right corner, "Alterar".

6.4 Discussion

The calendar visualization enabled us to represent the general behaviour of the Portuguese's consumption over time [**T1**]. Through the preprocessing of the data to an elementary time aggregation, we could represent the consumption variation values over two years in a limited space. For the representation of the deviations, we explored two approaches. In the first, we applied a linear scale and emphasised the deviations through coloured rectangles. In the second, we applied a range scale of 5 levels, highlighting the highest deviations through saturation and representing the outliers differently. This second approach improved the understanding of the different deviation ranges, thus, improving the insights that could be retrieved from the model. The overall exploration of the calendar enabled the understanding of yearly and monthly patterns as well as the detection of the impact that holidays and celebrations can have on consumption.

By representing the deviations from the week-based baseline, we were able to emphasise the atypical days [**T4**], solving the periodic weekly repetition that was intrinsic to this dataset. In summary, the calendar visualization was able to highlight the consumption deviations, emphasise weeks of greater importance [**T2**], and enable the comparison of deviations between days [**T3**]. Finally, our visualization model, despite being developed and created for the visualization of our particular dataset, can be applied to others, given that it is time-varying and uni-variate.

Our visualization aimed to fulfil the requirement to efficiently use the display space [279]. Nonetheless, we think there is room to

further investigate and improve the calendar structure. As such, we further explore space minimisation creating a radial calendar. With the radial model, we aim to be able to compare both years in the same visualization, thus improving Task 1, “enable the detection of yearly consumption patterns”. We also aim to enable the analysis of the evolution of both consumption values and deviations and to improve the interactions and analytical possibilities. These considerations led us to our second visualization model, presented in the next Chapter.

7

Calendar II—A Radial Approach

The SONAE's analysts required a visualization model to overview the data that could be easily understood and adopted by the team. For this reason, the visualization should be compact and simple, representing in a single image the whole time span and, at the same time, be able to show details about each specific day. Hence, our radial calendar aims at improving the previous visualization model in assisting the SONAE's analysts requirements in the search for information about the sales values, how they changed over time, and the impact of sales promotions and external events (e.g., Christmas). In addition to this, we aim to enable the analysis, comparison, and detection of the most relevant departments in the business. This second approach results in a web tool[1] in which we implemented two new views to enable a more detailed analysis and to fulfil the following new tasks:

[1] This web tool can be accessed through the following link: https://cdv.dei.uc.pt/cmacas/deviations_web/

T5 Analyse the weekly average consumption values in each month. For a general analysis of the consumption, the analysis of every day may be time-consuming and complex. Hence, enabling a statistical summary of each week will allow the analysts to faster analyse the data;

T6 Compare the product categories within the same product hierarchy level. In addition to the more detailed analysis of each individual product category, the analysts also need to be able to compare the different product categories from the same level.

In summary, in this new calendar visualization we implemented five views that aim to improve the analysis of the Portuguese consumption:

Product Categories shows each month's average week of consumption for every product category in the same level of the product hierarchy;

Consumption Values shows the consumption values of every day within the two years;

Deviation Values shows the deviations from the typical consumption values;

Consumption Average shows the average consumption value for each day of the week of every month;

Consumption Difference shows the difference between consumption values in the two years.

7.1 Data Analysis and Manipulation

As in the previous calendar, we applied a week-based baseline to compute the deviations. In contrast with the previous calendar, we computed this baseline by each individual month. This decision took into consideration the overall variation of consumption values within a given year (e.g., in December the consumption values tend to be higher than in May). These variations could influence the week-based baseline, making it difficult to accurately represent the consumption values and deviations in each month. It is also important to state that this calculation does not take into account the year, so, for instance, all the consumption values of June 2012 are computed together with the consumption values of June 2013. This decision was made under the understanding that the same month will have similar consumption values independently of the year. As such, to calculate the mean week-based baseline, we summed all consumption values made on each specific day of the week and month and divided it by that number of days. This way, the outlier values that may occur in other months do not influence the results when using the mean. Then we compute the deviations by subtracting a specific consumption value with the corresponding week-based baseline value of the same month.

To enable the visualization in the *Consumption Average View*, we calculated the mean consumption of all years in a certain calendar position, i.e. the first Monday of the first week of a certain month. To do so, we summed all consumption values in each position of the calendar and divided it by the number of occurrences. We intend to perceive if, in specific periods of time, certain consumption values are repeated independently of the year, enabling us to detect yearly patterns as, for example, the growth of consumption values during the Portuguese summer vacations, in August.

All these aggregations and calculations can then be seen separately in our visualization model through a set of buttons that appear in the web-application.

7.2 Radial Calendar

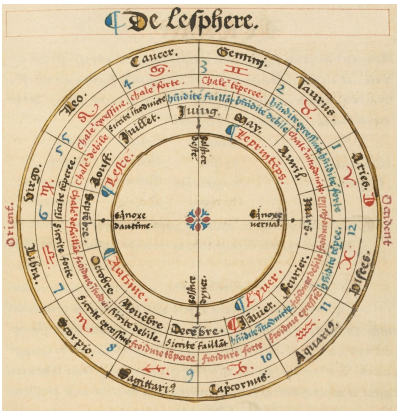

FIGURE 7.1: Astrological calendar from *Le Sphere du Monde*, 1549, by Oronce Fine.

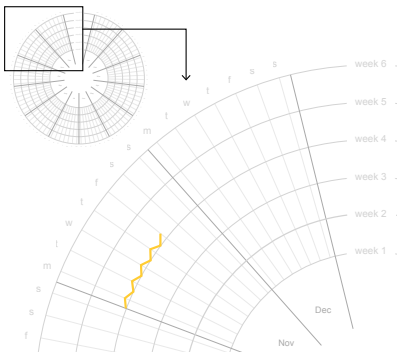

FIGURE 7.2: Structure of the radial calendar and respective labels positioning: month, in the inner part of the smallest ring for each wedge; day of the week, in the outer part of the biggest ring; and the number of the week, in the first wedge in the upper part of each ring. A yellowish zigzagging line will be positioned on the calendar to mark special events.

The use of radial calendars has a long tradition, especially in the representation of astrological data, such as the calendars from Oronce Finé in 1549 (Figure 7.1). We wanted to reinterpret this calendar model in new media to further explore this positioning of time and to emphasise weekly and seasonal behaviours that repeat over time. Also, we wanted to be able to augment the amount of data in a limited space without getting too much clutter.

For the implementation and design of our visualization, we divided the radial model into 12 equal wedges—representing the 12 months. As we wanted to represent and enhance the weekly consumption behaviour and its repetition through the different months, we opted to further divide each month wedge into seven parts—corresponding to the seven days of the week. Therefore, we divided each ring into a total of 84 wedges of the same size (12 months × 7 days of the week).

Since our goal is to understand how the data evolves during each month, we position separately each week of the month in rings with different radius (Figure 7.2). The first week is placed in the ring with the smallest radius, and the succeeding weeks are placed in the following rings. As a result, we have 6 rings which define the 6 possible weeks of a month (as months do not always start on a Monday, they can "occupy" from 4 to 6 different weeks). With this positioning, we can compare the evolution of the consumption values among every week of every month, and easily perceive weekly patterns.

We added labels to increase the readability of the calendar's values. These labels refer to the month represented in each wedge (positioned in the inner part of the smallest ring), the days of the week (positioned in the outer part of the biggest ring) and the number of the week (positioned between the first and last month wedges). We decided to visually separate the representation of the consumption values with the label of the week's number, to increase the distinction between the end and beginning of the year. In our first analysis of the data, we perceived a substantial difference in the consumption values between December and January. With this visual gap, we aim to emphasise this consumption difference and, at the same time, facilitate the identification of each week number.

With the calendar structure defined, we can position every value in the corresponding slot of the radial calendar. For every year, we place the consumption values in the respective slot, which means that

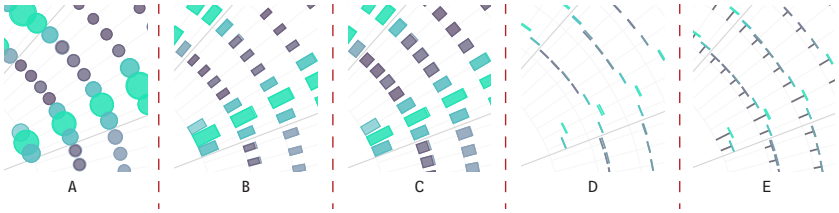

FIGURE 7.3: Different marks to represent the consumption values: (A) circles; (B) rectangles with width and height according to the consumption value; (C) rectangles with only the height representing the consumption value; (D) parallel lines; (E) parallel and perpendicular lines..

there may be more than one consumption value (of different years) in the same slot. Note that the first day of July 2012, for instance, may not be positioned in the same slot as the first day of July 2013. If those days do not occur on the same day of the week, they will be positioned in different slots, corresponding to the specific day of the week of each. Nevertheless, they will be positioned in the same ring, the one with the smallest radius, which represents the first week of the month.

7.2.1 *Day Mark Study*

For the representation of each consumption day, we explored the use of three different marks: circles, rectangles, and lines (Figure 7.3). For the circle marks, we used the area of the circle to represent the consumption value. For the line marks, we explored two different approaches: (i) we created a parallel line to the week ring with a distance equivalent to the consumption value; and, (ii) in addition to the previous parallel line, we created a perpendicular line with the length of the consumption value—this perpendicular line is intended to emphasise the perception and distinction of the consumption value. For the rectangle marks, we defined the length of the rectangle as the consumption value, and also explored two approaches to define its width: (i) we maintained it constant; and, (ii) we changed it according to the consumption value.

From these approaches, we promptly discarded the line marks as they were more difficult to read. Since the calendar is radial, it was difficult to compare the distances of the lines to the baseline. Even with the perpendicular line to aid in that task, it was still difficult to read as the area that these marks occupy is reduced. In the case of the rectangle and circle marks, we concluded that the creation of gaps and clutter with those marks emphasise the differences in consumption values and aid in the detection of higher consumption days (more clutter) and lower consumption days (more gaps between days). Both represent the consumption value with their size, but, as the circles are harder to compare, we opted to use the rectangles, whose length and width alter according to the consumption value, as our main mark for the radial calendar. However, as the values

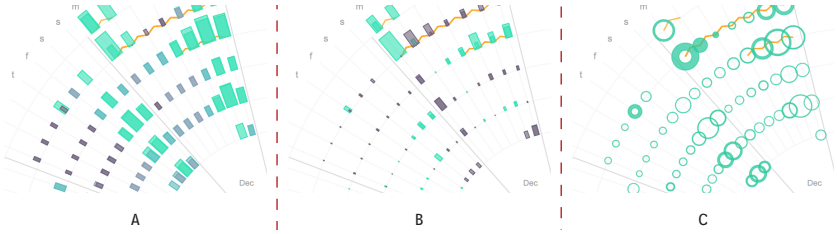

FIGURE 7.4: Three different visualizations: (A) the consumption values aggregated by day; (B) the deviations (size represents the deviation length and colour if it is negative or positive); and, (C) the differences between years (the thicker the doughnut, the highest the difference).

represented in each view differ, we hereafter explain how we applied different marks to represent the different values.

Consumption Values In this view, we visualise the consumption days of the dataset by a specific product category. We use the rectangle as a mark and the size represents the corresponding consumption value on a certain day of the time span. As stated before, both years are represented in the same model, which means that more than one rectangle can be represented in the same slot of the calendar structure. The values are normalised by the highest consumption value of the represented product category.

Deviation Values In this view, the deviations are also represented by a rectangle. We analyse the values of each individual product category. The size of the rectangle corresponds to the absolute deviation value of the week-based baseline (see Section 7.1). As in the previous view, both years are represented, therefore, two deviations can be represented in the same slot. All deviations are normalised by the highest deviation value in the two years in the represented product category.

Consumption Average In this view, and unlike the previous views, we visualise the average consumption value for each day of the week of every week of a certain month. This means that we can compare the typical consumption value in, for example, every last Sunday of a set of months. The rectangle is the mark and its size represents the average consumption value. The size is normalised by the maximum average value in the represented product category.

Consumption difference This view's day mark differs from the previous ones. In the mark experiments, we found that to show the differences in consumption from one year to the other, the use of circles was also a good choice, as they occupy more area and it is easier to understand by the analysts which days have the highest consumption. For this reason, we apply the circles to represent the difference in consumption values between both

years. We created a doughnut mark where the outer circle represents the maximum value of consumption, the internal circle represents the minimum value, and the area between them represents the difference in consumption. The thicker the doughnut, the larger the difference between consumption values (Figure 7.4). In this view, all values are normalised by the highest consumption value in the represented product category.

Product Categories In this view, the used mark is the rectangle and the structure of the calendar is slightly modified. We maintain the wedges representing the different months and the seven days of the week, but the rings of the calendar structure, instead of representing the number of weeks, represent the different product categories within a certain level. With this structure, we aim to represent the weekly average consumption values for each month in the different product categories. For this reason, the size of the rectangles represents the average consumption value of each day of the week in each month. All values are normalised by the highest average consumption value of a set of product categories.

As the size of a mark can be difficult to analyse [66], leading to misinterpretations of the data and wrong comparisons, we decided to use colour to emphasise the differences in consumption. As the consumption values can comprehend a large range of values (millions), and subtle differences in colour would also be imperceptible for the analyst, we opted to restrict the colour palette to four possibilities. This colour palette is defined through the calculation of the quartiles of the consumption values, which divide the dataset into four equal groups. Low values get a darker purple tone, and high values get a green bright tone[2]. This colour palette is applied in the majority of the views with the exceptions of the *Deviation Values* view, in which grey represents negative deviations, and green represents positive deviations, and in the *Consumption Difference* view, in which colour does not represent the value, as all circles are coloured in green.

[2]Four colour tones:

7.2.2 *Additional information*

To enable the analyst to have a better understanding of the weekly consumption behaviour in each product category, we added a 7 th line in the outer part of the visualization in which we draw the weekly average consumption value in each month for the product category being visualised. The rectangles in this line are mapped and coloured in the same way as the rectangles in the calendar structure.

To improve the understanding of the consumption patterns, we opted to give additional context about the data and marked on the calendar some of the principal festivities, e.g. Christmas and *São João* (one of the Portuguese traditional festivities). With the understanding that too much visual clutter may harm the legibility of the visualization, we opted to represent the periods in which the festivities occur with a yellowish zigzagging line (Figure 7.2). This line is intended to be visible but, at the same time, not intrusive. Therefore, it is drawn in a light colour and positioned behind the consumption marks. Note that these festivities may not be directly responsible for the growth or decay of the consumption values. This representation is an effort to aid the analyst to create possible associations between consumption and external events.

7.3 Graphical Interface

We implemented a web-application in javascript with a set of functions to enable the analyst to further explore the visualization and get an overall understanding of the dataset. A video of the web-application can be assessed in: `https://cdv.dei.uc.pt/radial-calendar/`.

The graphic interface is divided into two main areas: (i) the options area, placed on the left side of the page; and, (ii) the visualization area, placed on the right side (see Figure 7.5). In the first, the analyst has access to the title of the visualization, to a set of buttons that manipulate the visualization, and to an index with all Departments and respective lower levels of the product hierarchy. The title of the visualization enables the analyst to see which product category is being represented on the visualization, changing according to the product category that is being represented (e.g. "Analysis of Grocery").

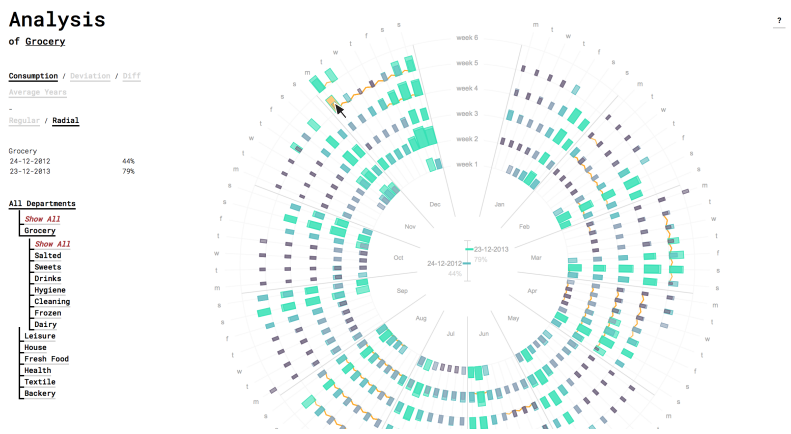

FIGURE 7.5: By hovering the mouse over the rectangles, the analyst can see the percentage of consumption on those specific dates. This information appears in two places: (i) in the centre of the calendar visualization; and, (ii) above the index of all Departments and Business Units.

Below the title, we positioned a set of buttons that enable the analyst to: (i) see the deviations from the normal consumption value; (ii) see the consumption value by day; (iii) see the mean consumption of the two years that occur on the same slot of the calendar; and, (iv) see the difference between the highest and lowest consumption values (Figure 7.5).

Also in the left area, the analyst can explore the product hierarchy through an index. In the beginning, the analyst only sees "All Departments", which means that, currently, the visualization is representing the values aggregated only by date, not distinguishing product categories. As the analyst clicks on "All Departments", a list of all Departments will appear below. In addition to the Departments, we also added a button entitled "Show All". By clicking on it, the analyst will see the mean week of consumption in each month for every sub-level (e.g., Department, Business Unit, Category). To distinguish it from the buttons of the other product categories, we used an italic font and coloured it in red.

When hovering over each consumption mark, the analyst gets additional information about the data (Figure 7.5). This information consists of the dates and percentages of the consumption values and appears: (i) in the left area, between the options and the index; and, (ii) inside the radial calendar, in a visual chart. This way, the analyst can analyse more precisely the differences in consumption on the same calendar's slot, perceiving by how much one consumption is bigger than the other and whether the consumption values are closer or further from the maximum and minimum values.

Finally, in the bottom right corner of the interface, there is a label, containing information about the used colours, the relation between size and consumption value, and the used representation for the festivities.

7.4 Usage Scenario

Having reviewed in detail the radial calendar and the graphic interface, we now present a usage scenario. We aim to illustrate how the radial calendar represents the Portuguese consumption data and how the interactive environment enables the exploration and analysis of its patterns, allowing a better understanding of the Portuguese's consumption patterns.

When the analysts access the web-application, they visualise the overall consumption values by day. In this first visualization, it is possible to identify a weekly pattern of consumption—the highest

consumption values are in most weeks situated near the end of the week (i.e., usually, customers do more shopping during the weekends than during the weekdays) [**T3**] (see Section 5.4). Furthermore, it is also possible to perceive that in December the consumption values tend to be higher and that during August there is a disruption in the weekly patterns, as there is no significant difference between the consumption values on weekends and weekdays [**T1, T3**]. These behaviours can be explained as follows: in December, the consumption grows due to the impact that Christmas has on the shopping habits of the Portuguese; and as in August most Portuguese are on vacations, the customers can do their shopping both on weekdays and weekends. Moving the mouse over each rectangle, the analysts can have a more detailed view of the consumption values, in each day.

By selecting the button "Average Years", the analysts get a less cluttered view of the consumption. If the analysts compare this view with the first one, they can see that the consumption values in both years have similar characteristics, since the rectangles do not change much in size from one view to the other. An exception is, for instance, on the Monday of the fifth week of December, where on 24 of December 2012 the consumption value is 52% of the maximum consumption value and on 23 of December 2013 the consumption value is 93% [**T5**]. This difference can be explained by the fact that on 24 the retail stores close earlier than on 23, and by the fact that most people have already done their Christmas shopping by December 24.

The analysts can then search through the hierarchy index and see how the consumption values are distributed over the different Departments and corresponding lower levels of the product hierarchy. By clicking on the button "All Departments", the index will expand and present all Departments of the product hierarchy. Additionally, a button named "Show All" will be visible. By clicking on it, the analyst can visualise a summary of the consumption values on all seven Departments in the same visualization (Figure 7.6) [**T6**]. In this view, each Department is represented by a single ring of consumption values, properly labelled with the name of the Department in the upper part of the ring. In each ring, we represent the mean week of consumption for each month in each specific Department. It is important to note that each consumption mark is mapped between the minimum and maximum mean consumption values of all present Departments, enabling the analyst to understand how the consumption values are distributed among Departments and detect which Department has the highest consumption value [**T6**]. In this case, the analyst can perceive that the *Grocery* and *Fresh Food* Departments are the ones with considerably higher consumption values.

FIGURE 7.6: Visualization of all Departments' monthly mean week. In this visualization, each ring represents one Department and each wedge one month.

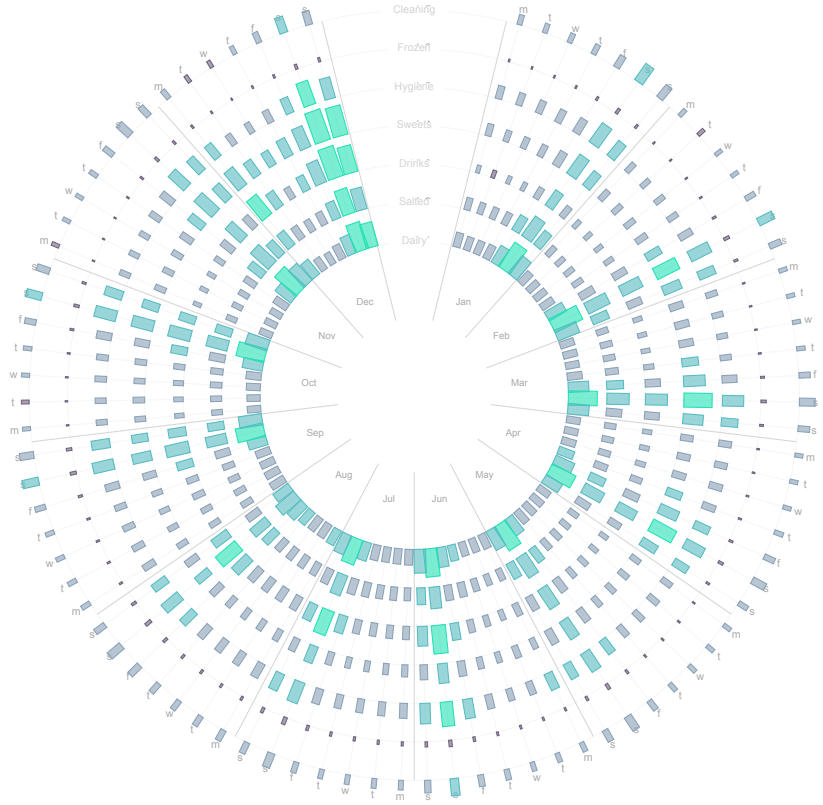

FIGURE 7.7: Visualization of all Business Units in the Grocery Department.In this visualization, each department is represented by the mean week of each month.

Furthermore, when the analyst clicks on a Department in the index, the index will also expand and show the corresponding Business Units. As in the previous step, a "Show All" button will appear, which has the same function of showing all Business Units of a certain Department in the same place, with all values mapped to the minimum and maximum mean values on those Business Units [**T6**]. For instance, if the analyst clicks on the *Grocery* Department in the index area, and then on the respective "Show All" button, the visualization will change and will represent all monthly means of consumption of the respective Business Units (Figure 7.7). The analyst can perceive that the consumption values of all Business Units are relatively well distributed, and the *Frozen* and *Cleaning* Business Units are the ones with lower consumption values [**T6**]. In this visualization, the analyst can also perceive that the *Dairy* Business Unit appears to have higher consumption values in the first days of the week, and in August most Business Units have equivalent consumption values throughout the whole week (compared to the majority of the other months) [**T5**]. The only exception to this is the *Sweets* Business Unit, which appears to have a similar week behaviour in all months, except in December, where it has superior consumption values, even when compared to other Business Units. These similarities in the consumption behaviour within all Business Units do not happen in all Departments. For example, in the *Textile* Department, the *Men* and *Women* Business Units have a similar consumption behaviour, whereas the *Kids* and *Baby* Business Units have a different consumption behaviour. These differences enable us to understand that there is a distinction between the preferred months to buy clothes for adults and kids.

The analyst can also analyse the difference in consumption values between the two years of the dataset at any level of the product hierarchy. For example, if the analyst clicks on the *Drinks* Business Unit and on the "Diff" button available on the left (Figure 7.8), it is possible to perceive that, in general, there are no big differences in the consumption values—the doughnuts are visually thin—which means that the consumption values have similar values along the years [**T1**]. However, on June 23 of 2012, there is a doughnut which is significantly thicker. With some research, it is possible to understand that this difference is caused by a promotion of 50% on beer which occurred only on that day. Thus, the analyst can understand that those promotions had a positive impact on sales.

When comparing all Business Units of the *Health* Department, one can perceive that the *Beauty* Business Unit is the one with the highest consumption values [**T6**]. By selecting in the index menu this Business Unit, the analyst can also understand how the consumption values

FIGURE 7.8: Visualization of the difference between the maximum and minimum consumption value of the Business Unit Drinks in the Grocery Department.

FIGURE 7.9: Consumption distribution along the years in the Beauty Business Unit.

FIGURE 7.10: Visualization of the deviations on the Women Business Unit of the Textile Department.

are distributed over the years [**T1**] (Figure 7.9). In this Business Unit, one can perceive that the highest consumption values occur in the middle of the months (i.e., second weeks) and from January to March the highest consumption values tend to occur between the third and fourth weeks [**T2**]. Then, there is a shift in the highest consumption values, which begin to occur between the second and third week from April to August. After this, the consumption values drop in the following months until December, where the second week of the month is the one with higher consumption. It is also interesting to note the periods of these peaks of consumption: in December, which coincides with the Christmas period, and in June, which coincides with the start of the summer vacation period. When verifying the promotions in June, the analyst can detect that the promotion on 17 of June of 2012 of 25% on *hygiene* and *hair products* and the promotion on 30 June of 2013 of 50% on *hair products* and *sun protectors* had a positive influence on consumption.

When seeing the deviations' visualization in the *Women* Business Unit of the *Textile* Department, one can perceive that there are no high deviations in most of the months, which is caused by the common low consumption values (Figure 7.10) [**T4**]. Nonetheless, in October, it is perceptible that the consumption values change throughout the month. At the beginning of the month, the consumption values

are below the "normal" consumption value, making the deviations negative. However, at the end of the month, a significant growth in consumption, especially during the weekend, is noticeable [**T3**]. It is also perceptible that this consumption growth is not shared by all years in the dataset. When clicking on the "Average Years" button, one can see that these high deviation values in the fourth week of October, decreases [**T5**]. These differences in consumption values were caused by one promotional event which occurred between 25 and 28 of October 2012, in which every clothing item had a 50% discount. This promotional event also occurred on 22 of June 2012, and in this visualization, its impact is also perceptible through the high consumption deviation of June.

7.5 User Study

We conducted a user study to compare our radial calendar to a regular calendar structure. The objective of this study is to explore the intuitiveness and effectiveness of the radial layout in the representation of consumption data. For this study, 30 participants tested the two visualization models concerning performance (i.e., efficiency and accuracy) and visualization quality—subjective evaluation of clearness, intuitiveness, and general attitude toward the visualization). The participants' ages were between 21 and 47 with an average of 27, and, on average, they ranked their expertise in visualization as "some understanding of the visualization domain". Each testing session took between 15 to 30 minutes.

7.5.1 *Method*

The study envisioned three different phases. In the first, we contextualised the used data, the visualization interface (presented in Section 7.3), and the two visualization models. In the second phase, and to reduce user fatigue, the participants had to fulfil 6 tasks:

Ta In which pair [month, week] the consumption value tends to be higher?

Tb Is there any week with recurring higher consumption than others?

Tc Which Business Unit has the highest consumption averages during the year?

Td Is the consumption on the first Saturday of January higher than the third Thursday of August?

Te Which pair [month, week] has the highest deviations?

Tf Which pair [month, week] has the highest differences between years?

Since we aimed to compare two alternative visualizations, instead of evaluating the interactive prototype itself, we used static images of the two visualizations. We also aimed for the users to be fully invested in analysing the visualization model and not distracted with interaction. In each test and for each task, we randomly used one visualization model to represent the same dataset.

All tests had a total of three tasks to be answered with the aid of the radial calendar and three other tasks to be answered with the regular one. The two visualization models were used in a balanced manner, each one of them being used fifteen times for each task. For each participant, the tasks and selected visualization model were randomly sorted. For each task, the participant starts by reading the task at hand and then the participant has to choose the correct answer from a multiple-choice answer. The time was counted from the moment the user ends reading the task until the completion of the task. After completing the task, we asked the participant to rate its difficulty. In the third phase, the participants were allowed to freely look at the prototype and interact with the visualizations. Then, they had five questions where they have to compare the two visualizations:

Q1 Which visualization was more aesthetically pleasing?

Q2 Which one aroused more curiosity?

Q3 Which one was easier to learn and navigate?

Q4 Which one did you find more useful?

Q5 Which one was more intuitive?

In the end, the participants could give subjective comments on both models[3].

[3] A PDF of the User Test can be found in the following link: https://cdv.dei.uc.pt/cmacas/RadialCalendar_UserTest.pdf

7.5.2 *Results*

To analyse the results of the user testing, we used the statistical software SPSS version 24. To check if our data followed a normal distribution we applied the *Kolmogorov-Smirnov* test with a level of significance $\alpha = 0.05$. The test revealed that the data did not follow a normal distribution. Based on this, we applied a set of non-parametric tests. We used the non-parametric multivariate version of the ANOVA, with a level of significance $\alpha = 0.05$. If we perform our analysis considering only the visualization model, the results revealed that

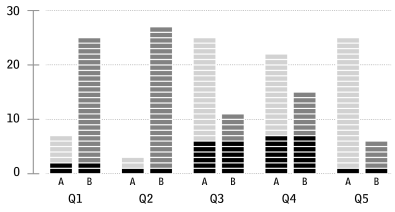

FIGURE 7.11: Distribution of answers to the questionnaire (each rectangle represents one answer). The lighter grey (columns A) represents the regular calendar, and the darker grey (columns B) represents the radial one. Black is used to indicate when a participant answered both models in a certain question.

in terms of time, the regular calendar is statistically different from the radial calendar (p-value= 0.019). In what concerns accuracy, there are no significant statistical differences. Additionally, if we consider the Visualization Model and Task variables together we see no statistically meaningful differences. By looking at these results, we could conclude that there is no superior model for answering all the questions. We also analysed the perceived difficulty of each Visualization Model and concluded that there is a small statistical difference between the models (p-value= 0.047). The p-value is very close to the significance level, meaning that to confirm this significance we would need to perform more tests.

For the five final questions of the test, we used a $\chi - squared$ for categorical variables. We found that there are statistically meaningful differences in all the questions (p-value= 0.000). Figure 7.11 illustrates the distribution of votes per question. As one can perceive, the majority of the participants found the radial visualization model more aesthetically pleasing [**Q1**] and it was the model which aroused more curiosity to explore the data [**Q2**]. The fact that the regular calendar was assigned to the most intuitive [**Q5**], is related to the fact that this structure is more familiar, as we are accustomed to interacting with this type of calendars daily. This fact also explains the tendency to choose the regular calendar as the most useful and easier to learn [**Q3**][**Q4**]. However, there was a higher number of participants who chose both visualizations in [**Q3**][**Q4**], revealing their uncertainty from which to choose.

In the comments section, most participants stated that, after some interaction, the radial visualization was easier to read. A participant also mentioned that the radial visualization condensed better the information and enabled the extraction of more insights. Although all participants stated that there is an unavoidable familiarity with the regular calendar, the radial calendar aroused more curiosity in them. One participant also stated that once learned, it was impossible not to prefer the radial visualization. Some participants also gave particular insights: one stated that the doughnut visualization of the differences was very intuitive and another participant referred to the regular visualization as a better alternative when comparing specific days of the month. To facilitate this last task, a user mentioned that interaction would improve the comparison of different days in the radial calendar.

Based on the above, we can conclude that the regular model is more intuitive at first, but as the user learns to navigate and read the radial calendar, it can become a better choice to represent the consumption data.

7.6 Discussion

To accomplish a more detailed analysis of the Portuguese consumption, we developed this second approach in which all consumption values of the dataset are aggregated by day and represented in a radial structure so the analyst can have a better overview of the data [**T1**]. Additionally, we aligned each week of the month, enabling the analyst to compare them faster [**T2**]. In our visualization model, we defined a set of views so the analyst can visualise: (i) the total consumption value in each day by product category [**T1**]; (ii) the deviation between the consumption value in a certain date and the "normal" consumption value for that day of the week [**T3, T4**]; (iii) the weekly consumption value by month [**T2**]; (iv) the weekly average consumption values for each month [**T5**]; and, (v) the difference between days of the same calendar position [**T1**]. Furthermore, the analyst can navigate through the product categories and analyse them individually, or compare the mean week of consumption of all product categories within the same level in one place [**T6**].

With the different views of our visualization, the analysts can explore and have a better understanding of the Portuguese consumption patterns through the years. With the disposition of the consumption values over our radial calendar, the analyst can understand the weekly patterns of consumption, both at a low level (how the consumption values distribute during the different days of the week) and at a high level (how the consumption values distribute in every week of the month). With the pre-computation of the mean consumption and deviations in all years, it is possible to have a less cluttered view of the data, enabling a more precise understanding of the values.

In the *Average Year* view, the consumption values that repeat over the years are emphasised, and the consumption behaviours that do not repeat themselves are toned down. With the *Deviation Values* view, the analyst can understand which month has higher deviations, meaning it has consumption values that deviate significantly from the normal consumption on specific days of the week, and which may require a further investigation on the causes of that disruption. Finally, with the visualization of the differences between the maximum and minimum values on the dataset, the analyst can better understand the variance between the years and detect the impact of promotional events.

After the analysis of the application of our model, we can point out its benefits and limitations, and compare the different views of our approach. Concerning the display of the consumption values

versus the deviations, we concluded that: (i) the deviations can be an aid in the identification of outliers; and, (ii) the visualization of the consumption values by day gives a good overview of the data as well as a detailed one. Nevertheless, the analysts would benefit if they could: (i) choose the way how the normal consumption value is calculated; and, (ii) change from a weekly mean to the mean of the month or year. This would improve the understanding of the deviations and facilitate understanding when a certain consumption value is higher or lower than the selected threshold. The implementation of the *Average Year* view is also an important feature since it enables a better understanding of the yearly consumption behaviour, enhancing behaviours that occur in the same period.

We also perceived that some festivities do not affect the consumption values in general. Most of the represented festivity periods are local and do not substantially affect the overall consumption in Portugal. The main exceptions are Christmas and New Year's Eve. Nonetheless, we think that the representation of the festivities is a good feature for the project, and, when representing lower levels of the product hierarchy, it may be more noticeable the impact of these festivities over the consumption values.

8

Conclusions

In this part of the thesis, we explored linear and radial calendar structures to support the analysis of time-series and the detection of periodic behaviours. We presented our design choices for the visualization models as well as for the definition of the graphic interface and interaction functionalities. These models were developed in the context of a partnership with SONAE—a Portuguese Retail Company—and were intended to be integrated into the workflow of the company's analysts. For the visualizations, the analysts' main aim was to promote an efficient analysis of the consumption variations over time. To test our models, we used a dataset with the consumption values within the SONAE's product hierarchy from May 2012 to April 2014.

In the linear calendar, we represented the deviations from what we characterised as the normal week of consumption. This decision was made due to the weekly and intrinsic repetition of similar consumption patterns that hid important deviations. By reducing this intrinsic behaviour we could highlight the real deviations and variations of consumption. With this model, the same days of the week (e.g., Monday, Thursday) are horizontally aligned and each week of the month is positioned from left to right. We explored two approaches for the representation of the daily deviations in which two scales were defined: a linear and a range scale. The latter improved the former by better highlighting both positive and negative deviations. By representing the deviations, the analysts could perceive faster how the consumption differed among the different days. However, we needed to represent the two years data in the same place, and to position the two years side by side occupied too much space, reducing the space available to represent the daily deviations. Also, the graphical interface lacked functionalities to allow the analysts to perform a more detailed analysis of the data.

The radial calendar web-application improved the previous model by allowing a more detailed analysis of the data and by using more

efficiently the canvas space. In this new structure, both years are represented simultaneously and the analysts have at their disposal a set of different views over the same data. In this calendar, we structure the months along the radius and align the different weeks of the month on the rings. Also, we represent the deviation values, the consumption values, the difference between both years, and the average consumption in each month. Additionally, we enable the analyst to visualise and compare the product categories at the same level. Overall, with this model, the analysts can have a more detailed analysis of the data. In our user study, we tested the ability of the radial model to represent the data. The results showed that, although it requires some learning and adaptation times, it is more intriguing than the regular calendar, and that it is easier for the analyst to explore the data and get insights from the model.

The presented work demonstrates how the calendar structure and the preprocessing of the data can be used to expand the analysis of time-oriented data. We could assess that, due to the familiarity to such structures, calendars facilitate the reading of the visualization values. Whereas the linear calendar was seen as more intuitive, the radial calendar aroused more curiosity in the users, improving their engagement and consequent exploration of the data. However, due to its novelty, the radial calendar initially required more time to be read. We could also perceive that these structures enable the use of more complex visual marks. However, due to the reduced space for each mark, the marks should not have too many details as they can hinder the comparison of values. To overcome this issue, interaction techniques that highlight similar values or show the respective values in each mark can be used. Although the presented visualization models were developed for a specific dataset and to comply with a previously defined set of tasks, we believe that they can be used with other time-oriented datasets.

We were able to establish that representing the deviations in reflection to the normal consumption highlights the disruptions in consumption values, and emphasises atypical values in cyclical data. With this approach, we were able to aid the analysts to detect periods of time in which the stocks should be reinforced, enabling a better understanding of how to manage their supplies. In sum, our main contributions include: (i) the characterisation of the Portuguese's consumption behaviours; (ii) the identification of patterns and periodic behaviours throughout the years with two adapted calendar structures; (iii) the highlight of weekly behaviours; and, (iv) a user study that compares a radial calendar layout to a regular one.

Part III

PATTERNS

9

Detecting Temporal Patterns of Fraud

"graphics should not simplify messages. They should clarify them, highlight trends, uncover patterns, and reveal realities not visible before."
— Alberto Cairo [43]

This part of the thesis concerns our investigation in the analysis and detection of fraudulent activities in finance. To enable the efficient and effective analysis of fraud, we developed two visualization models that simplify complex data structures and emphasise visually fraud patterns. These models resulted in two visualization tools that have two distinct goals: (i) the analysis of bank transactions over time; and, (ii) the detection of a specific fraud pattern—Account Takeover (ATO).

In this Chapter, we introduce the context and problems we aim to solve and the work related to fraud visualization and detection in finance. In the following Chapters, we present two visualization tools for the analysis of financial data and discuss the design rationale, the results of two validation methods, i.e., usage scenarios and user testing (Chapters 10 and 11), and our findings (Chapter 12).

9.1 Context

The analysis and detection of fraud is an important challenge that should be addressed with care [311]. Fraud can be defined as "*an uncommon, well-considered, time-evolving, carefully organised, and imperceptibly concealed crime*" that can affect singular people and large institutions from different domains [28, 166, 289, 311]. The analysis of financial data and the detection of fraud may prevent and suppress possible future losses for institutions and their clients, and for this reason, both tasks are of high importance. The management of fraud usually focuses on three main pillars: detection, prevention, and response [104]. Fraud detection implies a continuous monitoring system that measures and evaluates possible fraudulent activities. Fraud prevention are preventive methods that create barriers to fraud, discouraging fraudulent activities. Finally, fraud response can be

defined as a set of protocols that should be applied when fraud is detected [104]. In this part of the thesis, we focus on the first pillar and develop two visualization tools to enhance the analysis and detection of fraud.

Nowadays, experts in charge of fraud management base their analysis on tabular data, usually presented in the form of a spreadsheet and seldom supplemented with visualizations. With those methods, the inspection of irregularities and suspicious behaviours can be laborious, time-consuming, and arduous. Regarding the data to be analysed, it can be in a raw state or be the result of a previous analysis by Machine Learning (ML) systems trained to detect fraudulent behaviours. In both cases, the experts' current tools may be of little use for the analysis and overview of such complex data. Additionally, as technology evolves and the techniques applied to detect fraud become publicly available, fraudsters adapt and modify their ways of acting, making fraud an adversarial domain [28]. This may prevent existing ML models from correctly detecting all fraudulent transactions, and may lead to incorrect classifications. As such, investing only in ML algorithms for the detection of fraud may lead to undetected fraud cases.

To tackle the aforementioned problems, especially the lack of tools to analyse both raw data and the results from ML systems, Information visualization can be applied. Through visualization models that emphasise recurrent patterns, it is possible to make the detection of fraud more reliable, effective, and efficient [28, 79]. Also, through the combination of computational means with our visual cognitive intelligence [168] and by enabling the detailed analysis of suspicious behaviours that require careful investigation, visualization can facilitate the analysis of financial data and reveal new undetected fraudulent patterns.

The visualization models developed in this part of the thesis are the result of a partnership with Feedzai[1]. Feedzai is a world-leading company specialised in fraud prevention that owns a risk management platform powered by big data and ML. This platform is used mainly to identify fraudulent payment transactions and minimise risk in the financial industry (e.g., retail merchants and bank institutions). With their platform, Feedzai gives to its clients the possibility to analyse information to keep their customers' data and transactions safe. Also, Feedzai has its own fraud analysts whose goal is to provide a more detailed and humanised analysis of the data and, with the knowledge retrieved from it, improve their ML systems. In more detail, this improvement is achieved by manually analysing suspicious activities and validating the ML's classifications, complementing the ML sys-

[1]Feedzai (https://feedzai.com) is the market leader in fighting financial crime with Artificial Intelligence (AI). One of their main products is an advanced risk management platform.

tem. The collaboration between Humans, in this case, fraud experts, and Artificial Intelligence (AI) systems raises many questions that go beyond the traditional Human-Machine Interaction paradigm. Although an in-depth analysis of these issues is beyond the scope of our thesis, we highlight the need to design systems that take advantage of the complementarity between humans and machines, resulting in "humanised technologies" that surpass the current limitations of humans and machines, while being provably beneficial [200, 246].

We use two distinct datasets with two distinct problems that Feedzai aims to tackle: the analysis and characterisation of bank transactions and the emphasis on fraud patterns in the e-commerce domain. The first contains a set of transactions made by several clients of a specific bank. The second contains a set of e-commerce transactions made by different clients on a retail company's website. In both datasets[2], the fraudulent transactions are labelled. However, in the first dataset, fraud is labelled by the bank itself, and in the second dataset, fraud is labelled by Feedzai's ML system. In both cases, Feedzai's main goal is to provide to its analysts the visualization tools that enable the proper analysis and characterisation of different subsets of data. These analysts' level of expertise in Information Visualization is reduced and their experience in analysing fraud may vary. During their analysis process, there is a lack of tools to aid them. In fact, they only have at their disposal spreadsheet-style tools and a limited data analysis platform, which makes the analysis of fraud an overwhelming task. In this matter, the creation of visualization tools is intended to hasten their analysis process by giving information about data relations, such as time intervals between events, similarity, and recurring patterns, and to provide an overall sense of scale in the financial time-oriented data.

[2]Note that due to the high sensitivity of the datasets, the datasets were previously anonymised and encrypted, but retained the fraud patterns of the real datasets. This enables us to visually explore the data in real case scenarios, without compromising the users' anonymity.

9.2 Related Work

The majority of fraud prevention companies employ ML to detect patterns of fraud [28, 213, 224, 311], discarding events that are promptly classified as fraud by their systems. However, with the generalised growth of ground knowledge about existent fraud prevention approaches, fraudsters are evolving and adapting their fraud mechanisms to overcome nowadays security systems and perpetuate fraud. For that reason, fraud prevention companies employ fraud analysts to complement the ML system and manually analyse suspicious activities and validate the classifications attributed by the systems. Additionally, although they also adopt several AI-based drift detec-

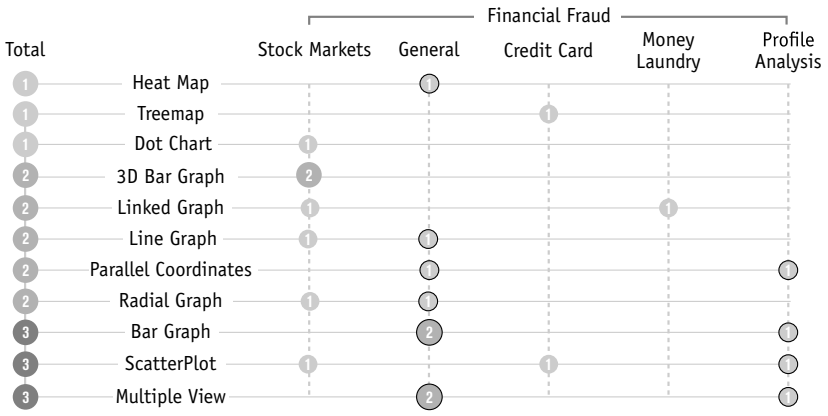

FIGURE 9.1: Distribution of the visualization techniques used for the detection of fraud. The outlined circles represent visualization models applied in multiple-views.

tion techniques, early drift detection and discovery involve human intervention (e.g., even when a new pattern is discovered by the AI model, human validation of the pattern is required).

Due to the size and heterogeneity of financial data, manual analysis can be difficult and time-consuming [155]. Currently, to evaluate financial events most analysts use spreadsheets and tabular forms which support various operations to extract more detailed information. Nevertheless, they are not effective at providing a clear representation of patterns, trends, and correlations hidden in the data [53]. To aid in such tasks, fraud analysts have recognised the relevance of Information visualization, as it enables them to get insights from the data more easily, draw conclusions more rapidly, and, consequently, improve decision-making [79].

The visual exploration of data has already proved its value concerning exploratory data analysis, as the user is directly involved in the exploration process, adjusting its goals along the analysis [143]. Additionally, the use of visual techniques to aid in the detection of fraud has already been explored in multiple domains and multiple surveys can be found in the literature. For example, in the work of Leite et al. [166], which does not focus solely on financial fraud, from the 40 surveyed approaches to detect fraud, 28 combine visualization techniques with other fraud detection mechanisms. The most used visualization techniques in fraud detection, regardless of the domain of application, are line and bar charts, and node-link diagrams. These techniques are used to represent changes over time, facilitate the comparison of categorical values, and represent networks and relationships, respectively. Focusing only on the financial domain, two surveys present a smaller set of projects, which apply techniques, such as parallel coordinate plots, scatterplots, and bar and line charts [85, 155]. In Figure 9.1, we summarise the most used visualization models from our research. For a more detailed description of the techniques

and the different taxonomies, please refer to the works [155, 166].

In financial fraud, most published papers focus on the description of statistical and data mining approaches [28, 213, 224, 311] and only 8 apply visualization techniques. For the representation of specific financial fraud patterns, six works can be found concerning the visualization of (i) *Stock Market Fraud*, which focuses on the analysis of abnormal changes in stock market values along time [131, 154]; (ii) *Profile Analysis*, which focuses on the analysis of personal bank transactions [167]; (iii) *Credit Card Fraud*, which focuses on the analysis of improper use of credit cards [248]; and, (iv) *Money Laundering*, which focuses on the analysis of the network of transactions [76, 77]. From these, four projects [76, 77, 167, 248] focus merely on the improvement of the respective automatic evaluation systems, not applying visualization for the manual analysis of fraud cases. Also, in the work of Sakoda et al. [248], they visualise directly the fraud labels given by the ML system, not giving further details of each transaction to enhance its analysis. Finally, from this subset, most tools apply more than one visualization technique in separate or multiple views.

From our research, we only found one visualization model related to the visualization of bank data. Wire Viz [53], is a coordinated visualization tool that aims to identify specific keywords within a set of transactions. Also, they apply different views to depict relationships over time. For example, they use a keyword network to represent relationships among keywords, a heatmap to show the relationships among accounts and keywords, and a time-series line chart to represent the transactions over time. Their goals are to give an overview of the data, provide the ability to aggregate and organise groups of transactions, and compare individual records [53].

With this research, we could conclude that the analysis of fraudulent activities through visualization is gaining popularity, but its use to detect specific types of fraud is uncommon. In the case of bank transactions, the only related work [53] uses a different type of dataset, which contains transactions to and from other banks, whereas, in the dataset made available by Feedzai, we only have access to the transactions made from the accounts of a specific bank. With this dataset, we are not able to follow the connections between different transactions, being our main aim to characterise the transactions of specific clients that may be referred to as suspicious cases. We argue that to properly understand the behaviours of a certain client, a more detailed analysis of their patterns of transactions must be conducted, so it is possible to distinguish atypical and suspicious transactions from common transactions.

Concerning the visualization of fraud in e-commerce, we argue that by focusing on the representation of a specific fraud pattern, it is possible to ease and reduce the time needed to detect fraud. From our research, we found no visualization tool specific for the detection of ATO or Bot Attack (BA) patterns, which are the most common patterns in e-commerce. Also, the majority of the analysed visualization tools are intended to be highly interactive, requiring a high amount of time to analyse the data, or are only being applied to improve fraud detection rules of automatic systems [167, 248]. We consider that, although visualization can be used for more time consuming and detailed analysis, the fraud analysts can also benefit if visualization is used for the quick identification of fraud. In this case, by diminishing the time of analysis, the time needed to take action and stop transactions from being approved is consequently reduced, improving the efficiency of the company. For this reason, we argue that the visualization models should enable the analyst to further explore the details in a more explorative manner, but at the same time, be able to give the overall patterns within the data at a first glance.

10

Bank transactions

The analysis of financial transactions and the characterisation of the overall behaviour of individual bank clients can be an overwhelming task, especially given that the analysts have at their disposal limited spreadsheet-style tools, which are complex, time-consuming, and tend to be inappropriate for complex analysis [53]. However, this task is of high importance as it may lead to the detection of suspicious behaviours which, in turn, may lead to the finding of fraudulent activities and enable the banks to take action. Also, this knowledge can then be used to improve ML systems through the specification of new behaviours that indicate fraud. To solve this problem, we developed *VaBank*, a visualization tool that aims to facilitate the analysis of bank transactions.

With VaBank, we aim to: ease the analysis of the distribution of bank transactions over time; the detection of the main characteristics and topology of those transactions; and, as such, aid in the detection of suspicious behaviours. These objectives were based on the goals of Feedzai's fraud analysts, which can be summarised as: (i) be able to inspect collections of transactions in a single visualization—usually, these collections are grouped by attributes, such as client ID or location IP; (ii) understand the overall behaviours indicated by a set of transactions; and, (iii) detect the most common types of transactions.

Our tool was implemented in Java and used Processing to render the visualization. A video of the tool can be accessed through: `https://cdv.dei.uc.pt/radial-calendar/`. VaBank is divided into two main areas, the visualization of the transactions' distribution over time and the visualization of the topology of the transactions. In the latter, we apply a Self-Organising Maps (SOM) algorithm to: represent the topology of a subset of transactions; and, enable the detection of the most common type of transactions thus characterising the client main behaviours. SOM have already proved their usefulness and robustness

for the analysis of large amounts of data [87]. The visualization of their results provide a visual summary of the data topology and can ease the interpretation of behaviours in a single image [70, 151]. We present the SOM's results through two visualization techniques: a matrix grid and force-directed projection. Both aim to represent the profiling of a group of transactions and enable the understanding of the characteristics of the most common transactions. Finally, the transaction's history visualization provides a set of analytical features, enabling the analyst to navigate, explore, and analyse the sequence of transactions over time.

Our main contributions are: (i) a user-centred visual tool, developed with the aid of fraud experts; (ii) a method that characterises the topology of the transaction through a SOM algorithm; (iii) the visual characterisation of transactions through complex glyphs; and, (iv) the assessment of the tool effectiveness through a usage scenario and a user study. Based on the analysts' feedback, we conclude that our tool can improve substantially their line of work which currently involves the time-consuming analysis of spreadsheets.

10.1 Self-Organising Maps

SOM take advantage of artificial neural networks to map high-dimensional data onto a discretised low-dimensional grid [156]. Therefore, SOM is a method for dimensionality reduction that preserves topological and metric relationships of the input data. SOMs are a powerful tool for communicating complex, nonlinear relationships among high-dimensional data through simple graphical representations. Although there are multiple variants, the traditional SOM passes through different stages that affect the state of the network [156]. In the first, all neurons are initialised with random values. Then, for each datum of the training data input, the so-called Best Matching Unit (BMU) is defined. This is done by computing Euclidean distances to all the neurons and choosing the closest one. Finally, the weights of the BMU and the neighbour neurons are adjusted towards the input data, according to a Gaussian function—which shrinks with time. This process is then repeated for each input vector for a predefined number of cycles.

Since the present work deals with mixed data, we present SOM algorithms that work with that type of data. The topological self-organising algorithm for analysing mixed variables was proposed by Rogovschi et al. [242]. In this approach, categorical data is encoded into binary variables. The algorithm uses variable weights to

adjust the relevance of each feature in the data. Hsu et al. [128, 129] proposed another method in which they use semantics between attributes to encode the distance hierarchy measure for categorical data. Similarly, Tai and Hsu [270] use semantic similarity inherent to the categorical data to describe distance hierarchy by a value representation scheme. Hsu [63] used distance hierarchies to unify categorical and numerical values, and measure the distances in those hierarchies. Finally, del Coso et al. [320] a frequency-based distance measure was used for categorical data and a traditional Euclidean distance for continuous values.

10.1.1 *Visualization of Self-organising Maps*

The visualization of SOMs is typically concerned with the projection of neurons into a 2D/3D grid. The most common projection is the Unified Distance Matrix (U-Matrix), in which neurons are placed in a grid and the Euclidean distances between neighbouring neurons are represented through a greyscale colour palette. This visual mapping can be used in the detection of clusters [157, 256] or in the definition of thresholds [220]. Additionally, hexagonal grids [199] can also be used [12], increasing neighbourhood relations, although not always resulting in more detailed insights [12]. The results of SOMs have also been used as data inputs for other visualization models. In most cases, researchers used SOMs to define clusters or characterise different behaviours and then represent such groups in the visualization models. Gorricha and Lobo [111] used a 3D SOM to define clusters categorised visually with colour, which later is applied in geographic areas with different characteristics. Morais et al. [208] used a SOM to define clusters in data, and then those clusters were represented through various visualization models, such as parallel coordinates and Chernoff faces. In fact, the use of Chernoff faces and glyphs, in general, characterises multiple works, which will be discussed later. Finally, Adrienko et al. [7] visualises the clusters resulting from the SOM algorithm through a two views visualization, consisting of the representation of the clusters on a map and in a temporal grid.

To improve the reading and understanding of each neuron, some works used complex glyphs. Furletti et al. [108] represented the neurons through a timeline, portraying the temporal profile of call logs, and, in the background, a circle is drawn with the size depending on the number of elements used to train each neuron. Schreck et al. [251] represented each neuron with a squared glyph coloured according to the quantisation error, and, inside each square, a line is drawn to represent a certain trajectory. Kameoka et al. [140] repre-

sented the neurons with a radar glyph which shows the consumption value of a specific product. Finally, Wehrens and Buydens [297] used a rose diagram to represent the weights of each feature of the SOM.

10.1.2 *Self-organising Maps in Finance*

Several projects have applied SOM algorithms to analyse transactional data. The majority of them uses SOMs to provide an analytical view on the financial market trajectories [250, 251, 252] and to analyse their stability and monitor multi-dimensional financial data [249]. Other works use SOM to better understand the stock market dynamics[263] or to analyse the financial performance of companies [87].

10.2 Data Analysis and Preprocessing

We worked with an anonymised dataset that contains only the transactions generated by the clients of a certain bank—there is no data about the bank transfers that each client received (in other words, if someone transfers money to the client's account this will not be visible, as such the only transactions will be the ones made by the client). Each transaction of the dataset is characterised by attributes corresponding to the: client (e.g., ID, IBAN), location (e.g.,Client IP, Country IP), monetary amount (e.g., amount, currency), transaction (e.g., type, descriptor, fraud label), beneficiary details (e.g., IBAN), and date. Each transaction can be of two types: online, corresponding to regular transactions; and business, corresponding to business transactions. All clients can have transactions of both types. The transactions also have a descriptor, composed of two or three acronyms, that characterise the transaction according to: (i) the interface used; (ii) the type of operation (e.g., national, international, loan); and, (iii) whether it is for a new beneficiary or not. These characteristics must be known by the analysts, so they can properly analyse the transactions. This task has a high level of difficulty as these descriptors can have different combinations. To enable a better understanding of the descriptor elements, we herein list them according to each type of transaction (Business or Online):

Business Transactions:

- Type of Interface: ATM Specific, Telephone, ATM, and Branch
- Type of Operations: Cash In and National
- Type of Beneficiary: New and Old.

Online Transactions:

- Type of Interface: Barc. Mobile, MBWay, Web, and App
- Type of Operations: Instant, International, National, Loan, Address change, and Agenda
- Type of Beneficiary: New and Old

Additionally, all transactions are labelled by the bank as fraudulent or not. For this project, we group dynamically the transactions of a certain subset in different range scales: time and monetary amount. These two ranges are then plotted into the x- and y-axis, respectively. Also, to be able to properly summarise the data, for each pair $[time, amount]$ we aggregate the transactions with the same characteristics (i.e., the same values for the attributes type, descriptor, and fraud label).

10.2.1 SOM *Algorithm*

We applied a variant of the Frequency Neuron Mixed Self-Organising Map (FMSOM), a SOM algorithm prepared to handle mixed data [320]. Our algorithm uses the SOM for handling the numerical variables and extends the neuron prototype with a set of category frequency vectors. The algorithm follows the traditional *competition, cooperation* and *adaptation* process. Since we focus on the visualization tier of the SOM and not on the algorithm, any other method could be used. However, the FMSOM model allowed us to adapt our SOM algorithm to define the dissimilarity between neurons, used in the visualization of the transaction's topology.

Features

First, we extracted the features for each input raw data. In our project, 7 features and their types were identified: *amount*, *day* of week, *month* of the year, *year*, *time passed* since the last and until the next transactions (in milliseconds), *fraud*, *transaction type*, *operation type*, *beneficiary*, and *interface channel*. The later five features were briefly described in Section 10.2 and cannot be fully revealed due to the specificity and sensitivity of the dataset. The *amount* is the amount of money involved in the transaction. From the date of a transaction, we extract the day of week $[1-7]$, the month of the year $[1-12]$, and the year. The features *time passed* since the last transaction and until the next transaction are previously calculated and are intended to capture the patterns of the transactional regularity.

Dissimilarity Metric

We applied different measures to compute the distances between neurons. We applied the traditional Euclidean distance for continuous values, and a measure based on probabilities (described in [320]) for categorical features. Ultimately, two types of dissimilarity measures were defined: one for the training of the SOM; another for the visualization.

Regarding the SOM, as in FMSOM [320], the dissimilarity measure between the neuron and the input feature vector consists of the following. Suppose that P is the number of input feature vectors $X_p = [x_{p1}, ..., x_{pF}]$, where F is the number of features in that vector. Also, suppose that n and k are the number of continuous and categorical features, respectively, where $[a_k^1, ..., a_k^r]$ is the set of categories of the k_{th} feature. Finally, suppose that the reference vector of the i_{th} neuron is $W_i = [W_{i1}, ..., W_{in}, W_{in+1}, ..., W_{iK}]$, where I is the number of the neurons in the network. The dissimilarity between an input vector and the reference vector of a neuron is defined as the sum of the numerical and categorical parts. The numerical part is calculated using Euclidean distance on normalised values. For the categorical dissimilarity measure the sum of the partial dissimilarities is calculated, i.e., the dissimilarity is measured as the probability of the reference vector not containing the category in the input vector. For more details on the FMSOM algorithm consult [320].

Regarding the visualization domain, the dissimilarity measure between two neurons is determined as follows. For the numerical part, the traditional Euclidean distance is applied $Dn(W_i, W_j) = \sqrt{\sum_{z=1}^{n}(W_{iz} - W_{jz})^2}$. For the categorical features the dissimilarity measure was defined as the Euclidean distance between the probabilities for each of the categories present in the reference vector $Dk(W_i, W_j) = \sqrt{\sum_{z=n}^{k}\sum_{m=1}^{r}(W_{iz}[a^m] - W_{jz}[a^m])}$. So, the final dissimilarity measure is given by $d(W_i, W_j) = Dn(W_i, W_j) + Dk(W_i, W_j)$.

10.3 Tasks and Requirements

Given our collaboration with the fraud detection company, we were able to hold several meetings with their analysts, which aided us to better define the domain and requirements for the analysis of the bank data. The analysts emphasised two main tasks: **[T1]** understand the transaction history, and **[T2]** detect the most common types of transactions. The latter is especially important as it enables the distinction between typical and atypical behaviours. Then, the analysts described their line of work, referring that their analysis tends to start

by grouping the data by a specific attribute. Usually, the analysts aggregate the transactions by a specific client ID to better analyse and characterise the client's transactions. Then, they search for groups of transactions with similar characteristics, especially the ones labelled as fraud. This task is particularly difficult using a spreadsheet, since when the transactions are not ordered, common attributes will not stand out. Of all attributes, the analysts referred to the amount spent, type of transaction, and fraud label, as the most relevant. They referred to the importance of detecting similar transactions and identifying the profile of the client or a subset of transactions. Based on the meetings, the analysts defined five requirements to which VaBank should comply:

R1 **Search by field.** The analysts usually sort the data by a certain field, such as client IBAN, client ID, or Country of IP, and analyse the transactions with common values on those fields. The creation of a mechanism that enables the analyst to easily select a field and choose a certain value to group the transactions is of utmost importance. This will speed up the analysis process and ease the analysis of all transactions with common values;

R2 **Distinguish amount values.** When dealing with bank transactions, the transacted amount can be a sign of fraudulent activity. Transactions with high amounts, or above a certain threshold are worthy of a more detailed analysis. The visual sorting of the transactions by their amounts can enhance the detection of suspicious transactions;

R3 **Distinguish transactions.** Visually characterising each transaction may help the analysts to distinguish the transactions and focus their attention on the ones of the same type. This may help to perceive the behaviours within the different types of transactions, facilitating the detection of atypical behaviours;

R4 **Search common fields.** When dealing with this data through spreadsheets, the analysts have difficulties in detecting transactions that share more than one attribute. This is of utmost importance when analysing fraudulent transactions which can share attributes with others. For this reason, it is important to implement a mechanism that enables the analyst to select an attribute and highlight all transactions with that same attribute;

R5 **Detect typical transactions.** Understanding the most common types of transactions may enhance the analysis of the data and aid the analyst in the detection of unusual transactions, which

can be related to fraudulent behaviours. Hence, it is important to characterise the space and facilitate the detection of typical transactions in a certain subset of the data.

10.4 VaBank Design

The tool is divided into three views: the *Transactions History*; the *Transactions Topology*; and the *Transactions Relationships*. The first view (Figure 10.1) aims to answer to task **[T1]** and arranges all transactions by time and amount (see Section 10.3). The last two views aim to answer to **[T2]** and display the results of the SOM algorithm (see Section 10.2.1) in a grid and through a force-directed graph, respectively. All three views share a common visual element, the transactions. We developed a glyph that serves to identify the type of transaction and its position in time and range values. We aim to facilitate the distinction between transactions with different characteristics and to provide coherence between views. In the following subsections, we present the design rationale of the glyph and the three views.

10.4.1 *Transaction Glyph*

To ease the distinction and visual characterisation of the transactions, we implemented a glyph, fulfilling **Requirement 3 [R3]** (see Section 10.3). The glyph is composed of three levels of visual detail. These levels were defined together with the company's analysts, according to the relevance of the types of attributes when analysing their bank data. First, the analysts aimed to distinguish online trans-

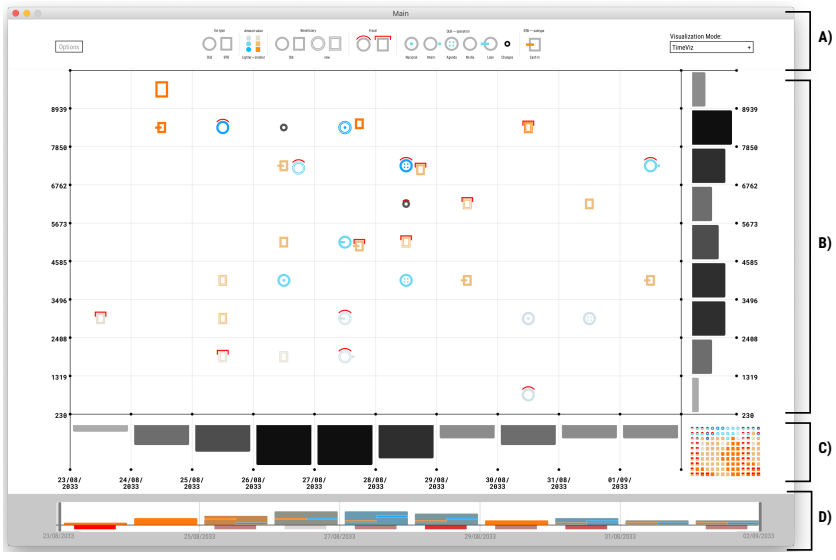

FIGURE 10.1: Transaction History view and its components, from top to bottom: GUI Panel (A); Matrix View and Amount Histogram (B); Time Histogram and Small SOM Matrix (C); and Timeline (D).

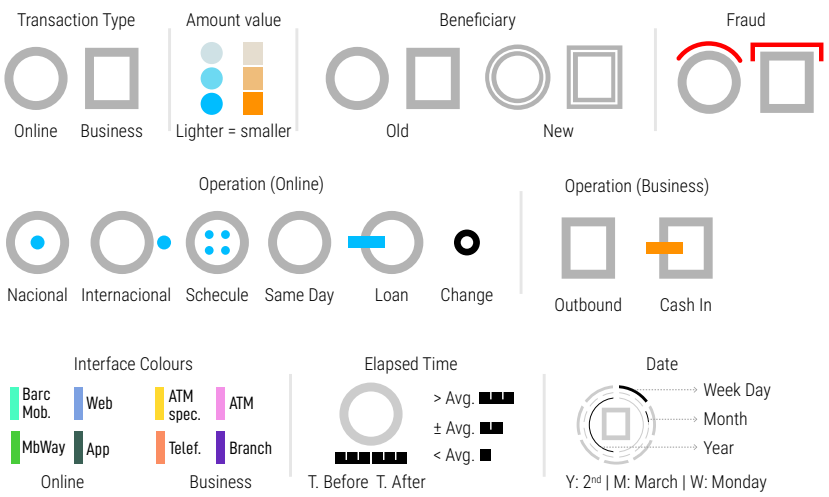

FIGURE 10.2: Glyph elements that characterise each transaction.

actions from business transactions. Then, the transaction amount and whether it was considered as fraud or not are analysed. These three characteristics represent the first level of visual impact. Then, the analysts want to drill down and distinguish between: (i) inbound and outbound bank transfers; and, (ii) new and old beneficiaries. These characteristics represent the second level of visual impact. Finally, the time characterisation of each transaction and the interface with which the transaction was made were defined as less important than the characteristics described above. For this reason, they are grouped into the third level of visual impact and should have a lower visual impact.

As colour has a high impact on visualization [191], we apply colour to emphasise the characteristics of the first visual level. We apply different hues to the types of transaction: orange for business; blue for online. Additionally, we use different shapes to emphasise this distinction between transaction types—a rectangle for business; and a circle for online. Then, we use saturation to represent the amount: the brighter the colour, the higher the amount. As small differences in saturation would be imperceptible to the human eye, we defined three levels of saturation to distinguish: low, medium, and high amounts. These levels are computed as follows. We compute the average amount $\overline{x}$, define a window w, and if the value is: below $\overline{x} - w$, we consider the amount as low; between $\overline{x} - w$ and $\overline{x} + w$, the amount is medium; and higher than $\overline{x} + w$, the value is high. The window w is a percentage of the average value that was defined in collaboration with the company's analyst. Finally, to represent a fraudulent transaction, we place a red line above the main shape (see Figure 10.2).

The transactions' shape is complemented with a set of symbols that

represent the types of operation. They are divided according to the directionality of the transaction, outbound or inbound[1]. The inbound is represented by the same symbol in online (i.e., Loan) and business (i.e., Cash In) transactions: a vertically centred horizontal rectangle positioned on the left. The outbound operations are represented as depicted in Figure 10.2, in which the business transaction only has one type, and the online transactions have five. As the new beneficiary characteristic is a binary value, we represent transactions for new beneficiaries by dividing the stroke of the main shape in two. If the beneficiary is not new, no change is made (Figure 10.2).

[1]Note that the inbound are transactions made only by the client, when asking for a loan (in online transactions) or when doing a deposit (in business transactions)

For the third level, we represent the year, month, and day of the week of the transaction. Each time variable is represented by a ring with a different radius centred in the main shape. The year is the smallest ring, and the day of the week the biggest one (Figure 10.2). To distinguish periods of time, we divide the ring into 7 wedges, for the days of the week; 12 wedges for the months; and, for the years, in the total number of years in the dataset. All wedges are coloured in light grey, except the wedge that marks the period of the transaction, coloured in black. The day of the week has a thicker stroke, as the analysts referred it is the most important time variable. We also represent the elapsed time between the current transaction and the previous and following transactions. We apply an equal rationale to represent these two-time distances. Like with the amount thresholds, we defined three levels of time distances that are computed in the same way. These three levels are represented as depicted in Figure 10.2. Note that for the sake of simplicity this data was aggregated, even though in the SOM we use absolute values. Finally, the interface of the transaction is represented by filling the elapsed time's shape with the corresponding interface colour (Figure 10.2).

The glyphs used in the views concerning the SOM's result make use of all representations described above. However, in the Transaction History view, we only represent the first two levels of visual detail, as time is already being represented in the x-axis.

10.4.2 *Transaction History View*

In this view, there is a set of visualization models that display different data aggregations. The main representation, which occupies more canvas space, is the *Transaction Matrix* (Figure 10.3, A). It divides the space in different ranges of monetary values on the y-axis and temporal values on the x-axis **[R2]**. The transactions' glyphs are then distributed by the cells of the matrix, according to their date and amount. If more than one transaction with the same characteristics

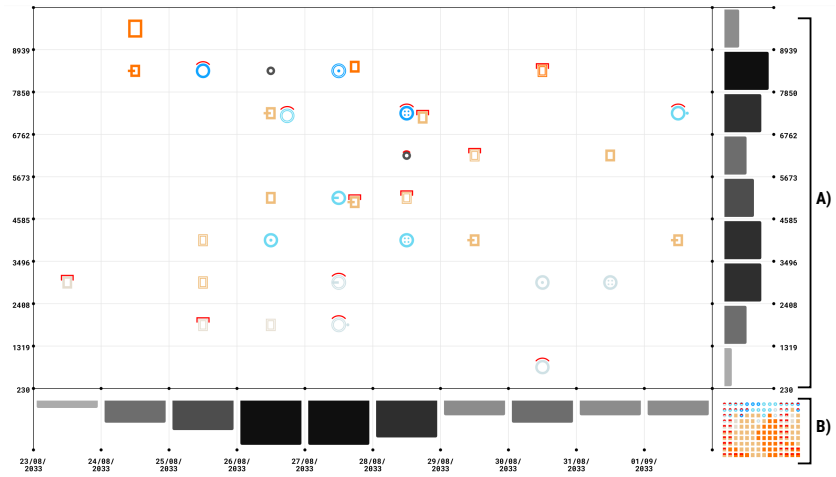

FIGURE 10.3: Zoom of the Transaction Matrix area of VaBank. In this zoomed image we can distinguish two areas: Matrix View and Amount Histogram (A); and, Time Histogram and Small SOM Matrix (B).

FIGURE 10.4: Zoomed image of the timeline area.

(defined in Section 10.4.1) occurs within the same cell, they are aggregated and the glyph grows in size. The placement within each cell is made through a circle packing algorithm which starts by placing the biggest glyph in the middle of the cell and the others around it.

In the bottom and right sides of the Transaction Matrix, histograms are drawn to show the total number of transactions per column and row, respectively (Figure 10.3). The histogram's bars are coloured according to the number of transactions: the darker the bar, the higher the number of transactions. To give a visual hint to the analyst about the distribution of the different transactions, enhancing the understanding of typical/atypical transactions **[R5]**, we draw a small matrix of glyphs. This matrix is positioned in the bottom right corner of the Transaction Matrix area, and it represents the result of our SOM algorithm, concerning three attributes: amount, transaction type, and fraud (Figure 10.3, B).

To allow the analyst to select different periods of time, we placed an interactive timeline at the bottom of the canvas (Figure 10.4). This timeline represents the entire time-span. To represent all data in the timeline, we applied a hierarchical time aggregation algorithm that semantically aggregates the transactions according to the space of the timeline (see Section 10.4.2). The timeline is divided horizontally into equal sections, representing different periods of time with the same duration. Each section of the timeline is vertically divided into two parts.

To represent which type of transaction occurs the most, we represent the number of transactions through a bar, in the upper part of the timeline. To put it briefly, each bar is drawn as follows: (i)

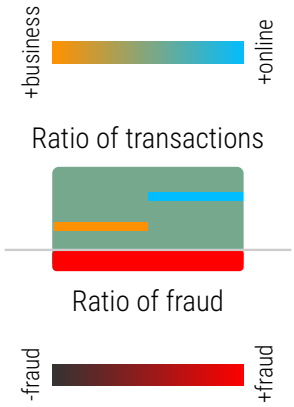

FIGURE 10.5: Timeline bar composition and respective colour ranges.

its height represents the total number of transactions; and, (ii) its main colour is defined by a gradient between blue and orange—the bluer, the higher the number of online transactions, the more orange, the higher the number of business transactions (Figure 10.5). To give a more detailed view, we also represent the quantity of each transaction type with two thin rectangles that are drawn inside the previous bar. They are placed horizontally according to the type of transaction—the business is the one on the left and the online the one on the right—and are placed vertically according to the percentage of occurrence. Additionally, they are coloured according to the transaction type.

In the bottom part of each timeline section, we place a rectangle with a predefined height. This rectangle is only visible if one or more fraudulent transactions occur. Then, it is coloured according to the percentage of fraudulent transactions in that specific period of time. The higher the number of fraudulent transactions, the brighter and redder the bar will be (Figure 10.5). If no fraud occurs, no bar is drawn.

Hierarchical Temporal Aggregation

Fixed timelines can create multiple problems (see for example [54, 219]). For example, different time spans can result in a cluttered timeline, a timeline with an uneven distribution of the time bars (e.g., one bar on the left and the other bars concentrated on the right), or a timeline that uses inefficiently the canvas space, due to the time granularity. With our algorithm, we intend to solve the problem of fixed timelines. The main goal is to allow the representation of any temporal range by adapting automatically the timeline's granularity and the size of the time bars.

Our adaptive timeline algorithm takes as arguments the available space for the timeline and the minimal width of a time bar. The algorithm follows an iterative top-down approach. We start at the biggest time unit existing in the computation systems (e.g., epoch), and descent, iterating over consecutive ISO time units (e.g., years, quarter years, months) until we find an optimal balance between time granularity and time bar's size. The algorithm has to meet one single criterion that is tested at each temporal resolution. Consider that T_i is the time tier currently evaluated, T_{min} and T_{max} are the minimal and maximal timestamp of the selected data subset, W_{min} is the minimum allowed width for the bars, and W_{total} is the width of the timeline. So, the criteria to determine the time resolution and the width of a bar is computed as follows: $W_{total}/T_{i+1}(T_{max} - T_{min}) < W_{min}$.

Note that we compute the width of bars at the $i+1$ temporal tier. If the bar width at the next tier is smaller than W_{min} we stop, and the current tier is the selected one. The left part of the expression is the width of bars, determined by our algorithm.

Interaction

To allow the analysis of the transactions in more detail, we defined a simple set of interaction techniques. In the Transaction Matrix, the analyst can hover each glyph to see more details—Country IP, amount, beneficiary, and the number of transactions. When the analyst clicks on a glyph, these details are fixed in the canvas. By doing so, the analyst can interact with each one of the attributes. If the analyst clicks on an attribute, the transactions which share that same attribute will be highlighted with a black ring. The goal is to enhance the understanding of the transactions that may share the same suspicious attributes **[R4]**. The user can also interact with each bar of the histograms. By hovering a bar, the total number of transactions is shown and the analysts can easily perceive the total number of transactions in a certain period of time (x-axis) or the total number of transactions within a certain range of monetary values (y-axis).

We also defined a set of interaction techniques for the timeline. The analyst can select different periods of time to visualise in the transaction matrix. To do so, the analyst drags two vertical bars which are positioned in the leftmost and rightmost parts of the timeline area. By selecting a shorter period of time, the transaction matrix becomes more detailed (i.e., the different periods of time in the x-axis will have shorter durations; the minimum duration is one day). This way, the analysts can see in more detail the distribution of the transactions over time. Also, the analyst can drag the horizontal line between the bars to maintain the selected duration, but change the initial and final periods of time. Finally, the analyst can hover each bar of the timeline. By doing so, a set of statistics become available. These concern: the total number of transactions in that period of time; the start and end dates; the percentage of online and business transactions; and the percentage of fraud.

10.4.3 *Transaction Topology View*

In this view, we visualise the result of the SOM algorithm defined in Section 10.2.1. To visualise it, we use the positions of the neurons in SOM's matrix to distribute the glyphs on the canvas within a grid with the same number of columns and rows (Figure 10.6, top). Addi-

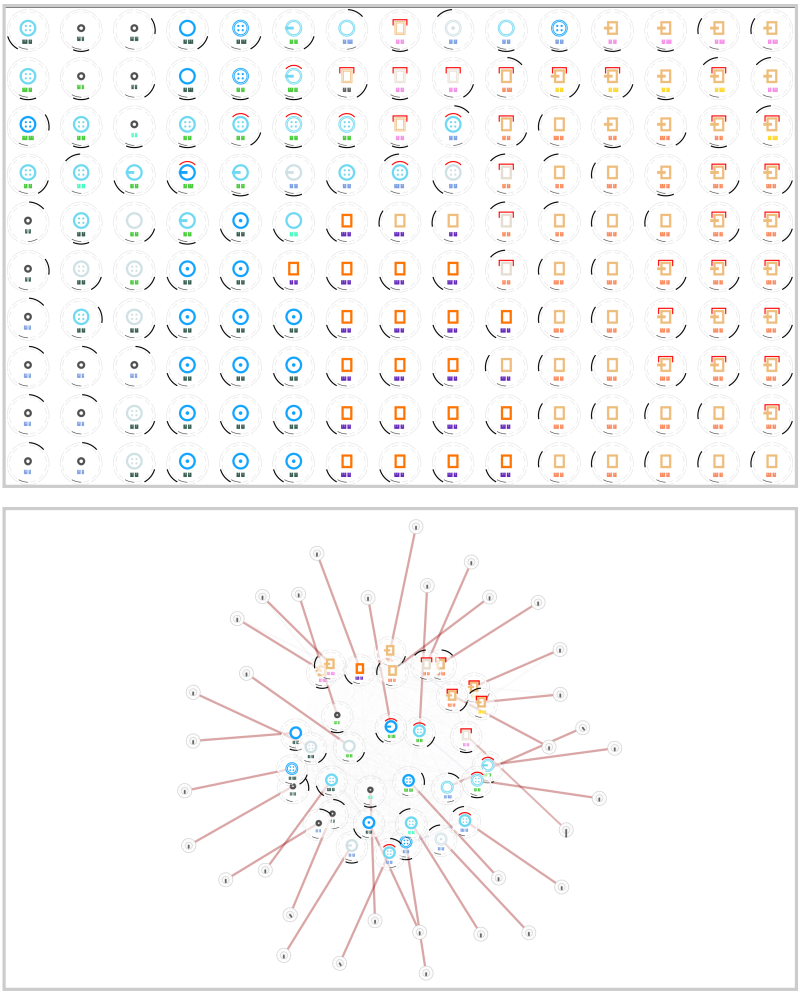

FIGURE 10.6: Projections of the SOM results for the same bank client through the matrix projection (top) and force-directed graph (bottom).

tionally, and as referred previously, we use the three levels of visual detail to represent each neuron (see Section 10.4.1).

This approach allows the analyst to visualise the most common types of transactions through the analysis of the distribution of the different glyphs (representing the transaction's characteristics) in the matrix[2]. However, this view lacks a more detailed representation of the dataset, which could enable, for example, the representation of how many transactions are related to each neuron and which neuron is the most representative of the dataset. The latter task is especially difficult to achieve when more than one feature is being represented in the glyphs, as it can hinder the comparison between glyphs. To overcome this, and to promote a better understanding of the client's profile, we implemented a second approach, in which we place the neurons within a force-directed graph and represent their relationships to the transactions.

[2]Note that, as in the analysis of any SOM, the number of glyphs in the canvas is not representative of the number of transactions within the dataset

10.4.4 *Transaction Relationship View*

For the force-directed graph, neurons and sets of transactions are represented as nodes, and are positioned within the canvas according to their similarity measure: the more similar two neurons are, the closer they will get (Figure 10.6, bottom). Our implementation of the graph layout is based on the *Force Atlas 2* algorithm [134]. All nodes have forces of repulsion towards each other so they do not overlap. However, only nodes whose similarity is above a predefined threshold have forces of attraction. This makes similar nodes to get closer to each other, generating clusters defined by the SOM topology. Additionally, we added a gravitational force that pulls all nodes towards the centre of the canvas. The higher the number of connections between nodes, the higher the gravitational force. Therefore, clusters that are more representative of the dataset will be in the centre of the canvas, and the ones representing atypical transactions in the periphery.

To avoid clutter, only neurons selected as BMU in the training process of the SOM are represented. We opted to filter the neurons with this method, as the neurons that are selected as BMU are the ones that are more similar to the transactions within the dataset, and for this reason, are the ones that are more representative. Also, the transactions which have the same neuron as BMU are aggregated and this aggregation is represented with a node. The force of attraction of these new nodes is defined by their average force of attraction to other neurons.

The nodes have distinct representations. The neurons are represented with the glyphs described in Section 10.4.1. For the groups of transactions, we use a circular chart that represents the number of transactions by month of occurrence. This representation is intentionally simpler since our main goal is to give more visual impact to the result of the SOM. Also, if these nodes are connected to a certain neuron, it means they share similar characteristics, being redundant to use the glyphs approach.

We used lines to connect the nodes. These lines are coloured: (i) in red if they connect a node representing a group of transactions and their BMU neuron; (ii) in light grey, if they connect a group of transactions and other neurons which are also similar to them, but are not their BMU; and, (iii) in blue, if they connect two similar neurons. These lines are represented to enhance the comprehension of the nodes' proximity, but as they should have less visual emphasis, their opacity and thickness diminish according to the similarity values.

FIGURE 10.7: Zoomed image of the Control Panel area, available on the top of VaBank tool.

10.5 Control Panel

To enable a better transition between views we created a *Control Panel* (Figure 10.7). By clicking on the "Options" button, on the upper left corner, the *Options Panel* is shown, containing a list of all unique attributes of a predefined field—client ID. This list is scrollable and is sorted in an ascending way, according to the number of transactions of each client. Also, each row contains a set of statistics concerning the grouped transactions: the total number of transactions, the maximum, minimum, and average amount values, and the percentage of fraudulent transactions. On the *Options Panel*, the analyst can also access a list of fields and select a different one to group the transactions **[R1]**. On the upper right corner of the *Control Panel*, there is a dropdown that enables the analyst to change between the three views. Finally, in the middle of the *Control Panel*, a caption is shown to describe the glyphs that represent the transaction's characteristics (Figure 10.7). This caption is especially important due to the complexity of the glyphs: with it, the analyst can easily read the glyph without needing to memorise or search for the caption anywhere else.

10.6 Usage Scenario

In this Section, we discuss three usage scenarios in which we analyse subsets of the dataset with fraudulent transactions. These subsets were selected according to their complexity over the entire dataset. We aim to highlight the efficiency and effectiveness of VaBank in enabling a detailed analysis of the data. In each scenario, we visualise the transactions made by a certain bank client during one month[3]. Due to the limited time range, all scenarios present a reduced number of transactions. However, we argue that this is not a limitation as our model is prepared to aggregate the data in different time ranges, and for that reason, the model would accommodate a larger dataset with ease. Notice that the number of transactions per period of time would not change significantly, meaning that wider time spans would only result in bigger time ranges in the timeline. Nonetheless, with our timeline, the user can select smaller periods of time which reduces the time span to be represented in the Transaction Matrix, enabling a more detailed analysis of each transaction.

[3] this small temporal range is due to the limited accessibility to the data

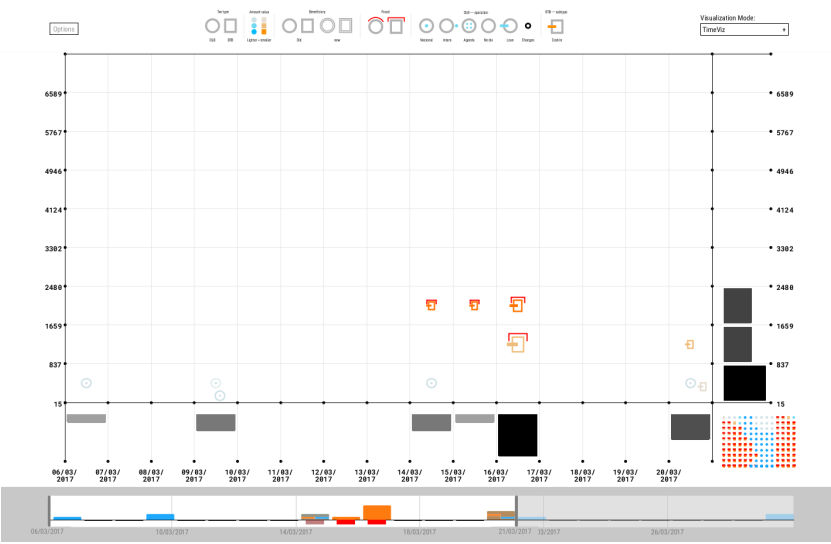

FIGURE 10.8: Transaction History View of client A. With this first view, it is possible to detect a group of business transactions.

► CLIENT A: In Figure 10.8, we can see the *Transactions History View* of *Client A*. We can instantly perceive, through the positioning of the first transactions, that the monetary value of those transactions is relatively low—in relation to the other ranges of values visible in the y-axis. Also, by looking at the bar chart in the timeline, it is possible to understand that the transactions tend to occur periodically—there is an initial set of transactions, then no transactions are made in the following three days, then another set of transactions occurs, and so on. On March 14, there was a business transaction with the highest amount values that was marked by the bank as fraud. We can also see that Client A tried, consecutively, to make that type of transaction on the two following days with similar and smaller values, but got the same result, a fraud label by the bank. All of these transactions are Cash In operations, which means that the client attempted to add money to his/her account. Later, on March 2nd, we can see the same type of transactions with smaller values, however, this time they were not labelled as fraud. By looking at the small matrix—generated from the SOM—it is possible to see that the majority of the business transactions were considered fraudulent, especially the ones with high values.

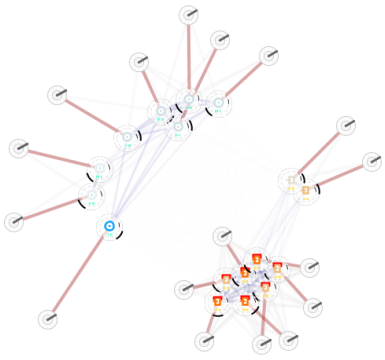

FIGURE 10.9: Transaction Relationship View of client A. Two major clusters can be seen, separating online from business transactions. Also, the business cluster is subdivided into fraudulent and non-fraudulent transactions.

When analysing the *Transactions Topology View* (Figure 10.10), we can verify the assumptions made previously and see that for the business transactions, Client A used mainly the ATM interface (yellow) and for the online transactions the interface used was the Barc. Mobile. Furthermore, we can see that the majority of the online transactions were of the national type and for new beneficiaries. Finally, by checking the *Transactions Relationship View*, we can see these clear distinctions between online and business transactions

FIGURE 10.10: Transaction Topology View of client A. Through this visualization it is possible to perceive that the majority of the business transactions with higher values are considered as fraudulent.

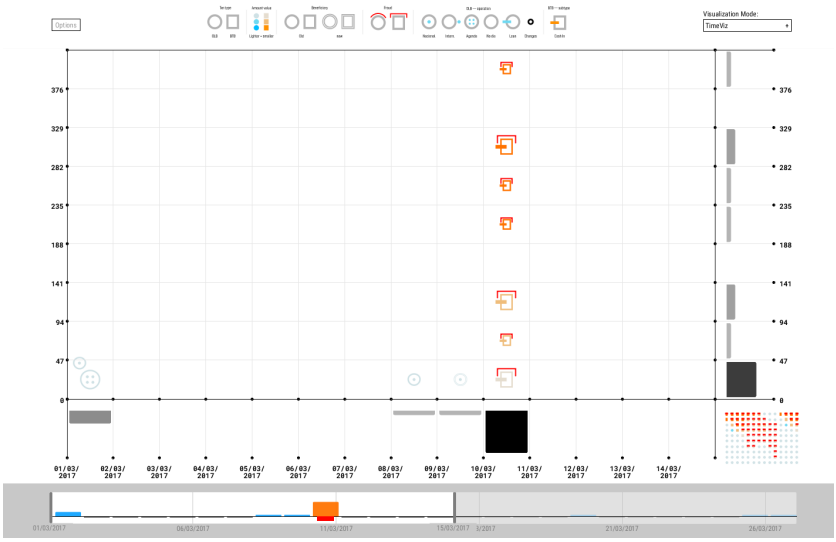

FIGURE 10.11: Transaction History View of client B. Similarly to client A, client B performs a set of fraudulent business transactions. This occurs on the same day with different amount ranges.

(Figure 10.9). In the cluster of business transactions, we can easily define two sub-clusters: the fraudulent transactions with high values and the ones with smaller values. Also, it is possible to see that the business transactions with low values were made on Tuesday, whereas the fraudulent ones occurred between Wednesday and Friday. For all these reasons, this client's actions are seen as suspicious.

- CLIENT B: In the second usage scenario, there were also visible fraudulent transactions (Figure 10.11). Although the values are low, in comparison to the previous client, this client attempted several transactions of the business type with different amounts and aimed to add money to the account. When comparing these business transactions with the rest of Client B transactions, which are usually placed below the €50 limit, the business transactions are of high value. Through the timeline, we can see that this client, after the peak of

FIGURE 10.12: Transaction Topology View of client B. The majority of the neurons are of the business type and are considered to be fraudulent.

business transactions on March 10, made far fewer transactions. By looking at the small SOM matrix, we can see that this client behaves similarly to the previous one: the majority of the business transactions are considered fraudulent, and the online transactions are all of small value.

All the previous assertions can be verified with the *Transactions Topology View* (Figure 10.12). The business transactions are, in the majority of the cases, considered as fraud, and the online transactions have, in their majority, smaller value ranges—in comparison to the business transactions. Additionally, online transactions are divided into two subtypes: the ones of the agenda type and the others of the national type. With the aid of the *Transactions Relationship View*, we can easily visualise these assumptions through the two well-defined clusters, one for the online transactions and the other for the business transactions (Figure 10.13). Like client A, this client's activities can be seen as suspicious.

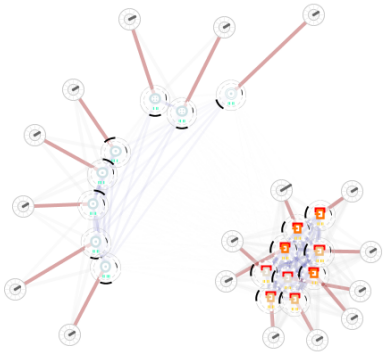

FIGURE 10.13: Transaction Relationship View of client B. In this visualization, two clusters can be found, one of the fraudulent business type (on the right) and the other of online transactions of small amounts (on the left).

▶ CLIENT C: When analysing the third client's data, we can see that it differs from the previous examples as the majority of the transactions are of the online type (Figure 10.14). The values in these transactions fall in their majority in two different ranges: in the lowest range, from €30 to €500, and in the highest range, above €4200. Also in both ranges, there are fraudulent transactions of the online type. The fraudulent transactions are common in highest values, but not so common in lower ranges. In this case, we can also see that, on March 5, there are fraudulent transactions in both ranges, which is uncommon. This client starts by doing an international transaction of low value and on the other day makes a few more

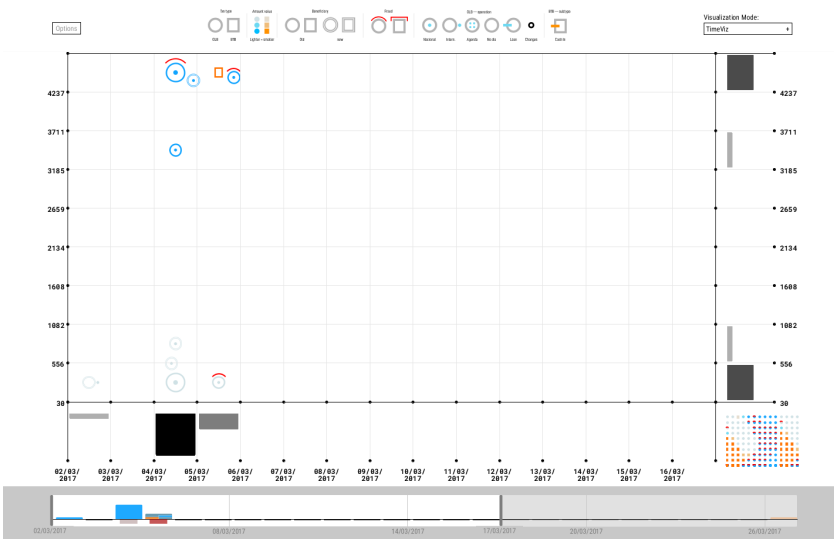

FIGURE 10.14: Transactions History View of client C. This client starts to perform a set of online transactions of small values and then performs a set of online transactions of high amounts, which are considered to be fraudulent.

FIGURE 10.15: Transaction Topology View of client C. In this view, we can see a more varied transaction matrix, when comparing client A and client B. This means that this client performs a more varied type of transactions.

national transactions to new beneficiaries of both high and low values. From eight transactions, only two were considered as fraudulent. On the next day, there is a high amount of business transactions and two national transactions. This can be defined as suspicious behaviour, since there are no more transactions in the following days, until March 26. By looking at the small SOM matrix, we can see a more varied SOM representation, in which fraud appears only on the national online transactions.

By looking at the *Transactions Topology View*, it is possible to see that the majority of the transactions have low-value ranges (Figure 10.15). Also, it is interesting to see that fraudulent transactions are made via Barc. Mobile (see Section 10.2) and non-fraudulent national transactions are made via the web. This may indicate a breach in one of the applications and should be analysed in more detail. Also, it is possible to perceive that the business transactions of

low ranges were made via ATM and the transactions with high values via Branch (see Section 10.2).

When analysing the *Transactions Relationship View*, we can see three main clusters and two outliers—which are the business transactions of low and high ranges (Figure 10.16). Also, in online transactions, we can see a distinction between fraudulent and non-fraudulent transactions. Among the non-fraudulent, we can distinguish four types of transactions: international, to new beneficiaries, national with low values, and national with high values. From these results, we can argue that, when comparing the *Transactions Topology View* versus the *Transactions Relationship View*, we were able to analyse more rapidly the different transactions and their relationships with the latter.

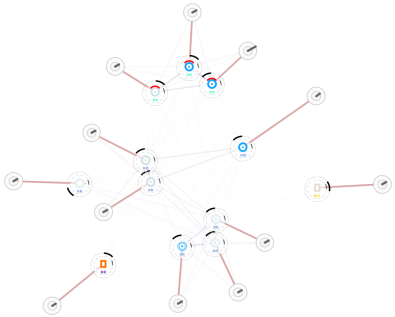

FIGURE 10.16: Transaction Relationship View of client C. In this visualization, it is possible to perceive two types of transactions that can be considered as outliers, the transactions of the business type of small and high amounts. Also, it is possible to distinguish two types of online transactions, fraudulent and non-fraudulent.

10.7 User Study

To evaluate the tool's usefulness and effectiveness in the analysis of bank transactions, we performed user testing with a group of fraud analysts from the Feedzai company that was not involved in the tool's development. In this user testing, the participants were asked to: (i) perform a set of specific tasks; (ii) analyse and characterise the transactions from two clients as fraudulent, non-fraudulent, or suspicious, through the interaction with the VaBank tool; and, (iii) give feedback on the aesthetics, interpretability, aid, and learning curve of each one of the three views. The tasks were defined to validate the models, and determine the effectiveness of the visual encodings. The second part—the analysis of a subset of transactions—was defined to assess the complete functionality of the VaBank tool as a whole and whether the analysts were able to retrieve insights from the visualizations, proving its usefulness in the analysis of bank transactions. The third part was defined so the opinions of the analysts could be registered and analysed.

10.7.1 *Participants*

The user testing was performed by five fraud analysts. These analysts worked for Feedzai, but have no a-priori knowledge about the VaBank tool. On average, they worked in fraud analysis for five years. The participant with the least experience was working as a fraud analyst for three years, and the one with the most experience was working with fraud for eight years. Three of the analysts had no experience with working with Information Visualization, and the other two had a reduced number of interactions with the field. Despite this being

TABLE 10.1: Tasks to be performed by the analysts. Note that tasks 1 to 10 are to be performed with the Transactions History View; tasks 11 to 14 are to be performed with the Transactions Topology View; and tasks 15 to 18 are to be answered with the Transactions Relationship View.

Number	Task
1	In what time period did BTB transactions have the greatest monetary value?
2	What is the time period with the most OLB transactions? How many?
3	Are there any temporal patterns in the histogram? Which?
4	How many time periods have less than 3 transactions?
5	Generally speaking, what pattern can be observed in BTB transactions?
6	In how many time periods do fraudulent transactions occur?
7	Which attributes appear in the time interval between 03/14/2017 and 03/15/2017?
8	How many national transactions took place between 03/09/2017 and 03/21/2017?
9	Generally speaking, what temporal pattern exists in terms of money spent?
10	Are there any attributes with a downward trend? If yes, which one?
11	Identify the glyphs that represent OLB transactions.
12	How many glyphs are there for the Scheduling attribute?
13	How many clusters are there with the fraud attribute?
14	What is the most predominant attribute in the SOM?
15	Identify the glyphs that represent OLB transactions.
16	How many glyphs are there for the Scheduling attribute?
17	How many clusters are there with the fraud attribute?
18	What is the most predominant attribute in the SOM?

a reduced number of participants, this user testing aimed at understanding the impact of a tool such as VaBank in the analysis process of fraud experts—which are more used to deal with spreadsheets. For this reason, we believe that this number of participants was sufficient to fulfil the test requirements and provide a general sense of the VaBank impact on their analysis process.

10.7.2 *Methodology*

The tests were performed as follows: (i) we introduced the glyphs of the transactions, the views of the tool, and respective interaction mechanisms; (ii) we asked the analysts to perform 18 tasks (Table 10.1) concerning: the *Transactions History View* (6), the interpretability of the glyphs (4), the *Transactions Topology View* (4),

and the *Transactions Relationship View* (4); (iii) then, the analysts analysed two clients in terms of fraudulent behaviours; and, (iv) the analysts were asked to give feedback on the models concerning aesthetics, interpretability, aid in the analysis, and learning curve. The second and third part of the tests were timed and, at the end of each task or analysis, the analysts were asked to rate the difficulty of the exercise and the certainty of their answers on a scale from 1 to 5—from low to high, respectively.

The 18 tasks of the user testing were divided into 4 groups, depending on the component they aimed to validate: **G1** Transaction History view; **G2** Transaction glyphs; **G3** SOM Matrix; and **G4** SOM Graph. In the Transactions History View, we tested the analysts' ability to understand temporal patterns and the transactions' distribution concerning time and amount values. In the views related to the SOM projections, we aimed to compare both views and perceive which one was more useful and efficient in solving tasks like counting clusters and identifying all glyphs from a certain attribute. For this reason, the tasks are equal for both views.

The third part of the test—which is concerned with the interaction with the VaBank tool and the analysis of two different clients' data—aims to understand the tool's usefulness and its ability to aid the analysts to detect suspicious patterns and possible frauds. During this part of the test, the analysts were asked to explore and analyse the visualization, explain out loud what they were seeing at each moment of their exploration, and refer to whether the client performed fraudulent, non-fraudulent, or suspicious actions.

The final part of the test was also intended to give the analysts the opportunity to express their opinions on the tool. Although such feedback might be subjective, it is an indicator of the tool's impact on the analysts' workflow and can give clues on its effectiveness and efficiency.

All tests occurred in the same room within the Feedzai installations and were performed under the same conditions (i.e., the participants had access to the same computer and performed the test in the same sequence)[4]. We recorded the audio from each test so we could analyse each session afterwards.

[4] A PDF of the User Test can be found in the following link: https://cdv.dei.uc.pt/cmacas/VaBank_UserTest.pdf

10.7.3 *Results*

In Figure 10.17, we summarise the results concerning difficulty, certainty, accuracy, and duration for each group of tasks. Hereafter, we further analyse each group of tasks, discuss the results of the third part of the test, and analyse the analysts' feedback.

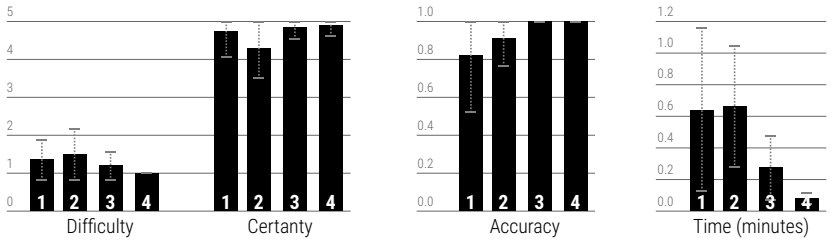

FIGURE 10.17: The difficulty, certainty, accuracy and time values for the 4 tasks groups. In general, the difficulty of the tasks was considered to be low, and the certainty and accuracy are considered to be high. With regard to time, the majority of the tasks were completed in less than one minute.

Tasks Analysis

The tasks related to the analysis of the *Transaction History View* (**G1**) and glyphs (**G2**) were the most difficult ones. Nonetheless, all values are low, considering that on average the difficulty was no higher than two (i.e., the second-lowest level of difficulty). Regarding the Transaction History View, the analysts had more difficulties in interpreting the positioning of the glyphs in the grid and the histograms. For example, for the task "In which period of time the business transactions had the highest amount?", some analysts started to look at the histogram on the right, which gives the total number of transactions for each range of amount values. However, as this was the first question of the test, they were still assimilating all the information and rationale of the tool. The analysts also had some difficulty in interpreting the glyphs, which made their certainty to be lower than the other groups. Nonetheless, on average, the certainty was no lower than four (i.e., the second-highest level of certainty). Also, an interesting point is that the accuracy of the analysts' answers for the glyphs tasks is higher than the accuracy for the Transaction History View tasks. Thus, since the glyphs are complex, the analysts were not certain if they were characterising all their attributes correctly, which caused the lower rates of certainty. Nonetheless, in the majority of the tasks related to the glyphs, their answers were accurate.

The groups of tasks related to the SOM's analysis—Transaction Topology View and Transaction Relationship View—took less time to perform (20 seconds, on average), had 100% of accuracy, and were the ones in which the analysts had more certainty in their answers and less difficulty in completing the tasks. Comparing both views, the Transaction Relationship View (**G4**) had the lowest duration and the difficulty of completion was also considered low. This can be explained by the fact that, as the graph is less complex (has fewer glyphs), the analysts were faster in their analysis of the glyphs and their relationships. These good results on both views also indicate the good acceptance of such models, and the ease with which the analysts interpreted the topology of the transactions.

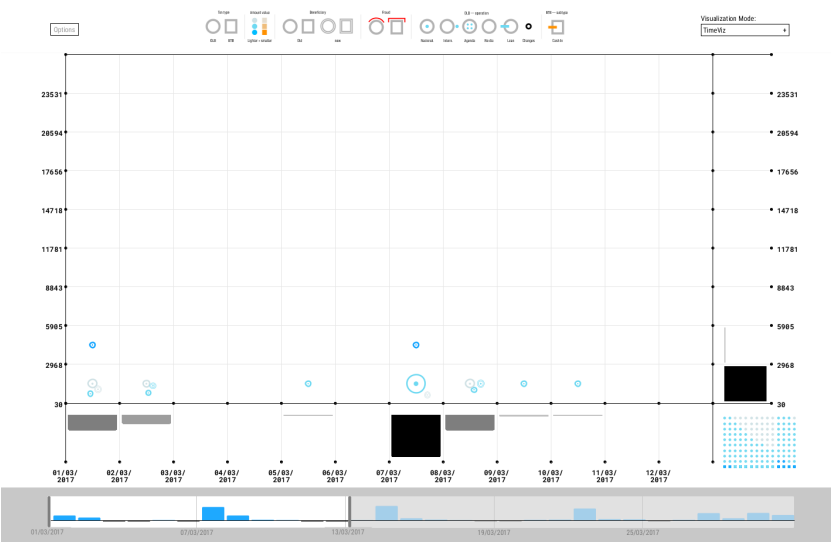

FIGURE 10.18: *Transactions History View* of client A, from the third part of user testing. This client performs a set of online transactions of small values.

VaBank Analysis

The third part of the test was concerned with the free exploration and analysis of the transactions of two clients. These clients have two different behaviours: client A has a suspicious behaviour at the end of his/her activities, and client B commits fraud at the beginning of his/her activities.

► CLIENT A: The majority of the analysts interacted with the tool in the same way. Hence, we hereafter summarise their interaction. In the beginning, the analysts spotted no fraud in the first period range and all detected a weekly periodicity of online transactions of the agenda type with low values (Figure 10.18). One analyst started to be suspicious when perceived that there were 13 transactions to different new beneficiaries. Then, the analysts scrolled on the timeline to see the other transactions of that month. By doing so, the analysts understood that there was a disruption of the initial pattern. At the end of the month, the transactions had no pattern, they were scattered along the last days, and the rate and value of the transactions increased. Through the interaction with the glyphs, one analyst noted that some transactions on the same day were made in different countries, which was considered suspicious. Also, by interacting with the glyphs, another analyst found that the beneficiary attribute changed in all transactions but the Internet Service Provider (ISP) was always the same. Additionally, after analysing the data through the Transaction Relationship View and Transaction Topology View, one analyst stated that the suspicious behaviours were more evident in the Transaction History View than in the other views.

In summary, the majority of the analysts identified the activities

as a suspicious case especially due to the pattern changes and the increase of amount and rate of transactions.

- CLIENT B: The analysts directly detected the fraudulent activities through the glyphs. In this subset, the client asked for a loan of high value that was considered as fraud and on the same day performed several online transactions of the agenda type, which were also considered as fraud. All analysts were intrigued by the fraudulent transactions, and all interacted with the glyph to try to understand what were the attributes of those transactions. By doing so, they could perceive that the transactions were made to different beneficiaries. The majority of the analysts referred to this type of behaviour as an external attack on a legitimate client account. Also, one analyst stated that it was also suspicious that client B tried so many transactions of high value, and then, after one week, made another transaction of a relatively small value.

  In summary, the activities were instantly classified as fraudulent, for his attempts of doing several transactions with high amounts for different accounts. Also, most analysts referred to client B as an account that might have been hacked. As this client had few transactions, the analysts could see every transaction in the transaction matrix, without needing to interact with the timeline.

Feedback

At the end of each test, the analysts rated each view in terms of aesthetics, interpretability, aid in the analysis, and learning curve. The Transaction History view got a better rate in terms of aesthetics and aid. Additionally, it was defined as easier to interpret, but had worse ratings in terms of the learning curve. This last rating may be caused by the complexity of this view, which included the histograms, the small SOM matrix, and the timeline. Thus it requires more experience.

Concerning the Transaction Topology View and the Transaction Relationship View, the analysts took more time to complete the tasks with the first and rated it with the highest values of difficulty. However, the Transaction Topology View was seen as a better aid for the analysis of the transaction patterns and was also defined as easier to learn, compared to the Transaction Relationship View. In fact, the Transaction Relationship View was the view with the lowest ratings in terms of aesthetic value and aid in the analysis of the data.

At the end of the tests, some analysts made some comments on the tool. They referred to the Transaction Topology View as a good auxiliary for their work and referred that with more practice

the glyphs would get easier to read and interpret. One analyst also suggested a new positioning of the glyphs in the transaction matrix: to place them in each cell radially so as to represent the hours at which the transaction was made.

10.8 Discussion

Through our interaction with the fraud analysts we were able to define the two main tasks to which VaBank should answer: (i) to enable the visualization of the transactions over time; and, (ii) to enable their profile characterisation. The analysts also aided us in the definition of the specific requirements for the tool which allowed us to define the visualization models and interaction mechanisms. These first steps revealed to be important for the development of a tool to be used for the analysis and detection of suspicious behaviours in bank transactions.

Through user testing, we could validate our tool in terms of efficiency, as most of the tasks were completed in reduced times—on average the tasks took less than 1.3 minutes to complete—and the exploration and analysis of the clients' data were also completed in a short time—it took on average 4 minutes, which the analysts referred to as a good time for the analysis in comparison to their current tools. We could validate the tool in terms of effectiveness, as all analysts were able to complete the tasks correctly. Through their interaction with VaBank, they were also able to analyse the details of the transactions, their main characteristics, and detect suspicious behaviours.

With the analysis of the tasks' results, we could assess the interpretability of the visualization models. For example, we could understand that despite the complexity of the glyphs, the three levels of visual impact achieved their purpose, as the analysts could focus on the first level (the type of transaction, amount, and fraud) and with a closer reading analyse the operations types and the rest of the transaction's characteristics. Although during the execution of the tasks the Transaction History View was seen as the most difficult, after the interaction, the analysts found it to be easier to interact with. Through the analysts' feedback, we could understand that this view was well-received by the analysts, which defined it as a good aid for their work. We could also perceive that, although the Transaction Relationship View was faster to analyse, defined as easier to learn, and the one in which the analysts were more certain about their answers, this view was also seen as less informative than the Transac-

tion Topology View. The Transaction Topology View was seen by the analysts as a better aid for the analysis of the transaction patterns and was also defined as easier to learn.

With the analysis of the results of the second part, we could conclude that all analysts understood the tool's interaction mechanisms and all were able to interact properly with the tool. Moreover, the analysts took a small amount of time to analyse and perceive the types of behaviours of the clients. Therefore, we can conclude that VaBank can aid in the detection of suspicious behaviours, which in turn, can improve the analysts' decisions.

Concerning the analysts' feedback, they stated that after the completion of the tasks they were more familiarised with the tool, and could easily use all interactive features. Additionally, they referred to the highlight of transactions with the same attribute as a good feature that was relevant for their line of work. This highlight aided in the creation of relationships between transactions and in the analysis of their attributes. In general, all visual elements were well received and understood. One analyst also referred that the timeline was an important asset as it enabled the visualization of different periods of time and the understanding of the types and amount of transactions on different time periods. Finally, the representation and highlight of fraud were well understood by every analyst.

11

Online Transactions

Fraudulent acts can use different techniques and follow different methods to perpetrate fraud. Nonetheless, they can be grouped and categorised into different patterns. For example, an ATO is a specific fraud pattern in the financial domain, more specifically in e-commerce. It is one of the major e-commerce fraud patterns and can be defined as the unauthorised use of another person's profile and corresponding credit card details [28]. In most cases, the fraudster exploits the stolen card as much as possible, before being detected [213]. Detecting this or any other kind of financial fraud can prevent large losses for companies and individuals, and for this reason, it is an increasingly relevant problem to solve [28, 166].

Currently, many fraud detection companies rely on ML approaches to capture fraudulent activities. However, as technology evolves and the techniques applied in fraud detection become publicly available, fraudsters adapt and modify their ways of acting [28]. This may prevent ML models to detect correctly all fraudulent transactions and may lead to their incorrect classification. To make the detection of fraud more reliable, effective, and efficient, Information visualization can be applied to aid fraud analysts in identifying fraudulent behaviours that may have passed undetected by the ML algorithms [28, 79]. Hence, visualization can enhance such automatic methods, by enabling the detailed analysis of each suspicious transaction that still needs to be carefully investigated, revealing new undetected fraudulent patterns.

In this Chapter, we present ATOVis, a visualization tool that aims to aid in the detection of ATO through the representation of e-commerce transactions. ATOVis was developed in collaboration with a world-leading company specialised in fraud prevention—Feedzai. In Feedzai's line of work, different groups of analysts are given different fraud patterns to study and tackle, making them experts in specific patterns. For this reason, and to fill the lack of visualization

tools for specific fraud patterns, ATOVis is aimed to speed up, ease, and improve the manual evaluation of fraudulent transactions which the company's ML model had difficulty classifying, assigning them a low confidence score. One of the major challenges of ATO detection, is to understand the client's typical transaction history and detect uncommon behaviours or patterns that can be related to fraud. With the aid of experienced fraud analysts, we tackle these challenges and enhance the visual representation of fraudulent patterns, and the overview of all transactions in a single tool. Hence, ATOVis is defined as a user-centred tool that improves decision-making tasks by providing a better understanding of each individual case.

ATOVis contributes to the state of the art as the first visualization tool highly specialised in ATO patterns. In particular, we discuss (i) a detailed task abstraction on ATO patterns derived from the interaction with expert fraud analysts; (ii) the design process and design decisions for the development of ATOVis; (iii) the design of a multiscale timeline which enables an overview of the data and, at the same time, a detailed view of its distribution over time; and, (iv) the findings and insights derived from the validation of our tool with experts and non-experts in fraud detection. Based on the feedback provided by the analysts, we could conclude that ATOVis is an efficient and effective tool in detecting specific patterns of fraud which can improve the analysts' productivity.

11.1 Account Take Over

In general, financial fraud can be subdivided into different types of fraudulent patterns. Based on our collaboration with the fraud prevention company, we could conclude that ATO and BA, which have similar ways of acting, are the most common patterns in e-commerce. An ATO is the illegal acquisition of legitimate clients' details to take over online accounts and use the stolen data (e.g., credit card details) to purchase products [284]. In general, it can be detected by constant changes in a client's details, an abnormal purchase rate, or an increase in monetary values. A BA is the use of one or more bots, i.e., software programs, to execute multiple attacks. In e-commerce, bots use stolen personal accounts of e-commerce websites to buy, illegally, the company's goods [93]. Most BAs can be detected by analysing the constant changes in the transaction's attributes (e.g., IP Address, Country) or by detecting high amounts of repetitive and periodic purchases. For this reason, this type of fraud is often connected to ATO, and by visualising specific patterns of ATO,

BA cases can also arise.

It is of utmost importance to quickly detect client accounts that have been compromised and stop fraudulent transactions to prevent large losses for the e-commerce company and the counterfeiting and selling of their products in black markets. The fraud prevention company already employs several of its analysts in the manual analysis of transactions with a low confidence score. We argue that the use of visualization as a tool to analyse specific patterns of fraud, can ease its detection, speed up the process of accepting/rejecting transactions, and, consequently, decrease the effort required for the task, and, potentially, increase the quality of the classifications, reducing the number of false positives and negatives. Through visualization, the analysts can have a broader overview of the transactions, which they cannot achieve with their current tools (i.e., spreadsheets). This would enable them to better understand the patterns of transactions, support their decisions, and detect new patterns that the ML algorithm may not have learned yet. More specifically, with our tool, we intend to: (i) represents the client's transaction patterns, emphasising consecutive changes in the transaction details; (ii) improve the understanding of the behaviour and periodicity of the transaction; and, (iii) enable the detection of similar fraud patterns (e.g., BA).

11.1.1 *Fraud Detection workflow*

To better understand the applicability of our visualization tool, we further describe the company's workflow. The fraud prevention company gives to its clients (e.g., an e-commerce company, a bank) a solution to detect and stop fraudulent transactions, which encompasses the following phases: (i) the automatic detection of fraudulent cases through a ML algorithm; (ii) the manual detection of fraud through human analysis of low confidence transactions; and, (iii) the analysis of older transactions, which may reveal new patterns of fraud. Phases two and three are independent of each other and commonly occur after phase one.

In phase one, the ML algorithm runs through all transactions made in the e-commerce company and assigns a score. This score describes the confidence in the transaction's labelling (as fraud or not). Although the ML algorithm is efficient in detecting fraudulent cases, fraudsters are constantly changing their ways of acting. To better fight fraud and to support drift detection the fraud prevention company manually analyses the ML results that are within a certain threshold range to identify false negatives or false positives.

In phase two, low confidence transactions are manually analysed

and the analysts must determine whether the transaction is fraudulent or not in a short time (no more than 10 minutes), so that actions can be taken to stop fraud. In this phase, our tool can aid the analysts by giving a temporal overview of all transactions from the same client and highlighting consecutive changes, positively shortening the analysis time.

In phase three, a group of analysts have more time to study uncommon scenarios and search for undetected frauds. With these new cases, they can take further actions (e.g., improve the ML algorithm, extend the blacklist of clients). In this phase, our visualization model can also be useful, as it enables the exploration of the data so the analyst can drill down on the details of the transactions to detect suspicious attributes. To enhance the detection and inspection of fraudulent activities, ATOVis highlights in red the transactions labelled as fraudulent by the ML algorithm and gives access to the confidence score and other information through a details area.

The current method used by the company in phases two and three relies mainly on spreadsheets, which are complex to analyse and hardly give an overview of the temporal patterns. They also have access to a simple interface in which they can see the details of each individual transaction in a textual form and do queries to find related transactions. They usually switch between spreadsheets and visual interface, which is a laborious task. The goal of our tool is to facilitate this process, by providing to the analysts all transactions grouped by client, the possibility to analyse the client's behaviours and possible fraud attempts, and the ability to get all details in a single tool.

11.2 Data

We had access to a dataset containing more than 4 million e-commerce transactions, made between November 2016 and February 2017. All transactions were previously processed by a ML algorithm for fraud detection, and are characterised by (i) a set of attributes specific to online purchases (e.g., ID, timestamp, amount, billing and shipping addresses); and, (ii) a set of attributes assigned by the ML algorithm (e.g., fraud label and confidence score). The dataset was properly anonymised, retaining the fraud patterns of the real dataset and enabling us to visually explore the data in real case scenarios, without compromising the clients' anonymity.

Each transaction of the dataset has a total of 32 attributes: 8 binary, 3 quantitative, and 21 nominal. For the visualization model, we focus on a subset of attributes, which were highlighted by the

analysts as being the most affected attributes by ATO cases and which can reveal behavioural changes with more accuracy. Those attributes are related to: (i) geographic locations—shipping/billing/account address and IP address; (ii) personal details—shipping/billing/account email and name; and, (iii) transactional attributes—card id, card country, and device. Note that the amount spent on each transaction is not represented in the visualization model, as it is not representative of an ATO case (i.e., the change in amount is not a fraudulent pattern by itself). Notwithstanding, these attributes can be altered at any given time, having no impact on the visualization model. ATO patterns are intrinsically connected to the consecutive changes of each client's purchase behaviour and/or account details. For this reason, we parsed the dataset by the individual client. In summary, our visualization model deals with each client separately, enabling the analysis of the client's transaction history both through overview and details on demand techniques.

11.3 Task Analysis

Unlike most visualizations in financial fraud detection, our project focuses on a specific fraud pattern: ATO. We interacted with fraud analysts to understand how they work with such sensitive data and which are their main requirements for the visualization. Before and during the design and implementation phases, four workshops were organised and three of the company's analysts who work on ATO cases shared their insights about their line of thought and the specificities of ATO patterns. This enabled us to understand the workflow used to detect ATO cases, contributing to a guided and user-centred development of the visualization models. Also, in the first workshop, the analysts showed us how they work to detect ATO and BA cases.

To analyse the e-commerce transactions, the analysts use spreadsheets and a web-based tool that gives access to all transactions. For each transaction, a dashboard containing their specific details is shown. This information is presented, mainly, in textual and tabular forms. For the detection of ATO patterns, they look for changes in the shopping behaviour and profile details of the client. However, to detect such changes, the analyst has to search and compare thoroughly the most important attributes (see Section 11.2) and create a mental image of the client's behaviour. This is a laborious, difficult, and time-consuming process. Considering that ATO is characterised by the changes in transactional attributes and behaviours, visually emphasising the change of attributes between transactions is an important

requirement, which became the basis for our work. The workshops, allowed us to promote and validate a set of tasks to be addressed:

T1 **Analyse transactional patterns.** The analyst needs to overview the transaction history to differentiate typical from atypical transaction patterns and perceive the changes;

T2 **Detect consecutive changes in attributes.** It is important to detect when the client changes his/her details and which ones. By highlighting transactions in which an attribute changed, the analyst will be able to easily detect suspicious behaviours. However, only attributes which are relevant in ATO detection should be emphasised (see Section 11.2);

T3 **Detect the reuse of attributes.** The analyst needs to compare the values between and within transactions. For example, it is important to highlight the use of different values for attributes of the same domain (e.g., account email, billing email, and shipping email), as it can be related to ATO.

T4 **Detect fraudulent transactions.** The analyst needs to instantly understand which transactions were identified as fraudulent by the ML model.

To improve the analysis of the transactions through ATOVis, details about the transactions must also be provided, enabling well-informed decision-making. Therefore, a set of secondary tasks was defined:

T5 **Inspect the values of the attributes.** The analyst should be able to visualise the sequence of values used in each field and determine if they can be considered dangerous.

T6 **See the transaction's details.** The analyst should be able to see all information about the transactions and detect values of risk.

11.4 Design Requirement

We defined most of our design requirements based on our interactions with the fraud analysts. However, from the related work on fraud visualization, we could retrieve one important requirement: the ability to compare different transactions. Although most fraud visualizations try to fulfil their requirements through interaction, we aim to propose a visualization that can represent the client's behaviour at a glance, requiring reduced to no interaction. Hence, we focus our

requirements on the direct representation of the data. We identified 3 main design requirements for ATOVis:

DR1 **Represent the transactional behaviour.** Due to the specificities of ATO, the analysts must have an overview of the sequence of changes throughout the transaction history. Hence, a design that sorts and emphasises all transactions in which changes occurred, assigning less visual emphasis to the ones with no change of attributes, may ease the detection of ATO patterns [**T2-4**]. To avoid visual clutter, we aggregate the transactions by day and type of transaction; consecutive patterns are further clustered;

DR2 **Summarise the temporal distribution of transactions.** The analysts need to understand the distribution of transactions over time, so they can understand and detect differences in periodicity, trends, and quantity of transactions. Thus, the visualization should represent chronologically the transactions and their characteristics, emphasising transactions annotated as fraud and with changes in attributes [**T1**, **T4**]. Also, the analyst should be able to select and filter specific periods of time to visualise. To allow this, a timeline is made available which is also a representation of the transactions' evolution;

DR3 **Summarise statistics and transaction details.** The visualization should support the analysts in the in-depth analysis of specific values used in each attribute, allowing them to detect suspicious attributes (e.g., IP, email domain) [**T5-6**]. Although we aim to give a complete understanding of each clients' transactions at a first glance—minimising interaction and analysis time—a set of interaction methods must be provided. They will allow detailed analysis and decisions that are more informed (e.g., showing the confidence score given by the company's algorithm);

11.5 ATOVis Design

In this Section, we describe the design of ATOVis (Figure 11.1). In summary, the visualization model is based on the realisation that when dealing with ATO the analysts focus their attention on the changes in the transaction's attributes. Consequently, we focus our visualization model on the representation of such changes and their periodicity. ATOVis is a functional application implemented in Java and using Processing to render the visualization. A video was recorded

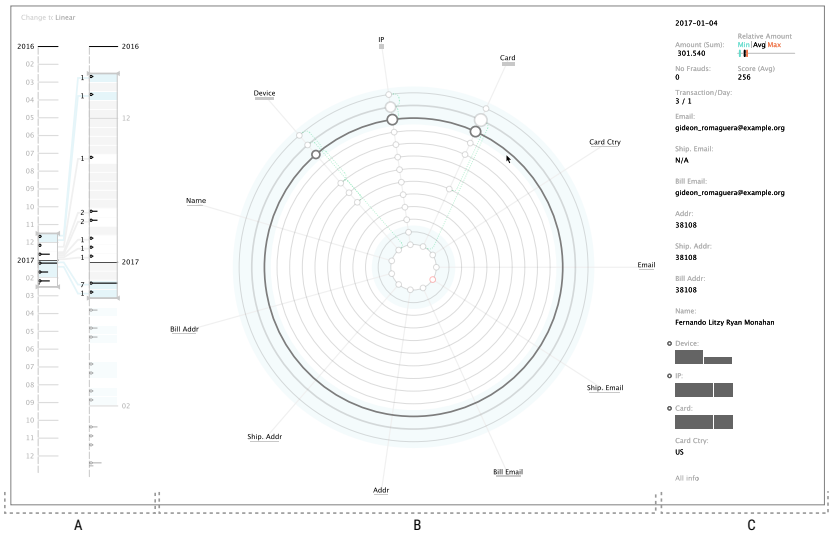

FIGURE 11.1: Screenshot of the ATOVis, capturing a non-fraudulent case. The analysts have access to the temporal arrangement of all transactions (horizontal bars) in a vertical timeline, where they can specify a desired period of time for further analysis (A). In the main area, the analysts can visualise all transactions—concentric circles—from the selected time interval (B). By hovering over the transaction(s)—darker concentric circle—, the analysts can further drill-down and get more details about their selection (C).

to exemplify the interaction with the application: https://cdv.dei.uc.pt/cmacas/ATOVis-video/ATOVis.mp4[1].

We defined three different areas in response to the design requirements: the *timeline*, the *main area*, and the *details area* (Figure 11.1). Through our workshops with the analysts, and following Dilla and Raschke [79], we can state that the process of discovering fraud usually involves detecting unusual patterns, drilling down into the data, and selecting individual items for further analysis. A similar procedure was proposed by Shneiderman [259]. For this reason, after selecting the period of interest from the timeline area, the analyst can visualise the filtered data, and analyse with more detail the client behaviours in the main area. If any transaction(s) raises suspicion, the analyst can further drill down and visualise, in the details area: (i) statistics on the selected transaction(s); and, (ii) their attributes placed in a tabular fashion. The order of the analytic steps described above and the natural western reading direction defined the layout of the elements, from left to right: 1st degree timeline; 2nd degree timeline; main area; and, details area.

All components of ATOVis have common design requirements: (i) the transactions, in both timeline and main view, must use similar representations, so the visualization as a whole is coherent; and, (ii) the use of colour to encode data must be as reduced as possible, so important attributes can be better highlighted from other elements [191]. As the detection of fraud is the main goal, we apply the red colour to highlight the transactions annotated as fraudulent [**T4**]. As a summary, we present the visual encoding of our visualization tool, organised according to the visual task taxonomy of Zhou and Feiner [317]. This taxonomy was selected due to its alignment with our tool's main tasks, as it is based on exploratory tasks which,

[1]To comply with the company's requirements, the video is based on a dataset generated with random values that follow the statistical properties of the original dataset. For the generation of images of the present Chapter, we used the anonymised data provided by Feedzai and detailed in Section 11.2

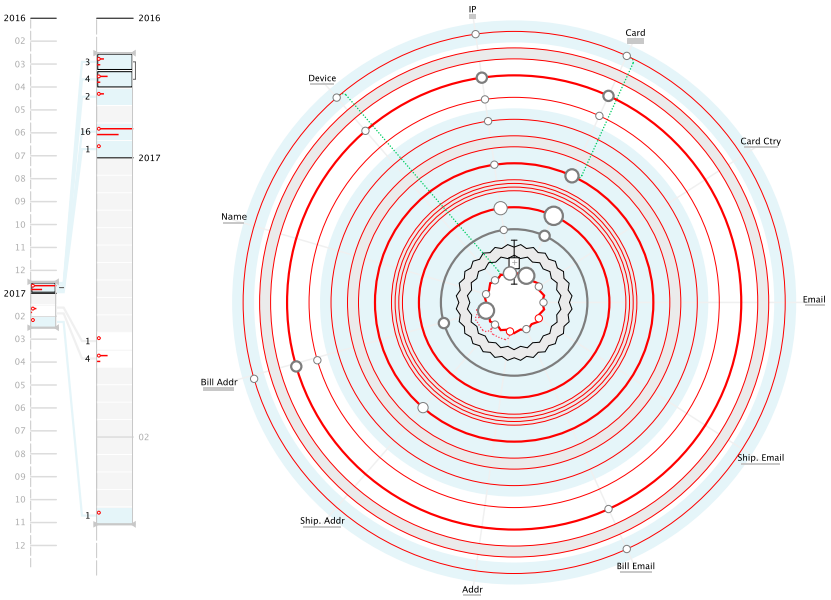

FIGURE 11.2: Screenshot of ATOVis, showing a case of fraudulent activities through the red colouring of the lines. On the left, the analyst can see the multiscale timeline and the distribution of the fraudulent transactions—red horizontal lines. In the 1st degree timeline (on the left), the analyst can see that, between 2016 and 2017, the transactions occur only between December 2016 and February 2017. In the 2nd degree timeline (on the right), the analyst can see in more detail the transactions' distribution within the period mentioned above. The circular visualization in the main area represents the transactions from the period selected in the 2nd degree timeline. In this visualization, every transaction with changes in attributes is grouped by day, and represented with a concentric circle. Red concentric circles represent days with fraudulent transactions.

in turn, are based on search and verification tasks:

Categorise we use two different shapes to represent the types of transactions: with and without change [**T1**].

Compare and Rank we use stroke thickness to represent, for each day, the number of transactions with change; use different shape complexities to represent the number of transactions with no change (Figure 11.4); and use the size of the circles to represent the number of changes of the attributes, during each day, allowing the analysts to perceive which attributes changed more [**T2**].

Distinguish and Emphasise through colour, we highlight fraudulent transactions [**T4**] (Figure 11.2).

Correlate we use dotted lines to connect attributes which are related between and within transactions [**T2-3**] (Figure 11.2).

Identify by hovering over any transaction, attribute, or block of time, their details will be shown in the details area [**T5-6**] (Figure 11.1).

Cluster we emphasise repeated transactional patterns by clustering them and representing those clusters through a zigzagging line (Figure 11.2). These clusters can be further expanded [**T1**].

11.5.1 *Main View*

The Main View is the principal focus of our project. It is centred on the interface area and occupies the larger part of the canvas. The

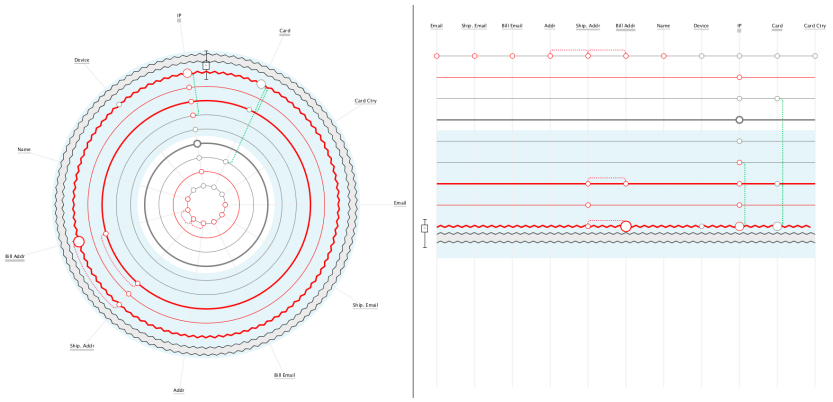

FIGURE 11.3: The ATOVis visualization models. Two layout explorations for the representation of changes in transactions: radial layout and linear layout.

main goal of this visualization is to give an overview of the transactional behaviours of each specific client, enabling the detection of transactions with changes and different transactional behaviours [**DR1**]. To achieve this, it is important to characterise and distinguish two types of transaction: *changed* and *normal*, i.e., if relevant attributes have changed or not concerning the previous transaction, respectively. As it is not reasonable to characterise a transaction as changed through, for example, the differences in the amount spent (i.e., not intrinsically connected to fraud), a set of attributes were defined with the aid of the analysts to discriminate between normal and changed transactions (see Section 11.2).

Transaction Representation

To plot the data we employed two spatial arrangements: *radial* and *linear* (Figure 11.3). The goal of the linear arrangement is to implement a visual system that is close to the analysts work base: the analysis of tables. With a visual system that resembles a table, we aim to ease the adaptation to the tool and improve the readability of the visualization. In this representation, all transactions are represented by a horizontal line and are ordered by time vertically, from top to bottom, with the most recent at the bottom. The attributes are horizontally arranged, from left to right, and equally spaced. For the radial arrangement, we aimed at understanding whether the representation of the transactions as a radial imprint of behaviours would enable a faster understanding of the transactions. According to the work of Draper et al. [83] and Diehl et al. [78], radial representations of data make more efficient use of space, which facilitate the comprehension of the visualization model and user interaction, thus enabling a faster understanding of the data being represented. To focus the analyst's attention on the sequence of attribute changes, we encode the attributes in the radial sectors [78]. In this encoding, the transactions are represented through concentric circles, with the

most recent in the exterior. The attributes are arranged around the circle, equally spaced.

Both approaches can be seen as a sequential timeline, in which all transactions are ordered chronologically, but where the space between them does not represent time. We implemented all the visualization methods for both spatial arrangements to compare their efficiency and effectiveness in terms of (i) detecting transactions and attributes with more changes; (ii) understanding the relationship of attributes between and within transactions; and, (iii) understanding the transactional behaviour of each client. The results can be consulted in Section 11.7.

The direct visualization of every transaction would lead to a cluttered visualization. For example, in BA cases it is common to see hundreds of transactions in a fraction of time. This would lead to a considerable increase in the visualization space, visual clutter, and complexity. To avoid this, we aggregate by day all transactions with changes and represent this aggregation through line thickness. The higher the number of transactions with changes in one day, the thicker the line. By doing so, we aim to enable the analyst to easily detect days with the higher amounts of transactions with changes, which are more suspicious, than days with lower amounts.

The lines representing transactions with change are complemented to represent which attributes changed [**T2**]. For every attribute that changed in relation to the previous transaction, we place a circle on top of the transaction's line and in the corresponding attribute's axis. If a line represents more than one transaction with change and an attribute changes more than once, the stroke thickness and size of the circle increase in proportion to the number of changes. In sum, we use stroke thickness in both line and circle to represent the number of transactions with change and emphasise the attributes with the highest numbers of changes through the circle's size.

The transactions without changes in attributes must also be represented, to detect differences in the transactional periodicity and behaviour. However, as they are not the main focus of our visualization, they must have a reduced visual impact. To this extent, we aggregate all transactions with no change that occurred between transactions with change, regardless of the days on which they occurred. To represent these aggregations and to enable the analyst to interpret them, we visually distinguish groups with low, medium, and high rates of transactions (Figure 11.4). To do so, we compute the average number of transactions by day, and represent 3 types of rates: average of transactions by day lower than 3; average between 3 and 10, inclusive; and higher than 10. Although these ranges

FIGURE 11.4: Three levels of detail to represent transactions without change, according to the average number ($\overline{x}$) of transactions (t) by day (d).

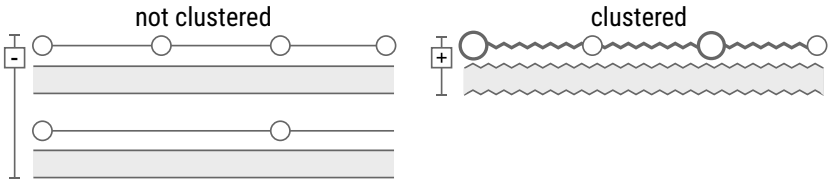

FIGURE 11.5: Difference between the representation of aggregated (left) and clustered (right) transaction patterns. On the left of each representation, there is a button to condense or expand the clusters.

were defined in collaboration with the fraud analysts, they can be modified at any time. We then represent the transactions with no changes as follows: first, we draw a rectangle with a length equal to the visualization space (linear arrangement), or a doughnut shape (radial arrangement), both with fixed height; then, to represent the different rates, we draw consecutive parallel lines inside the previous shape and adjust their density according to the rate it represents (Figure 11.4).

Finally, to emphasise fraudulent transactions, we colour the respective lines in red. To direct the attention of the analyst to problematic groups of transactions, we apply red to the aggregated transactions if at least one transaction in that group is marked as fraud (Figure 11.2).

Clustering transactional behaviours

During our workshops with the fraud analysts, it was pointed out that the visualization could be further simplified by grouping patterns of transactions [**T1**]. These patterns were defined by the occurrence of the following sequence: transaction (or group of transactions) with change followed by a transaction (or group of transactions) with no change. When this pattern repeats itself consecutively, those transactions can be clustered, so the number of lines is further reduced and a summary of the patterns of transactions can be better represented. The clustering algorithm only groups similar patterns, in the sense that a pattern of transaction with change, followed by a low rate of transactions with no change, will not be grouped with a pattern of transaction with change, followed by a high rate of transactions with no change. To visually emphasise the clustered patterns, and distinguish them from unclustered ones, we use zigzagging lines, instead of straight lines, as shown in Figure 11.5.

Relating attributes

We visually connect attribute's values that: (i) are reused in different transactions; and, (ii) are distinct but belong to the same attribute's domain (e.g., user email, shipping email, and billing email) in the same transaction [**T3**]. Visually, these connections are represented

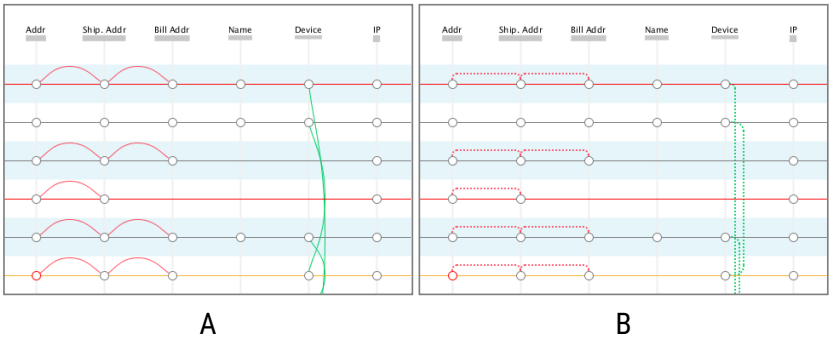

FIGURE 11.6: Comparison of the two approaches for connecting attributes: by using arcs (A); and straight dashed lines (B).

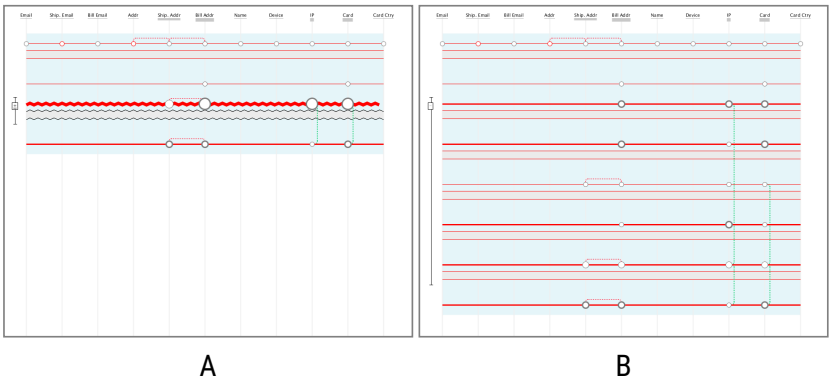

FIGURE 11.7: Cluster-expansion. Through interaction with the cluster (A), one can expand and visualise all transactions contained in the cluster (B).

similarly, through a dotted line, but distinguished with colour and angle. For the first type of connection, we use vertical green lines, and for the second type, we use horizontal red lines. An earlier approach for these connections was based on arcs. However, preliminary tests showed that this representation created more confusion and visual clutter. The comparison between the two approaches is displayed in Figure 11.6.

Interaction

We implemented a set of interaction techniques: details-on-demand, to obtain more information about the sequence of changes of a certain attribute and the transactions [**T5**]; and, cluster-expansion, to view the transactions inside clusters [**DR3**]. For the latter, we created an additional visual component that supports the detection of clusters and the interaction with them. For this visual component, we draw a vertical line that visually connects all transaction patterns inside the cluster. If the cluster is expanded, instead of the zigzagging lines that represent the overall pattern, we draw all transaction patterns inside it, without the zigzagging. Additionally, when the cluster is expanded, the vertical line accompanies all the transactions, and its height is defined by the number of transactions inside the cluster (Figure 11.7).

Regarding the details-on-demand, we provide to the analyst the

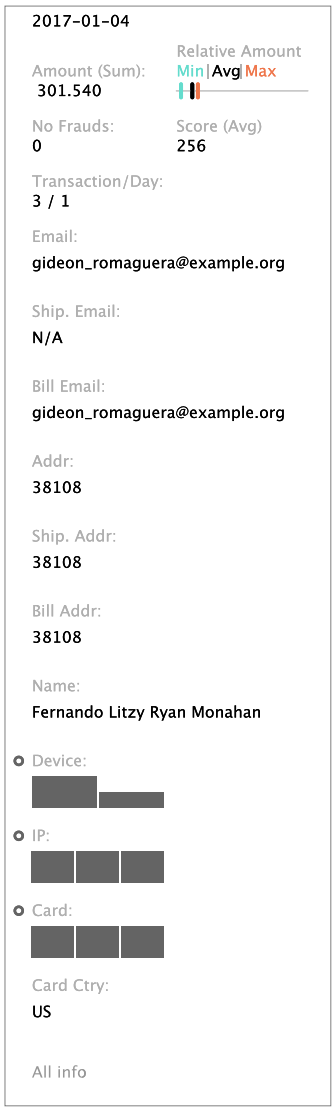

FIGURE 11.8: By hovering over the transaction(s), the analysts can further drill-down and get more details about their selection on the rightmost area of ATOVis.

details about each transaction and its attributes [**T5-6**]. The analyst can interact with the circles that represent a change of attribute. Through mouseover or click, the analyst can see, in the details area, the list of all used values in chronological order, starting with the first entry of the client's transaction history until the selected one. If the number of attributes exceeds the height of the panel, a scroll bar is made available. The analyst can also interact with the transaction lines. Through mouseover or click, the analyst can access the details and statistical information about the selected transaction (Figure 11.8). The attribute fields shown in the details area are the same, regardless of the number of transactions. If an attribute changes more than once in a set of transactions, a histogram is shown (Figure 11.8). The histogram depicts the number of different attributes used and the number of occurrences. In this area, we also show the amount spent and the ML score for the selected line. If the line represents more than one transaction, we show the average amount per transaction and average score. Additionally, we present the number of frauds and the average number of transactions by day. To contextualise the spent amount of the selected transaction(s), we added a graph that plots the average amount from the selected transactions, and the minimum and maximum computed from all client transactions (Figure 11.8, top right corner). To aid the analyst in getting all the details of the client's transaction(s), at the bottom of this area, a button gives access to the presentation of every transaction in tabular form [**T6**].

11.5.2 *Timeline Design*

To enable the analyst to navigate through all transactions, we implemented an adaptive multiscale timeline, whose main focus is to give an overview of the transactions' periodicity and to represent how they distribute over time [**DR2**]. This timeline can then be considered as a tool to highlight periods of interest.

The multiscale timeline consists of two vertical timelines with different time scales: the *overview* and the *detailed* view. In the first, an overview of all transactions is presented. In the second, a temporal closeup is shown so the analyst can get a more detailed view of the temporal distribution of the transactions. Both timelines are implemented using an adaptive algorithm. Depending on the time range, the algorithm adapts the timeline granularity so the time range can fit into the allocated space. Depending on the granularity, each month is subdivided into temporal blocks which can range from one (i.e., all transactions of the month are aggregated into one block) to 31 (i.e., the transactions are aggregated by day).

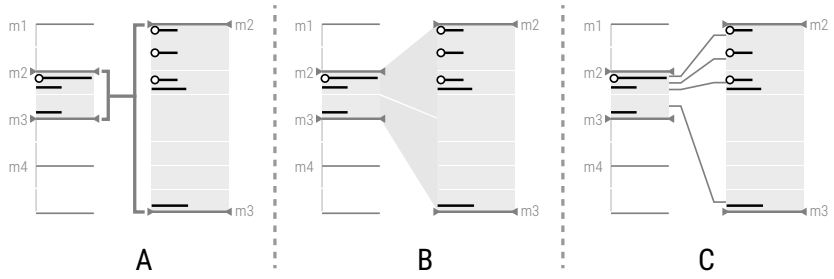

FIGURE 11.9: Comparison of the three different approaches for the visual connection between both timelines: connecting the ranges (A), using shades (B), and connecting with lines (C).

For each temporal block, we represent separately the number of transactions with and without changes. We aimed to maintain the visual rationale applied in the Main View, however, as the timeline has a smaller area, their distinction needs to be simplified. Both types of transactions are represented with a line and transactions with changes are complemented with a circle at the beginning of their lines. To represent quantity we use the length of the line. Also, we always draw the line representing transactions with change in the first half of the block and below we draw the line representing transactions without change. All lines are coloured in black unless one or more transactions of the same type in the corresponding block is considered fraudulent. In this case, it is coloured in red. We decided to apply the colour directly to the line, and not to visualise the percentage of fraudulent transactions with, for instance, a stacked graph, since we wanted to highlight every time block with fraud. For example, if in 20 transactions only one was considered as fraud, it would be imperceptible, since the area to draw the stacked graph is small. This would lead the analyst to miss such cases, which would negatively impact the detection of fraud.

To visually connect the overview and detailed timelines, we tested three different approaches empirically: (i) we isolated with straight lines the areas which connected both timelines; (ii) we connected the timeline's blocks through a shading area; and, (iii) we used a line to connect the corresponding blocks (Figure 11.9). In the tests, the analysts referred that the latter approach was more perceptible, and that it was easier to understand how a block from the overview timeline was subdivided into the blocks of the detailed timeline. Also, with these first tests, we could attest to the need to show the number of transactions (independently of the type) close to each block, so the analysts could easily relate different blocks (Figure 11.10).

The clusters of the main view are represented in the timeline with a line that connects every block where the clustered transactions occur. Additionally, we highlight the blocks within the cluster through a black outlined rectangle (Figure 11.2, detailed timeline top).

Finally, to enable the perception of the number of months being

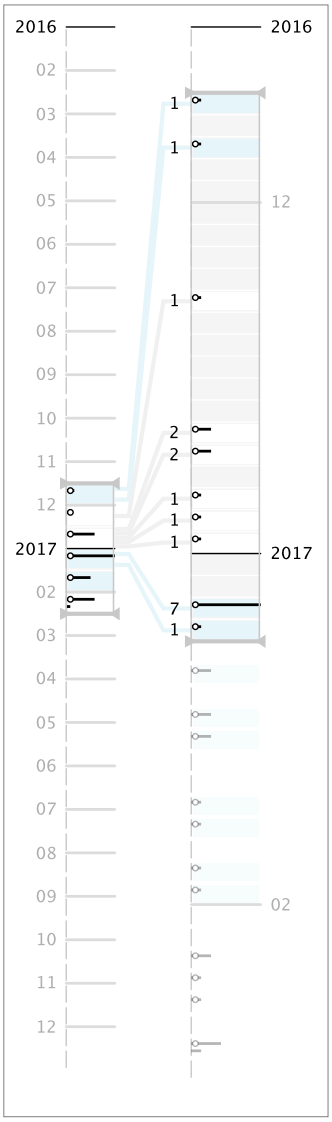

FIGURE 11.10: The analysts have access to the temporal arrangement of all transactions (horizontal bars) in two vertical timelines, where they can specify the desired period of time for further analysis.

represented and to ease the connection between the transactions in the timeline and their representation in the Main View (Figure 11.2), we enhanced the visual connection between the transactions in the main view and in the timeline by alternating between a shade of light blue and white under the transactions from different months.

Interaction

In both timelines, the analyst can select the time range to visualise. By defining the time range in the first timeline, the time granularity and range in the second timeline are adjusted accordingly. By defining the time range in the second timeline, the analyst directly defines the time period for the Main View. We implemented three mechanisms to manipulate the time range: (i) through an upper marker; (ii) through a lower marker; and, (iii) by dragging both markers. To avoid clutter in the Main View, we restricted the total number of transactions that can be visualised. This means that the time range in the second timeline is conditioned by the number of transactions that can be visualised in the Main View. To see the remaining transactions, the analyst must drag the time range to the intended time block. To emphasise in the timeline the transactions represented in the main view, all blocks which are not being represented have their saturation decreased.

The analyst can mouse over a block to get additional details concerning a specific temporal block, such as the period of time of the corresponding block, the total amount spent, the average amount spent, and the rate between fraudulent transactions and the total number of transactions.

11.6 Usage Scenario

In this Section, we aim to analyse how ATOVis can aid in the analysis of fraudulent transactions and the detection of suspicious behaviours. We also aim to describe how ATOVis can be used and how effective it can be in providing the analysts with an informative look at the data patterns. We choose three different cases so it is possible to perceive how the visualization model can represent different behaviours. In the first case, no fraudulent transaction was identified by the ML algorithm, and in the second and third cases, fraudulent transactions are highlighted by the ML algorithm. However, the third case differs from the second in terms of fraud pattern. Hereafter, we discuss the findings through the aid of ATOVis.

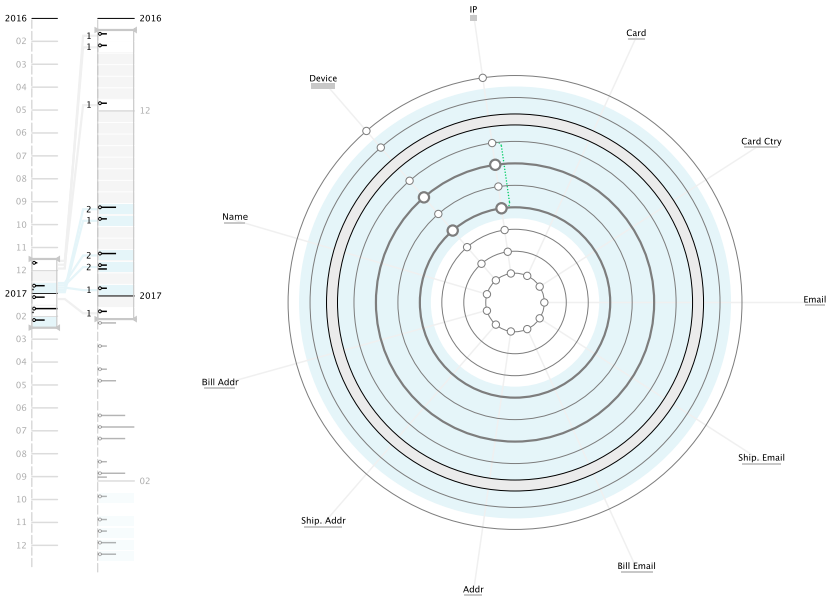

FIGURE 11.11: Screenshot of the client A visualization. It is possible to perceive that this client's transactions were not considered as fraud. Also, in the majority of transactions, the IP and Device are the only attributes that change, which is considered normal.

► CLIENT A: In the first usage scenario, we analyse a client who showed no fraudulent transactions, meaning that the ML system has not identified any fraudulent behaviour. By analysing the first timeline, we can see that this client shopping data occurs between middle November 2016 and middle February 2017.

When analysing the first transactions (Figure 11.11), we can see that the attributes are in their majority equal between transactions, with the exceptions of the device and IP—as can be seen by the circles positioned on that attribute's angle. These changes can be seen as a normal behaviour as clients tend to shop on either the smartphone, computer, or other devices. We can also see that there is a repetition of the IP number, perceived by the green dotted line connecting the two circles in the IP axis. This can also be a common behaviour as clients can shop in either their homes or workplaces. We can see in the visualization, a group of transactions with no attribute changes, represented by the grey thick circle. Through the analysis of its interior pattern, we can determine that there was a small number of transactions as the pattern is not complex. By hovering over that pattern we can see that there was only one transaction with no changes between the transactions with changes.

Given that most transactions occur in December this type of behaviour is common, as clients tend to shop more throughout that month. Also, this client reveals a small number of transactions. However, we can see through the timeline that in January there is a growth in products bought, probably due to returns, but in general, there was no suspicious behaviour in this client's transactions.

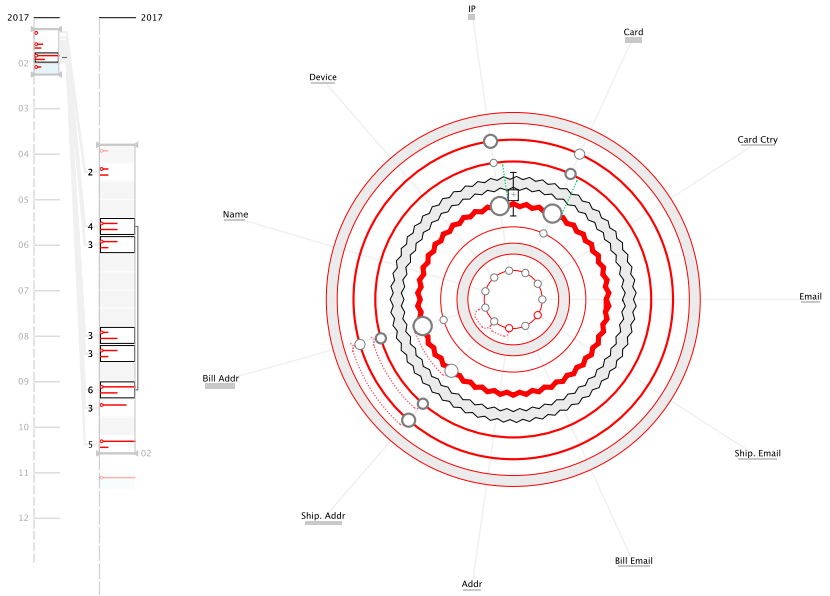

FIGURE 11.12: Screenshot of the client B visualization. With this visualization, it is possible to understand instantly that the client performs fraud in the majority of his/her transactions.

► CLIENT B: In the second usage scenario, we can easily perceive that the ML system labelled the majority of transactions as fraud—lines coloured in red (Figure 11.12). Also, with the aid of the timeline, we can see that all transactions occurred in the same month, January 2017. As in the previous example, there is not a large number of transactions. However, all types of transactions are marked as fraudulent, even the ones with no attribute changes, as we can see by the red lines in the second timeline. When analysing the visualization in the Main View and hovering the representation of a cluster of transactions, we can see in the details area, that in this cluster there are 16 transactions, that share the same pattern—there is a transaction with change, then a small number of transactions with no change, and then this pattern repeats itself. This type of behaviour can be visualised if the client clicks on the cluster to expand it. By looking at the Details Area, we can see that in 10 days the client performed 19 transactions, in which the attribute shipping address had only two distinct values, but the billing address, IP and Card had 7 distinct values. This type of behaviour can be seen as suspicious.

When clicking on the Details Area to visualise the table with the detailed information, we can see that the values of those transactions were around the same amount (€200) and their fraud scores were high (Figure 11.13). Also, all the transactions were declined, either automatically or manually. This is a probable case of ATO, in which the hacker maintained the shipping address but used several billing addresses and card numbers. This type of pattern may also indicate the testing of a bot which tries consecutively to buy products around the same price range with the account of a legitimate client. Finally,

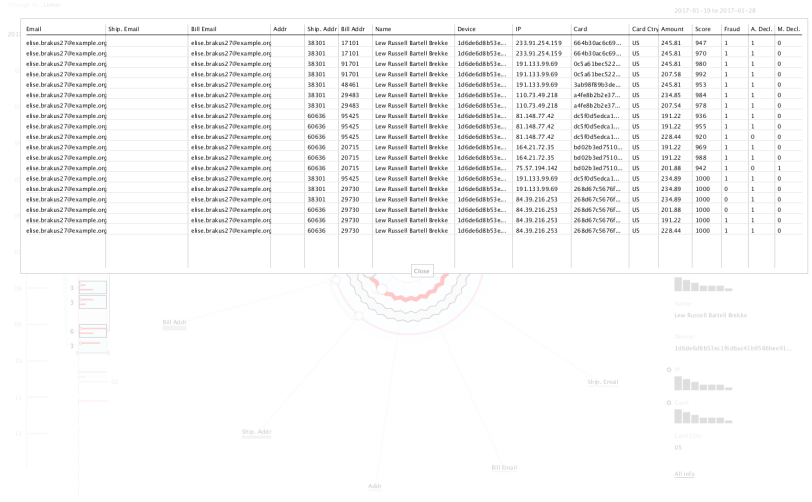

FIGURE 11.13: By clicking on the *all Info* button, it is possible to view all details arranged in a table. With this table, it is possible to analyse client B's continuous purchase of goods with similar values.

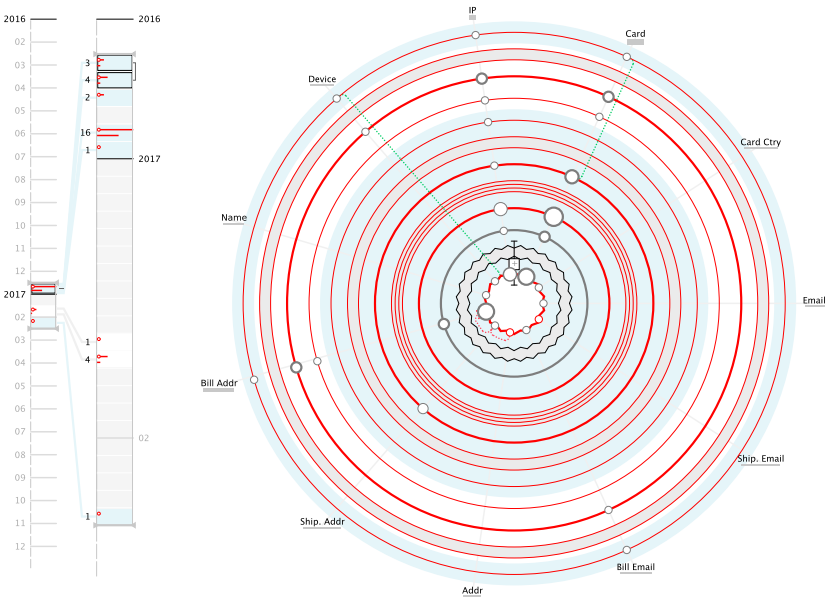

FIGURE 11.14: Screenshot of the client C visualization. This client also performs a set of fraudulent transactions. These transactions occur mainly during December.

with this usage scenario, we could perceive how the visual clustering of patterns of transactions can generate a compact visualization that is still able to summarise the transactions.

▶ CLIENT C: In the third usage scenario, the transactions occur between December 2016 and February 2017 (Figure 11.14). However, the majority of the transactions occur in December. By looking at the second timeline, we can see that nearer the end of the year the client makes sixteen transactions, which can be considered as a high number of transactions for a short period of time. By looking at the visualization, we can see that the client does not have a common and periodic behaviour, as in the previous case. This can be assessed by the reduced number of clusters. Also, there is only one small cluster in the beginning which comprises four transactions.

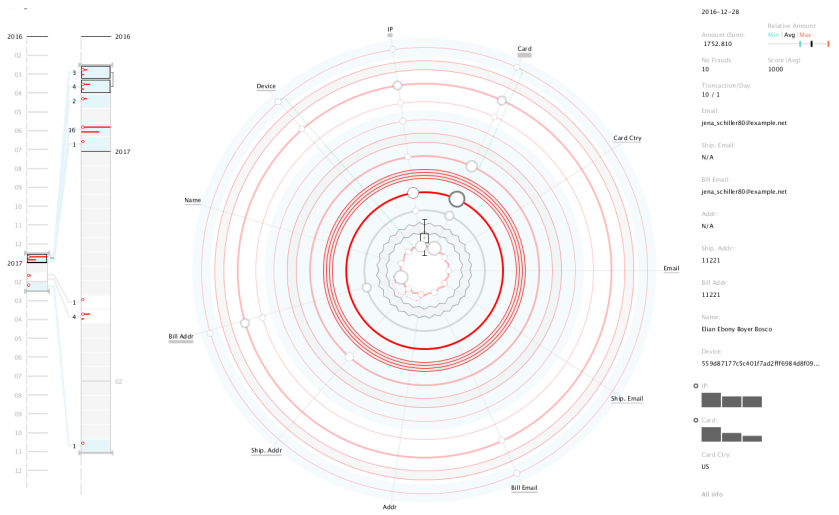

FIGURE 11.15: By analysing the Details of the clustered transactions, it is possible to see that client C performed 10 fraudulent transactions in the same day.

By looking at the visualization model (Figure 11.14), we can see that the client makes a reduced number of changes along the transactions. The card and IP are the attributes that change the most. By hovering over the day with the highest changes in the card attribute (i.e., with a bigger circle) we can see in the Details Area that in the five transactions that occurred on the same day, the client used three distinct cards. By analysing that day, in which there are transactions with and without changes, we can see that the client performed a total of ten transactions, all considered as fraud (Figure 11.15, details area top). When accessing the table through the Details Area, the transactions have three different amounts, which may indicate the attempt to buy three different objects. Also, we can see that the client attempted to buy each object at least three times, with different cards. This may represent a more manual attempt to improperly use the account of one client to test different cards and determine which one could be used to buy products.

11.7 User Study

We performed a combination of formative and summative evaluations [6, 89] to test the effectiveness and efficiency of the proposed visualization models, and gather feedback on how to improve the tool. Each individual session contained: (i) a scripted walk-through of the ATOVis application; (ii) a set of tasks to complete using screenshots of ATOVis; (iii) a qualitative study in which the participants were encouraged to describe the visualization models and interact with the tool; and, (iv) an open-ended questions about ATOVis. The tasks in (ii) were defined to validate the model and not the interface. They were designed to determine the usefulness of the system based

on its design—identifying the effectiveness of the models and visual encodings in the rapid detection of fraud, and assessing its suitability for the second phase of analysis (described in Section 11.1.1). The qualitative study aims to assess the complete functionality of filtering temporal intervals of interest, investigating the transactions in detail and assessing specific attributes through the selection of transactions and attributes. In short, we aim to analyse whether the analysts can make correct decisions and which insights can be acquired through the visualization models, assessing its suitability for the third phase of analysis (see Section 11.1.1).

11.7.1 *Participants*

Two groups of participants were involved in the test: experts in fraud analysis from Feedzai, who were not involved in the development of the tool; and experts from other fields of data science, recruited from the University of Coimbra, in Portugal. With the participants with no background in fraud detection, we aimed to study the self-explanatory aspects of the proposed models, and whether someone with no expertise could still identify fraudulent cases. This is particularly important as in Feedzai beginners at fraud detection have less knowledge on fraud patterns, and by testing the system with non-experts we also address the ATOVis effectiveness in representing ATO patterns. No participant, independently of the group, had previous knowledge of the tool or visualization models.

The group of experts consisted of 5 analysts. Their average working experience in fraud analysis is 5 years, and on average, the analysts had none to little interaction with Information Visualization on a daily basis. The second group consisted of 11 participants from diverse fields, such as ML, Information Visualization, and Design, with no background in fraud analysis. On average, these participants have more interaction with Information Visualization, two of whom work and interact with visualization every day. We performed the tests in similar ways with both groups, as we aimed to perceive if the visualization models could be interpreted both by experts and non-experts in fraud detection. The only difference between the tests is that the interaction with the tool was only performed by the analysts, as this is more focused on understanding the analyst rationale while searching for fraudulent patterns.

11.7.2 *Methodology*

At the beginning of each user testing, a small workshop, introducing the visual variables and visualization models, was held. It had a

duration of approximately 15 minutes. Afterwards, to enable the participants to get familiarised with the models, they were given a set of tasks to complete. These tasks, defined according to the tasks T1-4 (detailed in Section 11.3), concerned the comprehension of the models and were grouped depending on what they aimed to validate: **Group 1**– the distinction between transaction with and without changes [**T1**, **T4**]; **Group 2**– the relations between transactions through their attributes [**T2-3**]; **Group 3**– the clusters [**T1**]; and **Group 4**– the understanding of the timelines [**T1**]. All groups of tasks are balanced between tasks in which the participant has to count the number of occurrences and tasks of yes or no answers. This part had an average duration of 10 minutes, which includes the time of reading the task.

For the second part, four clients from the dataset were selected and their transactions were shown through ATOVis. Their data had different types of complexity and different patterns: (i) no fraud (Client 1); (ii) with fraud (Client 2); and, (iii) with fraud, but with few or no transactions coloured in red (Client 3 and 4, respectively). Only the expert analysts performed this part of the test. The analysts were asked to explore and analyse the visualization, explain out loud what they were seeing at each moment of their exploration, and refer if the client was fraudulent, non-fraudulent, or suspicious. The participants took on average 3.2 minutes to conclude each analysis. In the end, the participants were asked to rate the difficulty in the analysis and the certainty of their answer.

Finally, in the third part, the participants were asked to give feedback on the visualization, with special emphasis on the efficiency of ATOVis, its learning curve, and the differences between the radial and linear models. The non-experts group was asked to rate the visualization model as a whole, and the analysts were asked to rate the three components of the tool (i.e., timeline, radial, and linear models). We recorded the audio from each test so we could analyse each session afterwards. The user tests had an average duration of 37 minutes.

We created two different tests to compare the linear and radial models effectiveness. In test A, we started with the linear model, and in test B, we started with the radial. Then, we alternated between the two models for the following tasks. Both tests have the same tasks and order[2].

[2] A PDF of the User Test can be found in the following link: https://cdv.dei.uc.pt/cmacas/ATO_UserTest.pdf

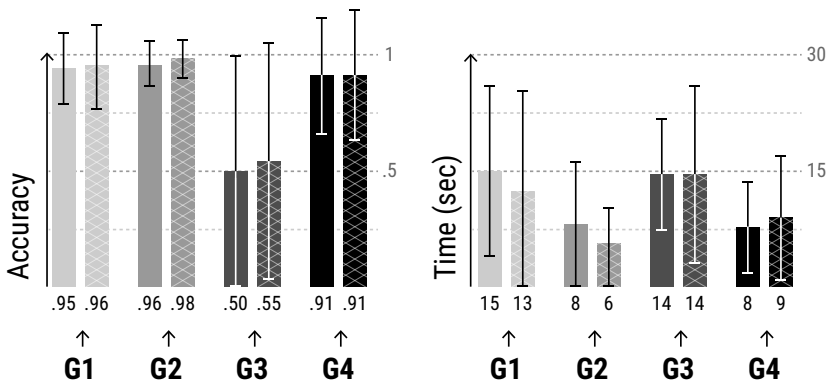

FIGURE 11.16: Accuracy, duration, and respective standard deviation values for all groups of tasks: transactions (**G1**); attributes' connections (**G2**); clusters (**G3**); and timeline (**G4**). The solid and scratched bars represent the experts and non-experts in fraud analysis, respectively.

11.7.3 *Results*

Hereafter, we describe the results of the three parts of the test, detailing the accuracy and time values for each group of tasks; the analysis of ATO patterns; and the feedback received.

Validation of the ATOVis visualization models

In Figure 11.16, it is possible to see the accuracy, duration, and respective Standard Deviation (SD) values for each one of the participant groups and groups of tasks. In general, most participants had no difficulty in completing the tasks and took on average less than 11 seconds to complete them. Although it is not possible to determine their statistical significance, due to the difference in groups dimension, our analysis revealed that, despite the different backgrounds and levels of expertise, both groups attained comparable performances in terms of accuracy and time. A possible interpretation for this result may be the following: on the one hand, the experts in fraud detection are used to analyse this type of data through their own tools, as such they may feel an initial unease and difficulty to adapt; on the other hand, non-experts have no preconceived notions regarding how this sort of data should be analysed and, on average, are more familiar to information visualization tools. Thus, the initial edge of the experts appears to be counterbalanced by the familiarity of non-experts with Information visualization and an unbiased approach. Due to the lack of statistical differences, and for the sake of parsimony in the presentation of the results concerning this issue, we analyse the aggregated results of both groups.

The group of tasks with the lowest accuracy refers to the cluster's group (**G3**). During the test, some participants had difficulties in understanding the concept of a cluster. As the participants were encouraged to think out loud and expose their difficulties, we could notice that some mistook the clusters with the aggregation of trans-

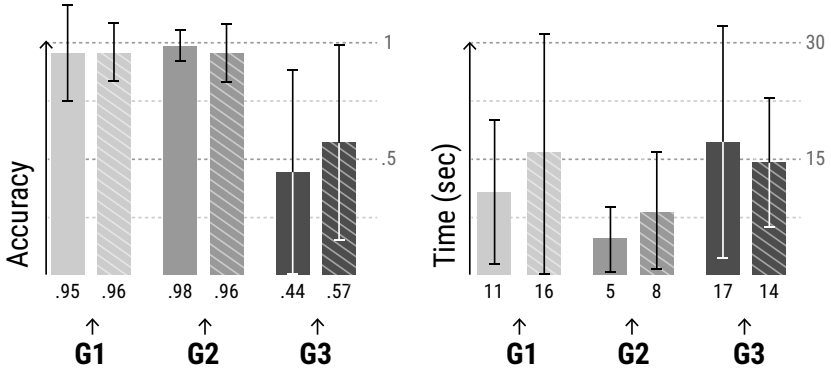

FIGURE 11.17: Differences in accuracy, duration, and respective standard deviation values for the tasks related to: transactions (**G1**), attribute's connections (**G2**), and clusters (**G3**). The solid and scratched bars represent the linear and radial approaches, respectively. **G4** is not represented as it is concerned with the timeline.

actions by day. More specifically, when they observed a sequence of transactions with change and transactions with no change, they thought of it as a cluster, although it should be seen as a single transactional behaviour. Additionally, when they were asked "How many clusters does the visualization have?", 45% of participants answered 2 clusters, while there was only one. In our understanding, the participants were interpreting the two types of transactions as two independent clusters.

For the other three groups of tasks (**G1**, **G2**, **G4**), the average accuracy is high. Concerning the duration, the group of tasks regarding the transactions (**G1**) is the one that took more time to complete. This can be explained by the time needed to answer the question "How many days with transactions with change are there in the visualization?". This task took on average 25 seconds to answer with a SD of 17 seconds. The following tasks in this group (**G1**) took on average 8 seconds. As the participants had to count the number of transactions in the visualization, this task proved to be more laborious and they needed more time. Additionally, this was the first task of the test, so the participants may have needed more time to familiarise themselves with all the visual encodings.

Concerning the connections between attributes (**G2**), the majority of the participants answered the questions with no difficulties and performed them in less time when compared to the other groups of tasks. The participants had no difficulty in completing the tasks about the timeline either (**G4**). The only tasks that added some difficulty were the ones related to the comparison of the lengths of the transaction lines and the identification of the type of transactions that occurred the most. Some participants mistook the latter with the former, lowering the overall accuracy for this group of tasks.

► LINEAR VS RADIAL Regarding the comparison between the two visualization models, Figure 11.17 shows the accuracy, duration, and

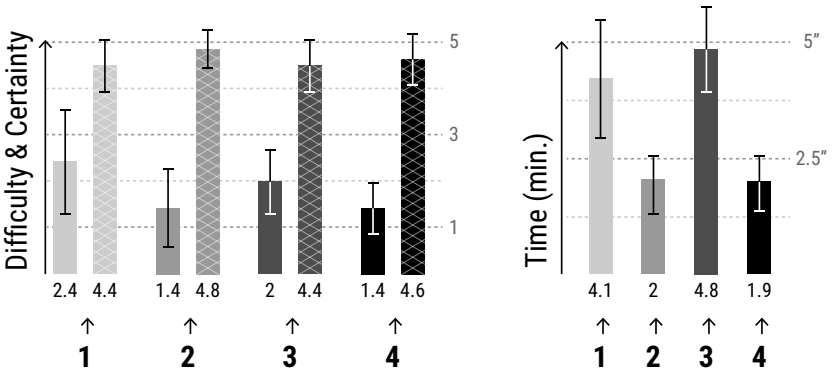

FIGURE 11.18: Ratings on difficulty, certainty, and duration (averages and standard deviations). Difficulty (bars with no pattern) and Certainty (bars with cross pattern) are rated from 1 to 5, 0% to 100% respectively. Time is mapped between 0 and 5 minutes.

respective SD values for each group of tasks related to both models. When comparing the accuracy in each group, both models have similar accuracy values, except for the group of clusters. Additionally, the majority of the participants took more time to complete the tasks related to the radial model. When comparing the tasks related to the clusters (**G3**), the accuracy in the radial model is highest and the elapsed time is lower than with the linear model. Concerning the tasks related to the attribute's connections (**G2**), most participants took more time to complete them with the radial model. Finally, most participants referred that the linear model was more familiar to them, as it could be compared to the way they analyse spreadsheets. This indicates that, in general, the more familiar the visualization model is, the faster is the performance of the participants, which is an expectable result.

ATOVis Patterns

In the second part of the test, fraud analysts had to analyse the transactions of 4 different clients through ATOVis (Figure 11.18). These analyses were performed with the radial and linear models. For both models, the average score given by the analysts concerning the difficulty of analysis was 1.8 (low difficulty). Although the time average of completion for the radial model is slightly lower than the linear (3.3 and 3.1 minutes, respectively), the results were not statistically significant (p-value: 0.76). As the differences in terms of difficulty, certainty, and duration between the models were not significant, and their analyses were performed in equal manners, we further discuss the results independently of the model.

From the selected clients, client 1 and client 3 were the ones who required more time of analysis and triggered the analysts to explore in more detail their transactions, in search of suspicious behaviours. As client 1 had no fraud patterns, the analysts were less certain of their answers and found it more difficult to analyse. This can be

explained by the bias that the test may have caused, as they were expecting to see fraudulent activities. Notwithstanding, all analysts correctly classified the case as non-fraudulent. In the case of client 3, the suspicious behaviours were not visible in the beginning. The analysts used the timeline to explore the rest of the dataset and found suspicious activities, such as the consecutive change of attributes within a single day, followed by a high rate of transactions with no changes. For client 2, the analysts could see instantly fraud, as most transactions were coloured in red. However, after some interaction with the tool, the analysts could be more precise about the fraud pattern. It was obvious the consecutive changes in the card details, which most analysts referred to as a carding pattern (constant test of stolen cards in new or stolen accounts). The case of client 4 was also rapidly classified as a BA pattern as it showed a high amount of transactions in a single day (through line thickness).

Through the analyst's interaction with ATOVis, it was possible to detect similar uses of the tool functionalities. Most analysts started their analysis by looking directly at the main model, discarding the timeline. When they finished their first analysis, they had a mental image of the initial behaviour of the client and searched the timeline for periods of time with: more transactions and/or fraud. If no period of time stood out in the timeline, they started to think of the case as non-fraudulent. However, they still looked for the main model while sliding the timeline. When confronted with fraudulent transactions, the analysts searched for attributes that had changed and analysed their values, especially the country of the card and the differences in attributes of the same domain (e.g., shipping/billing/account email).

From the attributes shown in the visualization, the less analysed was the IP, as it can have a high rate of changes in non-fraudulent cases. The analysts pointed out that a more relevant field would be the IP country. Also, to show the hours of each transaction would enhance the analysis, as it eases the distinction between a BA and a manual attack.

In summary, all fraud analysts could describe the clients' patterns correctly, explaining their transactional behaviours. Through these results, we can argue that the models are capable of representing ATO patterns, as well as other hidden fraud patterns, also represented by the consecutive change of attributes (e.g., BA, carding).

Feedback

The analysis of the feedback allowed us to conclude that ATOVis was well accepted and most participants considered it a relevant

tool to detect fraud and to facilitate the reading of e-commerce data. Regarding the learning curve, most participants stated that it was a lot of information to memorise at the beginning of the test, but as the test continued, the reading of the visualization got significantly easier and intuitive. Comparing the ratings of the radial and linear models, most analysts referred to the linear model as easier to learn and analyse the data. However, both models were rated equally in terms of their ability to aid in analysing the data (4 on a scale from 1 to 5).

During the tests, the analysts pointed out the functionalities that they considered the most important for their analysis process. The definition of a single area in the canvas to show the details was an important aspect of the tool, as they easily learned where to look when interacting with the visual elements to see more details. The functionality of showing the list of attributes used in each field also enabled them to perceive how many attributes were used and their specific values. The lines which connect each block of time from the overview timeline and the details timeline were referred to as essential for them to understand exactly how both timelines were connected. The timeline was also pointed out as a good feature to summarise the data and was defined as an important auxiliary tool to oversee the data.

11.8 Discussion

The interaction with fraud analysts, the specification of the task abstraction for ATO patterns and the design requirements aided in a better definition of which information to represent and which visual encodings to apply. This allowed us to derive the main pattern to look for in ATO, which is characterised by the consecutive change in the transactions' attributes, focusing on the visual highlight of such behaviours. This was seen by the analysts as a good feature to ease the detection of ATO.

Based on the user tests with two distinct groups (experts and non-experts in fraud detection), we can argue that, with ATOVis, both less and more knowledgeable participants can be equally accurate while performing judgement tasks, which diverged from the results of studies such as Cardinaels [46]. The most experienced could drill down faster in certain scenarios (e.g., discovering fraudulent patterns), while the less experienced reasoned more slowly about the transactions, but could arrive at similar conclusions. Independently of the expertise, most participants referred to multiple changes in the

IP attribute as a client shopping in multiple places, discarding fraud, whereas when seeing multiple changes in the card attribute followed by a high rate of transactions with no change, all participants referred to this as fraudulent testing of multiple cards. Therefore, we can state that, regardless of the expertise, all participants could focus on the analysis, understand the visualization, and easily detect suspicious transaction patterns.

The tasks defined for the first part of the tests enabled us to assess the interpretability of the visualization design and comprehension of the transactional behaviours and fraud patterns. We could attest how easily the participants understood the model and the transactional behaviours in a short period of time. As such, we can confirm the usefulness of ATOVis for the rapid comprehension of data, allowing the quick identification of fraud and the subsequent action to stop fraud. The interaction in the second part of the tests, allowed us to confirm the usefulness and intuitiveness of the tool's components. The analysts had no difficulty in interacting with the tool and used all functionalities during their exploration to drill down suspicious transactions and detect fraudulent activities. They also could arrive at conclusions in a short period of time (3.2 minutes on average). Although we could not test the timings concerning the company's current system, the analysts referred to their timings with ATOVis as a good improvement on their current tools (i.e., spreadsheets and web database), which commonly require more time to perform similar tasks. Also, the variety of behaviours within the different cases and the accuracy of the analysts' answers, allow us to conclude that the tool can aid in the detection of various suspicious cases, improving the analysts' decisions.

Concerning the comparison between the linear and radial models, the linear model was seen as more intuitive (due to the familiarity with tabular information), and the radial promoted more interest during the analysis. One analyst stated that the radial model was better suited to read the attributes changing over time, and after some interaction, it also got easier to read the overall behaviours. The radial model was also characterised as more informative, as it condenses more information. Therefore, we can conclude that the radial model is more suitable as an overview tool for phase 2 of the analysis workflow (see Section 11.1.1), and the linear model more suitable as an exploratory tool for phase 3.

Functionalities, such as the visual distinction between different months through interspersed blue shades, the use of red to highlight fraud, and the aggregation of transactions by day and type (easing the perception of general patterns), were well-received and under-

stood. In general, the visualization models were seen as visually appealing, which enticed the participants to read the visualization and extract meaningful information. On the other hand, the clusters of transactions and their representation fell short in what concerns readability. Nevertheless, most participants stated that, with more time to assimilate all information, it got easier to distinguish the representations, and the clusters were referenced as a good feature, to compact the visualization and, at the same time, to represent clearly similar behaviours.

The majority of the fraud analysts referred to the usefulness of ATOVis as an important aid in their daily work. They referred that the tool can give a complete picture of the transactional history, which does not happen with the tools they usually use. Additionally, the timeline is referred to as easy to use and as a functional mechanism that presents a clear overview of the distribution of the transactions over time. This overview can assist the analysts to synthesise the client's behaviour, highlighting the changes in behavioural patterns. In terms of time, and due to the limited access to the data and human resources (i.e., fraud analysts), it was impossible to measure properly how well ATOVis improves their current method. However, the analysts stated that, with our model it was much easier and faster to perceive the behaviours of specific clients and detect if a fraudulent transaction was an isolated event or a common behaviour of the client, than with their current tools.

As future work, some visual encodings can be improved, such as the highlighting of fraudulent attributes through the red filling of the circles, and the representation of clusters. The functionality to select different months and visualise them side by side should also be addressed. We intend to test the scalability of the presented models by measuring the minimum and the maximum number of attributes that can be used without impacting the overall effectiveness of the visualization. A future validation test to be conducted after the full implementation of the tool in the company's workflow is to analyse the number of positive and negative cases detected by the analysts in contrast to the outcomes of the analysis of the same cases through their current method (i.e., the use of spreadsheets) in order to test the detection accuracy. Overall, the present work allows us to conclude that ATOVis is effective and efficient in detecting fraudulent patterns.

12

Conclusions

In this part of the thesis, we explored visual solutions for the representation of temporal patterns in the finance domain. We presented our design choices for two visualization tools that aim to represent the typical behaviours and emphasise atypical ones in bank and e-commerce datasets. Both are user-centred visualization tools, were developed in the context of a partnership with Feedzai—a Portuguese Fraud Detection Company—, and were intended to be implemented in the workflow of the company's fraud analysts. The company's main aim for the visualizations was to promote an efficient analysis of the two datasets and emphasise suspicious behaviours, so that the analysts could detect them more rapidly.

In the VaBank tool, we represented the temporal patterns of bank transaction data. In collaboration with Feedzai, we were able to define the main requirements and tasks that would enable our tool to improve their analysis in comparison to their current tool—spreadsheets. Thus, our visualization models focus on: (i) the visual representation of the transaction's characteristics through a glyph visualization; (ii) the temporal visualization of the transactions; (iii) the characterisation of the transactions' topology through a SOM algorithm; and, (iv) the projection of the SOM results into a matrix and a force-directed graph.

We validated and compared the different visualization components of the tool through formative and summative evaluations with experts in fraud detection. Through these tests, we could assess the effectiveness of the tool on the characterisation of the transactions. The analysts were able to properly analyse the visualization and detect different behaviours in different bank clients. In summary, the results showed that the tool was well received by the analysts and that it could enhance their analysis.

Overall, we could perceive the impact that the analysts had in the development of VaBank. In the initial meetings, we could study

and understand their analysis process. In the following meetings, we could understand which tasks were the most important. From these meetings, we could define the main requirements and develop the tool accordingly. In VaBank, we also created complex glyphs that can represent nine different data attributes. To do so, the analysts aided us to rank each attribute by importance and define three levels of readability. The most relevant level represents the most relevant data attributes and uses the most impacting visual variables in terms of visual perception so they can be read instantly. The following levels have less importance and, therefore, can be only read after a more detailed analysis. With this system we were able to improve the comprehension of the transactional data. Nonetheless, complex glyphs have some limitations in terms of memorability and, for this reason, a caption should be provided to the users.

We contributed to the visualization domain in finance with a tool that focuses on the characterisation of bank transactions, on the representation of the topology of the transactions and, consequently, on the highlight of uncommon behaviours. By creating a tool that allows the simultaneous visualization of the transactions along time—emphasising the ones with higher amounts—and their topology—emphasising the typical behaviours—, we were able to promote better and faster analysis of atypical transactions and suspicious behaviours. In conclusion, the presented work demonstrates that VaBank is effective and efficient for the analysis of bank data and in the detection of suspicious behaviours. Also, the characterisation of transactions with complex glyphs can aid in the understanding of transaction patterns and facilitate the analysis of the overall data.

In the ATOVis tool, we aimed for the rapid analysis and detection of ATO. In collaboration with Feedzai, we were able to define the task abstraction for the detection of ATO cases. This led us to the definition of the main design requirements for the implementation of a visualization focused on consecutive changes in transaction attributes. We implemented two models to visualise the transactions and a multiscale timeline to enable the overview of all data in a temporal context. We validated ATOVis and compared both models through formative and summative evaluations, both with experts and non-experts in fraud detection.

The user tests showed that our approach to focus on the changes among transactions eased the finding of ATO characteristic behaviours. We could also attest that to analyse transactional data, visualization models that represent an overview of behaviours and emphasise their relations can enhance the detection of fraud. Such models can be used to detect single fraud patterns, as well as to be a basis for

improving ML algorithms with new types of patterns. Ultimately, our tests revealed that visualization models which enable analysts to have an overview of the data and are designed specifically for particular patterns of fraud can significantly improve current approaches of fraud detection within fraud prevention companies.

Overall, and similarly to the conclusions in Part II, the linear model was seen as more intuitive. However, the radial model triggered more the users' curiosity and interest and was seen as more informative. Additionally, we could assess that the radial model was also able to fasten the reading of attributes over time which are placed in different radial axes. Finally, with the user study, we were able to perceive that complex concepts, such as clustering techniques, were difficult to be completely understood in the beginning. However, we argue that such concepts can be easily learned if the users have more time to interact with the models.

We contributed to the visualization domain in financial fraud with a tool that focuses on the rapid analysis of a particular fraud pattern. In addition to our main contribution—ATOVis—we contributed with: the task abstraction, which was defined with the aid of fraud analysts; the visualization design to detect ATO patterns, as there is no other visualization to detect ATO cases; a multiscale timeline, created to enhance the detection of fraudulent periods of time; and the insights retrieved from the evaluation of ATOVis' readability, interpretability, and usefulness for detecting fraud.

Part IV

RHYTHMS

13

Aesthetics in Time-series

"Creating aesthetically appealing graphs is more than a quest for the beautiful—it has the practical aim of revealing underlying meaning and structure."
— Bennett et al. [19]

This part of the thesis describes our investigation concerning the use of aesthetics in Information Visualization. In this context, we aim to develop aesthetic and luring visualization models that explore the area between functionality and aesthetics. To do so, we base our explorations on the Portuguese consumption data presented in Section 5.3. We explore two different approaches. In the first, we create a visualization model which aims to be an aesthetic experience that lures the user to explore the data while preserving functionality. Additionally, we explore a multimodal visualization in which sound and image are applied to represent the rhythms of the Portuguese consumption. In the second, we further focus on aesthetics. We apply a swarm system to represent the consumption over time, and explore an Interactive Evolutionary Computation (IEC) system to enable the users to evolve artefacts that are more in tune with their preferences.

In this part, we begin by presenting the context of the projects herein presented (Chapter 13). Then, we describe the models (Chapters 14 and 15) and discuss our findings (Chapter 16).

13.1 Context

The ease of generating and storing large amounts of data increased the relevance of data analysis and enabled Information Visualization to gain recognition as an important aid in the analysis and understanding of complex datasets. Information Visualization, with its roots in scientific reasoning, emerged as an analytical tool [105, 197, 277, 278] to simplify, structure, and uncover patterns of interest, proving to be an important resource for business strategy, and scientific and social discovery [277]. However, with the popularisation, and consequent standardisation of visualization tools, the aesthetic aspect of the artefacts was sometimes relegated to a secondary plane.

We consider that aesthetics is an integral part of the design practice and arguably more closely linked to "functionality" than what the contemporary debate suggests. Especially in the context of information design, where communication is the main goal, the aesthetic experience plays a major role in how messages are received and absorbed by the audience [24, 126]. We argue that the quality of visualization is intrinsically related to its ability to amplify cognition (functionality) in a meaningful and enticing way (aesthetics). Although balancing functionality and aesthetics may not be an easy task, both dimensions should have the same level of importance. We also argue that a more experimental and aesthetic approach can be advantageous in luring the user towards the data and in enticing him/her to decipher the information.

It is in the context of an aesthetic exploration that the present works arise. In this Part, we explore the characteristics of the Portuguese consumption in the SONAE's super- and hypermarkets from May 2012 to April 2014—already detailed in Section 5.3 [1]. SONAE allowed us to freely explore more aesthetic approaches and create visual artefacts that can capture the interest of the general public. The main aim of SONAE is to lure the user to explore the data, to be involved, and to understand the impact of the company on society. Therefore, and as the visualization users/stakeholders are not the company's analysts, we intend to create artefacts that, through the simplification of complex datasets, can trigger the users' curiosity. In sum, we seek to adapt already known techniques with special emphasis on aesthetics and functionality, to present complex datasets in a visually engaging way. We intend to: (i) explore the functionality of aesthetics to properly represent patterns; (ii) explore the space between Data Visualization and Information Aesthetics; and, (iii) create models that lure the user to explore the visualization and to gather knowledge.

We explore two models to represent the Portuguese consumption evolution over time. In the first, we aim to emphasise the Portuguese shopping preferences and represent how they evolve through time. We focus on highlighting the rhythm of consumption and times of the year that disrupt the normal consumption patterns. To do so, we first apply a small-multiples technique, and then present it in static and animated ways. Then, we study the multimodal representation of the Portuguese consumption patterns. Here, we focus on the rhythmic nature of the data to create and discuss audio and visual representations that highlight disruptions and sudden changes in the normal consumption patterns. For this study, we present two distinct visual and audio representations, and discuss their strengths and

[1] We discarded the datasets of Part III containing fraud patterns in finance (see Chapters 10 and 11), due to our reduced accessibility to them and due to the data sensitivity, which would not enable us to explore the dataset in a more unconstrained way.

limitations.

For the second model, we apply a swarm system to represent the differences in consumption in the SONAE's departments. In this exploration, we focus on the ability of the emergent visualizations to communicate information while engaging the viewer with organic visuals. We argue that the introduction of Multi-Agent Systems (MAS) can aid in the creation of emergent patterns that convey meaningful information in an appealing way, exploring the boundaries between Data Visualization and Information Aesthetics. We further explore this concept by developing a framework to explore the aesthetic dimension of the swarm model. We rely on IEC to evolve the parameterisation of a visualization model, enabling the user to explore new possibilities to represent the consumption patterns. The developed system can create emergent visual artefacts that are intriguing and aesthetically appealing. Additionally, we show that the framework can create a wide and diverse set of solutions and explore different possibilities.

13.2 Related Work

Information Visualization is usually considered a field reserved for computer science or data analysis specialists [204]. It is a field that brings together the knowledge of distinct fields, such as statistics, HCI, and visual cognition research, to develop visualization models that can convey knowledge to the readers and enable them to detect hidden data patterns [163, 191, 204]. In most Information Visualization projects, the main goal is to optimise the model's functional requirements, such as effectiveness—the accuracy and completeness with which users achieve specific tasks—, and efficiency—the time and computational power needed to complete a task [204]. With this focus on solving functional requirements, some researchers argue that Information Visualization may have been neglecting the potentially positive influence of aesthetics [163, 205].

In the following subsections, we discuss the role of aesthetics in Information Visualization and present the field of Information Aesthetics which focuses on the beautification of information processing [193] to engage the viewers and lure them into decoding the visualization.

13.2.1 *Aesthetics*

As a discipline, aesthetics is concerned with the study of the form, and with the (aesthetic) experience resulting from the perception of the form. The term "aesthetics" is a well-known term and is usually

used to refer to anything visually pleasing. However, aesthetics does not refer uniquely to beauty or vision but to any combination of the senses that causes pleasure in the viewer [162]. A beautiful object may have little or no aesthetic value if it does not provoke thought or create new points of view [162]. Hence, there is functional and communicative value in the aesthetic experience [30].

The idea of a functional object as being completely disconnected from aesthetics and an aesthetic object completely disconnected from function is seen as somewhat problematic [30]. For example, Graham [112] highlights this issue of *form versus function* and explains that in architecture the form cannot easily be separated from its function. Hansson [116] proposes a theory of "aesthetic duality", where design objects can have aesthetic value both for their functional and aesthetic qualities. According to Jordan [137], people are wired to seek pleasure, and, in people's lives, objects that are not merely functional, but also meaningful, can convey pleasure. Findeli [98] refers that it is a designer's job to convey a symbolic value to an object. Through this, the designer "avoids banality and uselessness and makes the object more pleasurable and aesthetic". Hence, aesthetics can be used as a means of appealing to users that may have never considered visualization before, in order to attract their attention, encourage personal involvement, and allow for more profound, long-term impressions [163]. Thus, in Information Visualization, the higher the aesthetic value of the model, the more engaged the viewer tends to be in trying to decode its meaning [162].

Aesthetics has been studied previously [16, 41, 43, 97] and has already been discussed as a key factor in several subfields of Information Visualization, such as in ambient visualization [192], graph drawing [235], and user experience research [164]. In Information Visualization, there are some divergent opinions regarding the role of aesthetics in the Information Visualization field, typically focused on developing high-end applications for research and commercial enterprises, and in Information Aesthetic approaches, more focused on developing experimental visualizations for non-expert audiences [30, 204].

On one side, traditional practitioners may be concerned that "aesthetics" will undermine the functional or analytical goals of visualization artefacts [170]. In this perspective, aesthetics is seen as a mere subjective decoration that distracts the users from the analytical and objective purpose of information transfer [30]. In this matter, Jorge Frascara [103] referred that communication is the most important aspect in visualization and that the "aesthetic quality of a design does not determine its overall quality", making the aesthetics of the final

result to be placed in a secondary level. Roger Scruton [255] also defines aesthetics as the "choices remaining when utility is satisfied", with these choices relating mainly to the surface appearance of the object. In short, traditional practitioners believe that the medium should present the data as objectively and neutrally as possible, without focusing on the aesthetic side [30].

Other practitioners argue that the aesthetic principles of visual design should not be treated as superficial or less important, but rather be embraced as a necessary aid to improve the understandability and accessibility of information communication [30, 52]. Aesthetics is an integral aspect of the design practice, and for this reason, it is more closely linked to functionality than contemporary debate suggests [30]. Lau and Vande Moere [163] referred that by focusing solely on the effectiveness and functional part of the visualization, the positive influence of aesthetics on task-oriented measures is neglected. Ben Fry [106] also defends that the aesthetic principles of visual design should not be treated as superficial or less important in Information Visualization. This emphasis on the positive influences of aesthetics can also be seen in earlier works by theorists such as William Morris and John Ruskin, which emphasised the importance of beauty in design during the mid-nineteenth century [30]. In more recent years, Edward Tufte [278, 279] is another practitioner of the idea that "communication can be both beautiful and useful". Hence, aesthetics should be seen as a necessary aid for improving the understandability and accessibility of information communication [30].

Some researches even refer that the total absence of aesthetic concerns in more neutral approaches to data visualization may hinder critical engagement or reflection. For example, Borgmann [29] refers that an exaggerated emphasis on functionality may lead to less engaging artefacts and therefore less meaningful visualizations. Richard Wurman [307], also argues that absolute accuracy in data visualization is not strongly correlated to data understanding, which may be considered the ultimate aim of all information. In this sense, it is important to recognise that aesthetics are not superficial or non-functional, and that subjective expression might play a part in effective information visualization practice. Hence, to stimulate the understanding, aesthetics can be applied to trigger and retain interest and make information memorable and meaningful [30].

Due to the subjective nature of aesthetics and the difficulty in defining the aesthetic value of visualizations, measuring its influence on visualization might be challenging [30]. However, some studies have already been done and were able to conclude that aesthetics could improve the efficiency and effectiveness of task performance

by reducing the completion time and error rate [52, 162]. Also, aesthetics could improve the level of user patience, by making users more closely to attractive visualizations [52]. Therefore the more aesthetically a graphic is perceived, the longer the viewer will try to decode the meaning of it or extract a certain information [162]. This supports the theory of Norman [216]—originally formulated towards industrial design—that when the user finds a positive affection towards an object, our brains are encouraged to think creatively and solve any problem the object might present. In Information Visualization, this is reflected in the understanding and analysis of the data representation. Aesthetics were also tested in regards to its positive influence in diminishing the latency in task abandonment and the time taken for erroneous responses [52]. Recent empirical results also show that "visual embellishments" in the form of meaningful metaphors do not affect the interpretation accuracy, and influence positively the memorability of simple infographic charts [205].

Aesthetics has already proved its value that goes beyond the experiential or superficial, and that it is useful in supporting functionality [205]. In the context of visualization, aesthetics can be useful to lure the user to engage with the data, enabling a more effective communication of information [205]. In this sense, aesthetics plays an important role in the system's overall attractiveness, as a significant incentive for initial use. Thus, Information Visualization only benefits by embracing aesthetics as a persuasive medium [52].

13.2.2 *Information Aesthetics*

With the aid of the democratisation of the Information Visualization field, through the accessibility of data and easy-to-use visualization tools, designers started to explore the aesthetic components of Information Visualization as a means to create communicative and attractive visual artefacts, expanding the conceptual horizon of Data Visualization to an artistic practice [30, 205, 287]. This increased the interest in the aesthetic value of visualization and led to the emergence of *Information Aesthetics*—a subcategory of visualization with a strong focus on visual appeal [30]. Works on Information Aesthetics are at the intersection of a more functional Information Visualization and the Fine Arts [204]. Whereas in some Information Visualization works, the visual quality is considered secondary [106], in Information Aesthetics the visual engagement is considered the most important aspect and a more thorough understanding of the data a secondary one [163]. The main aim of Information Aesthetics is to freely explore different aesthetic models, sometimes more related

to artistic works, and focus less on the functional requirements of typical academic or expert visualization tools [204, 205].

Information Aesthetic's visualization techniques should facilitate both the detection and understanding of intrinsic data patterns and also the understanding of the extrinsic meaning underlying the data—conveying a more subjective, deeper meaning about what the data represents [163]. In this sense, information aesthetics artefacts should be interpreted, rather than be a means to facilitate tasks or be an exact imprint of a certain dataset [163]. In Information Aesthetics, and similarly to Information Visualization, the mapping techniques may be direct and accurate, but have a higher focus on stylistic and artistic techniques, as in Visualization Art. For this reason, and according to Andrea Lau and Vande Moere [163], Information Aesthetics can be analysed both from "an information visualization perspective, in terms of functionality and effectiveness" and from "visualization art, in terms of artistic influence and meaningfulness" [30, 163]. From this perspective, both scientific and artistic research communities would benefit from bridging the gap and opening a multi-perspective space for collaboration and exchange [138].

A thorough overview of the field of Information Aesthetics is beyond the scope of this thesis. Nevertheless, considering that the work presented herein embraces this field, we present a short survey of works in this area. In 2009, Jer Thorp created a visualization of the UK's National DNA Database (NDNAD). With the use of a Perlin Noise and the representation of each DNA with a single coloured dot, Thorp generated a single continuous strand that filled up a certain area size, creating tangle-like visuals [273]. The Project *Out of Statistics: Beyond Legal*, created by Rebecca Xu and presented at SIGGRAPH 2009, is a series of abstract digital prints based on the latest crime statistics in the United States. This project employs an aesthetic-oriented approach rather than a more conventional visualization technique. The final output is a set of images with a dual purpose: poetic/aesthetic, but with underlying information encoded. The authors expect that the pleasing visuals catch the eye of the viewers, engaging them sufficiently so that they can decode the visualization with the aid of the legend [308]. In 2011, Frederik de Wilde created a series of images, the *Numer1cal Recipe Series* [NRS], to emphasise the fact that we live in a time where large amounts of data are constantly presented to us in various ways. Wilde's intent is to study how the merging of art and data can provide a mirror of society by giving data "ears and eyes". Wilde refers to these series as digital landscapes, electronic shadows and subjective representations of our environment and our lives [75].

13.3 Data Analysis and Preprocessing

The data used in this part of the thesis consists of the consumption values in 729 Portuguese supermarkets and hypermarkets of the SONAE's chains, which cover the entire country (see Section 5.3). We choose this dataset due to its richness, size, quality and nature. We believe that the dataset is a valuable asset of the work, offering us the opportunity to transform the Portuguese consumption patterns into aesthetic artefacts, while exploring, highlighting and visualising their periodic nature.

We use all the transactions made on these supermarkets and hypermarkets from May 2012 to April 2014. Each transaction corresponds to one product bought and it has properties such as price, date, and time of its purchase. Each product is placed in the product hierarchy of the company, which has 6 levels. For this work, we aggregate all the purchases in 9 distinct categories: Grocery; Alcohol & Sweets; Health care; Beauty; Clothes; Furniture; House Care; Culture & leisure; Pets & Nature Care. Each transaction has the hour, minute, and second of the purchase. However, the representation with this degree of detail was not our main goal. Therefore, we choose to aggregate the consumption values by day.

14

The Rhythm of Consumption

With the high number of collected data about the Portuguese consumption, the need to analyse and make sense of it is of the utmost importance for SONAE. Information visualization is an invaluable asset in this process, and therefore, it is the main focus of our research. Concerning the analytical aspect, we developed a series of visualization tools to optimise SONAE's operations at different levels (e.g., enable the understanding of how the consumption values are distributed along their product hierarchy). Part II is our contribution to this matter. However, we believe that, through the visual representation and inherent simplification of complex datasets, we can enhance the understanding of the data by a more general public. Hence, we explore simple visualization models that can be defined as an aesthetic experience but still preserve functionality.

In this Chapter, we create a set of visualizations to represent the variation of consumption in different departments of the SONAE's hypermarket chain. Our main goal is to visualise the Portuguese shopping preferences and emphasise how the customers change their shopping habits over time. Also, we aim to highlight the rhythm of consumption and times of the year that disrupt the normal consumption patterns. For this reason, we focus on the animated visualization of the consumption and on its sonification to emphasise their rhythmic patterns. As referred before, our target audience is the general public, so our goal is to create a simple yet engaging visualization that empowers the viewers to get a more critical opinion on the data.

We use the data of every consumption made in the SONAE's super- and hypermarket chain during 24 months, from May 2012 to April 2014, described in Section 13.3. We grouped the product categories into groups of three by their type of consumption. The three types of consumption are defined as follows: essential (Grocery, Health Care, and Clothes); non-essential (Alcohol & Sweets, Beauty, and Culture & Leisure); and other (Furniture, House Care, and Pets & Nature Care).

Note that this aggregation was defined to simplify understanding and to categorise different types of consumption, highlighting their evolution and changes throughout time.

We divide our study into two different approaches: one more traditional, in which we apply a small-multiples technique, and another more explorative, in which we apply a multimodal approach, combining sound and image to represent the same data.

In the first approach, we apply a small-multiples technique to visualise the changes in each type of product over time and to enable the comparison of the consumption values on different days. We used two arrangements for the small-multiples display (i.e., animated and static) and explored their effect on the data analysis. We opted to use animation[1] due to its inherent ability to facilitate the representation and perception of changes over time [240]. The application of animation can be seen in a variety of projects such as the representation of transitions or trends [17, 23, 240]. Although animation can be challenging for detailed analysis, Robertson et al. [240] demonstrated that it is useful for the analysis of simple datasets and to make the representation of data more enjoyable and exciting for wider audiences. Also, due to its simplicity, it can be easily interpreted. However, Robertson et al. [240] also found that animation can be less effective than static small-multiples displays for a thorough analysis of the data, as small-multiples may lead to fewer errors. Additionally, small-multiples may provide a higher level of information, enabling the understanding of trends and progress. For these reasons, and given its ability to compel the user to compare and search for differences among objects [279], we also used a small-multiples grid display. Finally, to give to the user the ability to further analyse the visualizations, these arrangements are accessible through a web interface.

[1] Animation is a sequence of images used to convey the illusion of movement.

In our second approach, we further explore the data and its representation through a sonification technique, which is frequently presented as a suitable representation of time-varying data [158]. This second approach aims to create a multimodal experience that arouses the users' curiosity. We focus on the rhythmic nature of the Portuguese consumption to create visualization and sonification models that can be combined to produce appropriate multimodal representations. With this approach, we aim to highlight moments of greater importance. This study is extended to include a discussion on the relationship between visualization and sonification, namely when it comes to revealing different aspects of the data and to the understanding of some of the limitations associated with the proposed representations.

14.1 Time-series Sonification

Throughout the years, people enhanced their cognitive abilities, such as memory, thought, and reasoning, with the invention of external aids. One of the oldest and most decisive external aid is graphical representation [215]. Over time, the use of graphics for representing knowledge has revealed itself as an important and effective means for communicating quantitative information. Allied to our ability to perceive geometrical patterns, it enables and facilitates the detection of trends and relationships in data [67]. Nonetheless, our cognitive abilities to detect patterns in data can also be enhanced through sound. Sonification is the practice of turning data into sound. There are various sonification techniques, which can be broken down into the following categories: auditory icons, earcons, audification, parameter mapping sonification, and model-based sonification [123].

An example of sonification of time-varying data is *Quotidian Record*, a project created by Brian House, presented in 2012. In *Quotidian Record*, House represents through sound each place he had visited over one year. The basis for his project was the idea that our routines have inherent musical qualities and that through music, we can create an emergent portrait of each individual [127]. Other examples of sonification of time-varying data include *Climate Symphony* [236], a sonification of climate change based on the analysis of the chemical composition of an ice core drilled up in Greenland, and Living Symphonies [36], a musical installation that portrays the activity of the forest's wildlife, plants and atmospheric conditions.

Listen to Wikipedia is a project created by Stephen LaPorte and Mahmoud Hashemi, in 2013, to represent the most recent feed changes in Wikipedia articles [302]. These changes are represented by a bell when some entry is added, and by string plucks when someone removes an entry. The pitch is also manipulated depending on the size of the edit. In this project, the visualization was generated to not overshadow the sonification. To visually represent the editions, and to add more insight into the data, the authors draw green circles to show edits made from registered contributors, white circles to represent unregistered users, and purple circles to represent edits made by automated bots.

In 2014, a group from the IUAV University Communication Laboratory generated a sonification to represent how many animal species entered the IUCN Red List of endangered species. Each class (mammals, birds, amphibians, fishes, and reptiles) is represented by a note of a B minor chord and the number of species for the corresponding

class is represented by the number of repetitions of every note [109].

In 2014, the Office for Creative Research group created the Specimen Box, an interactive work that visualises and sonifies the Botnet activity. Botnets are distributed entities infected with malware that reach thousands of PCs. In this project, those networks are represented by their aggregated geographic position, creating visual and audible temporal patterns. The Office for Creative Research characterises this project as an exploratory tool that enables digital crime units' investigators to examine the different profiles of the Botnets through geographic position and time, and to understand their unique characteristics: their behaviour, how they propagate through PCs, and how they are adapting to the environment [265].

In 2015, Brian Foo created a musical sonification of income inequality on the New York City Subway's 2 Seventh Avenue Express, whose route includes three different boroughs: Brooklyn, Manhattan and the Bronx [100]. Foo's sonification is modelled after Steve Reich's *New York Counterpoint*, a minimalistic composition written for 11 clarinets and a bass clarinet that tries to capture the vibrant atmosphere of Manhattan. The sonification emulates a ride on the aforementioned train. At each moment, the quantity and dynamics of the instruments reflect the median household income of that area. For example, in wealthier areas, the instruments will increase in quantity, volume, and force. In addition to the sonification, Foo presents a simple visualization of the train route, enabling the viewer to connect the sound to the respective geographic area. Foo's main intent in this project was to recreate a representation of the vibrant energy and orderly chaos of the New York City Subway system.

Similarly to the majority of the works presented before, we apply a parameter mapping sonification, in which data is responsible for varying different parameters of an audio signal [285]. With our sonifications, we aim to amplify our cognitive abilities and emphasise moments of greater importance. As in the work of Brian Foo [100], we change the volume of the instruments according to a certain value. In our case, the higher the consumption value, the higher the volume of the instrument. Also, to represent different types of consumption, and similarly to the works of Brian House [127] and Stephen LaPorte and Mahmoud Hashemi [302], we use different instruments to sonify different types of consumption. In Section 14.3, we further detail our approach.

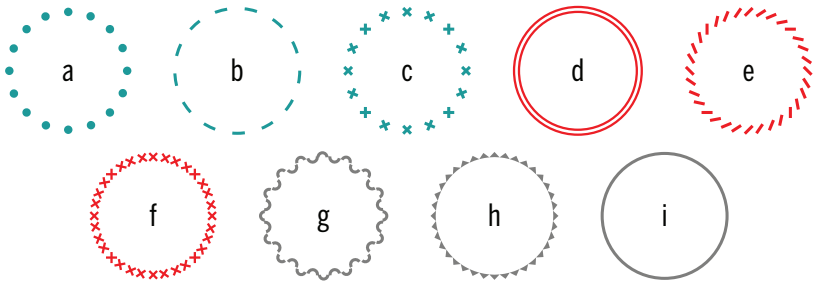

FIGURE 14.1: Representation of each category (shape) and their types (colour): (a) Clothes, essential; (b) Health Care, essential; (c) Grocery, essential; (d) Culture & Leisure, non-essential; (e) Beauty, non-essential; (f) Alcohol & Sweets, non-essential; (g) Pets & Nature Care, other; (h) House Care, other; (i) Furniture, other.

14.2 Visual Approach

With the representation of the customers' shopping behaviours, we aim to emphasise the differences in types of consumption over time. For this reason, we defined three different colours, red, green, and grey, to represent the three distinct types of consumption, non-essential, essential, and other, respectively. Then, to enable the viewer to understand the different categories of products that belong to each type of consumption, we defined different shapes. These shapes are intended to have less visual impact than the three categories mentioned above, but still enable the distinction between products. We defined nine different shapes for each product category as depicted in Figure 14.1. These shapes are then coloured according to the type of consumption to which they belong. Lastly, these shapes change in size according to the consumption values in the different days of the dataset, enabling the visualization of the consumption changes over time.

To represent all days, we apply a small-multiples technique. We use a grid where each row represents one month, and each cell of that row represents a day of that month. For the vertical alignment of the days, we defined two different arrangements. In the first, all months start in the leftmost column, enabling the comparison of the consumption by quarters of the month, meaning we can compare whether in the different quarters of the month and between months, the customers' behaviour is similar, or not. For the second arrangement, the days of the month are aligned by the day of the week. This enables the comparison of the different weekdays and the understanding of the impact of the weekend on the customers' behaviours.

In Figure 14.2, we can see the representation of the 24 months of our dataset. In this Figure, we align all days of every month by weekday. We can easily perceive that the type of consumption that has the highest consumption is the non-essential, especially in the month of December of 2012 and 2013 (8^{th} and 20^{th} lines). We also can see that the highest consumption in the Grocery category occurs

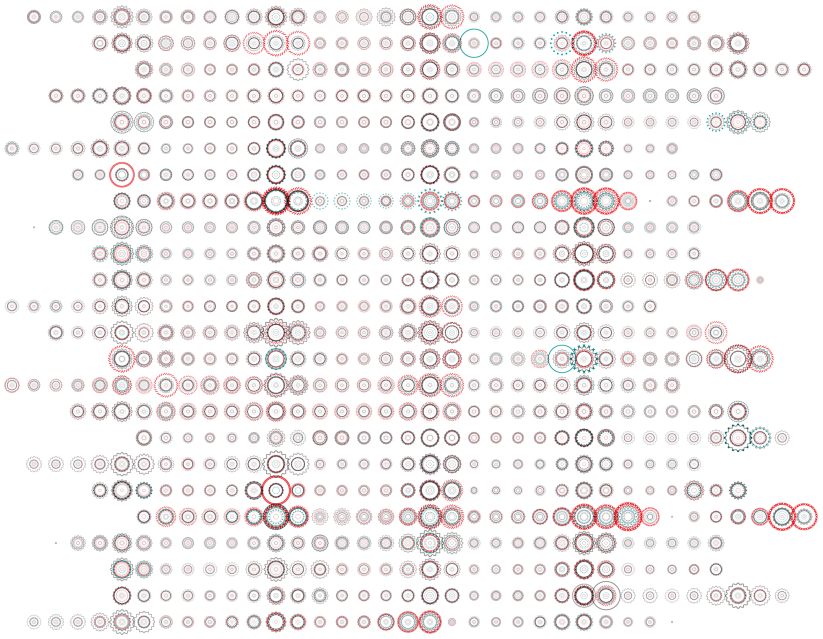

FIGURE 14.2: Representations of the 24 months, from April 2012 (top row) to May 2014 (bottom row). All days are aligned by the weekday and the values are normalised by the maximum value of the corresponding category.

in June 2013 (14th line).

In addition to the small-multiples arranged in a grid, we created an animation with the cells of the small-multiples. In this approach, the shapes' size adapts according to the consumption value of the day that is being represented at each time. The two approaches have two different objectives. The first aims for a more analytical representation, in which we can see how the consumption is affected by special events and vacations, and easily compare different days that are distant from each other. The second—the animation—aims for a dynamic visual and entertaining representation, in which we can emphasise the difference between the days which are close to each other.

We explore two different normalisations: one global, where all the values are mapped between zero and the maximum value of all categories; and one local, where each category is mapped between zero and its maximum value. With the global normalisation, we can compare the consumption values of all categories. With the local normalisation, we can compare the behaviours of the consumption values within each category, and perceive whether they tend to grow or decline through time.

In Figure 14.3, we can perceive the differences between the two normalisations. In the top image, we apply the local normalisation and we can perceive more differences in more categories than in the bottom image, where we apply the global normalisation. These two normalisations imply two different readings. In the first, we are not able to compare if one value of a certain category is higher

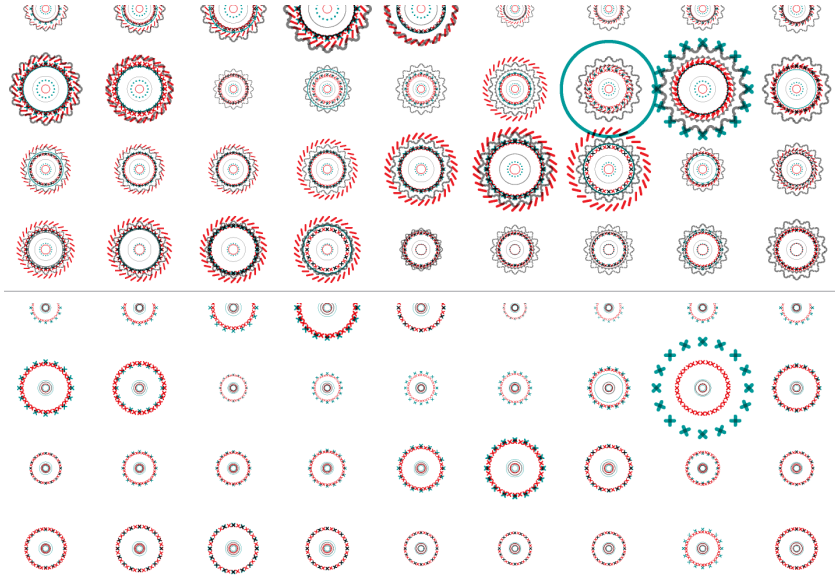

FIGURE 14.3: Comparison between the two normalisations: local normalisation (top) and global normalisation (bottom).

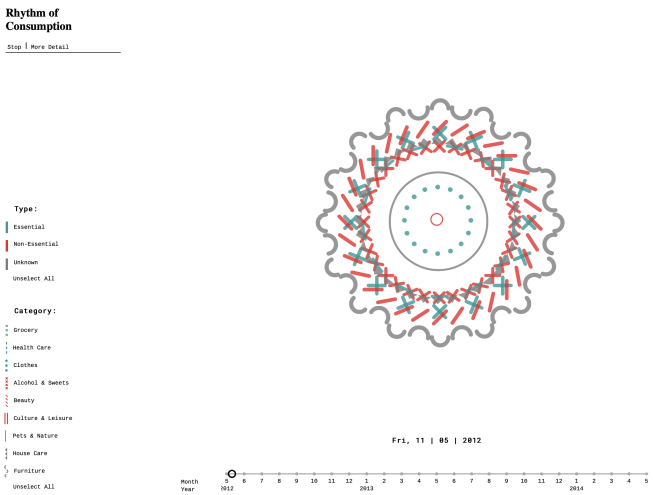

FIGURE 14.4: Landing Page. On the left, the user has access to a set of functionalities, such as to play or stop the animation, change the view to the small-multiples matrix, or view/hide a certain category. On the bottom of the canvas, the user has access to a timeline.

than the value from another category, but can compare how they evolve over time and on which day a certain category has its highest value. In the second, we can compare the values of every category, as they are mapped on the same scale. However, there are big differences in the consumption values between categories, which causes the visualization to emphasise the categories with the highest consumption—Grocery and the Alcohol & Sweets—and give less detail on the categories with lower consumption values.

14.2.1 *Graphic Interface and Interaction*

To enable the visualization and interaction with our two approaches, we implemented a web application[2] in Javascript and using Raphael.js, an open-source graphical library, to render the visualization (Figure 14.4). In this graphic interface, the viewer can interact with

[2]This web application can be accessed through the following link: https://cdv.dei.uc.pt/cmacas/rhythm/. Additionally, a video of the interface can be accessed through: https://cdv.dei.uc.pt/the-rythm-of-consumption/

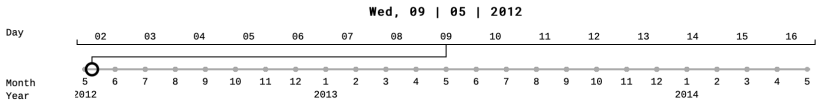

FIGURE 14.5: Detail of the timeline expanded. by clicking on the *stop* button, the user has access to this expanded timeline and can choose a specific day to be visualised.

the models and change between the grid visualization and the animated version of the small-multiples. These two approaches can be seen on two different pages. However, to maintain the coherence between pages, they share two elements: an interactive left panel, that contains all functionalities common to both approaches, and an interactive timeline, positioned at the bottom of the canvas.

In the left panel, the user has access to the "Play/Stop" button, which enables the user to start or stop the animations, and the "More/Less" Detail, which enables the user to switch between the animated and the grid visualization. These two buttons are positioned on the top of the panel. At the bottom of the panel, the user has access to an interactive caption of the visualization. This allows the user to distinguish the colours used for each type of consumption, and match the shapes with the product categories that they represent. This caption is interactive as the user can select and deselect different categories or types of consumption. By doing so, the user is defining which categories or types of consumption are represented in the visualizations.

In the timeline, at the bottom of the canvas, the user can visualise the date that is being represented in the animation, and, when the user is analysing the grid, define the starting date of the small-multiples visualization. The timeline has numbered thicks representing the different months, and in the middle-upper part of the timeline, the user can visualise the current date in a textual form.

The start page of this web application is the animation of the small-multiples. This decision of putting first the animation was due to its ability to make a subject more approachable, which may induce the user to explore and analyse in more detail the data [240]. By entering the page, the animation starts automatically showing the coloured shapes changing size as time passes. By clicking on the "Start/Stop" button—positioned on the left panel—the animation will stop. Additionally, this interaction triggers the appearance of a second timeline showing a range of 14 days, seven before and seven afterwards the current day (Figure 14.5). This detailed timeline is positioned above the first timeline to enable the user to select specific dates (close to the current date) and visualise the changing consumption on those dates.

By clicking on the button "Less/More detail", the user will be

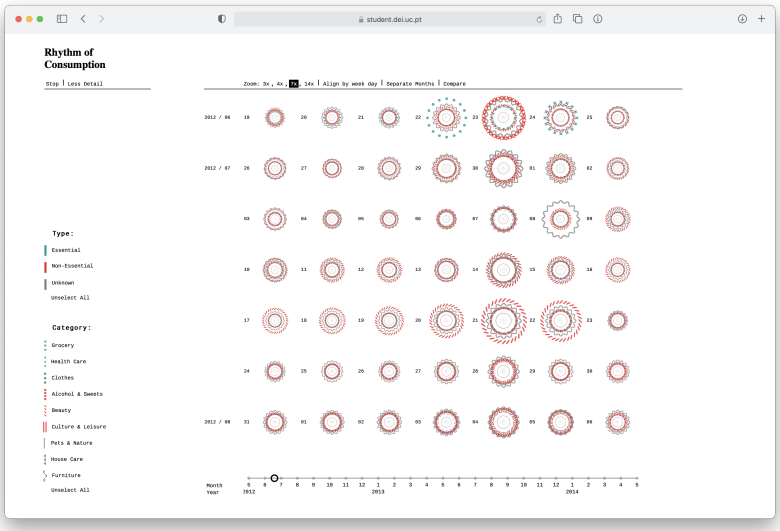

FIGURE 14.6: Small-multiples view. In this view, the user can see in more detail each day. The user still has access to the general functionalities and the timeline. By clicking on the timeline the user is selecting the start date for the small-multiples.

redirected to the small-multiples grid page (Figure 14.6). The only difference between this page and the previous one is that, instead of visualising the animation, now, the user will be able to visualise the grid with the frames of the animation dispositioned from left to right, up to bottom. To aid the user in a more detailed analysis of the data, we enabled a set of features that change the organisation of the grid's cells. The user can change the zoom level, opting to show 3, 4, 7 or 14 cells in each row. The lower the number of cells, the bigger and more detailed each cell will be. In terms of vertical organisation, the user can align the cells by weekday or maintain the sequence of days with no semantic vertical alignment. Also, the user can choose to separate the months. By choosing this option, all new months will start in a new row. Finally, by clicking on the "Compare" button, the user can compare the consumption in different non-consecutive periods of time. When this feature is selected, the canvas is divided vertically into two sections by a red horizontal line and two grid visualizations appear on each division of the canvas (Figure 14.7). Also, a red and a black circle are made available in the timeline on the bottom of the canvas. Each circle represents the starting time of each visualization. The user can drag the black and red circles to different time positions to change the starting periods of the visualizations.

14.2.2 *Usage Scenario*

By entering the web application, the user visualises the rhythmic pulse of the consumption values, as they grow and decline along the days of the week. By paying attention to the timeline, the user may recognise that those peaks coincide mainly with the weekend periods.

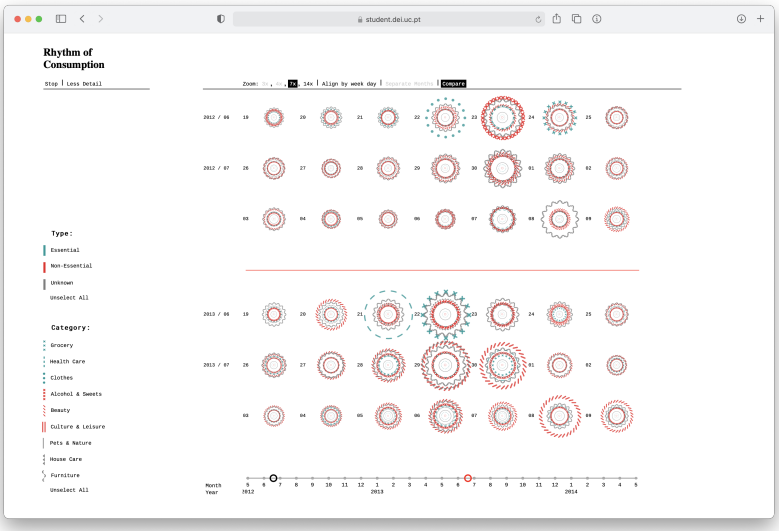

FIGURE 14.7: Small-multiples view, comparison functionality. In this mode, the user can compare the consumption in two different period ranges. In this example, we can compare the period from June 19 to July 9 in 2012 and 2013.

As the animation is being played, it is possible to notice big disruptions, especially in the month of June, as the type of consumption with the highest consumption values is always changing between the non-essential and the essential consumption. As the animation gets closer to August the consumption changes less, and the previous frenetic rhythms of June are replaced by smaller waves of consumption (i.e., smaller differences in consumption values and overall smaller consumption values) and less change between the different types of consumption.

From September to November, and in the majority of the days, the non-essential type of consumption has the highest values. However, on the first days of December, this behaviour changes and the user can visualise a higher consumption value on the essential type of consumption. This rapidly changes again, and near the end of the year, the non-essential type of consumption is again the one with the highest consumption values. With the aid of the caption positioned on the left side of the canvas, the user can perceive that the category Alcohol & Sweets is the one with the highest consumption values.

To visualise in more detail the different consumption categories, the user can click on the “More Detail” button. By doing so, the user can visualise the different days arranged in a grid. By default, the days are not organised by day of the week, and the user may fail to detect any weekly pattern. However, it is possible to see the repetition of three consecutive high consumption values of the non-essential type of consumption that repeat from June to August. By clicking on the “Align by week day” button, the user can visualise two weeks by row and see that the three days of higher consumption match the Friday, Sunday and Saturday. Also, it is possible to visualise the

highest consumption values in December, especially at the end of the month.

It is also possible to see the disruptions of those patterns. For example, on June 18 of 2012, the category Health Care was the one with the highest consumption values, which was an isolated event in this year. If the user clicks on the "Compare" button, the user will be able to analyse if this event repeats in both years of the dataset. By doing so, the user can perceive that in 2013 there is also a high consumption value, but it appears on a Friday (June 21 of 2013) instead of a Monday (June 18 of 2012). Nevertheless, in both years the second last weekend of June has high consumption values of the essential type. In 2012, this behaviour only occurs on Friday and Saturday, for the Clothes and Grocery categories of products, and in 2013, it is the Health Care and Grocery categories that have the highest consumption values.

Also, by comparing December 2012 and December 2013, we can see different behaviours. In December 2012, the consumption values of the essential type, especially of the Grocery category, were considerably higher than the other types of consumption. In 2013, this behaviour did not happen. The Furniture category of the Other type, and the Beauty category of the Non-essential type, were the ones with the highest consumption values. Finally, if we only select the Grocery and Health Care categories, we can also understand the differences in consumption values of these two categories. In 2012, there were higher consumption values than in 2013. However, the behaviours are similar, as, in the majority of the days, the Grocery category has higher consumption values, except for the first Sunday of December where the Health Care category has higher values in both years.

14.2.3 *Discussion*

In this first attempt at the beautification of an analytically inclined dataset, we could create simple visualizations that can represent the rhythms of consumption of the Portuguese data. With the animation, we were able to emphasise big changes in consumption during the weeks, as the weekends could be visually highlighted through the bigger sizes of the consumption representations. By attributing colour to each category of consumption, one is instantly able to detect the differences in types of consumption over time. Then, by clicking on the "Stop" button, it is possible to analyse in more detail which product category is the one with the highest consumption. The user can have an overview of all data through the visualization of the

animation, and have a more detailed view through the analysis of each frame of the animation.

Also, with the grid representation, a more sustained analysis can be made, as all frames are placed side by side. By making available mechanisms to compare different dates, to align the frames by day of the week, separate months vertically, and by enabling the user to zoom in and out the grid, we give the user the necessary tools for more thorough analysis. Hence, this second view improves the analysis of the data by facilitating the comparison and analysis of the different frames. Furthermore, in the grid representation, we can see how the consumption values are affected by special events and vacations, and easily compare days that are distant from each other. In the animation, we can, above all else, perceive the rhythmic patterns of consumption, marked by the weekends, and the disruptive effects caused by special events. In summary, these two visualization models can highlight the consumption variations, emphasise weekly periodic patterns, and enable the comparison of different types of consumption over time.

The exploration of both representations enabled the understanding of yearly and monthly rhythms and behaviours of consumption as well as the detection of the weekend impact in the consumption values and the impact of the festivities during celebration periods. Although the animation was readable, we think there is space to make it more representative of the data and make a more engaging experience for the user. For a more entertaining experience, we argue that sound could enable us to add another layer of experience. The following Section focuses on this multimodal experience in which sonification is added to the visualization.

14.3 Sonification Approach

Complex time-series can be represented both through visualization and sonification. While visualization tends to be regarded as a more effective representation for most types of data, sonification is frequently presented as another suitable representation of time-varying data, as it provides two dimensions for the representation: the sound itself and the idea of time [158]. Additionally, the transformation of regular repeated patterns in time-series into sound favours the construction of rhythmical musical representations, which tend to be pleasant and entertaining to the listener.

After visually exploring the rhythms of consumption, we intend to explore other sensory modality and create visualization and sonifi-

cation models that can be further combined to produce appropriate multimodal representations. In short, we add a second level of representation to the previous visualization, emphasising and exploring the rhythmic nature of the consumption data beyond visuality. These multimodal representations are intended to characterise the Portuguese consumption patterns—caused by the periodic growth and decrease of consumption during the week—and highlight the disruptions and sudden changes in the normal consumption patterns.

We present two distinct visual and audio representations[3] that demonstrate how the consumption in the different categories and types of consumption change over time, and discuss their strengths and limitations in revealing different aspects of the data.

[3] which can be accessed through the following link: https://cdv.dei.uc.pt/consumption-as-a-rhythm/

14.3.1 *Percussive Sonification*

In this Section, we define the main concept behind each sonification and how each attribute from our dataset is sonified.

The two percussive sonifications presented herein are not intended to highlight the rhythmical nature of the data—the rhythms are already explored through visualization. Instead, the sonification assumes the presence of such a feature and exploits it to primarily emphasise the sudden changes and disruptions in consumption patterns. In terms of instruments, both sonifications are built around Taiko drums, a wide range of traditional Japanese drums, whose existence is speculated to date back from Ancient Japan. Taiko drums have served many purposes: for example, they were a common element in Buddhist and Shinto rituals or in the battlefield either to alarm the enemies or to issue commands [18].

Each type of Taiko drum has its unique timbre and pitches. Despite the differences, the sound of these instruments can be characterised as powerful and dramatic, especially when bigger drums are used, such as the *odaiko* or the *chu-daiko*. We believe that these characteristics make Taiko drums an interesting basis for the sonification of the consumption data. First, and as previously mentioned, the consumption patterns tend to present a cyclic behaviour over time, i.e., there is a regular repeated pattern, which favours the idea of a rhythmic representation. Secondly, sudden changes and disruptions can be signalled with intensity changes, which, due to the stirring sound of this type of percussion, are expected to effectively inform the listener about the significant data modifications. Finally, the variety of Taiko drums allows us to assign different types of consumption, or even categories, to the various groups of drums. Note that our rhythmic compositions do not strictly follow the rules of Taiko drumming, as

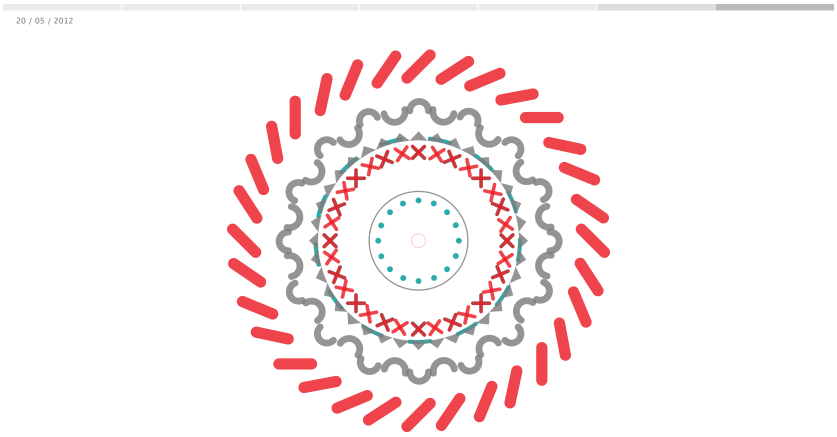

FIGURE 14.8: Screenshot of the first visualization layout. In the centre of the canvas, the visualization model described in the previous Section, and on top, a visualization to contextualise the day of the week is represented.

the main idea is to explore the characteristics of the timbre rather than using typical Taiko-based compositions.

The proposed sonifications follow a parameter mapping model, with the volume (gain) of the percussion instruments reflecting the numbers of transactions within the different types of consumption. For each type of consumption, we normalise the number of transactions to let the values vary between 0 and 127. Those values are then rounded to integers to set the volume of the drumbeat. In both sonifications, a beat represents a day.

As our goal is to explore a multimodal representation, we tried to introduce a certain complementarity between the visual representations and the auditory. The sonifications models disregard the consumption data at the category level. Furthermore, according to our model, only some types of consumption can be considered per beat. We followed this strategy not only to introduce complementarity between the different representations but also to study how informative or evocative this type of rhythmic sonifications can be.

The sonification process was developed in Max/MSP. A Max patcher is responsible for reading the data and generating MIDI notes according to the sonification model. The MIDI notes are then sent to Ableton Live. The instruments are played through Ableton Live using the AIR Xpand!2 VST plug-in, which includes a wide range of percussion instruments. It should be noted that the plug-in does not use the typical Japanese terminology to classify the different types of Taiko drums it provides. We use three types of Taiko drums: Taiko Big drums 1 and Taiko Big drums 2 (similar to *chu-daiko*), and Taiko drums (similar to *okedo-daiko*).

14.3.2 *First Approach*

For the first visual representation (Figure 14.8), our main goal was to represent the different categories and to visually perceive their behaviour over the two years. To explore the data, we created two different approaches. The first one was a classical small-multiples representation, where each day is placed in a grid, horizontally organised by day of the week and vertically by month. The second approach was an animation of each cell of the grid previously described (see Section 14.2). For this exploration, we focus on the second approach, as we intend to assemble the visual representation with the sonification. Hereupon, we briefly describe the latter (for more details see Section 14.2).

For the representation of the consumption in each category and due to the sparse consumption values (i.e., some categories had a range of values too small when compared with other categories), we decided to normalise each consumption value by category independently. This enabled us to represent the peaks of consumption in each category, instead of focusing on the comparison among them.

Each category is represented by a different circular shape, which is coloured depending on the type of consumption it represents: red, for non-essential, green for essential, and grey for other (Figure 14.1). To represent the differences in consumption values over time, we used the size of each shape. As the days go by, each shape grows or decreases in size depending on the consumption value. Every shape is centred in the middle of the canvas and morphs its size between days, so we can have a continuous perception of each category behaviour.

To allow the analysis of the consumption behaviour over time, we added the date in the upper left corner of the canvas. The user will not be able to analyse the consumption volumes of each day precisely but will be able to compare the consumption behaviour at different times of the year. Furthermore, as the consumption rhythm is mainly caused by the differences in consumption values between weekdays and weekends, we visually represent the day of the week on the upper side of the canvas. This representation is composed of a row of seven equal rectangles that occupy the whole length of the canvas. Each rectangle represents one day of the week—the first position of this row represents Mondays, the second position represents Thursdays, and so forth.

The sonification that complements the visualization has a time signature of $\frac{7}{4}$ and, as previously mentioned, each pulse represents a day of the week. Figure 14.9 presents the basic beat defined for this sonification: Taiko Drums Big 1 corresponds to essential goods,

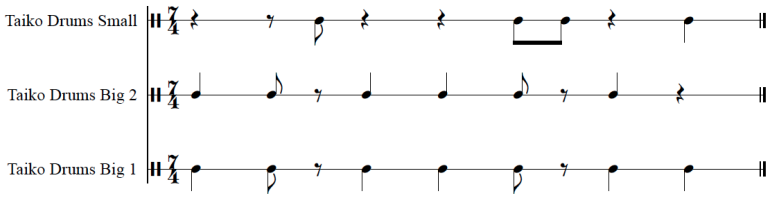

FIGURE 14.9: Basic Taiko drums beat for Representation #1.

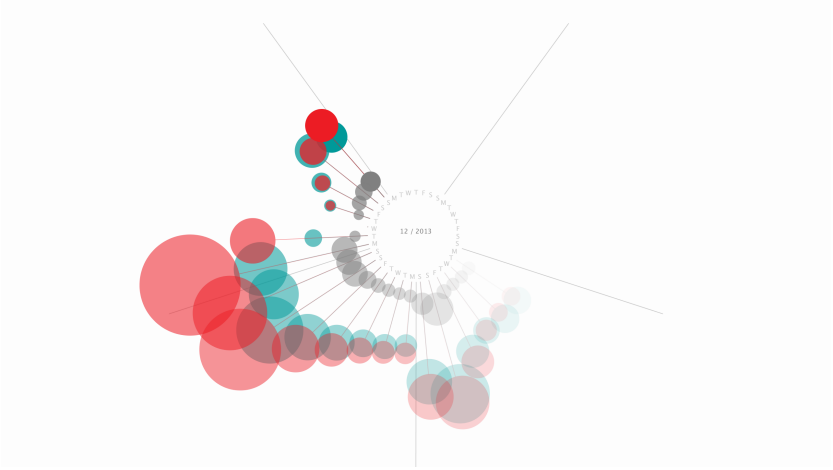

FIGURE 14.10: Screenshot of the second visualization layout. In this animation, the circles go around the middle circle to represent the consumption values over time, one lap represents one month, and the grey lines separate the different weeks of the month.

Taiko Drums Big 2 corresponds to non-essential goods, whereas the transactions from the remaining categories are associated with an ensemble of four Taiko drums.

As Figure 14.9 suggests, this sonification focuses on the transactions of essential and non-essential goods. To accentuate even more the changes in consumption patterns, we added other instruments to the sonification: the transactions of essential goods are also associated with thunder drums; an orchestral snare drum is used for transactions of non-essential goods. The remaining goods are not associated with any other instrument.

In this approach, we intended to use sonification to represent a different level of the data: the sonification of the types of consumption (essential, non-essential, other), instead of the sonification of the categories, as in the visualization. Through this, we aimed to enhance the understanding of the consumption patterns. However, as the sonification only sonifies the types of consumption and the visualization only represents the consumption by category, it was difficult to perceive the relation between the two representations. For this reason, we developed a second approach, in which we represent the same level of aggregation.

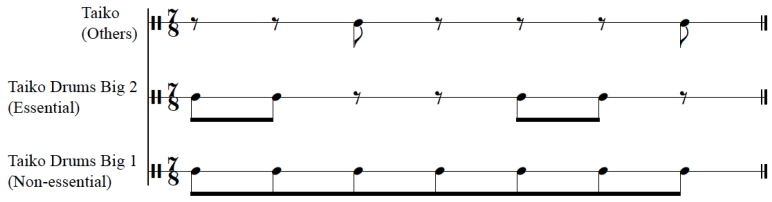

FIGURE 14.11: Basic Taiko drums beat for Representation #2.

14.3.3 *Second Approach*

For the second visual representation (Figure 14.10), our main goal was to highlight the differences between types of consumption. In the previous approach, it was not possible to compare the consumption values among themselves, as the values were normalised by each category independently. As such, for this representation, we summed all the categories in each type of consumption and normalised those values by the maximum value of all three types.

In this visual approach, each type of consumption is represented by a coloured circle. As in the previous approach, we used green, red, and grey to represent, respectively, the essential, non-essential and other consumption. The size of the circles varies depending on the consumption value of each type of consumption on a given day. To emphasise the different consumption values, we also distance each circle from the centre of the canvas. The lower the summed consumption value, the closer the circle is to the centre of the canvas.

Apart from the use of size and colour to distinguish the types of consumption, we wanted to explore the movement and positioning of the different elements to increase the amount of information represented. Since it was difficult to distinguish the different days in the previous approach, we aimed to improve its legibility in the second one. Hence, as time passes, the circles change position in an invisible circle. Every 360 degrees represents a month, and we visually divided the lap into five slices (as can be seen in 14.10), representing five distinct weeks. Each one of these five slices is further divided into seven days of the week. At the beginning of each month, the circle's start position is in the upper side of the circle (in the first slice) and as time passes it moves around the centre of the canvas. To facilitate the understanding of which month is being represented, we placed a caption in the centre of the canvas. Also, to better associate each circle to the respective day of the week, a line is drawn connecting the centre of the circle and the centre of the canvas. Finally, as time passes, we maintain some of the previous day's representations, with decreasing transparency, to create a sense of movement and to be able to give some context to the sonification.

The second sonification, which was designed to focus on the transactions of non-essential and other goods, has a time signature of $\frac{7}{8}$. The drums beat for this sonification is depicted in Figure 14.11: Taiko Drums Big 1 corresponds to non-essential goods, Taiko Drums Big 2 corresponds to essential goods, and the transactions from the remaining categories are associated with a group of two Taiko drums.

As in the previous sonification, other percussion instruments were added to the model: essential goods are associated with timpani, non-essential goods are associated with thunder-drums, other goods are associated with chimes, which are played at every beat. Chimes were introduced to represent the consumption patterns in terms of pitch, using the same strategy as the one that we defined for the gain of the instruments.

In this second approach, we could enhance the growth of consumption between the three types of consumption. As the data used for the visualization and sonification were the same, now the "reader" can easily establish relationships between the representations and be more attentive.

14.3.4 *Discussion*

The multimodal representations described herein stress, first and foremost, the rhythmic pace of the data. The visual artefacts can arguably be regarded as more informative, as they translate consumption to finer detail. Nevertheless, the proposed multimodality was created with the intent of becoming a richer experience in terms of understanding. The auditory counterpart contributes to a better comprehension of consumption peaks over time. While the sonification model was designed to not consider all types per pulse, the final representations become sufficiently informative as the sound and the visual artefacts complement each other.

This study suggests that percussive sonifications are an interesting approach to represent the type of data that we worked with. The percussion can portray the hustle and recurrence associated with shopping habits. Hence, we can consider that the use of sonification in addition to the visualization model may be suitable to represent and intensify the reading and understanding of information. However, both models (of sound and image) should share and/or represent similar data attributes, so the user can relate them. Sonification also has some limitations, especially in terms of information retrieval, if the data is sufficiently complex. A possible solution to circumvent this problem is to use pitched percussion instruments and consider pitch as a parameter to be mapped.

15

Swarming Consumption

Creating appropriate new visual models and new ways to manipulate big datasets are fundamental challenges for Information Visualization and Information Aesthetics. Another challenge is the development of visualization models that can continuously readjust to the changes of dynamic data and to the intentions of the user [133]. To address these challenges, Information Visualization can absorb the knowledge from other fields and adopt some of their techniques, as for example the appropriation of MAS from the AI field. MAS is a sub-field of computer science and can be described as a computerised system composed of multiple interacting agents within an environment [306]. This technique can be applied through the implementation of swarming agents to represent datasets changing over time [141, 218, 283]. This emergent and dynamic behaviour of MAS is ideal to explore the ever-changing consumption values of the Portuguese.

With the data of the Portuguese's consumption routines and the dynamic agents, we aim to create emergent visual artefacts that are driven and influenced by the Portuguese patterns. The SONAE's dataset is rich in daily, weekly, and monthly repetitions of consumption patterns which allows for the creation of visual artefacts with ornamented characteristics, enhancing their aesthetic value. Moreover, the data offer us the opportunity to transform the consumption patterns of the Portuguese into pleasing artefacts, while exploring, highlighting and visualising their periodic nature. Also, we can use the models to mirror the consumption patterns, as well as to give a socio-economic portrait of the country, since SONAE is one of the biggest Portuguese retail companies that cover the entire Portuguese territory.

In this Section, we expand upon the state of the art in Information Visualization through the creation of a novel model. To that end, we apply a swarm system to create emergent visual artefacts that represent the data and increase curiosity in the user. The goals for this

work are: (i) to visually explore the consumption evolution over time; (ii) to detect possible periodic behaviours; and, (iii) to explore the boundaries between Data Visualization and Information Aesthetics. We present two explorations of the swarming system. In the first, the user has to define the parameterisations of the system to create a balance between functional and aesthetically intriguing visualization. In the second, the parameterisation is performed through the use of an EA, opening the possibilities to create a wider range of visual solutions. Note that in the second phase of the project, the main concern is to create artefacts that are aesthetically appealing to the user, and not artefacts placed in the functional spectrum. Also, both explorations were implemented in Java and using Processing.

15.1 Swarms in Visualization

The earlier Information Visualization models were static representations but, with the advance of technology, they started to be generated through computational processes and to be more dynamic and interactive. Information Visualization uses different techniques to present data. Some of these techniques are based on MAS and are commonly used in interactive visualizations. With MAS, it is possible to create a diverse series of visualizations, from simple graphs to more sophisticated agent-based geographic representations [239]. The simulation of swarming and flocking behaviours [238] stimulated the interest of scientists, designers, and artists for two main characteristics: self-organization and emergence [136].

The use of swarm systems in an artistic context was also explored by artists and researchers such as Mauro Annunziato, Jon McCormack, Tim Barrass, Daniel Shiffman, Alice Eldridge, Gary Greenfield, Christian Jacob, Penousal Machado, Leonel Moura, Nicolas Monmarché, Paulo Urbano and Yann Semet [113], for the creation of a wide variety of static and interactive artworks. For instance, in *Swarm* by Daniel Shiffman[258]—presented at SIGGRAPH 2004—the organic paths created by the virtual *boids* were used to produce a non-photorealistic rendering of live video input. Swarming techniques were also used for visualization purposes. In *In-Formation Flocking* [283], Andrew Vande Moere and Andrea Lau used a Swarming System to group and visualise similar data entities without the need for supervision. *In-Formation Flocking* can create dynamic patterns and can represent volatile and chaotic time-varying datasets while sustaining a comprehensible representation at a general level as well as revealing more detailed patterns. Michael Ogawa and Kwan-Liu Ma [218] created

a swarm visualization application. In this visualization, the authors generated a series of visualization videos representing the history and evolution of software development. Finally, *We Feel Fine* [141] aggregates signifiers of emotion in social media, and visualises emotional statements as animated coloured particles. *We Feel Fine* conveys the general mood through the colour of the particles.

15.2 Swarming Consumption

To show consumption patterns and detect periodical behaviours, we applied a swarm system with the aim to stimulate the user to explore the visualizations and detect the periodicity of the consumption behaviours from the visual inspection of the swarm patterns. Additionally, we intend to create emergent visualizations of data that convey meaningful information and, at the same time, explore the boundaries between Data Visualization and Information Aesthetics. Although we focused on the application of a swarm system in the Portuguese consumption, our approach can be used with any other time-dependent dataset. In this context, we are not applying any algorithmic technique to automatically detect or highlight periodic behaviours, these patterns should naturally emerge and be represented by the swarming behaviour.

To visualise the consumption habits over time, the swarm system simulates the behaviour of multiple boids in an environment that reacts to the evolution of consumption over time (Figure 15.1). Each boid is characterised by properties such as velocity, position, and size, and follows three basic rules—cohesion, separation, and alignment [238]. In the following subsections, we detail the forces and rendering possibilities for the boids.

15.2.1 *Swarm Forces*

The swarm system simulates the behaviour of multiple boids. Each boid is described by properties such as velocity, position, size, and colour. This last property identifies to which Department from the SONAE's product hierarchy the boid belongs to (Figure 15.2). As the boids wander through the canvas, they leave an imprint of their shape, enabling the user to see their path, and consequently, the consumption values.

Based on the work of Reynolds [238], each boid follows three basic rules: (i) cohesion, which means that they should remain close to nearby boids; (ii) separation, which makes them avoid collisions with nearby boids and other objects; and, (iii) alignment, which means

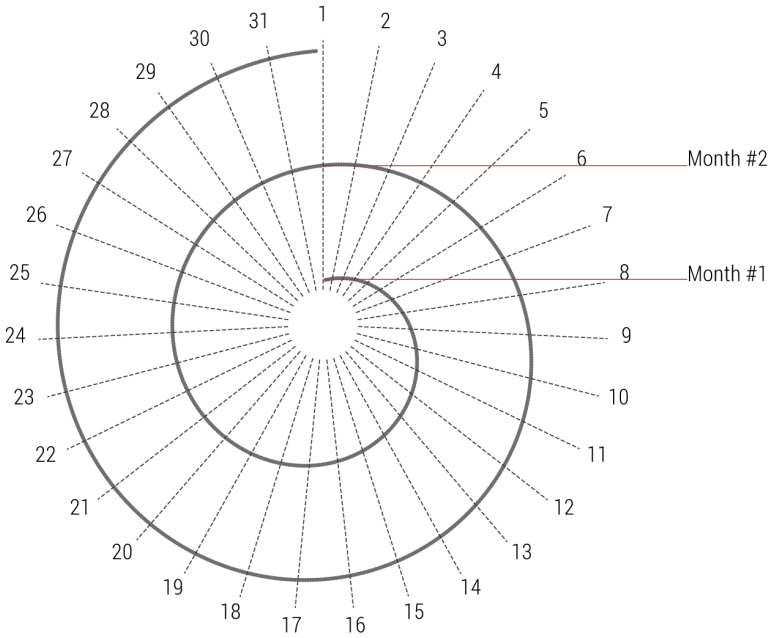

FIGURE 15.1: Spiral and its time division. The swarms movement is clock-wise and each lap represents one month.

FIGURE 15.2: Definition of the Departments' colour.

that they should match the velocity of the nearby boids. To properly explore the system, we applied different values to each force, so it was possible to create different outputs. Moreover, to enhance the distinction between different Departments, we apply small variations of the defined force values according to the neighbour's characteristics. If two neighbouring boids are from the same Department, we apply higher forces of attraction and lower forces of separation, but, if they are from different Departments, the attraction force is lower than the previously defined, and the separation force higher. If the neighbour's type is equal to the current boid's type, the separation is smaller, and the alignment and cohesion bigger. These rules make the boids seek for others of the same type and keep slightly away from boids of a different department, clustering by the department. This approach can be seen as an extension and generalisation of the size-specific swarming technique introduced by Eldridge [88]. Additionally, the separation forces between two boids are directly proportional to the sum of their separation radii, which prevents boids from overlapping in space.

To prevent the boids from randomly moving on the canvas (i.e., the environment) and to enhance their periodic behaviour, we defined a target boid that all the boids should look for. Hence, in addition to the previously described forces, all boids are under the influence of an attraction force towards this moving target. This target starts from

the centre of the canvas and swirls around, creating a spiral with equal distances between each lap. The target boid is not represented on the final visualizations so it is not confused with the other boids. Since we want to represent time-oriented data, the representation of time must be added to the visual artefacts. To do so, we consider that each lap represents one month. Then, we divide each lap by 31 days multiplied by 12 hours ($2\Pi \div (31 \times 12)$), since, as previously stated, the data is aggregated in intervals of two hours. Note that we define all laps to have 31 days. By doing so, all months start with the same angle (at the top of the circle), and, if they have less than 31 days, the consumption values during those nonexistent days are null (see Figure 15.1).

15.2.2 *Swarm Rendering*

Each boid represents the consumption value of a certain Department at a certain moment. Hence, to represent the consumption values, we vary the size of the boid according to the consumption value in the current time mark. The radius is mapped to a predefined minimum and maximum radii, that can represent the minimum and maximum sale of each individual Department (local normalisation), or the minimum and maximum sale of all Departments (global normalisation). With the global normalisation, the reader can compare the consumption values (i.e., boids' radius) between the different Departments, and see which ones have the highest consumption values. With the local normalisation, it is not possible to compare consumption values between Departments, but it is easier to see in which moments a certain Department has its highest sales.

To visually represent the path made by each boid, we define a set of properties that can be modified to achieve different visualizations with different levels of expressiveness. These properties change how each boid is represented and, consequently, how the consumption data is depicted. First, we define a different colour for each boid so the viewer can associate it with the different Departments (Figure 15.2). Then, although the system enables the fine-tuning of several rendering and behavioural options, which allows the creation of numerous alternative compositions, we focus on two specific rendering approaches.

In the first approach, we represent each boid with a circle. The system allows this circle to be coloured in two different ways: (i) it uses the colour of the corresponding Department to fill the circle area; or (ii) it applies the colour to the stroke line. Furthermore, the circles have different radii according to the consumption at a specific

time. Since the circles may overlap, we resort to transparency to enable the identification of the different shapes. We also implement a mechanism that sorts the circles in depth according to their radii, so that the smaller circles are drawn over the larger ones. Thus, the smaller circles are never hidden by the larger ones. The second approach consists of connecting with lines the centres of the boids that represent the same Department. These lines are also coloured according to the respective Department. In this approach, we can visualise the space between the boids.

In addition to these two approaches, the system allows the modification of the forces of separation, alignment, and cohesion depending on the radius of each boid, and, consequently, depending on the consumption value. Changing these values changes how the boids behave and represent the data. If we define strong forces of separation, the boids will distance themselves more. This movement will then be associated with a moment of high consumption values, which will be easily recognised by the viewer. If the boids have strong forces of separation and alignment, they will create a zigzagging path, considering that they will try to separate and, at the same time, be closer to each other. The modification of these forces enables us to explore different compositions, ones more concerned with the aesthetics of the final result, and others more concerned with the readability of the represented data.

Finally, to help the viewer to analyse the visualizations, we draw 31 dashed lines that divide the spirals into 31 days. These lines may be hidden if the user wants to.

15.2.3 *Experimental Results*

As previously mentioned, the system allows the modification of behavioural and visual properties. This leads to numerous different possibilities to represent the same data. Also, the system is prepared to represent different periods of time and different sets of Departments. To facilitate the exploration of our system, we also implemented a web application in Javascript and Raphael.js, in which the user can change the boid's parameters (e.g., forces and renderings). This web application can be accessed through the following link: https://cdv.dei.uc.pt/cmacas/swarming/.

In this Section, we summarise the experimental process, wherein different parameter settings and approaches were designed, applied, and analysed. During these experiments, we studied the impact of the parameters, as well as of the swarm rules, exploring the equilibrium between readability and aesthetics.

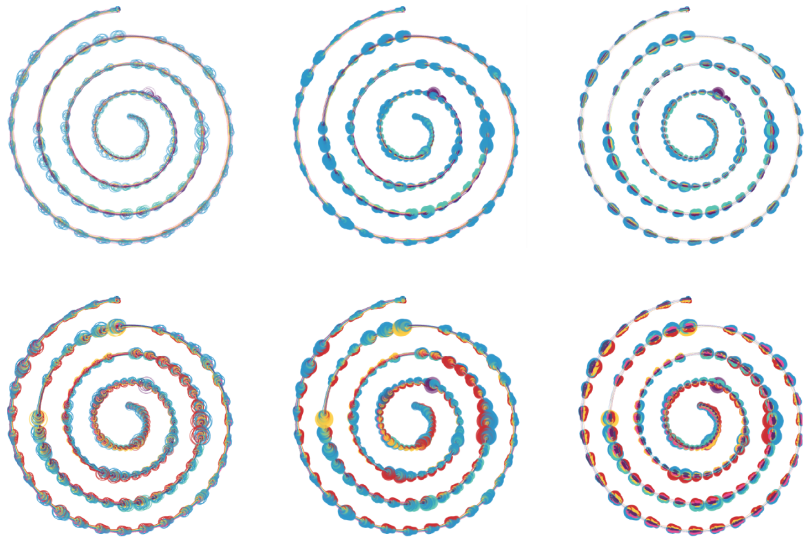

FIGURE 15.3: Circles approach with three different renderings: stroke (left), filled (middle), and sorted (right). In the top line, we use a global normalisation. In the top line, we use a normalisation by Department (local normalisation).

To illustrate this experimentation, we present some explorations of our system. To promote some coherence between the explorations we defined a fixed period of time of 4 months, from October 2012 to January 2013. We chose this period of time because it contains moments of high consumption, such as at Christmas, and more calm periods of time, such as in January.

We begin our exploration with a simple example where there are no forces of separation, alignment, or cohesion. The only rule for each boid is to follow the target boid and change the radius depending on the consumption values. This first exploration is intended to show the different possibilities of visual representation using circles (Figure 15.3)[1]. In Figure 15.3 (top) the consumption values are normalised according to the global minimum and maximum consumption value, and in Figure 15.3 (bottom), the consumption values are normalised according to each individual Department. We can easily perceive the differences between these two types of visualization. In the first, we see that *Grocery* and *Fresh Food* are the Departments with the highest consumption since their colours are more predominant. This type of normalisation enables the reader to compare the consumption in the different Departments, and to see which are the Departments with the highest consumption.

The second type of normalisation gives the reader the possibility to understand how the consumption values behave through time. It is not possible to compare the consumption values between Departments, but we can see the moments in which certain Departments have more sales. Despite these main differences, as it would be expected, we can also observe some similarities, resulting from the fact that the data being explored is the same. For instance, we can see that the consumption behaviour throughout the days is similar: fewer

[1] Large scale renderings of all the visualizations presented in this thesis can be found at http://cdv.dei.uc.pt/swarmviz/.

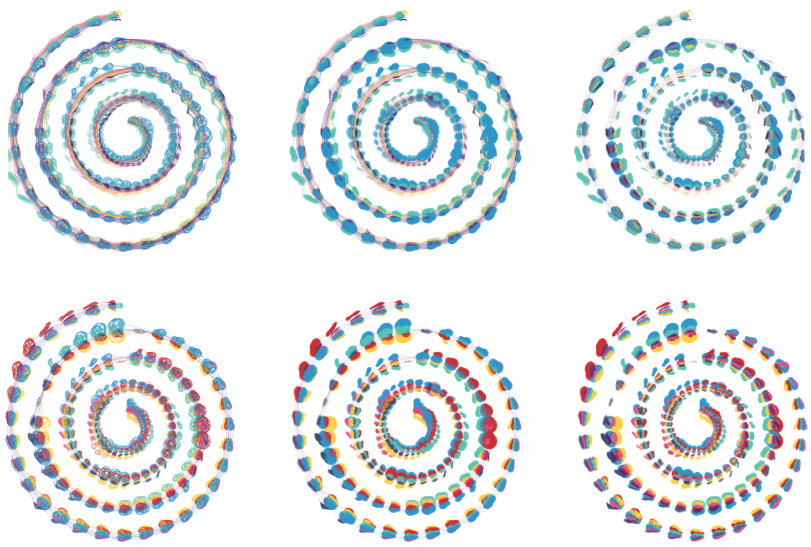

FIGURE 15.4: Circles approach with three different renderings: stroke (left), filled (middle), and sorted (right). In the top line, we use a global normalisation. In the top line, we use a local normalisation. The parameterisation is similar to the one used in the previous examples, but here we apply flocking rules.

sales in the first part of the day, and more consumption during the evening—this can be perceived by the "peanut" shape left by the boids in each day. Then, at the beginning of November, we can see an atypical consumption in the *Leisure* Department. This atypical consumption is caused by a discount made every year in a product category of this Department—Toys. Finally, as it was expected, we can also see higher consumption in December, and discern the impact of weekends in the consumption patterns—Fridays, Saturdays, and Sundays have higher consumption values than the remaining days of the week.

In a second exploration, we analyse the impact of the flocking rules (separation, cohesion and alignment) in the visualization. We apply higher forces to the separation rule, define that the cohesion rule is just applied to the boids of the same Department, and apply higher alignment forces among boids of the same Department than for the boids of different Departments. The results of this exploration are relatively similar to the ones of the previous exploration. However, in this visualization, the boids are more separated, and we can see that when the consumption grows the boids tend to break away from the main path (Figure 15.4).

Then, to further explore this effect of separation, we maintain the same separation forces but we define different minimum distances between the boids. If the boids belong to the same Department, the distance between the centres of the boids can be equal to the radius of each boid. If the boids represent different Departments, the minimum distance between them is equal to the sum of the two radii. In Figure 15.5, it is possible to see that the boids are too separated, making it harder to differentiate between the paths of different months. This exploration creates a cluttered graph that is

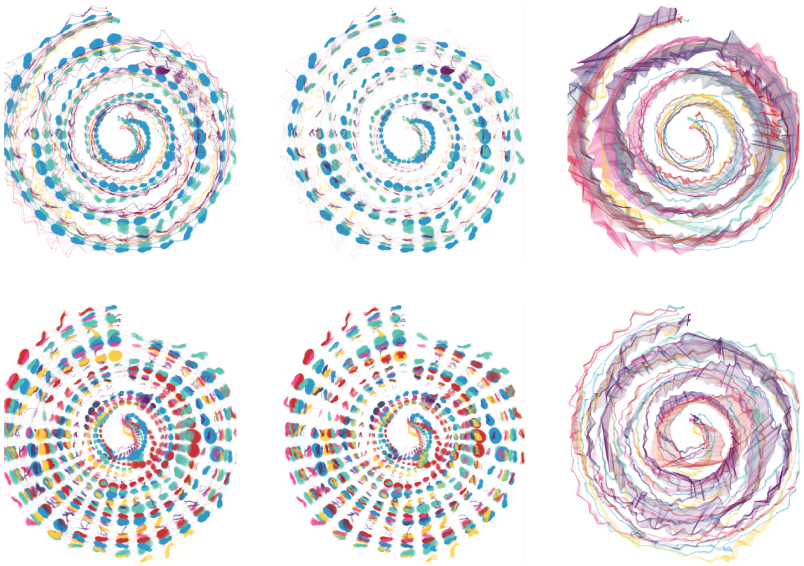

FIGURE 15.5: Circles approach with three different renderings: filled (left), sorted (middle), and lines (right). In the top line, we use a global normalisation. In the top line, we use a local normalisation. The parameterisation is similar to the one used in the previous examples, but with a higher repulsion between boids.

hard to read. From the different visual possibilities, the one with the sorted filled circles is the most readable. In this case, for example, we can distinguish the high consumption in the *Leisure* Department, represented in purple. In this setup, and since the boids are so apart from each other, we wanted to test the visualization of the connections between the boids of the same Department, as a way to increase the readability of the patterns.

The rightmost image of Figure 15.5 illustrates the results obtained when drawing lines to connect the boids of the same Department. As can be observed, since some boids are far apart, possibly because they cannot reach the boids of the same Department, these types of visualizations create big areas corresponding to Departments that do not have big consumption values. Thus, in this case, the emphasis given to a Department tends to be inversely proportional to the number of sales. Admittedly, this type of visualization is difficult, perhaps impossible, to interpret, and violates the expectations of the viewer. Nevertheless, and to some extent because of that, we find them intriguing and aesthetically pleasing. Thus, we consider that although they would be hard to justify in a purely functional Information Visualization project, they have a place in the context of Information Aesthetics, where the artefacts should also be seen as a form of self-expression.

One way to get a less cluttered visualization is to have bigger alignment forces so that the boids do not disperse so much. For this exploration, we apply higher forces to the alignment rule than to the separation rule. Additionally, we used different alignment forces depending on the neighbour's Department. If the neighbour boid is from the same Department, the alignment force is bigger than if the

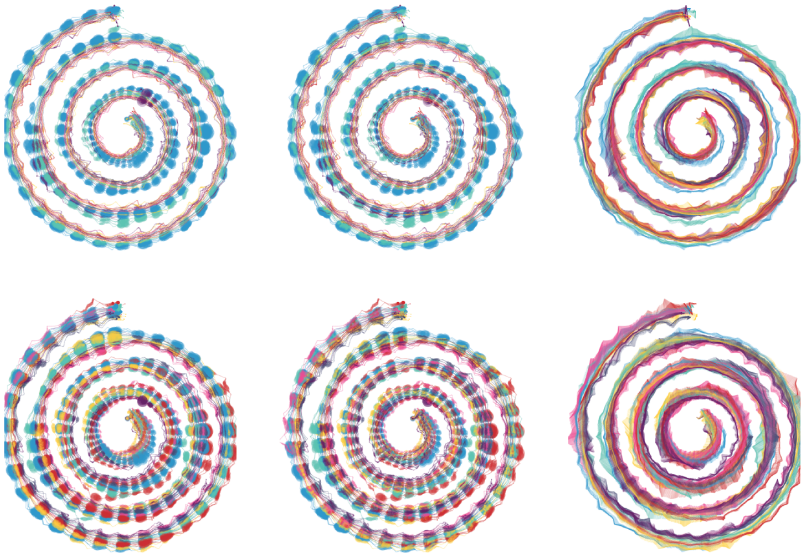

FIGURE 15.6: Circles approach with three different renderings: filled (left), sorted (middle), and lines (right). In the top line, we use a global normalisation. In the top line, we use a local normalisation. The parameterisation is similar to the one used in the previous examples, but with a higher alignment between boids.

boid is from a different one. This setup results in a clearer visualization. In Figure 15.6 (top), where the values are normalised with the global values, the boids corresponding to Departments with the highest consumption, the Grocery and Fresh Food Departments, tend to go to the "outside" of the spiral. Thus, the rules induce an emerging ordering of the boids, making those associated with Departments with lower consumption to approach the centre, while those associated with Departments with high sales are pushed to the periphery. This effect is particularly visible when lines are drawn among the boids of the same Department, as is the case in the rightmost image of Figure 15.6 (top). It is also interesting to observe the disruption of this pattern during Christmas, which is associated with a change in consumption habits.

When inspecting the visualizations using the normalised values for each Department (Figure 15.6, bottom), we can see the same boid's behaviour, where the boids with higher consumption values tend to go to the "outside" of the spiral. In this case, we do not see one or two predominant Departments, but we can perceive the moment at which a certain Department has the biggest consumption value.

As final experimentation, we vary the separation forces over time making them proportional to the associated consumption values. Some examples are presented in Figure 15.7. We also altered the force that makes the boids go after the target boid (see Figure 15.8). We consider that the results are aesthetically appealing in both cases. However, these visualizations become hard to interpret and lack functionality.

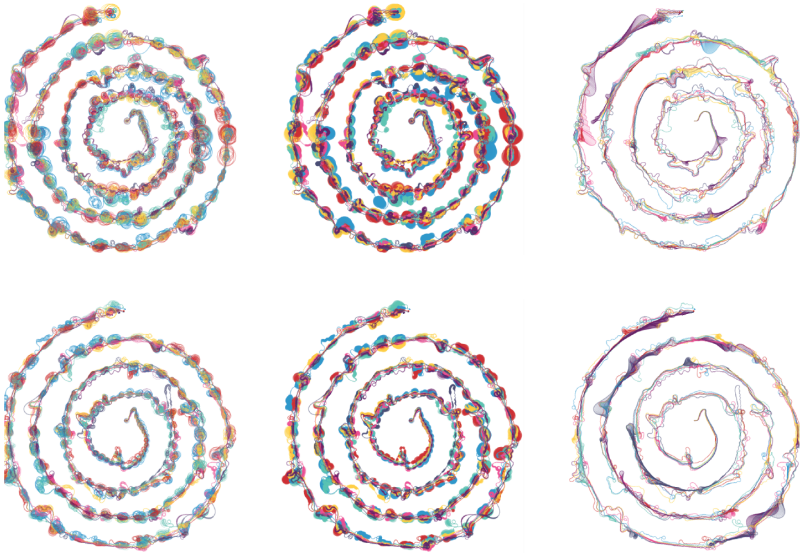

FIGURE 15.7: Circles approach with three different renderings: filled (left), sorted (middle), and lines (right). In the top line, we use a global normalisation. In the top line, we use a local normalisation. The parameterisation is similar to the one used in the previous examples, but varies the separation forces according to the consumption values.

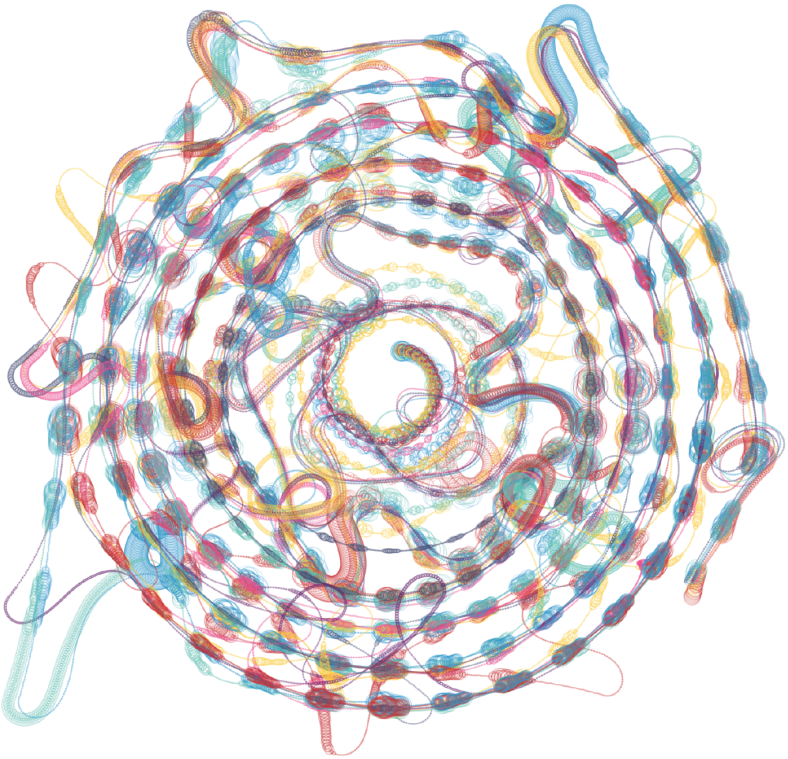

FIGURE 15.8: Circles approach with normalisation by Department: Stroke approach. In this experimentation, we altered the forces of the boids so their speed is higher than the speed of the target boid. This makes the boids to swirl around the target boid, creating uncommon patterns.

15.2.4 *Discussion*

In this project, we generated multiple visual representations of data to get a qualitative view. This enabled us to have an overview of how the consumption values behave through time. We focused on the ability of this emergent system to communicate information while engaging the viewer with organic visuals [136]. Through the experimental results, we could conclude that MAS, particularly swarming systems, can be used in Information Visualization. It enables the representation of patterns and, despite being an inherently fuzzy method, it gives to the reader an overview of how data evolves over time. Additionally, with different parameterisations, we were able to create a set of visual artefacts with different levels of legibility and attractiveness that constitute a form of self-expression.

In the following Section, we will explore in more depth the aesthetics of the system, and validate the results through user evaluation. We intend to further investigate this type of representations, exploring new behavioural rules that enable the boids to represent the consumption values in a wider range of solutions. Also, to make these solutions more appealing for a certain user, we will explore the application of a IEC system, enabling the user to easily guide the creation of new artefacts. Our main goal is to create new, diverse, and surprising visual artefacts, which, despite not being completely functional, are engaging and entertaining.

15.3 Evolving Swarm Artefacts

To further explore the aesthetic dimension and to improve the swarm system detailed in Section 15.2, we couple the swarm system with an EA. By doing so, we increase the degrees of freedom in the creation of visual models, enabling a diverse range of solutions. Note that, instead of being interested in creating a balanced artefact between aesthetics and functionality, we are now only concerned with matching the user's goals and aesthetic preferences. These new artefacts are intended to engage the viewers and lure them into decoding the visualization, or, at least, to entice them to further explore other analytical solutions. Additionally, our intention with this new approach is to develop visual artefacts that are continuously readjusting to the intentions of the user. Nevertheless, we also aim to enable the users to explore artefacts not imagined by them. In this way, SONAE can deploy the system to different audiences and the system will evolve and adapt to the different aesthetic preferences.

15.3.1 *Evolutionary Algorithm*

The visualization model described in Section 15.2 is easy to understand and use, yet it requires the definition of several parameters to create visual artefacts. Therefore, the understanding of these parameters is an important asset to maintain the balance between functionality and aesthetics. If they are not selected with care, we might end with a model that is unreadable and/or visually uninteresting.

To aid in the task of parameterising the visualization models, we propose a framework based on EA [86]. This framework will relieve the burden of searching for parameters that match the goals and preferences of the user. To evolve the swarm system, we will be searching for the best combination of the following system's parameters: (i) the separation, alignment, and cohesion forces; (ii) the minimum and maximum radius; (iii) the use of a global or local normalisation; and, (iv) the representation modes (lines, circles, transparency, sorted circles). In the following subsections, we present the parameters used to evolve the visual artefacts and the used genetic operators.

Representation

The genome of an EA solution is encoded as a set of values that correspond to the number of parameters needed by the swarm system. In concrete, we have 10 different parameters that are required by the visualization model:

Separation Force a real value between 0 and 3;

Alignment Force a real value between 0 and 3;

Cohesion Force a real value between 0 and 3;

Boid Type one of the following: lines, circles and filled circles;

Transparency a boolean value that enables transparency of the boids;

Size Ordered a boolean value. If this value is true, the visualization model sorts the boids by radius, i.e., that the boids with smaller radius will be in front of boids with larger radius;

Mapped a boolean value. If it is true, it indicates that the separation force is mapped depending on the radius of the boids;

Normalisation a boolean value that enables normalisation based on the maximum sales values;

Maximum Radius a real value between 30 and 80 that corresponds to the maximum radius of the boids;

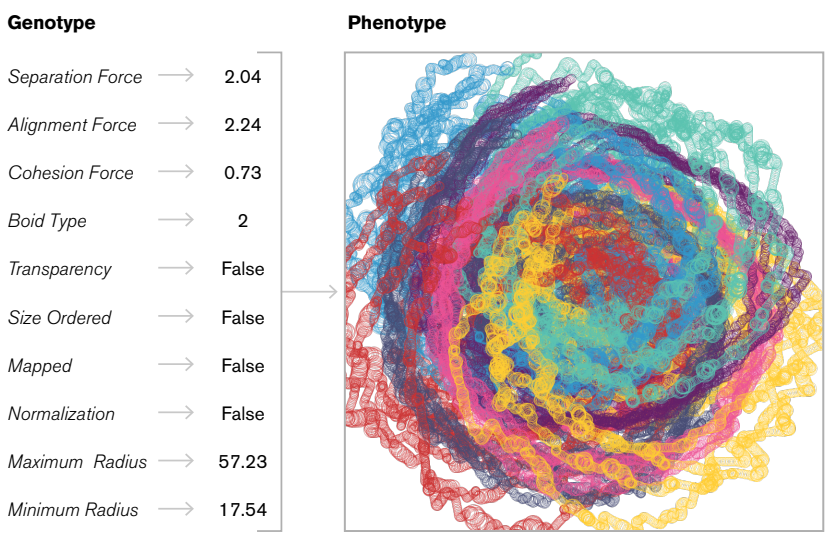

FIGURE 15.9: On the left, the genotype of a possible solution; On the right, the resulting phenotype, i.e., the visualization model.

Minimum Radius a real value between 0.1 and 15 that corresponds to the minimum radius of the boids.

An example of a possible genotype alongside its phenotype representation is depicted in Figure 15.9.

Genetic Operators

To promote the evolution and the proper exploration of the search space we rely on two operators: recombination and mutation. The recombination operator is the uniform crossover and combines two solutions by creating a random mask of the same size as the genotype, and then swap the genetic material according to the previously generated mask.

Regarding the mutation operator, we apply a per gene mutation to the candidate solutions. This allows the algorithm to change, from generation to generation, a significant percentage of the genes to other valid ones. We apply Gaussian mutation to real valued genes, with mean=0 and standard deviation=0.85, bit flip mutation to boolean genes, and random mutation to discrete ones. The mutation probability is set to 15% per gene, which is high when compared with conventional EAs due to the need to present a diverse set of phenotypes to the user.

Fitness Evaluation

The main goal of the EA is to promote the creation of solutions based on the user's preferences. To accomplish that, practitioners often resort to IEC systems, which ask for the user's input to rank the solutions that are being evolved by their preferences [271]. However,

TABLE 15.1: Experimental parameters.

Parameter	Value
Number of runs	30
Population size	20
Number of generations	50
Crossover rate	90%
Mutation rate	15%
Elite size	3 individual
Tournament Size	3 individuals

and to find if the EA is working properly, we also defined a simple metric to evaluate the capacity of the algorithm to traverse the search space. In concrete, we defined a fitness function in which the quality of the individual is proportional to the PNG file size, after the rendering of a solution. The larger the file size, the better the individual.

15.3.2 *Experimental Results*

We will use the EA detailed above to conduct two different experiments. First, we will address the ability of the automatic fitness function to promote evolution towards visualization models that result in images with large sizes in terms of storage. Next, we will evolve visualization models based on user preferences through IEC. We use the same period of time as in the previous experiments.

Experimental Setup

Table 15.1 details the EA parameters used in the experiments conducted in the following Sections. All the values, except the number of generations, are kept fixed throughout the experiments. When performing tests using interactive fitness, the number of maximum generations is not fixed, as the stopping criteria depend on the user preferences. We use high mutation and crossover rates so that the individuals in each generation have noticeable differences, promoting a diverse set of visualization models. This enables a faster convergence towards feasible solutions, i.e., the emergence of individuals that are considered of high quality by the user. Additionally, parent selection is performed using tournament selection.

Automatic Fitness Results

The first set of experiments focuses only on the use of the automatic fitness component. As such, we will analyse the capacity of the fitness function to guide the EA towards solutions with visualization models that take a large amount of space when stored on the hard drive.

FIGURE 15.10: Automatic Fitness Results.

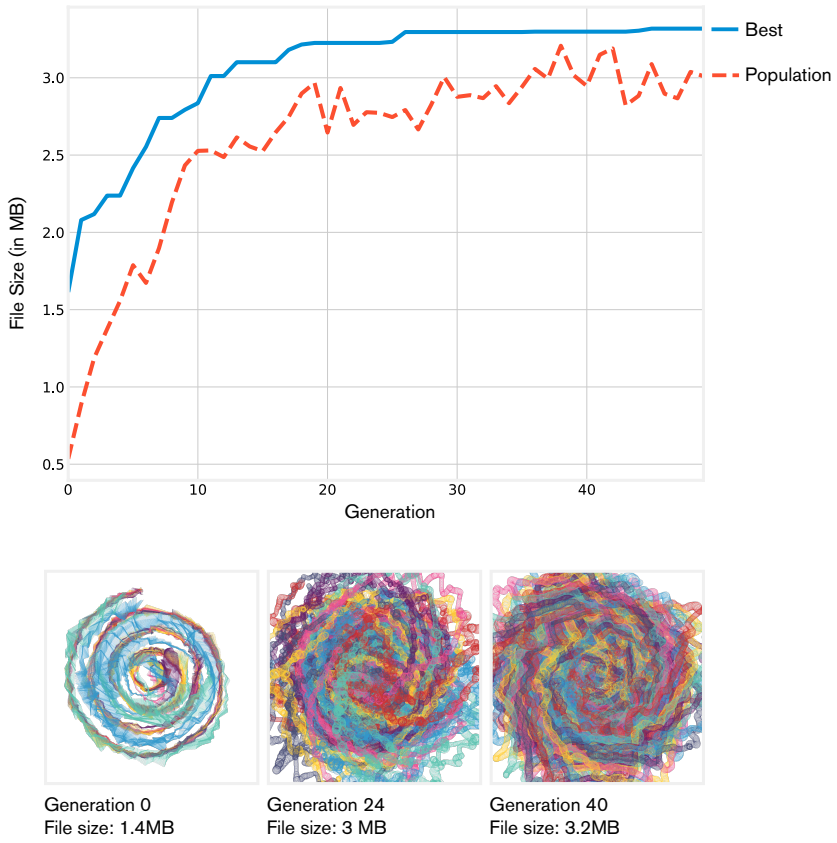

Figure 15.10 shows the evolution of the quality of the solutions, i.e., the size of the files in Megabytes (MB), across 50 generations. The results are averages of 30 runs. An inspection of the results shows that the system can converge towards regions of the search space with visualization models that have large file sizes. Looking at the fitness of the best solutions (solid line), it is possible to see that they have the smallest size in the first generation. Between the 1st and 10th generation, we see a rapid increase in the size, which stagnates around the 20th. After this point, evolution is still occurring, but at a slower pace. The fitness of the population follows the same trend, i.e, it increases rapidly in the first generations, and roughly after the middle of the evolutionary process, it stagnates between files with a size of 2.6 and 3.1 MB.

Looking at the phenotype of the solutions (Figure 15.10 bottom panel) it is possible to see that they start with something that has redundant information since there are not many differences in the colours of the pixels (e.g. many parts of the image are white). Since the PNG uses the Deflate algorithm to compress the image, to increase the file size we need to reduce the compression rate, and that is done by increasing the number of different pixel colours. This is precisely what is happening during the evolutionary process. In the

first generations, the system explores a wide variety of solutions, based on a small set of colours, and with a cleaner canvas. As the evolutionary process progresses, the individuals start to be more visually cluttered (increasing the density of the coloured pixels) and use a wider palette of colours. This increases the file size of each individual since the Deflate algorithm is not able to compress the information (i.e. pixel colours) if they are not redundant. Note that we are not changing the palette colours, to represent the different Departments. These differences emerge from the combination of the different parameters in the genotype such as normalisation and transparency which influence the used colours.

15.3.3 *User Guided Results*

In addition to the previous results, we explored the evolutionary system based on user preferences. These explorations were based on three different objectives and different types of users. In the first exploration, the user is asked to evolve solutions with specific parameterisation attributes: the boids must be represented with filled circles, use the local normalisation, and have a zigzagging pattern. In the second, the user is asked to lean towards solutions that must have a functional dimension, enabling the readability of the artefacts. Finally, for the third exploration, there is no predefined objective, so the solutions depend on the user's preferences. For the first two explorations, the user has some experience with how the system works, its parameters, and the data. In the last, the user has no experience with the data nor with the parameterisations, having no previously defined objective. In this last case, the user only wants to explore the possibilities and create artefacts that match their own preferences. As the generations evolve, the user chooses only the visual artefacts that visually intrigue, amaze, and/or correspond with their preferences.

To avoid user fatigue, we used a reduced number of individuals per generation. For each exploration, the number of individuals is set to 20 and the maximum generations to 10 (this value can be increased). With this parameterisation, we aim to enable the creation of various solutions per generation, giving the user multiple possibilities, without causing a weariness sensation. Additionally, the users can stop the evolutionary process at any given time.

In the first exploration, and as the user already had a target solution in mind, it was possible to perceive that the evolutionary system was evolving correctly towards the user's predefined goal (Figure 15.11). As the individuals were being created, the user selected

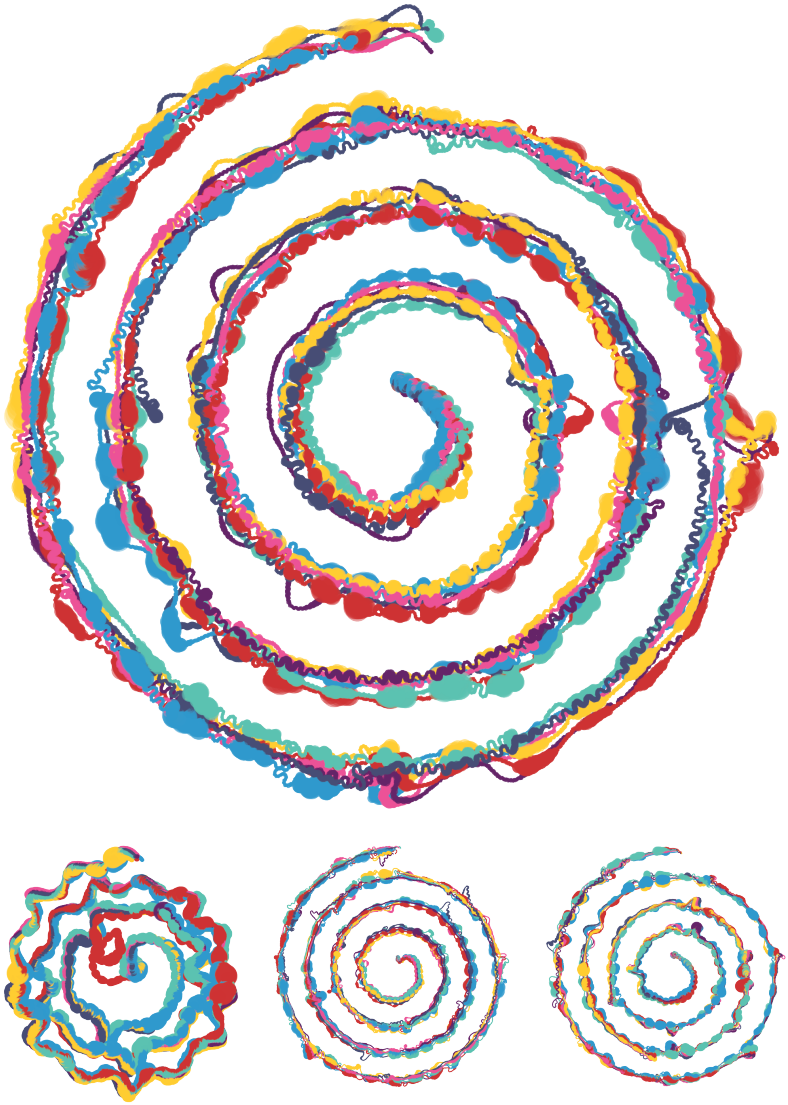

FIGURE 15.11: Chosen individuals by the user during the interactive evolution. In this example, the user aimed to create artefacts that use circles and have a balanced colour palette. The top individual is the final choice past the 10 generations. In the row below, the first artefact is the first choice in the first generation, the individual in the middle was gathered in the third generation, and the rightmost, the chosen one in the sixth generation.

only the ones with filled circles. The user also chose only the ones where all colours appeared balanced, meaning the user chose only artefacts with the local normalisation. When using the global normalisation, the artefacts colours consist mainly of blues and greens—the Departments represented by those colours are the ones with higher consumption values, and thus, the ones with more visual presence. With the local normalisation, each consumption value is mapped between zero and the maximum consumption value of the represented Department, enabling a more balanced distribution of colours.

For the second exploration, and intending to guide the evolutionary system to generate artefacts that can be aesthetic and at the same time functional, the system also proved to be capable of evolving good solutions (Figure 15.12). In the last generations, the majority of the individuals were readable. This means that they have less clutter and lower separation forces, which enables the boids to have

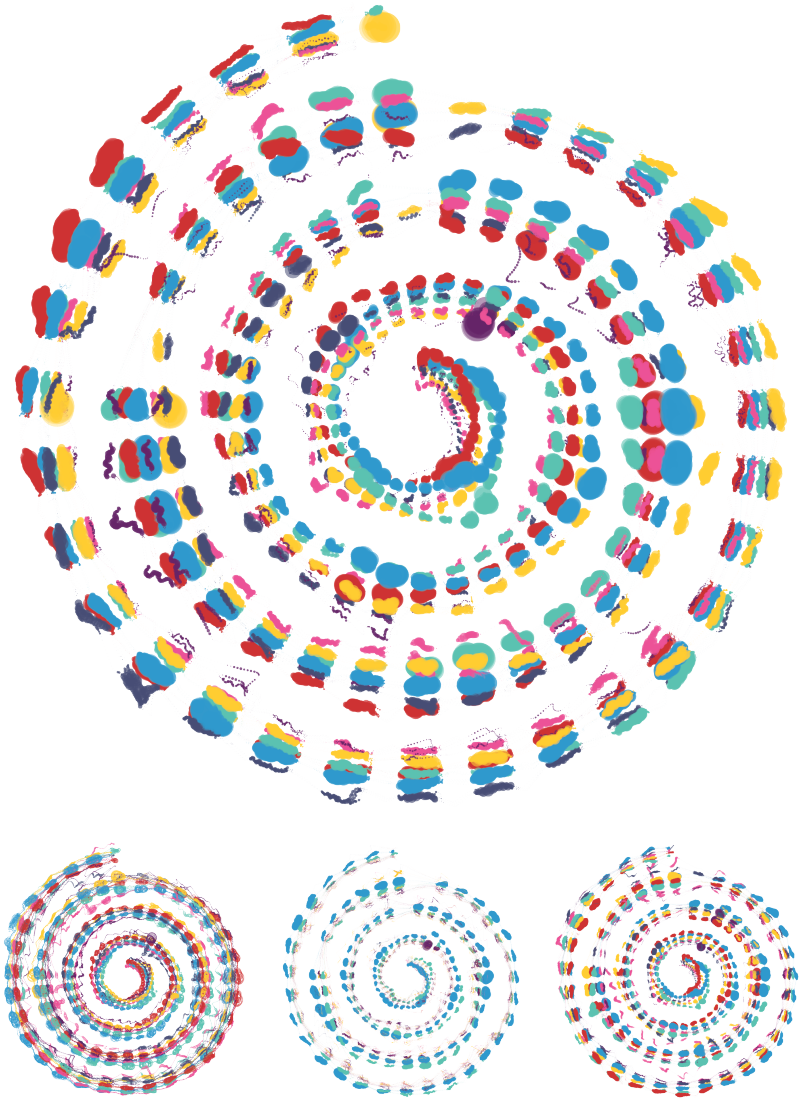

FIGURE 15.12: Chosen individuals by the user during the interactive evolution. In this example, the user aimed to create artefacts that have a functional dimension. The top individual is the final choice after 10 generations. In the row below, the first artefact is the first choice in the first generation, the individual in the middle was gathered in the third generation, and the rightmost, the chosen one in the sixth generation.

a more strict behaviour as they swirl, not deviating from the spiral path. Additionally, the majority of the generated artefacts are similar to the explorations made previously (see Section 15.2). For the majority of the generations, some cluttered visualizations still appeared, eliminating the functionality goal, but introducing some degree of novelty. Furthermore, when the user continued to further explore the system, these non-functional artefacts could contribute to new findings over the data. In this exploration, it was also possible to see an artefact that was never created by the previous system, but that could have a functional dimension (Figure 15.13). In this artefact, the user can visualise the moments in which each Department has its higher consumption values. This is accomplished by the predominant colour throughout the spiral. For example, in the second lap, we can see a wide purple circle, which corresponds to the Leisure Department and a promotional discount on 3 and 4 of November 2012 in which

FIGURE 15.13: An individual with balanced functionality and aesthetics. This solution was retrieved from the second exploration, in which the user aims to create functional visualizations.

all toys had a 50% discount. It is also possible to perceive the highest days of consumption in the Food & Bakery Department on the 24 and 31 of December. These growths of consumption are explained by the festive seasons of Christmas and New Year. This outcome is an interesting example of how the evolutionary system can work to aid in the design of functional artefacts.

For the last exploration, the user had no predefined objective. As the artefacts were being created and presented, the user chose freely the preferred ones. In the beginning, there was no guidance as the user had no previous knowledge of the system and did not know the possible outcomes and how to achieve them. This way, the user chose only the ones which he/she found visually intriguing. Usually, in the end, the user tends to settle on a specific style, e.g., the use of lines to represent the Departments (Figure 15.14). In this case, as the solutions appeared, the user constantly opted for the ones with lines and with swarming forces which make the boids deviate from the spiral path, as can be seen especially in the bottom right image of Figure 15.14.

We created a video to show how the system works and to show how the solutions evolve through time. The video can be accessed in: `https://cdv.dei.uc.pt/ie-of-swarms-in-visualisation/`.

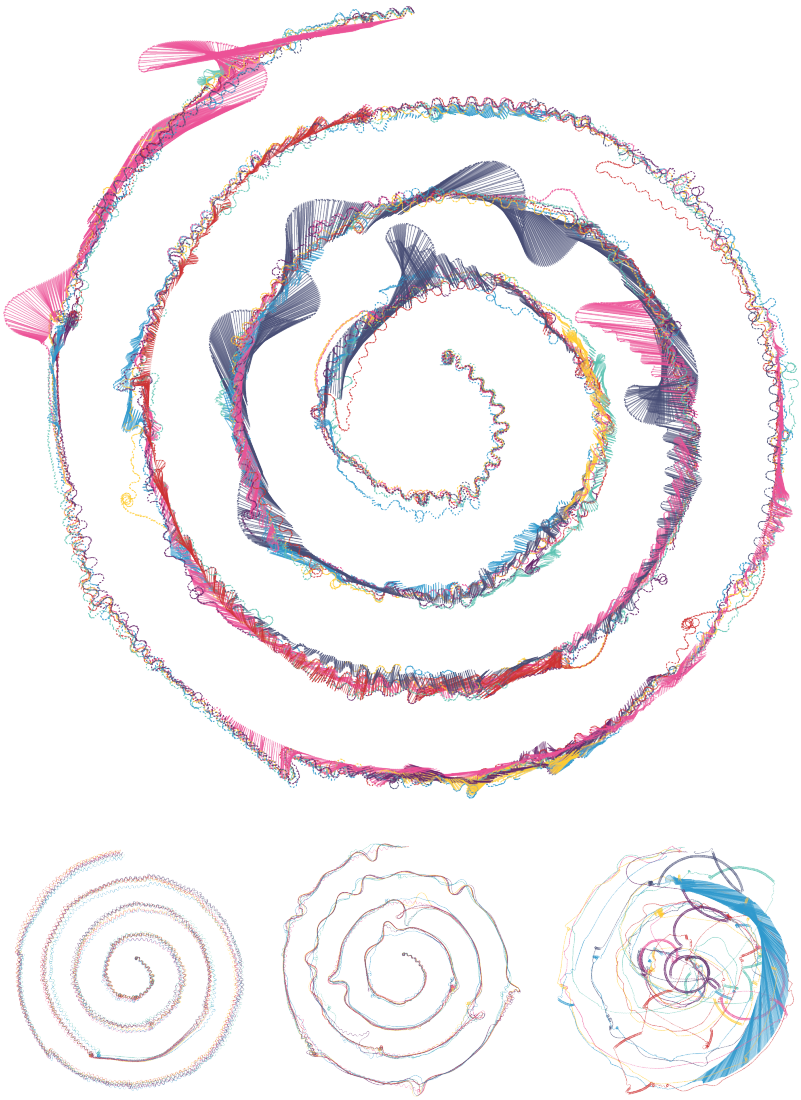

FIGURE 15.14: Chosen individuals by the user during the interactive evolution. In this example, the user has no predefined target solution, choosing only the solutions which fit his/her tastes. The top individual is the final choice past the 10 generations. In the row below, the first artefact is the first choice in the first generation, the individual in the middle was gathered in the third generation, and the rightmost, the chosen one in the sixth generation.

15.3.4 *User Study*

To validate our system with real participants we performed a user study. The objective of this study is to validate the ability of the EA to create visual artefacts that are aesthetically appealing, i.e., that are visually pleasing, and to create a diverse set of artefacts. The study was conducted using an online survey, where the participants were asked to perform three main tasks.

In the first, the participants were asked to select between two images: one that was generated using the evolutionary framework and another that was taken from the previous system, in which the parameterisation is defined manually to explore the functional domain of the artefacts. The used images were randomly paired. To select a set of visual artefacts of the previous system's artefacts, we defined that they should be functional, which resulted in artefacts with fewer

FIGURE 15.15: A small set of the variety of visual artefacts that the evolutionary system can create. Depending on the user guidance this set can be extended to other visual solutions not presented in this thesis.

deviations from the spiral path. This part encompasses 11 questions (**Q1—Q11**).

In the second task, we asked the participants to grade, from 1 to 5 the diversity of a set of images (Figure 15.15), where 1 corresponds to low diversity and 5 to high diversity (**Q12**). Additionally, we asked the participants to rate the set of images depending on their aesthetics (**Q13**) and their ability to captivate the user (**Q14**), both on a scale from 1 to 5, with 1 corresponding to not aesthetically pleasing or not captivating, and 5 to very aesthetically pleasing or very captivating[2].

Finally, we asked the participants if they would use the images as an avatar (**Q15**). Our main purpose with this last question was to understand if the images were interesting enough for the participants, so they would use them as a virtual public display of their preferences. This is a first step in perceiving the applicability of the system in a

[2] A PDF of the User Test can be found in the following link: https://cdv.dei.uc.pt/cmacas/Swarming_UserTest.pdf

real-world situation.

The study sample was composed of 56 participants with ages between 21 and 66 years old, with a median of 27. Their background ranged from Computer Science, Design, Psychology, and Economy.

Results

After gathering the questionnaire results, we analysed the answers using the statistical software SPSS version 24. To analyse the questions we used the χ^2 test since all of our data is categorical. We considered a significance level of $\alpha = 0.05$. The distribution of the answers per question is depicted in Figure 15.16. The black portion of the bars corresponds to the percentage of participants that selected the image generated by the evolutionary algorithm. As we can see by the results there are 6 questions (**Q3**, **Q6**, **Q7**, **Q9**, **Q10**) where the participants preferred the image that was not generated by the EA. On the contrary in **Q1**, **Q2**, and **Q8** the participants favoured the image that was generated by the proposed approach. In both situations, the results have statistical meaningful differences. In all other cases, there are no clear preferences. Based on these outcomes, we analysed the differences between the images presented in the questions and analysed why such differences existed. In the first task, the participants tend to select images that are more packed and coherent. This is evident by the answers to **Q6**, **Q7**, and **Q10**. However, in **Q8**, both images had patterns that deviated from the normal spiral path, and the participants had a preference for the EA artefact.

Concerning the second task, where the participants were asked to evaluate how diverse a set of images was (Figure 15.17), the results show that, in general, the participants found the images diverse, which indicates the capacity of the EA to generate novel images. Then, and to understand if the images were appealing, we showed the same set of images and asked the participants to grade, from 1 to 5, the pleasantness of the images. Results are depicted in Figure 15.18. The results obtained in this question, together with the ones from **Q12**, indicate that the EA is capable of generating diverse images and at the same time are visually interesting.

Moreover, Figure 15.19 shows the results for the question about how captivating the set of images was. Looking at the results, it is possible to see that the majority of the participants think that the images are captivating. This leads us to hypothesise that these visual artefacts would also have value if used outdoors, publicising the retail company, as they would captivate the passersby to look and analyse the outdoors.

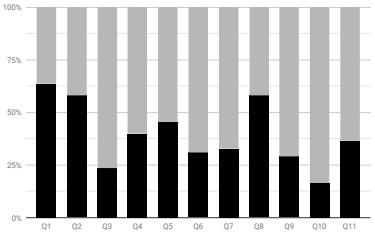

FIGURE 15.16: Distribution of answers in the first task of the questionnaire. The black portion of the bar corresponds to the percentage of participants that selected the artefact created with the EA.

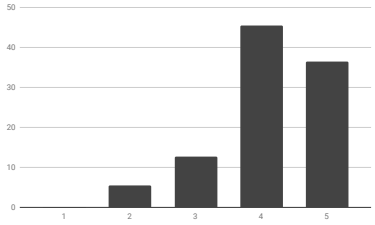

FIGURE 15.17: Distribution (in percentage) of the answers given to the question about the diversity of the images in Q12, grouped from 1 to 5.

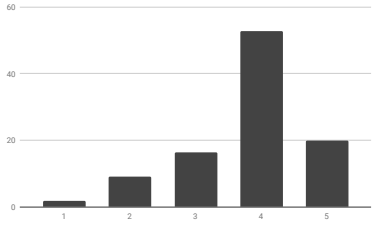

FIGURE 15.18: Distribution (in percentage) of the answers given to the question about the aesthetics in the images in Q13, grouped from 1 to 5.

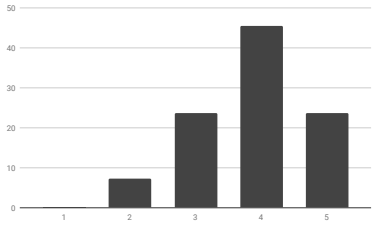

FIGURE 15.19: Distribution (in percentage) of the answers given to the question about how captivating the images in Q14 are, grouped from 1 to 5.

Finally, in **Q15**, we asked the participants if they would use the images as their avatar, for example in a social network or web site. The majority of the participants (61.9%) said that they would use the images as an avatar, whilst (38.1%) say that they would not. Note that the participants were not given the information that the images were generated using data about consumption in retail. If they were allowed to use their data, and generate an image that reflected their own consumption habits, we are convinced that the number of negative answers would decrease substantially, as the artefacts would gain more personal value.

15.3.5 *Discussion*

We started this project by applying a swarm-based system as a method to create emergent visualizations of the consumption values with the intent to convey meaningful information and, at the same time, explore the boundaries between Data visualization and Information Aesthetics. The application of swarms systems to visualise data can also be seen in [283][218]. We focused on the ability of this emergent system to communicate information while engaging the viewer with organic visuals [136]. Additionally, with different parameterisations, we were able to create a set of renderings with different levels of legibility and attractiveness. We used this swarming system to develop a framework that was able to evolve the configuration of the visualization model through the use of an EA. Our goal was to create new, diverse, and surprising visual artefacts, which, despite not being completely functional, could be engaging and entertaining.

We test the validity of the system in three ways: (i) we used an automatic function to evaluate each solution; (ii) we relied on the users' preferences as an input, to guide the solution towards what they find attractive; and, (iii) we validated the generated artefacts with a user study with the main aim of perceiving how the general public reacts to the diversity of the artefacts. We tested three different approaches for the user guidance, in which the final results fitted the three different objectives. As a summary, we found that the system was able to discover artefacts similar to the ones created by human designers, but also to find new ones. Additionally, it was capable of creating a new functional artefact, that was not conceived previously by human designers. Concerning the user study, we could also prove the suitability of our system in real use scenarios, such as the use of our visual artefacts as avatars or its use in outdoors to increase the curiosity of the general public.

One of the important aspects that we looked for in the previous

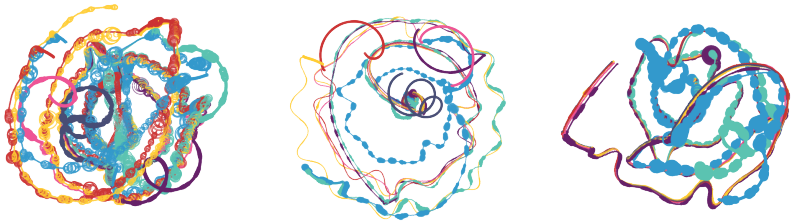

FIGURE 15.20: Individuals that deviate from the spiral path. This is caused by the swarming forces.

FIGURE 15.21: Different individuals with similar parameterisations.

experiments was the capacity of the system to generate and evolve a diversified range of visual artefacts.

The random initialisation of the visualization models creates a wide variety of behaviours. For example, if the forces are too strong, the boids will deviate from the spiral path, creating "random" zigzagging patterns (Figure 15.20). Similarly, these forces can also cause the boids to stagnate their position in the centre.Figure 15.21, illustrates how the system generates a set of individuals, based on a model that was selected by the user. At the start of the evolutionary process, the boids are trapped, swirling around without distancing themselves from the centre of the canvas. This behaviour emerges due to the forces that are applied to the boids. As the generations go by, the parameters change, which will result in the modification of the boids' behaviour. This is caused by the fact that as the user selects the individuals, they have higher chances of being chosen to produce descendants for the next generation. These descendants will inherit genetic traits, and hence visual characteristics, from their parents. The recombination operators promote the exploitation of such traits, while mutation will foster the exploration of the search space.

As the system evolves the parameters, without the direct intervention of the user, it can generate solutions that the user was not expecting and has not seen before. Even an experienced user, that knows the parameters and the system itself, can be surprised by the solutions found. As it can be seen in Figure 15.15, the developed system can create a diverse set of emergent visual artefacts. For example, if the system chooses the lines to represent the Departments, the visual artefact can even lose its swirl effect, as the boids of the same department can be distant from each other, generating long lines

throughout the canvas, and eliminating the spiral path (artefacts in the bottom and upper right corners of Figure 15.15). This last figure presents only a small set of the possible variety of visual solutions that our system can create. As the users' objectives change, the system evolves other visual solutions, creating an undetermined number of different artefacts.

In conclusion, our explorations showed that our framework can be guided by the users to show artefacts similar to the ones the users chose previously but, at the same time, it can give to the users other visual possibilities to enhance their choices. Also, our framework can be used by experienced users that understand the parameters and the possibilities of the visualization model, as well as by users that have no knowledge of the system and only selects artefacts visually relevant to their own preferences.

16

Conclusions

Over the last few years, there was an exponential increase in data in all sectors of business. Companies have seen this as an opportunity to improve their operations and increase the satisfaction of their customers. However, with such large amounts of data, there is a need for efficient tools to allow the effective extraction of knowledge. In Parts II and III, we presented a set of visualization applications that can improve the analysis of data through the representation of intrinsic time-series patterns.

In this part of the thesis, we aimed to represent how the customers of SONAE shop throughout the year in a simplified manner, improving the attractiveness of the visualization for a broader audience. Whereas in the previous parts, we aimed to create analytic tools that enabled the companies analysts to explore the data, in this part, we aimed to make the visualizations more aesthetically appealing to captivate the attention of the visualization reader.

The projects herein presented explore different approaches for the fulfilment of the same goal: to represent the changes in consumption behaviours and analyse the types of products more bought. Moreover, they explore different methods to visualise the data, with different levels of aesthetic explorations. In Chapter 14, we aimed to create an application which, although focused on the beautification and aesthetic side of the consumption data, also fulfilled the functional requirements of the visualization. We applied a small multiples technique to represent the different consumption patterns over time. The small multiples were presented in two different modes: animation and static grid. With the animation of the visualization, we aimed to make the users feel more engaged with the data, augmenting their interest to see the data in more detail through the grid representation of the small multiples.

Knowing in advance that this data has an intrinsic rhythm associated with the differences in consumption during the weeks, we

also explored a multimodal visualization, through sound and image. To represent this data visually, we presented two visualizations, one more concerned with legibility, and the other—which was more exploratory—more concerned with the representation of information through movement. To better emphasise sudden changes or disruptions in consumption patterns, we complemented the visual representations with two percussive sonifications, which were built around Taiko drums. Both sonifications adopted a parameter mapping model, with the volume of the percussion instruments reflecting the numbers of transactions within the different types of consumption. Pitch was also considered as a parameter in the second sonification. In this work, we could perceive that a sonification model that allows filtering or emphasising certain data attributes is something worth to be explored. Additionally, at the visual and auditory levels, the use of temporal zoom, i. e., the representation of data grouped by different units of time (e.g. hour, day, week or month), is also another feature worth exploring.

Overall, we could perceive that through animation it is possible to create a more enjoyable analysis of the data, which is a positive aspect when dealing with wider audiences. However, animation can hinder the overview of the data, as it is difficult to memorise all frames. To overcome this issue, small-multiples can be used to overview the data in a single space and to enable a better comparison between values which are close to each other. Finally, the application of sonification can improve the readability of an animation and highlight moments of greater importance. However, it is important that the data attributes being sonified are the same as the ones represented in the visualization.

In Chapter 15, we presented a swarm-based system as a method to create emergent visualizations of data, exploring the boundaries between Data Visualization and Information Aesthetics. Our contribution extends the literature by applying swarm techniques to the field of Data Visualization and Information Aesthetics, namely to the visualization of consumption patterns. Although not entirely new, the use of swarming techniques in this context is relatively unexplored, particularly for the production of static artefacts. The use of AI techniques in this field are uncommon, which contributes to the novelty of the presented work, and opens opportunities to explore the intersections between Information Visualization, Information Aesthetics and AI.

We extended the swarming visualization to be able to automate the construction of visual artefacts. We explored an automatic manipulation of a swarm system through the application of an EA. These

artefacts can have two different dimensions, one analytical and the other aesthetic. Nonetheless, with this work, our main goal was to explore the aesthetic dimension of the consumption data provided by SONAE. This work was also seen as a step forward to strengthen the relationship between the customer and the SONAE company. By creating appealing visual artefacts with their own data, SONAE can enrich the customers' knowledge on the Portuguese habits of consumption, trigger their curiosity, and lure them to further explore their data more analytically.

Overall, we were able to develop a MAS to create emergent and aesthetic experiences. Then, by implementing an IEC system, we were able to evolve visual artefacts with a higher aesthetic value for the user, as the artefacts result from their guidance. However, some of the final artefacts resulted in less functional visualizations. With these explorations, we could assess the feasibility of implementing techniques from other domains to create visualization models that can represent the data functionally and aesthetically. Nonetheless, in our model, it was difficult to have a more detailed analysis of the consumption values by day without a caption. Hence, the model as it is does not fit for more analytic purposes, where accuracy and precision are needed.

As future work, we intend to improve the IEC system by enabling it to learn the previous choices of the user and guide the evolution based on those parameters. This would allow us to augment the number of individuals per generation and, to diminish user fatigue, only show the fittest (based on previous choices) for the user to choose. Furthermore, this can enhance the capabilities of the system to generate a wider range of different visual artefacts.

Part V

THE END

17

Conclusions

"A graphic is never an end in itself; it is a moment in the process of decision making."
— Jacques Bertin

Information Visualization has its roots in scientific reasoning and is usually seen as an analytical tool. However, with the emergence of programming languages and frameworks focused on the design communities (e.g. JavaScript, Processing), along with the democratisation of data, the conceptual boundaries of Information Visualization, practitioners, and audiences expanded to a more exploratory and user-oriented field [105]. Furthermore, with the growth of collectable data within big enterprises, the interest in Information Visualization tools has increased, and companies from all areas of work gained interest in exploring their data to improve their knowledge over their own business sectors [142]. Hence, Information Visualization became a core area in which knowledge can be applied in different domains and contexts, proving its suitability and importance for businesses and research.

Due to the relevance of time in our daily lives, time-oriented data is one of the most common types of data. Its analysis enables us to better understand the past, present, and future events. In this thesis, we developed a set of visualization tools to improve the analysis of time-oriented data in the business domain. More specifically, and due to the relevance of the representation of trends and patterns in time-varying data [44], we focused our research on the representation of patterns to synthesise complex real-world data, and to facilitate the acquisition of information for *Business Intelligence*. Furthermore, we addressed the role of aesthetics in the representation of time-oriented patterns, and its ability to lure the user into exploring and reading the visualization, enabling the reader to have a more subjective and personal experience.

In this thesis, we investigated diverse approaches for the representation of time-oriented data. We focused on creating visualization models, motivated by specific real-world problems, and on applying

them to real-use scenarios where the visualization of time patterns was essential for the understanding of the data. We divided our research into three main problems in the visualization of time-oriented data: (i) the representation of deviations to highlight atypical values hidden in cyclic time-series (Part II); (ii) the representation of temporal patterns to overview and highlight consecutive events that should not be missed (Part III); and (iii) the representation of temporal rhythms to represent cyclic time-series and capture important trends (Part IV). Although these approaches answer different problems, their main goal is the same: to develop a visualization tool that facilitates the analysis and understanding of time-series data through the highlight of patterns. Hence, our research results in visualization tools that demonstrate the feasibility of representing temporal patterns to highlight specific behaviours in the business domain.

To achieve our goals, in Part I, we reviewed the literature. We started by introducing key topics of Information Visualization, such as visual perception and visual variables, and overviewing the tasks, interaction techniques, and evaluation methods commonly applied in Information Visualization (Chapter 2). We also introduced and characterised the visualization of time-series, by describing its representation methods, analysis goals, interaction mechanisms, and taxonomies specific to time-oriented data (Chapter 3). Finally, we provided a historical context on the visualization of time-series from the pre-eighteenth century to the most recent times (Chapter 4). The study of all this information allowed us to identify standard visualization techniques, and was a source of inspiration for the visualization models presented in the succeeding parts of this thesis.

In Part II, we explored the representation of deviations to facilitate the representation of the Portuguese consumption habits. This research was possible due to the involvement in a research project funded by SONAE. Due to this partnership, we had access to a high volume of data on Portuguese consumption in hundreds of their retail stores. The aim was to research, propose, and develop visualization models to optimise their operations by improving the understanding of how the consumption values are distributed along their product hierarchy (Chapter 5). We started by developing a linear calendar structure (Chapter 6) which provides the mechanisms to: (i) visually explore the consumption evolution over time, within the SONAE's product hierarchy; (ii) detect periodic behaviours; and, (iii) enable the comparison between different days. Then, we explored a radial model of the calendar structure (Chapter 7). The goals were the same as the previous structure but we also aimed to provide the analysts more tools to explore the data further. In the end, we were able to

study the use of space in a radial model, and how it improved the readability of the data. Both calendar views were able to represent the daily consumption deviations from a weekly baseline, highlighting deviations over time, eliminating their intrinsic periodic behaviour, and emphasising moments of disruption. Also, we could demonstrate that the pre-analysis and formulation of statistical values, such as deviations, can contribute to a better summarisation of the time-series.

Overall, in Part II, we contributed with two visualization tools that enabled the representation of the deviations in consumption over time. Additionally, we contributed with a user study that validates the effectiveness of our models.

In Part III, we explored the representation of fraudulent actions in finance. This work was developed in the context of a research project funded by Feedzai, which allowed us access to their data. The aim of this research was to develop visualization tools that could be integrated into their analysis workflow, and enable their analysts to study and detect possible cases of fraud more efficiently (Chapter 9). We were able to explore two distinct sources of information: bank transactions (Chapter 10) and online shopping transactions (Chapter 11). In the first, we aimed to improve the analysis and characterisation of each individual bank client's transactions. In our visualization tool, we represented the time positioning of bank transactions' timestamp along the x-axis and highlighted the typical transactions of a certain client through a Self-Organising Maps (SOM). In the second, our research focused on a specific fraud pattern in finance: Account Takeover (ATO). By analysing its main characteristics and defining the most important behaviours, the consecutive change of personal attributes, we developed a visualization model that highlights such patterns. Through this model, we were able to facilitate the understanding of a wide range of transactions in a single view, and the detection of other related fraud patterns, such as Bot Attack (BA).

Overall, in Part III, we contributed with two visualization tools for the domain of fraud visualization in finance. In both tools, we apply a multiscale timeline technique that aims to: enable the interaction with the data; and, overview the behaviours over time. Furthermore, we developed two profiling systems to enable the distinction and understanding of the behaviours of each individual client. Finally, we also performed a user study with fraud analysts to assess the efficiency of both tools.

In Part IV, we investigated the aesthetic dimension of data visualizations on Portuguese consumption. The research developed for this Part aimed to be a step towards captivating the interest of the big public through aesthetic experiences, luring them to further explore

the data. Hence, our goal was to enable SONAE to communicate with a wider public (Chapter 13). Both works focus on the representation of the consumption values over the period of two years. Although they are intended to explore the aesthetics of visualization, in the first (Chapter 14), we were also concerned with functionality and tried to respect the legibility requirements. In Chapter 14, we explored how visualization can highlight the rhythms and disruptions of the consumption patterns through animation and small-multiples techniques. Additionally, we explored a multimodal representation, adding sonification to the animation of the consumption values over time.

For the second work in Part IV, we applied a swarm-based system as a method to create emergent and organic visualizations of the consumption values with the intent to convey meaningful information and, at the same time, explore the boundaries between Information Visualization and Information Aesthetics (Chapter 15). Additionally, with different parameterisations, we were able to create a set of renderings with different levels of legibility and attractiveness. Then, we developed a framework to evolve the configuration of the visualization model through the use of Evolutionary Algorithm (EA) and Interactive Evolutionary Computation (IEC). With this system, we were able to create new, diverse, and surprising visual artefacts, which, despite not being completely functional, are engaging and entertaining for the user.

Overall, in Part IV, we contributed to the aesthetic domain in time-oriented data. We developed two visualization models which generate visual rhythms, concerning the consumption values in retail. We explored a multimodal approach, and implemented a swarming algorithm to represent the consumption over time. Finally, we performed a set of validation tests to study the aesthetics of these visual abstractions.

The research hypothesis of this thesis is that the visual highlight of temporal patterns and trends can be used to create valuable visualization tools that promote a better analysis and comprehension of the characteristics of time-oriented data. To respond to the research hypothesis, we focused our work on the visual exploration of intrinsic data patterns in different real case scenarios. Through these explorations, we were able to respond to the research questions.

To answer the research question “Which visualization models exist that can be used to structure time?” we focused on the investigation of the state of the art in time-oriented visualizations. First, Chapter 3 enabled us to overview how time is perceived and can be structured for better data analysis. Then, in Chapter 4, we could assess that time is

more commonly represented as a linear structure. However, applying a radial structure to time-dependent data also has its advantages, such as the highlight of periodical data, the reduction of canvas space, and the ease of interactions.

Concerning the research question "Can the representation of time-oriented patterns and their disruptions be useful and valuable for the analysis of real-world data?", we can base our answer on the visualization tools presented in Parts II and III. All tools were tested and assessed according to their efficiency and efficacy, and proved to be useful for the analysis of real-world data. Furthermore, the representation of specific patterns and pre-analysed data enabled us to perceive its utility in synthesising the data and facilitating its comprehension. To answer "which mechanisms should be applied to structure, summarise and emphasise intrinsic data patterns?" we can say that, depending on the goal of the project, calendar and timeline structures can be applied when the temporal distance between events is relevant for the analysis. A sequence of events can be used if the relationship between those events is more important than the timings. Additionally, to answer to "whether such mechanisms should be applied in the pre-analysis of the data or during the construction of the visualization model?" we can refer that to do a pre-analysis of the data and, for example, calculate the deviations (Chapters 6 and 7), can be advantageous when the data is periodic. With this pre-analysis, it is possible to highlight atypical values. On the other hand, applying aggregation mechanisms in the visualization (e.g., represent aggregated or clustered days, as in Chapter 11) can better synthesise the data, reduce the space needed for the visualization, and enhance the appearance of different patterns.

To answer the research question "Can the positioning of the variable time influence the reading of time-oriented patterns and enhance the detection of anomalies, producing clear and valuable insights?", we can focus on the visualization tools presented in Parts II to IV. By structuring the time variables in a calendar manner (Chapters 6 and 7), we were able to synthesise the data, enabling a faster comprehension of the most common patterns due to the familiarity of most users to such structure. Also, time can be mapped in a continuous timeline (Chapters 10 and 15), or be represented as consecutive events (Chapter 11). Whereas in the first, the understanding of the time between events can be more clear, in the second, the behaviours intrinsic to the data can be better highlighted and understood. Additionally, time can be mapped into time itself (Chapter 14), providing a more direct understanding of the evolution of time-dependent variables. All of these findings also answer the questions: "how should

time variables be mapped into the visual space?" and "which are the most appropriate ones for different tasks?".

Finally, concerning the research question "Can temporal patterns from business datasets be beautified and simplified for wider audiences?" we focus on the works developed in Part IV. Overall, we were able to represent the data from SONAE in a simplified way, and enhance its aesthetic value. In both works (Chapters 14 and 15), we were able to maintain the functional part of the visualization. However, in Chapter 15 by enabling the users to adapt the visual models to their visual taste, the functionality of the artefacts was not a goal. Hence, and in response to "can aesthetics be applied without reducing functionality?", we argue that when focusing only on aesthetics, the functionality of the visual artefact may be reduced (Chapter 15). However, such models are usually more enticing for the users, and may increase their curiosity to explore the data. Aesthetics and functionality can also be used in a balanced way (Chapter 14), which may suggest that Information Visualization should not discard aesthetics. Finally, to be able to answer the question "can the users be able to adapt the visual models to their visual taste?" we evolved a set of visual artefacts through IEC. This allowed us to conclude that it is possible to generate visual artefacts based on the user's preferences. However, such models may hinder the functional part of the visualization model.

In what concerns the dissemination of our research, most of the contributions have been presented at international conferences and published in international journals. The resulting publications are listed below, sorted by each Part of the thesis.

The works presented in Part II were published in:

[178] C. Maçãs et al. "Time-series Application on Big Data - Visualization of Consumption in Supermarkets". In: *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015*. Ed. by J. Braz, A. Kerren, and L. Linsen. SciTePress, 2015, pp. 239–246. DOI: 10.5220/0005307702390246. URL: https://doi.org/10.5220/0005307702390246

[185] C. Maçãs and P. Machado. "Radial Calendar of Consumption". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 96–102. DOI: 10.1109/iV.2018.00027. URL: https://doi.org/10.1109/iV.2018.00027

The publication [185] was also exhibited in *Romaria Cultural*, Gouveia, Portugal, in 2018.

The works presented in Part III were published in:

[188] C. Maçãs, E. Polisciuc, and P. Machado. "VaBank: Visual Analytics for Banking Transactions". In: *24th International Conference Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*. 2020, pp. 336–343. DOI: 10.1109/IV51561.2020.00062. URL: https://doi.org/10.1109/IV51561.2020.00062

The work presented in Chapter 11 was exhibited in *Porto Design Biennale*, Porto, Portugal, in 2019.

The works presented in Part IV were published in:

[179] C. Maçãs, P. Cruz, P. Martins, and P. Machado. "Swarm Systems in the Visualization of Consumption Patterns". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Q. Yang and M. J. Wooldridge. AAAI Press, 2015, pp. 2466–2472. URL: http://ijcai.org/Abstract/15/349

[183] C. Maçãs and P. Machado. "The Rhythm of Consumption". In: *5th Joint Symposium on Computational Aesthetics, Sketch-Based Interfaces and Modeling, and Non-Photorealistic Animation and Rendering, Expressive 2016 - Posters, Artworks, and Bridging Papers, Lisbon, Portugal, May 7-9, 2016, Proceedings*. Ed. by E. Akleman, L. Bartram, A. Çamci, A. G. Forbes, and P. Machado. Eurographics Association, 2016, pp. 11–12. DOI: 10.2312/exp.20161259. URL: https://doi.org/10.2312/exp.20161259

[184] C. Maçãs and P. Machado. "The Rhythm of consumptions". In: *IEEE VIS Arts Program*. Baltimore, EUA, 2016

[186] C. Maçãs, P. Martins, and P. Machado. "Consumption as a Rhythm: A Multimodal Experiment on the Representation of Time-Series". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 504–509. DOI: 10.1109/iV.2018.00093. URL: https://doi.org/10.1109/iV.2018.00093

[181] C. Maçãs, N. Lourenço, and P. Machado. "Interactive Evolution of Swarms for the Visualisation of consumptions". In: *Interactivity, Game Creation, Design, Learning, and Innovation - 7th EAI International Conference, ArtsIT 2018, and 3rd EAI International*

Conference, DLI 2018, ICTCC 2018, Braga, Portugal, October 24-26, 2018, Proceedings. Ed. by A. L. Brooks, E. Brooks, and C. Sylla. Vol. 265. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2018, pp. 101–110. DOI: 10.1007/978-3-030-06134-0_11. URL: https://doi.org/10.1007/978-3-030-06134-0_11

[182] C. Maçãs, N. Lourenço, and P. Machado. "Evolving visual artefacts based on consumption patterns". In: *Int. J. Arts Technol.* 12.1 (2020), pp. 60–83. DOI: 10.1504/IJART.2020.107693. URL: https://doi.org/10.1504/IJART.2020.107693

The work in [183, 184] was also exhibited in *IEEE VIS Arts Program*, Baltimore, EUA, in 2016, and the work [181] was featured as a cover artwork at *SIGEVOlution*, in 2019.

Additionally, during the thesis, other works covering Information Visualization, although not time-oriented, were published. For example, during our collaboration with SONAE, the following publications were published:

[231] E. Polisciuc et al. "Arc and Swarm-based Representations of Customer's Flows among Supermarkets". In: *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015*. Ed. by J. Braz, A. Kerren, and L. Linsen. SciTePress, 2015, pp. 300–306. DOI: 10.5220/0005316503000306. URL: https://doi.org/10.5220/0005316503000306

[180] C. Maçãs, P. Cruz, E. Polisciuc, H. Amaro, and P. Machado. "Isoedges for the Geovisualization of consumptions". In: *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP, Rome, Italy, February 27-29, 2016*. Ed. by N. Magnenat-Thalmann et al. SciTePress, 2016, pp. 222–229. DOI: 10.5220/0005785702200227. URL: https://doi.org/10.5220/0005785702200227

[232] E. Polisciuc, P. Cruz, H. Amaro, C. Maçãs, and P. Machado. "Flow Map of Products Transported among Warehouses and Supermarkets". In: *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP, Rome, Italy, February 27-29, 2016*. Ed. by N. Magnenat-Thalmann et al. SciTePress,

2016, pp. 179–188. DOI: 10.5220/0005787301770186. URL: https://doi.org/10.5220/0005787301770186

[233] E. Polisciuc, C. Maçãs, F. Assunção, and P. Machado. "Hexagonal gridded maps and information layers: a novel approach for the exploration and analysis of retail data". In: *SIGGRAPH ASIA 2016, Macao, December 5-8, 2016 - Symposium on Visualization*. Ed. by W. Chen and D. Weiskopf. ACM, 2016, 6:1–6:8. DOI: 10.1145/3002151.3002160. URL: https://doi.org/10.1145/3002151.3002160

[187] C. Maçãs, E. Polisciuc, and P. Machado. "GlyphSOMe: Using SOM with Data Glyphs for Customer Profiling". In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 3: IVAPP, Valletta, Malta, February 27-29, 2020*. Ed. by A. Kerren, C. Hurter, and J. Braz. SCITEPRESS, 2020, pp. 301–308. DOI: 10.5220/0009178803010308. URL: https://doi.org/10.5220/0009178803010308

Concerning our exploration with force-directed graphs and glyph representation, seen in Chapter 10, the following publications were published:

[189] C. Maçãs, A. Rodrigues, G. Bernardes, and P. Machado. "MixMash: A Visualisation System for Musical Mashup Creation". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 471–477. DOI: 10.1109/iV.2018.00088. URL: https://doi.org/10.1109/iV.2018.00088

[190] C. Maçãs, A. Rodrigues, G. Bernardes, and P. Machado. "MixMash: An Assistive Tool for Music Mashup Creation from Large Music Collections". In: *Int. J. Art Cult. Des. Technol.* 8.2 (2019), pp. 20–40. DOI: 10.4018/IJACDT.2019070102. URL: https://doi.org/10.4018/IJACDT.2019070102

In summary, the results of this research show that representing temporal patterns and their disruptions can enhance the understanding of the data, and increase the knowledge for business intelligence. Hence, this thesis demonstrates the feasibility of highlighting temporal patterns to promote better data analysis.

From this research, it was possible to perceive that it is important to understand the specific characteristics of each time-series data (e.g.,

patterns, disruptions, trends). To create appropriate visualization models, these characteristics should be considered and emphasised, avoiding repetition and clutter that would hinder the analysis of important information (e.g., atypical values and patterns). Furthermore, the use of position to determine the time data point can facilitate the comprehension of the data distribution over time, and enable the use of other visual variables to represent more complex time-dependent variables. In relation to the positioning of time data points, it was possible to establish that analysts have more ease in reading linear or tabular visualizations as they are closer to their common models. However, radial models provoke more interest in the users, resulting in visualization models which promote exploration and thus, may improve the acquisition of information.

This thesis focused on several visualization approaches, raising awareness for the need for a deeper study on the introduction of visualization in the business domain, and the necessity to create specific models for specific problems and contexts. For example, it was possible to analyse the importance of the stakeholders' background when defining the visual models to be used. While for the general public, visualization models can provide only an overview of the data, for analysts, visualization models should provide an overview of the data but also provide the details. Additionally, we could assess the impact of the level of visual complexity of the visualization model in readability. High visual complexity can have a higher aesthetic impact for the users, but more difficult to read. In contrast, low visual complexity can be easier to read, especially for inexperienced users, but tends to have less aesthetic impact than high visual complexity.

We also consider that widening the range of visualization techniques of time-oriented data will open new horizons in representing complex data in intelligent and human-centred ways, promoting efficient and enjoyable experiences. Additionally, to solve the requirements of each specific problem and develop our visualization models, our research focused on multiple areas, including AI, ML, HCI, design, and statistics. Hence, we argue that Information Visualization is a multidisciplinary domain, and will only benefit from the collaboration of areas both from Science and Design.

As for future work, we aim to expand this research to more domains and, in this way, investigate how Information Visualization can improve data analysis in other contexts. In addition, we also intend to research the combination of Information Visualization and AI. In the present thesis, we emphasised the application of AI techniques to represent data—through a MAS in Chapter 15—and the use of IEC to evolve visual artefacts according to user preferences. Additionally, we

used ML models to retrieve the transactions topology (Chapter 10) and visualised the ML results in finance fraud (Chapter 11). However, we aim to deepen our research in both areas.

AI and ML research is rapidly gaining importance and expanding its domains of application. Some of the most notable ML techniques are black-box methods. Moreover, even when the techniques do not fall into what is typically considered black-box approaches (e.g. when random forests are used, as in Feedzai's case), the complexity of the models makes them a *de facto* black-box. In other words, in practice, once trained, it tends to be unfeasible to understand how and why a given model produces a given output.

The black-box nature of the models raises two kinds of problems: (i) it is difficult for the researcher to understand, and thus enhance, the model; and, (ii) end-users are put in a difficult position, not knowing when to accept or question the outcome of the model. In what concerns the use of Information Visualization in ML, our main goal sheds light on black-box methods, by developing a set of Information Visualization tools to assist researchers and end-users. We argue that the knowledge retrieved from this thesis on time-series visualization can be applied to enable the users to perceive how the models evolve through time, and which are the patterns to highlight in ML black-boxes. By these means, we will promote the development of better models and assist decision-making by end-users, ultimately contributing to superior and informed decisions. We will build upon previous work to develop visualization tools to: (i) analyse the training, i.e. learning, thus helping researchers to fine-tune their algorithms and obtain better models; (ii) visualise the classification process, thus assisting the end-user in the decision making process by providing insights about the inner working of the models; and, (iii) contribute to the assessment and promotion of fairness and transparency.

Lastly, in what concerns the application of AI techniques to Information Visualization, we are particularly interested in the automatic selection and parametrisation of visualization models by AI techniques. In the present thesis, we showed the feasibility of using AI techniques to parameterise a visualization model according to the users' preferences. However, in our work, the users needed to manually select every artefact so the AI model could learn their visual taste. Hence, we aim to take a further step and develop tools that automatically adapt to the nature of the data, users' needs, preferences, and resources. In this context, the integration of EA techniques into the design process of Information Visualization will also be explored.

Ultimately, we will continue exploring the intersection of Information Visualization, Aesthetics, Visual Analytics, and AI.

Bibliography

[1] W. Aigner, A. Rind, and S. Hoffmann. "Comparative Evaluation of an Interactive Time-Series Visualization That Combines Quantitative Data with Qualitative Abstractions". In: *Comput. Graph. Forum* 31.3pt2 (June 2012), pp. 995–1004. ISSN: 0167-7055. DOI: 10.1111/j.1467-8659.2012.03092.x. URL: https://doi.org/10.1111/j.1467-8659.2012.03092.x (cit. on p. 63).

[2] W. Aigner. *Interactive Visualization and Data Analysis: Visual Analytics With a Focus on Time*. Habilitation Thesis. 2013. URL: http://publik.tuwien.ac.at/files/PubDat_227076.pdf (cit. on pp. 15, 27, 30, 31, 35–39).

[3] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. "Visualizing time-oriented data—a systematic view". In: *Computers & Graphics* 31.3 (2007), pp. 401–409 (cit. on pp. 4, 6, 41, 45, 85).

[4] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011 (cit. on pp. 5, 6, 16, 23, 30, 33–40, 43, 50–56, 58–61, 72).

[5] R. Amar, J. Eagan, and J. Stasko. "Low-Level Components of Analytic Activity in Information Visualization". In: *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*. INFOVIS '05. USA: IEEE Computer Society, 2005, p. 15. ISBN: 078039464x. DOI: 10.1109/INFOVIS.2005.24. URL: https://doi.org/10.1109/INFOVIS.2005.24 (cit. on p. 25).

[6] K. Andrews. "Evaluating Information Visualisations". In: *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*. BELIV '06. Venice, Italy: Association for Computing Machinery, 2006, pp. 1–5. ISBN: 1595935622. DOI: 10.1145/1168149.1168151. URL: https://doi.org/10.1145/1168149.1168151 (cit. on p. 172).

[7] G. Andrienko, N. Andrienko, P. Bak, S. Bremm, D. Keim, T. von Landesberger, C. Pölitz, and T. Schreck. "A framework for using self-organising maps to analyse spatio-temporal patterns, exemplified by analysis of mobile phone usage". In: *Journal of Location Based Services* 4.3-4 (2010), pp. 200–221. DOI: 10.1080/17489725.2010.532816 (cit. on p. 127).

[8] N. V. Andrienko and G. L. Andrienko. *Exploratory analysis of spatial and temporal data - a systematic approach*. Springer, 2006. ISBN: 978-3-540-25994-7. DOI: 10.1007/3-540-31190-4. URL: https://doi.org/10.1007/3-540-31190-4 (cit. on p. 25).

[9] M. Ankerst, D. A. Keim, and H. peter Kriegel. "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets". In: 1996 (cit. on p. 65).

[10] G. Arthur Van, F. Staals, M. Loffler, J. Dykes, and B. Speckmann. "Multi-Granular Trend Detection for Time-Series Analysis." eng. In: *IEEE Trans Vis Comput Graph* 23.1 (2017), pp. 661–670. ISSN: 1941-0506 (Electronic); 1077-2626 (Linking). DOI: 10.1109/TVCG.2016.2598619 (cit. on p. 62).

[11] A. S. Association et al. "Joint committee on standards for graphic presentation". In: *Publications of the American Statistical Association* 14.112 (1915), pp. 790–797 (cit. on p. 60).

[12] C. A. Astudillo and B. J. Oommen. "Topology-oriented self-organizing maps: a survey". In: *Pattern Analysis and Applications* 17.2 (2014), pp. 223–248. DOI: 10.1007/s10044-014-0367-9. URL: https://doi.org/10.1007/s10044-014-0367-9 (cit. on p. 127).

[13] B. Bach, P. Dragicevic, D. Archambault, C. Hurter, and S. Carpendale. "A Review of Temporal Data Visualizations Based on Space-Time Cube Operations". In: *EuroVis - STARs*. Ed. by R. Borgo, R. Maciejewski, and I. Viola. The Eurographics Association, 2014. ISBN: 978-3-03868-028-4. DOI: 10.2312/eurovisstar.20141171 (cit. on p. 4).

[14] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. "Time curves: Folding time to visualize patterns of temporal evolution in data". In: *IEEE transactions on visualization and computer graphics* 22.1 (2015), pp. 559–568 (cit. on pp. 4, 67, 72).

[15] K. Bale, P. Chapman, N. Barraclough, J. Purdy, N. Aydin, and P. Dark. “Kaleidomaps: A New Technique for the Visualization of Multivariate Time-Series Data”. In: *Information Visualization* 6.2 (Jan. 2007), pp. 155–167. DOI: 10.1057/palgrave.ivs.9500154. URL: https://doi.org/10.1057/palgrave.ivs.9500154 (cit. on pp. 64, 67).

[16] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. “Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 2573–2582. ISBN: 9781605589299. DOI: 10.1145/1753326.1753716. URL: https://doi.org/10.1145/1753326.1753716 (cit. on p. 192).

[17] B. B. Bederson and A. Boltman. “Does animation help users build mental maps of spatial information?” In: *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis’99)*. 1999, pp. 28–35. DOI: 10.1109/INFVIS.1999.801854 (cit. on p. 198).

[18] S. Bender. *Taiko Boom: Japanese Drumming in Place and Motion*. 1st ed. University of California Press, 2012. ISBN: 9780520272415. URL: http://www.jstor.org/stable/10.1525/j.ctt1ppx45 (cit. on p. 209).

[19] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch. “The Aesthetics of Graph Visualization”. In: *Computational Aesthetics in Graphics, Visualization, and Imaging*. Ed. by D. W. Cunningham, G. Meyer, and L. Neumann. The Eurographics Association, 2007. ISBN: 978-3-905673-43-2. DOI: 10.2312/COMPAESTH/COMPAESTH07/057-064 (cit. on p. 189).

[20] S. Berkovich and D. Liao. “On Clusterization of "Big Data" Streams”. In: *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications*. COM.Geo ’12. Washington, D.C., USA: Association for Computing Machinery, 2012. ISBN: 9781450311137. DOI: 10.1145/2345316.2345320. URL: https://doi.org/10.1145/2345316.2345320 (cit. on p. 71).

[21] L. Berry and T. Munzner. “BinX: Dynamic Exploration of Time Series Datasets Across Aggregation Levels”. In: *InfoVis* (2004), pp. 215.2–215.2 (cit. on p. 61).

[22] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. ISBN: 978-1-589-48261-6. Redlands: Esri Press, 2010. ISBN: 978-1-589-48261-6 (cit. on pp. 3, 15, 16, 19, 21, 24, 25, 30, 35, 60).

[23] M. Bianchi, M. Boyle, and D. Hollingsworht. “A comparison of methods for trend estimation”. In: *Applied Economics Letters* 6.2 (1999), pp. 103–109. DOI: 10.1080/135048599353726. eprint: https://doi.org/10.1080/135048599353726. URL: https://doi.org/10.1080/135048599353726 (cit. on p. 198).

[24] D. Bihanic, ed. *New Challenges for Data Design*. Springer London, 2015. DOI: 10.1007/978-1-4471-6596-5. URL: https://doi.org/10.1007%2F978-1-4471-6596-5 (cit. on p. 190).

[25] D. Bihanic. *New challenges for data design*. Springer, 2015 (cit. on p. 4).

[26] A. Black, P. Luna, O. Lund, and S. Walker. *Information design: research and practice*. Taylor & Francis, 2017 (cit. on pp. 16–19, 27–30, 44, 47, 48, 50–54, 57, 72).

[27] M. Bögl, P. Filzmoser, T. Gschwandtner, T. Lammarsch, R. A. Leite, S. Miksch, and A. Rind. “Cycle Plot Revisited: Multivariate Outlier Detection Using a Distance-Based Abstraction”. In: *Comput. Graph. Forum* 36.3 (June 2017), pp. 227–238. ISSN: 0167-7055. DOI: 10.1111/cgf.13182. URL: https://doi.org/10.1111/cgf.13182 (cit. on p. 63).

[28] R. J. Bolton and D. J. Hand. “Statistical fraud detection: A review”. In: *Statistical science* (2002), pp. 235–249 (cit. on pp. 119–121, 123, 153).

[29] A. Borgmann. “The Depth of Design”. In: *Discovering design: explorations in design studies*. Ed. by R. Buchanan and V. Margolin. Vol. 13. 22. University of Chicago Press, 1995 (cit. on p. 193).

[30] A. Bowie. “Aesthetics versus functionality: challenging dichotomies in information visualisation”. In: *Image & Text: a Journal for Design* 2011.18 (2011), pp. 64–81 (cit. on pp. 192–195).

[31] I. Boyandin, E. Bertini, P. Bak, and D. Lalanne. “Flowstrates: An Approach for Visual Exploration of Temporal Origin-Destination Data”. In: *Proceedings of the 13th Eurographics / IEEE—VGTC Conference on Visualization*. EuroVis’11. Bergen, Norway: The Eurographs Association & John Wiley & Sons, Ltd., 2011, pp. 971–980. DOI: 10.1111/j.1467-8659.2011.01946.x. URL: https://doi.org/10.1111/j.1467-8659.2011.01946.x (cit. on p. 75).

[32] U. Brandes, B. Nick, B. Rockstroh, and A. Steffen. "Gestaltlines". In: *Computer Graphics Forum* 32.3pt2 (2013), pp. 171–180. DOI: 10.1111/cgf.12104. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12104. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12104 (cit. on p. 63).

[33] M. Brehmer and T. Munzner. "A Multi-Level Typology of Abstract Visualization Tasks". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2376–2385 (cit. on p. 26).

[34] P. J. Brockwell, R. A. Davis, and S. E. Fienberg. *Time series: theory and methods: theory and methods*. Springer Science & Business Media, 1991 (cit. on p. 38).

[35] A. Buja, D. Cook, and D. F. Swayne. "Interactive high-dimensional data visualization". In: *Journal of computational and graphical statistics* 5.1 (1996), pp. 78–99 (cit. on p. 41).

[36] J. Bulley and D. Jones. *Living Symphonies*. 2015. URL: http://www.livingsymphonies.com/ (cit. on p. 199).

[37] P. Buono, C. Plaisant, A. Simeone, A. Aris, B. Shneiderman, G. Shmueli, and W. Jank. "Similarity-Based Forecasting with Simultaneous Previews: A River Plot Interface for Time Series Forecasting". In: *2007 11th International Conference Information Visualization (IV '07)*. 2007, pp. 191–196 (cit. on pp. 61, 62).

[38] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. "Interactive pattern search in time series". In: *Visualization and Data Analysis 2005*. Ed. by R. F. Erbacher, J. C. Roberts, M. T. Grohn, and K. Borner. SPIE, Mar. 2005. DOI: 10.1117/12.587537. URL: https://doi.org/10.1117/12.587537 (cit. on p. 67).

[39] P. Buono, M. F. Costabile, and R. Lanzilotti. "A Circular Visualization of People's Activities in Distributed Teams". In: *J. Vis. Lang. Comput.* 25.6 (2014), pp. 903–911. ISSN: 1045-926X. DOI: 10.1016/j.jvlc.2014.10.025. URL: https://doi.org/10.1016/j.jvlc.2014.10.025 (cit. on pp. 64, 76, 77).

[40] M. Burch and S. Diehl. "TimeRadarTrees: Visualizing Dynamic Compound Digraphs". In: *Computer Graphics Forum* 27.3 (May 2008), pp. 823–830. DOI: 10.1111/j.1467-8659.2008.01213.x. URL: https://doi.org/10.1111/j.1467-8659.2008.01213.x (cit. on p. 64).

[41] V. Burgio and M. Moretti. "Infographics as Images: Meaningfulness beyond Information". In: *Proceedings* 1.9 (2017). ISSN: 2504-3900. DOI: 10.3390/proceedings1090891. URL: https://www.mdpi.com/2504-3900/1/9/891 (cit. on p. 192).

[42] L. Byron and M. Wattenberg. "Stacked Graphs – Geometry Aesthetics". In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1245–1252. DOI: 10.1109/TVCG.2008.166 (cit. on pp. 74, 75).

[43] A. Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. Voices That Matter Series. New Riders, 2013. ISBN: 9780321834737. URL: https://books.google.pt/books?id=BiT1ugAACAAJ (cit. on pp. 119, 192).

[44] M. Card. *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999 (cit. on pp. 4, 15, 41, 247).

[45] S. K. Card and J. Mackinlay. "The structure of the information visualization design space". In: *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*. IEEE. 1997, pp. 92–99 (cit. on p. 41).

[46] E. Cardinaels. "The interplay between cost accounting knowledge and presentation formats in cost-based decision-making". In: *Accounting, Organizations and Society* 33.6 (2008), pp. 582 –602. ISSN: 0361-3682. DOI: https://doi.org/10.1016/j.aos.2007.06.003. URL: http://www.sciencedirect.com/science/article/pii/S0361368207000463 (cit. on p. 179).

[47] J. V. Carlis and J. A. Konstan. "Interactive Visualization of Serial Periodic Data". In: *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*. UIST '98. San Francisco, California, USA: Association for Computing Machinery, 1998, pp. 29–38. ISBN: 1581130341. DOI: 10.1145/288392.288399. URL: https://doi.org/10.1145/288392.288399 (cit. on pp. 65, 75, 76).

[48] M. S. T. Carpendale. *Considering Visual Variables as a Basis for Information Visualisation*. Tech. rep. Calgary, AB: University of Calgary, 2003 (cit. on pp. 19–23).

[49] S. Carpendale. "Evaluating Information Visualizations". In: *Information Visualization: Human-Centered Issues and Perspectives*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 19–45. ISBN: 9783540709558. URL: https://doi.org/10.1007/978-3-540-70956-5_2 (cit. on p. 30).

[50] S. M. Casner. "Task-Analytic Approach to the Automated Design of Graphic Presentations". In: *ACM Trans. Graph.* 10.2 (Apr. 1991), pp. 111–151. ISSN: 0730-0301. DOI: 10.1145/108360.108361. URL: https://doi.org/10.1145/108360.108361 (cit. on p. 25).

[51] E. Catmull and R. Rom. "A Class of Local Interpolating Splines". In: *Computer Aided Geometric Design*. Ed. by R. E. B. R. F. Riesenfield. Academic Press, 1974, pp. 317 –326. ISBN: 978-0-12-079050-0. DOI: https://doi.org/10.1016/B978-0-12-079050-0.50020-5. URL: http://www.sciencedirect.com/science/article/pii/B9780120790500500205 (cit. on p. 80).

[52] N. Cawthon and A. V. Moere. "The effect of aesthetic on the usability of data visualization". In: *2007 11th International Conference Information Visualization (IV'07)*. IEEE. 2007, pp. 637–648 (cit. on pp. 4, 193, 194).

[53] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. "WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions". In: *2007 IEEE Symposium on Visual Analytics Science and Technology*. 2007, pp. 155–162 (cit. on pp. 44, 62, 122, 123, 125).

[54] R. Chang, A. Lee, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. "Scalable and Interactive Visual Analysis of Financial Wire Transactions for Fraud Detection". In: *Information Visualization* 7.1 (2008), pp. 63–76. DOI: 10.1057/palgrave.ivs.9500172 (cit. on p. 136).

[55] D. Chankhihort, B.-M. Lim, G.-J. Lee, S. Choi, S.-O. Kwon, S.-H. Lee, J.-T. Kang, A. Nasridinov, and K.-H. Yoo. "A Visualization Scheme with a Calendar Heat Map for Abnormal Pattern Analysis in the Manufacturing Process". In: *International Journal of Contents* 13.2 (2017), pp. 21–28 (cit. on p. 65).

[56] C. Chen. "Top 10 Unsolved Information Visualization Problems." In: *IEEE Computer Graphics and Applications* 25.4 (2005), pp. 12–16 (cit. on p. 15).

[57] C.-h. Chen, W. K. Härdle, and A. Unwin. *Handbook of data visualization*. Springer Science & Business Media, 2007 (cit. on pp. 43, 45–47, 51, 52, 54, 56–61).

[58] X. Chen. *Sky Color of 10 Chinise Cities*. 2011. URL: http://www.xiaoji-chen.com/2011/sky-color-of-10-chinese-cities/2011/sky-color-of-10-chinese-cities (cit. on p. 76).

[59] Q. Chengzhi, Z. Chenghu, and P. Tao. "Taxonomy of visualization techniques and systems–Concerns between users and developers are different". In: *Asia GIS Conference*. Vol. 35. 2003, p. 37 (cit. on p. 41).

[60] E. H.-h. Chi. "A taxonomy of visualization techniques using the data state reference model". In: *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*. IEEE. 2000, pp. 69–75 (cit. on p. 41).

[61] E. H.-h. Chi and J. T. Riedl. "An operator interaction framework for visualization systems". In: *Proceedings IEEE Symposium on Information Visualization (Cat. No. 98TB100258)*. IEEE. 1998, pp. 63–70 (cit. on p. 41).

[62] M. C. Chuah and S. F. Roth. "On the semantics of interactive visualizations". In: *Proceedings IEEE Symposium on Information Visualization'96*. IEEE. 1996, pp. 29–36 (cit. on pp. 26, 41).

[63] Chung-Chian Hsu. "Generalizing self-organizing map for categorical data". In: *IEEE Transactions on Neural Networks* 17.2 (2006), pp. 294–304. DOI: 10.1109/TNN.2005.863415 (cit. on p. 127).

[64] A. Claudia and L. O. Arlindo. "Temporal Data Mining: an Overview". In: *Proceedings of KDD Workshop on Temporal Data Mining*. 2001, pp. 1–13 (cit. on p. 38).

[65] W. C. Cleveland and M. E. McGill. *Dynamic Graphics for Statistics*. 1st. USA: CRC Press, Inc., 1988. ISBN: 053409144X (cit. on p. 28).

[66] W. S. Cleveland. *The Elements of Graphing Data*. USA: Wadsworth Publ. Co., 1985. ISBN: 0534037305 (cit. on pp. 73, 102).

[67] W. S. Cleveland and R. McGill. "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods". In: *Journal of the American Statistical Association* 79.387 (1984), pp. 531–554. ISSN: 01621459. URL: http://www.jstor.org/stable/2288400 (cit. on p. 199).

[68] W. S. Cleveland and R. McGill. "An Experiment in Graphical Perception." In: *Int. J. Man Mach. Stud.* 25.5 (1986), pp. 491–501 (cit. on p. 24).

[69] E. Costanza, B. Bedwell, M. Jewell, J. Colley, and T. Rodden. "'A bit like British Weather, I suppose' Design and Evaluation of the Temperature Calendar". In: *CHI 2016 (12/05/16)*. 2016. URL: https://eprints.soton.ac.uk/385035/ (cit. on p. 65).

[70] A Costea, A Kloptchenko, B Back, I Ivan, and I Rosca. "Analyzing economical performance of central-east-European countries Using neural networks and cluster analysis". In: *Proceedings of the Fifth International Symposium on Economic Informatics*. Bucharest, Romania. 2001, pp. 1006–1011 (cit. on p. 126).

[71] S. B. Cousins and M. G. Kahn. "The visual display of temporal information". In: *Artificial Intelligence in Medicine* 3.6 (1991). Medical Temporal Reasoning, pp. 341 –357. ISSN: 0933-3657. DOI: https://doi.org/10.1016/0933-3657(91)90005-V. URL: http://www.sciencedirect.com/science/article/pii/093336579190005V (cit. on p. 62).

[72] P. Craig and N. Roa-Seïler. "A Vertical Timeline Visualization for the Exploratory Analysis of Dialogue Data". In: *2012 16th International Conference on Information Visualisation*. 2012, pp. 68–73 (cit. on p. 62).

[73] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. J. Gao, X. Tong, and H. Qu. "TextFlow: towards better understanding of evolving topics in text." eng. In: *IEEE Trans Vis Comput Graph* 17.12 (2011), pp. 2412–2421. ISSN: 1941-0506 (Electronic); 1077-2626 (Linking). DOI: 10.1109/TVCG.2011.239 (cit. on pp. 62, 63).

[74] T. N. Dang, N. Pendar, and A. G. Forbes. "TimeArcs: Visualizing Fluctuations in Dynamic Networks". In: *Computer Graphics Forum* 35.3 (2016), pp. 61–69. DOI: 10.1111/cgf.12882. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12882. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12882 (cit. on p. 62).

[75] F. de Wilde. *[NRS] Vectors 4 [UN]Certainty*. 2011 (cit. on p. 195).

[76] W. Didimo, G. Liotta, F. Montecchiani, and P. Palladino. "An advanced network visualization system for financial crime detection". In: *2011 IEEE Pacific Visualization Symposium*. 2011, pp. 203–210. DOI: 10.1109/PACIFICVIS.2011.5742391 (cit. on p. 123).

[77] W. Didimo, G. Liotta, and F. Montecchiani. "Vis4AUI: Visual Analysis of Banking Activity Networks." In: *GRAPP/IVAPP*. 2012, pp. 799–802 (cit. on p. 123).

[78] S. Diehl, F. Beck, and M. Burch. "Uncovering Strengths and Weaknesses of Radial Visualizations—an Empirical Approach". In: *IEEE Transactions on Visualization and Computer Graphics* 16.6 (2010), pp. 935–942. DOI: 10.1109/TVCG.2010.209 (cit. on pp. 64, 76, 162).

[79] W. N. Dilla and R. L. Raschke. "Data visualization for fraud detection: Practice implications and a call for future research". In: *International Journal of Accounting Information Systems* 16 (2015), pp. 1 –22. ISSN: 1467-0895. DOI: https://doi.org/10.1016/j.accinf.2015.01.001. URL: http://www.sciencedirect.com/science/article/pii/S1467089515000020 (cit. on pp. 120, 122, 153, 160).

[80] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara. "Interactive focus+ context analysis of large, time-dependent flow simulation data". In: *Simulation* 82.12 (2006), pp. 851–865 (cit. on p. 39).

[81] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou. "LeadLine: Interactive visual analysis of text data through event identification and exploration". In: *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 2012, pp. 93–102 (cit. on p. 62).

[82] P. Dragicevic and S. Huot. "SpiraClock". In: *CHI'02 extended abstracts on Human factors in computing systems - CHI 2002*. ACM Press, 2002. DOI: 10.1145/506443.506505. URL: https://doi.org/10.1145/506443.506505 (cit. on p. 65).

[83] G. M. Draper, Y. Livnat, and R. F. Riesenfeld. "A Survey of Radial Methods for Information Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 15.5 (2009), pp. 759–776. DOI: 10.1109/TVCG.2009.23 (cit. on pp. 64, 73, 75, 76, 162).

[84] J. Drucker. *Graphesis: Visual Forms of Knowledge Production*. MetaLABprojects Series. Harvard University Press, 2014. ISBN: 9780674724938. URL: https://books.google.pt/books?id=zO0vnwEACAAJ (cit. on p. 3).

[85] M. Dumas, M. J. McGuffin, and V. L. Lemieux. "Financevis. net-a visual survey of financial data visualizations". In: *Poster Abstracts of IEEE Conference on Visualization*. Vol. 2. 2014 (cit. on p. 122).

[86] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, 2015, pp. 1–258. ISBN: 978-3-662-44874-8 (cit. on p. 227).

[87] T. Eklund, B. Back, H. Vanharanta, and A. Visa. "Assessing the Feasibility of Using Self-Organizing Maps for Data Mining Financial Information". In: *Proceedings of the 10th European Conference on Information Systems (ECIS) 2002*. Ed. by S. Wrycza. Vol. 1. AIS, 2002 (cit. on pp. 126, 128).

[88] A. Eldridge. "You Pretty Little Flocker: Exploring the Aesthetic State Space of Creative Ecosystems". In: *Artif. Life* 21.3 (Aug. 2015), pp. 289–292. ISSN: 1064-5462. DOI: 10.1162/ARTL_a_00169. URL: https://doi.org/10.1162/ARTL_a_00169 (cit. on p. 218).

[89] G. Ellis and A. Dix. "An Explorative Analysis of User Evaluation Studies in Information Visualisation". In: BELIV '06. Venice, Italy: Association for Computing Machinery, 2006. ISBN: 1595935622. DOI: 10.1145/1168149.1168152. URL: https://doi.org/10.1145/1168149.1168152 (cit. on p. 172).

[90] M. E. Elmer. "Symbol considerations for bivariate thematic maps". In: *Proceedings of 26th International Cartographic Conference* (2013) (cit. on p. 24).

[91] O. J. Espinosa, C. Hendrickson, and J. Garrett. "Domain analysis: a technique to design a user-centered visualization framework". In: *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis' 99)*. IEEE. 1999, pp. 44–52 (cit. on p. 41).

[92] J. Eyler. "The changing assessments of John Snow's and William Farr's cholera studies". In: *Sozial- und Präventivmedizin* 46 (2005), pp. 225–232 (cit. on p. 76).

[93] B. Farinelli. *Fraud Bots: Is Your Business at Risk?* URL: https://blog.clear.sale/fraud-bots-is-your-business-at-risk (cit. on p. 154).

[94] P. Federico, S. Hoffmann, A. Rind, W. Aigner, and S. Miksch. "Qualizon Graphs: Space-Efficient Time-Series Visualization with Qualitative Abstractions". In: *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. AVI '14. Como, Italy: Association for Computing Machinery, 2014, pp. 273–280. ISBN: 9781450327756. DOI: 10.1145/2598153.2598172. URL: https://doi.org/10.1145/2598153.2598172 (cit. on p. 63).

[95] D. Feeney. "Synchronizing Times I: Greece and Rome". In: *Caesar's Calendar: Ancient Time and the Beginnings of History*. 1st ed. University of California Press, 2007, pp. 7–42. ISBN: 9780520251199. URL: http://www.jstor.org/stable/10.1525/j.ctt1pndn8.6 (cit. on p. 44).

[96] D. Fernöndez-Prieto, C. Naranjo-Valero, J. T. Hernández, and H. Hagen. "STRAD Wheel: Web-Based Library for Visualizing Temporal Data". In: *IEEE Computer Graphics and Applications* 37.2 (2017), pp. 99–105 (cit. on p. 64).

[97] S. Few. "The Chartjunk Debate A Close Examination of Recent Findings". In: *Visual Business Intelligence Newsletter* (2011) (cit. on p. 192).

[98] A. Findeli. "Ethics, Aesthetics, and Design". In: *Design Issues* 10.2 (1994), pp. 49–68. ISSN: 07479360, 15314790. URL: http://www.jstor.org/stable/1511628 (cit. on p. 192).

[99] D. Fisher, S. Drucker, and M. Czerwinski. "Business Intelligence Analytics [Guest editors' introduction]". In: *IEEE Computer Graphics and Applications* 34.5 (2014), pp. 22–24. DOI: 10.1109/MCG.2014.86 (cit. on p. 72).

[100] B. Foo. *Data-Driven DJ | Two Trains*. 2015. URL: https://datadrivendj.com/tracks/subway (cit. on p. 200).

[101] A. U. Frank. "Different types of "times" in GIS". In: *Spatial and temporal reasoning in geographic information systems* (1998), pp. 40–62 (cit. on p. 33).

[102] T. C. Franke. *Enabling New Perspectives on the Weather Through Data Visualizations*. URL: https://www.cleverfranke.com/work/weather-charts (cit. on p. 76).

[103] J. Frascara. "Graphic Design: Fine Art or Social Science?" In: *Design Issues* 5.1 (1988), pp. 18–29. ISSN: 07479360, 15314790. URL: http://www.jstor.org/stable/1511556 (cit. on p. 192).

[104] "Fraud Prevention, Detection, and Response". In: *Essentials of Forensic Accounting*. John Wiley & Sons, Ltd, 2017. Chap. 8, pp. 211–243. ISBN: 9781119449423. DOI: 10.1002/9781119449423.ch8 (cit. on pp. 119, 120).

[105] M. Friendly. "A Brief History of Data Visualization". In: *Handbook of Data Visualization* (2008), pp. 15–56. DOI: 10.1007/978-3-540-33037-0_2. URL: http://dx.doi.org/10.1007/978-3-540-33037-0_2 (cit. on pp. 3, 15, 46, 189, 247).

[106] B. J. Fry. "Computational information design". PhD thesis. Massachusetts Institute of Technology, 2004 (cit. on pp. 8, 193, 194).

[107] S. Fu, J. Zhao, H. F. Cheng, H. Zhu, and J. Marlow. "T-Cal: Understanding Team Conversational Data with Calendar-Based Visualization". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13. ISBN: 9781450356206. DOI: 10.1145/3173574.3174074. URL: https://doi.org/10.1145/3173574.3174074 (cit. on p. 67).

[108] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. "Identifying Users Profiles from Mobile Calls Habits". In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. UrbComp '12. Beijing, China: Association for Computing Machinery, 2012, pp. 17–24. ISBN: 9781450315425. DOI: 10.1145/2346496.2346500 (cit. on p. 127).

[109] R. Gioria, A. Nogare, and A Rao. *VANISHING / Digital / Data sonification.* 2014. URL: https://www.behance.net/gallery/18558125/VANISHING-Digital-Data-sonification (cit. on p. 200).

[110] I. A. Goralwalla, M. T. Özsu, and D. Szafron. "An object-oriented framework for temporal data models". In: *Temporal Databases: Research and Practice*. Springer, 1998, pp. 1–35 (cit. on p. 41).

[111] J. M. L. Gorricha and V. J. A. S. Lobo. "On the Use of Three-Dimensional Self-Organizing Maps for Visualizing Clusters in Georeferenced Data". In: *Information Fusion and Geographic Information Systems: Towards the Digital Ocean*. Ed. by V. V. Popovich, C. Claramunt, T. Devogele, M. Schrenk, and K. Korolenko. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 61–75. ISBN: 978-3-642-19766-6. DOI: 10.1007/978-3-642-19766-6_6 (cit. on p. 127).

[112] G. Graham. *Philosophy of the Arts: An Introduction to Aesthetics*. Taylor & Francis, 2005. ISBN: 9781134271221. URL: https://books.google.pt/books?id=cwa0QGyrfLcC (cit. on p. 192).

[113] G. Greenfield and P. Machado. "Swarm Art: Curator's Introduction". In: *Leonardo* 47.1 (2014), pp. 5–7. DOI: 10.1162/LEON{_}a{_}00695. URL: http://dx.doi.org/10.1162/LEON_a_00695 (cit. on p. 216).

[114] H. Gruendl, P. Riehmann, Y. Pausch, and B. Froehlich. "Time-Series Plots Integrated in Parallel-Coordinates Displays". In: *Comput. Graph. Forum* 35.3 (June 2016), pp. 321–330. ISSN: 0167-7055 (cit. on p. 63).

[115] Y. Han, A. Rozga, N. Dimitrova, G. D. Abowd, and J. Stasko. "Visual Analysis of Proximal Temporal Relationships of Social and Communicative Behaviors". In: *Computer Graphics Forum* 34.3 (2015), pp. 51–60. DOI: 10.1111/cgf.12617. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12617. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12617 (cit. on p. 62).

[116] S. O. Hansson. "Aesthetic functionalism". In: *Contemporary Aesthetics* 3 (2005). QC 20120221 (cit. on p. 192).

[117] M. C. Hao, Umeshwar Dayal, D. A. Keim, and T. Schreck. "Importance-driven visualization layouts for large time series data". In: *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 2005, pp. 203–210 (cit. on pp. 63, 67).

[118] M. C. Hao, U. Dayal, D. A. Keim, and T. Schreck. “Multi-resolution techniques for visual exploration of large time-series data”. In: *EUROVIS 2007*. 2007, pp. 27–34 (cit. on pp. 66, 67).

[119] M. C. Hao, M. Marwah, H. Janetzko, U. Dayal, D. A. Keim, D. Patnaik, N. Ramakrishnan, and R. K. Sharma. “Visual exploration of frequent patterns in multivariate time series”. In: *Information Visualization* 11.1 (2012), pp. 71–83. DOI: 10.1177/1473871611430769. eprint: https://doi.org/10.1177/1473871611430769. URL: https://doi.org/10.1177/1473871611430769 (cit. on p. 62).

[120] H. Hauser, F. Ledermann, and H. Doleisch. “Angular brushing of extended parallel coordinates”. In: *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.* IEEE. 2002, pp. 127–130 (cit. on p. 39).

[121] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. “ThemeRiver: visualizing thematic changes in large document collections”. In: *IEEE Transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 9–20 (cit. on pp. 60, 63).

[122] J. Heer, N. Kong, and M. Agrawala. “Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations”. In: *ACM Human Factors in Computing Systems (CHI)*. 2009. URL: http://vis.stanford.edu/papers/horizon (cit. on pp. 62, 74).

[123] T. Hermann, A. Hunt, and J. G. Neuhoff, eds. *The Sonification Handbook*. Berlin, Germany: Logos Publishing House, 2011, pp. 1–586. ISBN: 978-3-8325-2819-5 (cit. on p. 199).

[124] A. R. Hevner. “A three cycle view of design science research”. In: *Scandinavian journal of information systems* 19.2 (2007), p. 4 (cit. on p. 7).

[125] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. “DNA visual and analytic data mining”. In: *Proceedings. Visualization 1997 (Cat. No. 97CB36155)*. 1997, pp. 437–441 (cit. on p. 64).

[126] R. E. Horn. “Information design: Emergence of a new profession”. In: *Information design* 2 (1999) (cit. on p. 190).

[127] B. House. *Brian House, Quotidian Record.* 2012. URL: http://brianhouse.net/works/quotidian_record/ (cit. on pp. 199, 200).

[128] C. Hsu and C. Kung. “Incorporating unsupervised learning with self-organizing map for visualizing mixed data”. In: *2013 Ninth International Conference on Natural Computation (ICNC)*. 2013, pp. 146–151. DOI: 10.1109/ICNC.2013.6817960 (cit. on p. 127).

[129] C.-C. Hsu and S.-H. Lin. “Visualized analysis of mixed numeric and categorical data via extended self-organizing map.” eng. In: *IEEE Trans Neural Netw Learn Syst* 23.1 (2012), pp. 72–86. ISSN: 2162-237X (Print); 2162-237X (Linking). DOI: 10.1109/TNNLS.2011.2178323 (cit. on p. 127).

[130] D. Huang, M. Tory, and L. Bartram. “A Field Study of On-Calendar Visualizations”. In: *Proceedings of the 42nd Graphics Interface Conference*. GI ’16. Victoria, British Columbia, Canada: Canadian Human-Computer Communications Society, 2016, pp. 13–20. ISBN: 9780994786814 (cit. on p. 65).

[131] M. L. Huang, J. Liang, and Q. V. Nguyen. “A Visualization Approach for Frauds Detection in Financial Market”. In: *2009 13th International Conference Information Visualisation*. 2009, pp. 197–202. DOI: 10.1109/IV.2009.23 (cit. on p. 123).

[132] Instituto Nacional de Estatística. *Estimativas Provisórias de População Residente : Portugal, NUTS II, NUTS III e Municipios : 2007*. Lisboa, 2008. URL: https://www.ine.pt/xurl/pub/ (cit. on p. 77).

[133] S. Ishizaki. “Multiagent Model of Dynamic Design: Visualization as an Emergent Behavior of Active Design Agents”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’96. Vancouver, British Columbia, Canada: Association for Computing Machinery, 1996, pp. 347–354. ISBN: 0897917774. DOI: 10.1145/238386.238566. URL: https://doi.org/10.1145/238386.238566 (cit. on p. 215).

[134] M. Jacomy, T. Venturini, S. Heymann, and M. Bastian. “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software”. In: *PLOS ONE* 9.6 (June 2014), pp. 1–12. DOI: 10.1371/journal.pone.0098679. URL: https://doi.org/10.1371/journal.pone.0098679 (cit. on p. 139).

[135] Jinsong Zhang, Yan Chen, and Taoying Li. "Opportunities of innovation under challenges of big data". In: *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*. 2013, pp. 669–673. DOI: 10.1109/FSKD.2013.6816280 (cit. on p. 71).

[136] D. Jones. "Swarm Aesthetics: A Critical Appraisal of Swarming Structures in Art Practice". MA dissertation. Middlesex University, Sonic Arts, 2007. URL: https://www.erase.net/files/pubs/swarm-aesthetics.pdf (cit. on pp. 216, 226, 238).

[137] P. W. Jordan. "Human factors for pleasure seekers". In: *Design and the social sciences*. CRC Press, 2002, pp. 26–40 (cit. on p. 192).

[138] G. Judelman. "Aesthetics and inspiration for visualization design: bridging the gap between art and science". In: *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.* 2004, pp. 245–250. DOI: 10.1109/IV.2004.1320152 (cit. on p. 195).

[139] R. Kamaleswaran, C. Collins, A. James, and C. McGregor. "PhysioEx: Visual Analysis of Physiological Event Streams". In: *Computer Graphics Forum* (2016). ISSN: 1467-8659. DOI: 10.1111/cgf.12909 (cit. on pp. 63, 67).

[140] Y. Kameoka, K. Yagi, S. Munakata, and Y. Yamamoto. "Customer segmentation and visualization by combination of self-organizing map and cluster analysis". In: *2015 13th International Conference on ICT and Knowledge Engineering (ICT Knowledge Engineering 2015)*. 2015, pp. 19–23. DOI: 10.1109/ICTKE.2015.7368465 (cit. on p. 127).

[141] S. D. Kamvar and J. Harris. "We Feel Fine and Searching the Emotional Web". In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. WSDM '11. Hong Kong, China: Association for Computing Machinery, 2011, pp. 117–126. ISBN: 9781450304931. DOI: 10.1145/1935826.1935854. URL: https://doi.org/10.1145/1935826.1935854 (cit. on pp. 215, 217).

[142] D. Keim, H. Qu, and K. Ma. "Big-Data Visualization". In: *IEEE computer graphics and applications* 33 4 (2013), pp. 20–1 (cit. on pp. 72, 247).

[143] D. A. Keim. "Information visualization and visual data mining". In: *IEEE transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 1–8 (cit. on pp. 26, 41, 122).

[144] D. A. Keim and H.-P. Kriegel. "Visualization techniques for mining large databases: A comparison". In: *IEEE Transactions on knowledge and data engineering* 8.6 (1996), pp. 923–938 (cit. on p. 41).

[145] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. "Visual Analytics: Scope and Challenges". In: *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Ed. by S. J. Simoff, M. H. Böhlen, and A. Mazeika. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 76–90. ISBN: 978-3-540-71080-6. DOI: 10.1007/978-3-540-71080-6_6. URL: https://doi.org/10.1007/978-3-540-71080-6_6 (cit. on pp. 26, 27, 79).

[146] D. A. Keim, T. Nietzschmann, N. Schelwies, J. Schneidewind, T. Schreck, and H. Ziegler. "A spectral visualization system for analyzing financial time series data". In: *Eurographics/IEEE TCVG Symposium on Visualization*. 2006, pp. 195–202 (cit. on p. 66).

[147] D. A. Keim, J. Schneidewind, and M. Sips. "CircleView: A New Approach for Visualizing Time-Related Multidimensional Data Sets". In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '04. Gallipoli, Italy: Association for Computing Machinery, 2004, pp. 179–182. ISBN: 1581138679. DOI: 10.1145/989863.989891. URL: https://doi.org/10.1145/989863.989891 (cit. on pp. 64, 76).

[148] D. Keim. "Scaling visual analytics to very large data sets". In: *Workshop on Visual Analytics, Darmstadt*. 2005, pp. 114–125 (cit. on p. 39).

[149] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. "Mastering the information age: solving problems with visual analytics". In: (2010) (cit. on pp. 16, 28, 30).

[150] T. Keller and S.-O. Tergan. "Visualizing Knowledge and Information: An Introduction". In: *Knowledge and Information Visualization* (2005), pp. 1–23. ISSN: 1611-3349. DOI: 10.1007/11510154_1. URL: http://dx.doi.org/10.1007/11510154_1 (cit. on p. 15).

[151] M. Y. Kiang and A. Kumar. “An Evaluation of Self-Organizing Map Networks as a Robust Alternative to Factor Analysis in Data Mining Applications”. In: *Information Systems Research* 12.2 (2001), pp. 177–194. ISSN: 10477047, 15265536. URL: http://www.jstor.org/stable/23011078 (cit. on p. 126).

[152] B. Kim, B. Lee, S. Knoblach, E. Hoffman, and J. Seo. “GeneShelf: A Web-based Visual Interface for Large Gene Expression Time-Series Data Repositories”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 905–912 (cit. on p. 63).

[153] R. Kincaid and H. Lam. “Line Graph Explorer: Scalable Display of Line Graphs Using Focus+Context”. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI ’06. Venezia, Italy: Association for Computing Machinery, 2006, pp. 404–411. ISBN: 1595933530. DOI: 10.1145/1133265.1133348. URL: https://doi.org/10.1145/1133265.1133348 (cit. on p. 63).

[154] J. D. Kirkland, T. E. Senator, J. J. Hayden, T. Dybala, H. G. Goldberg, and P. Shyr. “The NASD Regulation Advanced-Detection System (ADS)”. In: *AI Magazine* 20.1 (1999), p. 55. DOI: 10.1609/aimag.v20i1.1440. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1440 (cit. on p. 123).

[155] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert. “A Survey on Visual Analysis Approaches for Financial Data”. In: *Computer Graphics Forum* 35.3 (2016), pp. 599–617. DOI: https://doi.org/10.1111/cgf.12931. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.12931. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12931 (cit. on pp. 122, 123).

[156] T. Kohonen. “The self-organizing map”. In: *Proceedings of the IEEE* 78.9 (1990), pp. 1464–1480. DOI: 10.1109/5.58325 (cit. on p. 126).

[157] E. Koua. “Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets”. In: *Proceedings of 21st int. cartographic renaissance (ICC)* (2003), pp. 1694–1702 (cit. on p. 127).

[158] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. H. Flowers, N. Miner, and J. Neuhoff. *Sonification report: Status of the field and research agenda*. International Community for Auditory Display, 1999 (cit. on pp. 198, 208).

[159] M. Krstajic, E. Bertini, and D. Keim. “CloudLines: Compact Display of Event Episodes in Multiple Time-Series”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2432–2439 (cit. on p. 62).

[160] N. Kumar, V. N. Lolla, E. Keogh, S. Lonardi, C. A. Ratanamahatana, and L. Wei. “Time-series Bitmaps: a Practical Visualization Tool for Working with Large Time Series Databases”. In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Apr. 2005. DOI: 10.1137/1.9781611972757.55. URL: https://doi.org/10.1137/1.9781611972757.55 (cit. on p. 66).

[161] T. Lammarsch. “Facets of Time - Making the Most of Time’s Structure in Interactive Visualization”. eng. PhD thesis. Vienna, Austria, 2010. URL: http://publik.tuwien.ac.at/files/PubDat_217966.pdf (cit. on pp. 20–22, 24, 28, 33, 34, 41).

[162] A. Lang. “Aesthetics in information visualization”. In: *Trends in information visualization* 8 () (cit. on pp. 192, 194).

[163] A. Lau and A. Vande Moere. “Towards a Model of Information Aesthetics in Information Visualization”. In: *2007 11th International Conference Information Visualization (IV ’07)*. 2007, pp. 87–92. DOI: 10.1109/IV.2007.114 (cit. on pp. 191–195).

[164] B. Laurel. *Computers as Theatre*. Addison-Wesley, 1993. ISBN: 9780201550603. URL: https://books.google.pt/books?id=LtwfAQAAIAAJ (cit. on p. 192).

[165] S. Laxman and P. S. Sastry. “A survey of temporal data mining”. In: *Sadhana* 31.2 (2006), pp. 173–198 (cit. on p. 38).

[166] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. “Visual analytics for event detection: Focusing on fraud”. In: *Visual Informatics* 2.4 (2018), pp. 198 –212. ISSN: 2468-502X. DOI: https://doi.org/10.1016/j.visinf.2018.11.001. URL: http://www.sciencedirect.com/science/article/pii/S2468502X18300548 (cit. on pp. 119, 122, 123, 153).

[167] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner. "Visual Analytics for Fraud Detection: Focusing on Profile Analysis". In: *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: Posters*. EuroVis '16. Groningen, The Netherlands: Eurographics Association, 2016, pp. 45–47 (cit. on pp. 123, 124).

[168] V. L. Lemieux, B. W. Shieh, D. Lau, S. H. Jun, T. Dang, J. Chu, and G. Tam. "Using visual analytics to enhance data exploration and knowledge discovery in financial systemic risk analysis: The multivariate density estimator". In: *iConference 2014 Proceedings* (2014) (cit. on p. 120).

[169] I. Letunic and P. Bork. "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy". In: *Nucleic Acids Research* 39 (Apr. 2011), W475–W478. ISSN: 0305-1048. DOI: 10.1093/nar/gkr201. eprint: https://academic.oup.com/nar/article-pdf/39/suppl_2/W475/7623907/gkr201.pdf. URL: https://doi.org/10.1093/nar/gkr201 (cit. on p. 76).

[170] M. Lima. *Information Visualization Manifesto*. 2009. URL: http://www.visualcomplexity.com/vc/blog/?p=644 (cit. on p. 192).

[171] J. Lin, E. Keogh, and S. Lonardi. "Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases". In: *Information Visualization* 4.2 (Apr. 2005), pp. 61–82. DOI: 10.1057/palgrave.ivs.9500089. URL: https://doi.org/10.1057/palgrave.ivs.9500089 (cit. on pp. 61, 67).

[172] J. Liu, J. Li, and W. Li. "Temporal Patterns in Fine Particulate Matter Time Series in Beijing: A Calendar View". In: *Scientific Reports* 6.1 (Aug. 2016). DOI: 10.1038/srep32221. URL: https://doi.org/10.1038/srep32221 (cit. on p. 65).

[173] S. Liu, Y. Wu, E. Wei, M. Liu, and Y. Liu. "StoryFlow: Tracking the Evolution of Stories". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2436–2445 (cit. on p. 62).

[174] D. Luo, J. Yang, M. Krstajic, W. Ribarsky, and D. A. Keim. "EventRiver: visually exploring text collections with temporal references." eng. In: *IEEE Trans Vis Comput Graph* 18.1 (2012), pp. 93–105. ISSN: 1941-0506 (Electronic); 1077-2626 (Linking). DOI: 10.1109/TVCG.2010.225 (cit. on p. 62).

[175] J. E. MCGRATH. "Methodology matters: Doing research in the social and behavioural sciences". In: *Readings in Human–Computer Interaction*. Ed. by R. M. BAECKER, J. GRUDIN, W. A. BUXTON, and S. GREENBERG. Interactive Technologies. Morgan Kaufmann, 1995, pp. 152 –169. ISBN: 978-0-08-051574-8. DOI: https://doi.org/10.1016/B978-0-08-051574-8.50019-4. URL: http://www.sciencedirect.com/science/article/pii/B9780080515748500194 (cit. on p. 30).

[176] A. M. MacEachren. *How Maps Work - Representation, Visualization, and Design*. Guilford Press, 2004. ISBN: 978-1-57230-040-8 (cit. on pp. 20, 21, 23).

[177] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. "Visual Semiotics & Uncertainty Visualization: An Empirical Study". In: *IEEE Trans. Vis. Comput. Graph.* 18.12 (2012), pp. 2496–2505. DOI: 10.1109/TVCG.2012.279 (cit. on pp. 21, 23).

[178] C. Maçãs, P. Cruz, H. Amaro, E. Polisciuc, T. Carvalho, F. Santos, and P. Machado. "Time-series Application on Big Data - Visualization of Consumption in Supermarkets". In: *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015*. Ed. by J. Braz, A. Kerren, and L. Linsen. SciTePress, 2015, pp. 239–246. DOI: 10.5220/0005307702390246. URL: https://doi.org/10.5220/0005307702390246 (cit. on pp. 9, 252).

[179] C. Maçãs, P. Cruz, P. Martins, and P. Machado. "Swarm Systems in the Visualization of Consumption Patterns". In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Q. Yang and M. J. Wooldridge. AAAI Press, 2015, pp. 2466–2472. URL: http://ijcai.org/Abstract/15/349 (cit. on pp. 11, 253).

[180] C. Maçãs, P. Cruz, E. Polisciuc, H. Amaro, and P. Machado. "Iso-edges for the Geovisualization of consumptions". In: *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP, Rome, Italy, February 27-29, 2016*. Ed. by N. Magnenat-Thalmann, P. Richard, L. Linsen, A. Telea, S. Battiato, F. H. Imai, and J. Braz. SciTePress, 2016, pp. 222–229. DOI: 10.5220/0005785702200227. URL: https://doi.org/10.5220/0005785702200227 (cit. on p. 254).

[181] C. Maçãs, N. Lourenço, and P. Machado. "Interactive Evolution of Swarms for the Visualisation of consumptions". In: *Interactivity, Game Creation, Design, Learning, and Innovation - 7th EAI International Conference, ArtsIT 2018, and 3rd EAI International Conference, DLI 2018, ICTCC 2018, Braga, Portugal, October 24-26, 2018, Proceedings*. Ed. by A. L. Brooks, E. Brooks, and C. Sylla. Vol. 265. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2018, pp. 101–110. DOI: 10.1007/978-3-030-06134-0_11. URL: https://doi.org/10.1007/978-3-030-06134-0_11 (cit. on pp. 11, 12, 253, 254).

[182] C. Maçãs, N. Lourenço, and P. Machado. "Evolving visual artefacts based on consumption patterns". In: *Int. J. Arts Technol.* 12.1 (2020), pp. 60–83. DOI: 10.1504/IJART.2020.107693. URL: https://doi.org/10.1504/IJART.2020.107693 (cit. on pp. 12, 254).

[183] C. Maçãs and P. Machado. "The Rhythm of Consumption". In: *5th Joint Symposium on Computational Aesthetics, Sketch-Based Interfaces and Modeling, and Non-Photorealistic Animation and Rendering, Expressive 2016 - Posters, Artworks, and Bridging Papers, Lisbon, Portugal, May 7-9, 2016, Proceedings*. Ed. by E. Akleman, L. Bartram, A. Çamci, A. G. Forbes, and P. Machado. Eurographics Association, 2016, pp. 11–12. DOI: 10.2312/exp.20161259. URL: https://doi.org/10.2312/exp.20161259 (cit. on pp. 11, 253, 254).

[184] C. Maçãs and P. Machado. "The Rhythm of consumptions". In: *IEEE VIS Arts Program*. Baltimore, EUA, 2016 (cit. on pp. 11, 253, 254).

[185] C. Maçãs and P. Machado. "Radial Calendar of Consumption". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 96–102. DOI: 10.1109/iV.2018.00027. URL: https://doi.org/10.1109/iV.2018.00027 (cit. on pp. 9, 252, 253).

[186] C. Maçãs, P. Martins, and P. Machado. "Consumption as a Rhythm: A Multimodal Experiment on the Representation of Time-Series". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 504–509. DOI: 10.1109/iV.2018.00093. URL: https://doi.org/10.1109/iV.2018.00093 (cit. on pp. 11, 253).

[187] C. Maçãs, E. Polisciuc, and P. Machado. "GlyphSOMe: Using SOM with Data Glyphs for Customer Profiling". In: *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2020, Volume 3: IVAPP, Valletta, Malta, February 27-29, 2020*. Ed. by A. Kerren, C. Hurter, and J. Braz. SCITEPRESS, 2020, pp. 301–308. DOI: 10.5220/0009178803010308. URL: https://doi.org/10.5220/0009178803010308 (cit. on pp. 10, 255).

[188] C. Maçãs, E. Polisciuc, and P. Machado. "VaBank: Visual Analytics for Banking Transactions". In: *24th International Conference Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*. 2020, pp. 336–343. DOI: 10.1109/IV51561.2020.00062. URL: https://doi.org/10.1109/IV51561.2020.00062 (cit. on pp. 10, 253).

[189] C. Maçãs, A. Rodrigues, G. Bernardes, and P. Machado. "MixMash: A Visualisation System for Musical Mashup Creation". In: *22nd International Conference Information Visualisation, IV 2018, Fisciano, Italy, July 10-13, 2018*. Ed. by E. Banissi et al. IEEE Computer Society, 2018, pp. 471–477. DOI: 10.1109/iV.2018.00088. URL: https://doi.org/10.1109/iV.2018.00088 (cit. on p. 255).

[190] C. Maçãs, A. Rodrigues, G. Bernardes, and P. Machado. "MixMash: An Assistive Tool for Music Mashup Creation from Large Music Collections". In: *Int. J. Art Cult. Des. Technol.* 8.2 (2019), pp. 20–40. DOI: 10.4018/IJACDT.2019070102. URL: https://doi.org/10.4018/IJACDT.2019070102 (cit. on p. 255).

[191] J. Mackinlay. "Automating the Design of Graphical Presentations of Relational Information". In: *ACM Trans. Graph.* 5.2 (Apr. 1986), pp. 110–141. ISSN: 0730-0301. DOI: 10.1145/22949.22950. URL: https://doi.org/10.1145/22949.22950 (cit. on pp. 20, 24, 28, 30, 133, 160, 191).

[192] J. Mankoff, A. K. Dey, G. Hsieh, J. Kientz, S. Lederer, and M. Ames. "Heuristic Evaluation of Ambient Displays". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. Ft. Lauderdale, Florida, USA: Association for Computing Machinery, 2003, pp. 169–176. ISBN: 1581136307. DOI: 10.1145/642611.642642. URL: https://doi.org/10.1145/642611.642642 (cit. on p. 192).

[193] L. Manovich. *Information and Form*. 2000. URL: http://manovich.net/index.php/projects/information-and-form (cit. on p. 191).

[194] G. C. Mariano, V. G. Staggemeier, L. P. C. Morellato, and R. da S. Torres. "Multivariate cyclical data visualization using radial visual rhythms: A case study in phenology analysis". In: *Ecological Informatics* 46 (July 2018), pp. 19–35. DOI: 10.1016/j.ecoinf.2018.05.003. URL: https://doi.org/10.1016/j.ecoinf.2018.05.003 (cit. on pp. 64, 67).

[195] B. H. McCormick. "Visualization in Scientific Computing". In: *SIGBIO Newsl.* 10.1 (Mar. 1988), pp. 15–21. ISSN: 0163-5697. DOI: 10.1145/43965.43966. URL: https://doi.org/10.1145/43965.43966 (cit. on pp. 3, 15).

[196] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. "LiveRAC: Interactive Visual Exploration of System Management Time-Series Data". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '08. Florence, Italy: Association for Computing Machinery, 2008, pp. 1483–1492. ISBN: 9781605580111. DOI: 10.1145/1357054.1357286. URL: https://doi.org/10.1145/1357054.1357286 (cit. on pp. 62, 67).

[197] I. Meirelles. *Design for Information: An Introduction to the Histories, Theories, and Best Practices Behind Effective Information Visualizations*. Rockport Publishers, 2013. ISBN: 9781610589482. URL: https://books.google.pt/books?id=RFb0AwAAQBAJ (cit. on pp. 3, 189).

[198] M. Meyer, T. Munzner, A. DePace, and H. Pfister. "MulteeSum: a tool for comparative spatial and temporal gene expression data." eng. In: *IEEE Trans Vis Comput Graph* 16.6 (2010), pp. 908–917. ISSN: 1077-2626 (Print); 1077-2626 (Linking). DOI: 10.1109/TVCG.2010.137 (cit. on pp. 44, 62).

[199] M. Milosevic, K. M. Valter McConville, E. Sejdic, K. Masani, M. J. Kyan, and M. R. Popovic. "Visualization of trunk muscle synergies during sitting perturbations using self-organizing maps (SOM)." eng. In: *IEEE Trans Biomed Eng* 59.9 (2012), pp. 2516–2523. ISSN: 1558-2531 (Electronic); 0018-9294 (Linking). DOI: 10.1109/TBME.2012.2205577 (cit. on p. 127).

[200] M. Mitchell. *Artificial intelligence: A guide for thinking humans*. Penguin UK, 2019 (cit. on p. 121).

[201] W. J. T. Mitchell. "Spatial Form in Literature: Toward a General Theory". In: *Critical Inquiry* 6.3 (1980), pp. 539–567. ISSN: 00931896, 15397858. URL: http://www.jstor.org/stable/1343108 (cit. on p. 44).

[202] T. Mitsa. *Temporal data mining*. CRC Press, 2010 (cit. on p. 38).

[203] A. V. Moere. "Time-Varying Data Visualization Using Information Flocking Boids". In: *IEEE Symposium on Information Visualization*. 2004, pp. 97–104 (cit. on pp. 66, 67).

[204] A. V. Moere. "Aesthetic Data Visualization as a Resource for Educating Creative Design". In: *Computer-Aided Architectural Design Futures (CAADFutures) 2007*. Ed. by A. Dong, A. V. Moere, and J. S. Gero. Dordrecht: Springer Netherlands, 2007, pp. 71–84. ISBN: 978-1-4020-6528-6 (cit. on pp. 191, 192, 194, 195).

[205] A. V. Moere and H. Purchase. "On the role of design in information visualization". In: *Information Visualization* 10.4 (2011), pp. 356–371. DOI: 10.1177/1473871611415996. eprint: https://doi.org/10.1177/1473871611415996. URL: https://doi.org/10.1177/1473871611415996 (cit. on pp. 191, 194, 195).

[206] M. J. Mohammadi-Aragh and T. J. Jankun-Kelly. "MoireTrees: Visualization and Interaction for Multi-Hierarchical Data". In: *EUROVIS 2005: Eurographics / IEEE VGTC Symposium on Visualization*. Ed. by K. Brodlie, D. Duke, and K. Joy. The Eurographics Association, 2005. ISBN: 3-905673-19-3. DOI: 10.2312/VisSym/EuroVis05/231-238 (cit. on p. 76).

[207] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. "Temporal Event Sequence Simplification". In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2227–2236 (cit. on p. 62).

[208] A. M. M. Morais, M. G. Quiles, and R. D. C. Santos. "Icon and Geometric Data Visualization with a Self-Organizing Map Grid". In: *Computational Science and Its Applications – ICCSA 2014*. Ed. by B. Murgante, S. Misra, A. M. A. C. Rocha, C. Torre, J. G. Rocha, M. I. Falcão, D. Taniar, B. O. Apduhan, and O. Gervasi. Cham: Springer International Publishing, 2014, pp. 562–575. ISBN: 978-3-319-09153-2 (cit. on p. 127).

[209] Muller and Schumann. "Visualization methods for time-dependent data - an overview". In: *Proceedings of the 2003 Winter Simulation Conference, 2003*. Vol. 1. 2003, 737–745 Vol.1 (cit. on pp. 6, 43, 61, 67).

[210] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2015. ISBN: 9781498759717. URL: https://books.google.de/books?id=NfkYCwAAQBAJ (cit. on p. 75).

[211] T. Munzner. "Process and Pitfalls in Writing Information Visualization Research Papers". In: *Information Visualization: Human-Centered Issues and Perspectives*. Ed. by A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 134–153. ISBN: 978-3-540-70956-5. DOI: 10.1007/978-3-540-70956-5_6. URL: https://doi.org/10.1007/978-3-540-70956-5_6 (cit. on p. 8).

[212] T. Munzner. "A nested model for visualization design and validation". In: *IEEE transactions on visualization and computer graphics* 15.6 (2009), pp. 921–928 (cit. on pp. 30, 41).

[213] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun. "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature". In: *Decision Support Systems* 50.3 (2011). On quantitative methods for detection of financial fraud, pp. 559 –569. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2010.08.006. URL: http://www.sciencedirect.com/science/article/pii/S0167923610001302 (cit. on pp. 121, 123, 153).

[214] F. Nightingale. *Notes on matters affecting the health, efficiency, and hospital administration of the British army: founded chiefly on the experience of the late war*. Harrison and Sons, St. Martin's Lane, WC, 1987 (cit. on p. 76).

[215] D. Norman and T. Dunaeff. *Things That Make Us Smart: Defending Human Attributes In The Age Of The Machine*. Basic Books, 1994. ISBN: 9780201626957. URL: https://books.google.pt/books?id=g05KXUljlcAC (cit. on p. 199).

[216] D. Norman. *Emotional Design: Why We Love (or Hate) Everyday Things*. Basic Books, 2004. ISBN: 9780465051359. URL: https://books.google.pt/books?id=z2jvRlqhdlwC (cit. on p. 194).

[217] C. North. "Toward measuring visualization insight". In: *IEEE Computer Graphics and Applications* 26.3 (2006), pp. 6–9 (cit. on p. 30).

[218] M. Ogawa and K.-L. Ma. "code_swarm: a design study in organic software visualization." eng. In: *IEEE Trans Vis Comput Graph* 15.6 (2009), pp. 1097–1104. ISSN: 1077-2626 (Print); 1077-2626 (Linking). DOI: 10.1109/TVCG.2009.123 (cit. on pp. 215, 216, 238).

[219] J. Olsson and M. Boldt. "Computer forensic timeline visualization tool". In: *Digital Investigation* 6 (2009). The Proceedings of the Ninth Annual DFRWS Conference, S78 –S87. ISSN: 1742-2876. DOI: https://doi.org/10.1016/j.diin.2009.06.008. URL: http://www.sciencedirect.com/science/article/pii/S1742287609000425 (cit. on p. 136).

[220] D. Olszewski. "Fraud detection using self-organizing map visualizing the user profiles". In: *Knowledge-Based Systems* 70 (2014), pp. 324 –334. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2014.07.008 (cit. on p. 127).

[221] S. E. Palmer. "Common region: A new principle of perceptual grouping". In: *Cognitive Psychology* 24.3 (1992), pp. 436 –447. ISSN: 0010-0285. DOI: https://doi.org/10.1016/0010-0285(92)90014-S (cit. on p. 19).

[222] R. Peng. "A Method for Visualizing Multivariate Time Series Data". In: *Journal of Statistical Software, Code Snippets* 25.1 (2008), pp. 1–17. ISSN: 1548-7660. DOI: 10.18637/jss.v025.c01. URL: https://www.jstatsoft.org/v025/c01 (cit. on pp. 63, 67).

[223] D. Pfitzner, V. Hobbs, and D. Powers. “A unified taxonomic framework for information visualization”. In: *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24*. Australian Computer Society, Inc. 2003, pp. 57–66 (cit. on p. 41).

[224] C. Phua, V. Lee, K. Smith, and R. Gayler. “A comprehensive survey of data mining-based fraud detection research”. In: *arXiv preprint arXiv:1009.6119* (2010) (cit. on pp. 121, 123).

[225] W. A. Pike, J. Stasko, R. Chang, and T. A. O’Connell. “The Science of Interaction”. In: *Information Visualization* 8.4 (2009), pp. 263–274. DOI: 10.1057/ivs.2009.22. eprint: https://doi.org/10.1057/ivs.2009.22. URL: https://doi.org/10.1057/ivs.2009.22 (cit. on p. 26).

[226] P. Pirolli and S. Card. “The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis”. In: (2005), pp. 2–4 (cit. on pp. 16, 17).

[227] C. Plaisant. “The Challenge of Information Visualization Evaluation”. In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI ’04. Gallipoli, Italy: Association for Computing Machinery, 2004, pp. 109–116. ISBN: 1581138679. DOI: 10.1145/989863.989880. URL: https://doi.org/10.1145/989863.989880 (cit. on p. 30).

[228] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. “LifeLines: Visualizing Personal Histories”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’96. Vancouver, British Columbia, Canada: Association for Computing Machinery, 1996, pp. 221–227. ISBN: 0897917774. DOI: 10.1145/238386.238493. URL: https://doi.org/10.1145/238386.238493 (cit. on p. 62).

[229] W Playfair. “The commercial and political atlas (London)”. In: (1786) (cit. on pp. 52, 53).

[230] W. Playfair. *The statistical breviary; shewing the resources of every state and kingdom in Europe*. J. Wallis, 1801 (cit. on p. 53).

[231] E. Polisciuc, P. Cruz, H. Amaro, C. Maçãs, T. Carvalho, F. Santos, and P. Machado. “Arc and Swarm-based Representations of Customer’s Flows among Supermarkets”. In: *IVAPP 2015 - Proceedings of the 6th International Conference on Information Visualization Theory and Applications, Berlin, Germany, 11-14 March, 2015*. Ed. by J. Braz, A. Kerren, and L. Linsen. SciTePress, 2015, pp. 300–306. DOI: 10.5220/0005316503000306. URL: https://doi.org/10.5220/0005316503000306 (cit. on p. 254).

[232] E. Polisciuc, P. Cruz, H. Amaro, C. Maçãs, and P. Machado. “Flow Map of Products Transported among Warehouses and Supermarkets”. In: *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 2: IVAPP, Rome, Italy, February 27-29, 2016*. Ed. by N. Magnenat-Thalmann, P. Richard, L. Linsen, A. Telea, S. Battiato, F. H. Imai, and J. Braz. SciTePress, 2016, pp. 179–188. DOI: 10.5220/0005787301770186. URL: https://doi.org/10.5220/0005787301770186 (cit. on p. 254).

[233] E. Polisciuc, C. Maçãs, F. Assunção, and P. Machado. “Hexagonal gridded maps and information layers: a novel approach for the exploration and analysis of retail data”. In: *SIGGRAPH ASIA 2016, Macao, December 5-8, 2016 - Symposium on Visualization*. Ed. by W. Chen and D. Weiskopf. ACM, 2016, 6:1–6:8. DOI: 10.1145/3002151.3002160. URL: https://doi.org/10.1145/3002151.3002160 (cit. on p. 255).

[234] J. Priestley. *A Description of a Chart of Biography: By Joseph Priestley....* Vol. 1. Printed at Warrington, 1764 (cit. on p. 51).

[235] H. C. Purchase. “Metrics for Graph Drawing Aesthetics”. In: *Journal of Visual Languages & Computing* 13.5 (2002), pp. 501–516. ISSN: 1045-926X. DOI: https://doi.org/10.1006/jvlc.2002.0232. URL: http://www.sciencedirect.com/science/article/pii/S1045926X02902326 (cit. on p. 192).

[236] M. Quinn. “Research set to music: The climate symphony and other sonifications of ice core, radar, DNA, seismic and solar wind data”. In: *Proceedings of the 7th International Conference on Auditory Display (ICAD 2001)*. Georgia Institute of Technology. 2001 (cit. on p. 199).

[237] A. Rajpurohit. “Big data for business managers — Bridging the gap between potential and value”. In: *2013 IEEE International Conference on Big Data*. 2013, pp. 29–31. DOI: 10.1109/BigData.2013.6691794 (cit. on pp. 71, 72).

[238] C. W. Reynolds. “Flocks, Herds and Schools: A Distributed Behavioral Model”. In: *SIGGRAPH Comput. Graph.* 21.4 (Aug. 1987), pp. 25–34. ISSN: 0097-8930. DOI: 10.1145/37402.37406. URL: https://doi.org/10.1145/37402.37406 (cit. on pp. 216, 217).

[239] N. Roard and M. W. Jones. “Agents Based Visualization and Strategies”. In: *Full Papers Proceedings of WSCG*. Jan. 30, 2006, pp. 63–70. ISBN: 80-86943-03-8 (cit. on p. 216).

[240] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. “Effectiveness of Animation in Trend Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 14.6 (2008), pp. 1325–1332. DOI: 10.1109/TVCG.2008.125 (cit. on pp. 198, 204).

[241] P. K. Robertson. “A methodology for choosing data representations”. In: *IEEE Computer Graphics and Applications* 3 (1991), pp. 56–67 (cit. on p. 41).

[242] N. Rogovschi, M. Lebbah, and Y. Bennani. “A self-organizing map for mixed continuous and categorical data”. In: *Int. Journal of Computing* 10.1 (2011), pp. 24–32 (cit. on p. 126).

[243] B. E. Rogowitz and L. A. Treinish. “Data visualization: the end of the rainbow”. In: *IEEE Spectrum* 35.12 (1998), pp. 52–59 (cit. on p. 22).

[244] D. Rosenberg and A. Grafton. *Cartographies of time: A history of the timeline*. Princeton Architectural Press, 2013 (cit. on pp. 44–49, 51–58).

[245] R. Roth. “Visual Variables”. In: Jan. 2017, pp. 1–11. DOI: 10.1002/9781118786352.wbieg0761 (cit. on pp. 19–24).

[246] S. Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019 (cit. on p. 121).

[247] M. J. Saary. “Radar plots: a useful way for presenting multivariate health care data”. In: *Journal of Clinical Epidemiology* 61.4 (2008), pp. 311–317. ISSN: 0895-4356. DOI: https://doi.org/10.1016/j.jclinepi.2007.04.021. URL: http://www.sciencedirect.com/science/article/pii/S0895435607003320 (cit. on p. 76).

[248] C. Sakoda, A. Nagasaki, T. Itoh, M. Ise, and K. Miyashita. “Visualization for assisting rule definition tasks of credit card fraud detection systems”. In: *IIEEJ Image Electronics and Visual Computing Workshop*. 2010 (cit. on pp. 123, 124).

[249] P. Sarlin. “Sovereign debt monitor: A visual Self-organizing maps approach”. In: *2011 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*. 2011, pp. 1–8. DOI: 10.1109/CIFER.2011.5953556 (cit. on p. 128).

[250] P. Sarlin and T. Eklund. “Fuzzy Clustering of the Self-Organizing Map: Some Applications on Financial Time Series”. In: *Advances in Self-Organizing Maps*. Ed. by J. Laaksonen and T. Honkela. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 40–50. ISBN: 978-3-642-21566-7 (cit. on p. 128).

[251] T. Schreck, J. Bernard, T. Tekusova, and J. Kohlhammer. “Visual cluster analysis of trajectory data with interactive Kohonen Maps”. In: *2008 IEEE Symposium on Visual Analytics Science and Technology*. 2008, pp. 3–10. DOI: 10.1109/VAST.2008.4677350 (cit. on pp. 127, 128).

[252] T. Schreck, T. Tekušová, J. Kohlhammer, and D. Fellner. “Trajectory-Based Visual Analysis of Large Financial Time Series Data”. In: *SIGKDD Explor. Newsl.* 9.2 (Dec. 2007), pp. 30–37. ISSN: 1931-0145. DOI: 10.1145/1345448.1345454. URL: https://doi.org/10.1145/1345448.1345454 (cit. on p. 128).

[253] H. Schulz, T. Nocke, M. Heitzler, and H. Schumann. “A Design Space of Visualization Tasks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2366–2375 (cit. on pp. 25, 26).

[254] H. Schumann and W. Müller. *Visualisierung : Grundlagen und allgemeine Methoden*. Berlin; Heidelberg; New York; Barcelona; Hongkong; London; Mailand; Paris; Singapur; Tokio: Springer, 2000. ISBN: 3540649441 9783540649441. URL: http://www.worldcat.org/search?qt=worldcat_org_all&q=3540649441 (cit. on p. 28).

[255] R. Scruton. “In Search of the Aesthetic”. In: *The British Journal of Aesthetics* 47.3 (July 2007), pp. 232–250. ISSN: 0007-0904. DOI: 10.1093/aesthj/aym004. eprint: https://academic.oup.com/bjaesthetics/article-pdf/47/3/232/68344/aym004.pdf. URL: https://doi.org/10.1093/aesthj/aym004 (cit. on p. 193).

[256] Z. Shen, M. Ogawa, S. T. Teoh, and K.-L. Ma. “BiblioViz: a system for visualizing bibliography information”. In: *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60*. Australian Computer Society, Inc. 2006, pp. 93–102 (cit. on p. 127).

[257] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu. “RankExplorer: Visualization of Ranking Changes in Large Time Series Data.” eng. In: *IEEE Trans Vis Comput Graph* 18.12 (2012), pp. 2669–2678. ISSN: 1941-0506 (Electronic); 1077-2626 (Linking). DOI: 10.1109/TVCG.2012.253 (cit. on pp. 61, 63, 66).

[258] D. Shiffman. *Swarm*. 2004. URL: http://shiffman.net/projects/swarm/ (cit. on p. 216).

[259] B. Shneiderman. “The eyes have it: a task by data type taxonomy for information visualizations”. In: *Proceedings 1996 IEEE Symposium on Visual Languages*. 1996, pp. 336–343. DOI: 10.1109/VL.1996.545307 (cit. on pp. 26, 27, 31, 39, 41, 53, 160).

[260] B. Shneiderman and B. B. Bederson. *The Craft of Information Visualization: Readings and Reflections*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003. ISBN: 1558609156 (cit. on p. 16).

[261] S. F. Silva and T. Catarci. “Visualization of linear time-oriented data: a survey”. In: *Proceedings of the First International Conference on Web Information Systems Engineering*. Vol. 1. 2000, 310–319 vol.1 (cit. on p. 61).

[262] R. Sims. “Interactivity: A forgotten art?” In: *Computers in Human Behavior* 13.2 (1997), pp. 157 –180. ISSN: 0747-5632. DOI: https://doi.org/10.1016/S0747-5632(97)00004-6. URL: http://www.sciencedirect.com/science/article/pii/S0747563297000046 (cit. on p. 27).

[263] K. Šimunić. “Visualization of Stock Market Charts”. In: *In Proceedings from the 11th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen-Bory (CZ)*. 2003 (cit. on p. 128).

[264] D. Skau and R. Kosara. “Arcs, Angles, or Areas: Individual Data Encodings in Pie and Donut Charts”. In: *Comput. Graph. Forum* 35.3 (June 2016), pp. 121–130. ISSN: 0167-7055 (cit. on p. 76).

[265] G. J. Smith. *Specimen Box—Analyzing Botnet Activity*. 2014. URL: http://www.creativeapplications.net/news/specimen-box-analyzing-botnet-activity/ (cit. on p. 200).

[266] P. Sokol, L. Kleinová, and M. Husák. “Study of attack using honeypots and honeynets lessons learned from time-oriented visualization”. In: *IEEE EUROCON 2015 - International Conference on Computer as a Tool (EUROCON)*. 2015, pp. 1–6 (cit. on p. 65).

[267] I. Spence. “William Playfair and the psychology of graphs”. In: *Proceedings of the American Statistical Association, Section on Statistical Graphics*. 2006, pp. 2426–2436 (cit. on p. 76).

[268] M. Steiger, J. Bernard, S. Mittelstädt, H. Lücke-Tieke, D. Keim, T. May, and J. Kohlhammer. “Visual analysis of time-series similarities for anomaly detection in sensor networks”. In: *Computer graphics forum*. Vol. 33. 3. Wiley Online Library. 2014, pp. 401–410 (cit. on p. 65).

[269] M. Suntinger, H. Obweger, J. Schiefer, and M. E. Groller. “The Event Tunnel: Interactive Visualization of Complex Event Streams for Business Process Pattern Analysis”. In: *2008 IEEE Pacific Visualization Symposium*. 2008, pp. 111–118 (cit. on pp. 64, 67, 72).

[270] W.-S. Tai and C.-C. Hsu. “Growing Self-Organizing Map with cross insert for mixed-type data clustering”. In: *Applied Soft Computing* 12.9 (2012), pp. 2856 –2866. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2012.04.004. URL: http://www.sciencedirect.com/science/article/pii/S1568494612001731 (cit. on p. 127).

[271] H. Takagi. “Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation”. In: *Proceedings of the IEEE* 89.9 (2001), pp. 1275–1296. DOI: 10.1109/5.949485 (cit. on p. 228).

[272] J. Thomas and K. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE CS Press, 2005 (cit. on p. 16).

[273] J. Thorp. *WIRED UK, JULY '09 – VISUALIZING A NATION'S DNA*. 2009. URL: http://blog.blprnt.com/blog/blprnt/wired-uk-july-09-visualizing-a-nations-dna (cit. on p. 195).

[274] C. Tominski, J. Abello, and H. Schumann. "Axes-Based Visualizations with Radial Layouts". In: *Proceedings of the 2004 ACM Symposium on Applied Computing*. SAC '04. Nicosia, Cyprus: Association for Computing Machinery, 2004, pp. 1242–1247. ISBN: 1581138121. DOI: 10.1145/967900.968153. URL: https://doi.org/10.1145/967900.968153 (cit. on p. 65).

[275] C. Tominski and H. Schumann. "Enhanced interactive spiral display". In: *SIGRAD 2008. The Annual SIGRAD Conference Special Theme: Interaction; November 27-28; 2008 Stockholm; Sweden*. 034. Linköping University Electronic Press. 2008, pp. 53–56 (cit. on pp. 65, 76).

[276] C. Tominskietal, W. Aigner, S. Miksch, and H. Schumann. "Images of time". In: *Information design: research and practice*. Routledge, 2017, pp. 39–58 (cit. on pp. 15, 33, 35–37).

[277] M. Tory and T. Moller. "Rethinking visualization: A high-level taxonomy". In: *IEEE Symposium on Information Visualization*. IEEE. 2004, pp. 151–158 (cit. on pp. 36, 41, 189).

[278] E. R. Tufte. *The visual display of quantitative information*. Vol. 2. Graphics press Cheshire, CT, 2001 (cit. on pp. 3, 15, 45, 54, 58, 74, 189, 193).

[279] E. R. Tufte, N. H. Goeler, and R. Benson. *Envisioning information*. Vol. 126. Graphics press Cheshire, CT, 1990 (cit. on pp. 15, 16, 45, 46, 60, 73, 74, 95, 193, 198).

[280] J. W. Tukey. "The future of data analysis". In: *The annals of mathematical statistics* 33.1 (1962), pp. 1–67 (cit. on p. 59).

[281] L. Tweedie. "Characterizing interactive externalizations". In: *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*. 1997, pp. 375–382 (cit. on pp. 26, 41).

[282] J. J. Van Wijk and E. R. Van Selow. "Cluster and calendar based visualization of time series data". In: *Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)*. 1999, pp. 4–9 (cit. on pp. 65–67, 75).

[283] A. Vande Moere and A. Lau. "In-Formation Flocking: An Approach to Data Visualization Using Multi-agent Formation Behavior". In: *Progress in Artificial Life*. Ed. by M. Randall, H. A. Abbass, and J. Wiles. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 292–304. ISBN: 978-3-540-76931-6 (cit. on pp. 215, 216, 238).

[284] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li. "AIˆ2: Training a Big Data Machine to Defend". In: *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*. 2016, pp. 49–54. DOI: 10.1109/BigDataSecurity-HPSC-IDS.2016.79 (cit. on p. 154).

[285] P. Vickers, B. Hogg, and D. Worrall. "The Aesthetics of Sonification". In: *Body, Sound and Space in Music and Beyond:Multimodal Explorations*. Ed. by C. Wöllner. New York, USA: Routledge, 2017. Chap. 6, pp. 89–109. ISBN: 978-1-315-56962-8 (cit. on p. 200).

[286] F. B. Viegas, D. Boyd, D. H. Nguyen, J. Potter, and J. Donath. "Digital artifacts for remembering and storytelling: posthistory and social network fragments". In: *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of the*. 2004, 10 pp.– (cit. on pp. 43, 65, 67, 71, 75).

[287] F. B. Viégas and M. Wattenberg. "Artistic Data Visualization: Beyond Visual Analytics". In: *Online Communities and Social Computing*. Ed. by D. Schuler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 182–191. ISBN: 978-3-540-73257-0 (cit. on p. 194).

[288] F. B. Viégas, M. Wattenberg, and K. Dave. "Studying Cooperation and Conflict between Authors with History Flow Visualizations". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '04. Vienna, Austria: Association for Computing Machinery, 2004, pp. 575–582. ISBN: 1581137028. DOI: 10.1145/985692.985765. URL: https://doi.org/10.1145/985692.985765 (cit. on pp. 62, 63, 74).

[289] V. V. Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens. "GOTCHA! Network-Based Fraud Detection for Social Security Fraud". In: *Management Science* 63.9 (2017), pp. 3090–3110. DOI: 10.1287/mnsc.2016.2489 (cit. on p. 119).

[290] T. D. Wang, C. Plaisant, B. Shneiderman, N. Spring, D. Roseman, G. Marchand, V. Mukherjee, and M. Smith. "Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison". In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), pp. 1049–1056 (cit. on p. 63).

[291] M. Ward and J. Yang. "Interaction Spaces in Data and Information Visualization". In: *Proceedings of the Sixth Joint Eurographics - IEEE TCVG Conference on Visualization*. VISSYM'04. Konstanz, Germany: Eurographics Association, 2004, pp. 137–146. ISBN: 390567307X (cit. on p. 26).

[292] M. O. Ward, G. Grinstein, and D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Applications, Second Edition - 360 Degree Business*. 2nd. USA: A. K. Peters, Ltd., 2015. ISBN: 1482257378 (cit. on pp. 16, 20–23, 30).

[293] C. Ware. *Visual Thinking: For Design*. Morgan Kaufmann Series in Interactive Technologies. Amsterdam: Morgan Kaufmann, 2008 (cit. on p. 17).

[294] C. Ware. "Information Visualization: Perception for Design". In: Morgan Kaufmann, 2013 (cit. on pp. 16–19).

[295] M. Weber, M. Alexa, and W. Muller. "Visualizing time-series on spirals". In: *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.* 2001, pp. 7–13 (cit. on pp. 65, 67, 75, 76).

[296] S. Wehrend and C. Lewis. "A problem-oriented classification of visualization techniques". In: *Proceedings of the First IEEE Conference on Visualization: Visualization90*. IEEE. 1990, pp. 139–143 (cit. on pp. 25, 41).

[297] R. Wehrens and L. M. C. Buydens. "Self- and Super-organizing Maps in R: The kohonen Package". In: *Journal of Statistical Software, Articles* 21.5 (2007), pp. 1–19. ISSN: 1548-7660. DOI: 10.18637/jss.v021.i05. URL: https://www.jstatsoft.org/v021/i05 (cit. on p. 128).

[298] M. Wertheimer. "Laws of organization in perceptual forms". In: *A Source Book of Gestalt Psychology*. Ed. by W. Ellis. London: Routledge and Kegan Paul, 1938, pp. 71–88 (cit. on pp. 16, 17).

[299] T. White. "Symbolization and the Visual Variables". In: *Geographic Information Science & Technology Body of Knowledge* 2017.Q2 (2017). DOI: 10.22224/gistbok/2017.2.3. URL: http://dx.doi.org/10.22224/gistbok/2017.2.3 (cit. on pp. 20–24).

[300] B. Wilkins. "Meld : a pattern supported methodology for visualisation design". In: 2003 (cit. on p. 8).

[301] G. Wills. *Visualizing time: Designing graphical representations for statistical data*. Springer Science & Business Media, 2011 (cit. on pp. 6, 17, 36, 43, 53, 54, 58, 61).

[302] M. Wilson. *Listen To The Orchestra Of Users Updating Wikipedia.* 2013. URL: http://listen.hatnote.com/ (cit. on pp. 199, 200).

[303] M. Wolter, I. Assenmacher, B. Hentschel, M. Schirski, and T. Kuhlen. "A time model for time-varying visualization". In: *Computer Graphics Forum*. Vol. 28. 6. Wiley Online Library. 2009, pp. 1561–1571 (cit. on p. 35).

[304] K. Wongsuphasawat and D. Gotz. "Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization". In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), pp. 2659–2668 (cit. on p. 62).

[305] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. "LifeFlow: visualizing an overview of event sequences". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2011, pp. 1747–1756 (cit. on p. 62).

[306] M. Wooldridge. *An Introduction to MultiAgent Systems*. 2nd. Wiley Publishing, 2009. ISBN: 0470519460 (cit. on p. 215).

[307] R. Wurman, L. Leifer, M. Nolan, D. Sume, and K. Whitehouse. *Information Anxiety 2*. Hayden/Que. Que, 2001. ISBN: 9780789724106. URL: https://books.google.pt/books?id=KjhjQYP3I08C (cit. on p. 193).

[308] R. R. Xu and S. H. Zhai. *Out of Statistics : Beyond Legal*. 2009. URL: http://floatingcube.org/beyondlegal/ (cit. on p. 195).

[309] Yarden Livnat, J. Agutter, S. Moon, R. F. Erbacher, and S. Foresti. “A visualization paradigm for network intrusion detection”. In: *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop*. 2005, pp. 92–99 (cit. on p. 64).

[310] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko. “Toward a Deeper Understanding of the Role of Interaction in Information Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 13.6 (Nov. 2007), pp. 1224–1231. ISSN: 1077-2626. DOI: 10.1109/TVCG.2007.70515. URL: https://doi.org/10.1109/TVCG.2007.70515 (cit. on pp. 25–27, 31, 39).

[311] Yufeng Kou, Chang-Tien Lu, S. Sirwongwattana, and Yo-Ping Huang. “Survey of fraud detection techniques”. In: *IEEE International Conference on Networking, Sensing and Control, 2004*. Vol. 2. 2004, 749–754 Vol.2. DOI: 10.1109/ICNSC.2004.1297040 (cit. on pp. 119, 121, 123).

[312] J. Zhang. “A representational analysis of relational information displays”. In: *International journal of human-computer studies* 45.1 (1996), pp. 59–74 (cit. on p. 41).

[313] Y. Zhang, K. Chanana, and C. Dunne. “IDMVis: Temporal Event Sequence Visualization for Type 1 Diabetes Treatment Decision Support”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 512–522 (cit. on pp. 61, 62, 67).

[314] J. Zhao, N. Cao, Z. Wen, Y. Song, Y. Lin, and C. Collins. “FluxFlow: Visual Analysis of Anomalous Information Spreading on Social Media”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1773–1782 (cit. on pp. 44, 63, 67).

[315] J. Zhao, F. Chevalier, E. Pietriga, and R. Balakrishnan. “Exploratory Analysis of Time-Series with ChronoLenses”. In: *IEEE Transactions on Visualization and Computer Graphics* 17.12 (2011), pp. 2422–2431 (cit. on pp. 62, 67).

[316] J. Zhao, F. Chevalier, and R. Balakrishnan. “KronoMiner: Using Multi-Foci Navigation for the Visual Exploration of Time-Series Data”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, pp. 1737–1746. ISBN: 9781450302289. DOI: 10.1145/1978942.1979195. URL: https://doi.org/10.1145/1978942.1979195 (cit. on pp. 64, 67, 76, 77).

[317] M. X. Zhou and S. K. Feiner. “Visual Task Characterization for Automated Visual Discourse Synthesis”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’98. Los Angeles, California, USA: ACM Press/Addison-Wesley Publishing Co., 1998, pp. 392–399. ISBN: 0201309874. DOI: 10.1145/274644.274698. URL: https://doi.org/10.1145/274644.274698 (cit. on pp. 25, 160).

[318] Y. Zhu. “Measuring Effective Data Visualization”. In: *Proceedings of the 3rd International Conference on Advances in Visual Computing - Volume Part II*. ISVC’07. Lake Tahoe, NV, USA: Springer-Verlag, 2007, pp. 652–661. ISBN: 3540768556 (cit. on p. 30).

[319] Y. Zhu, J. Yu, and J. Wu. “Chro-Ring: a time-oriented visual approach to represent writer’s history”. In: *The Visual Computer* 32.9 (Mar. 2016), pp. 1133–1149. DOI: 10.1007/s00371-016-1213-4. URL: https://doi.org/10.1007/s00371-016-1213-4 (cit. on p. 64).

[320] C. del Coso, D. Fustes, C. Dafonte, F. J. Nóvoa, J. M. Rodríguez-Pedreira, and B. Arcay. “Mixing numerical and categorical data in a Self-Organizing Map by means of frequency neurons”. In: *Applied Soft Computing* 36 (2015), pp. 246 –254. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2015.06.058. URL: http://www.sciencedirect.com/science/article/pii/S1568494615004512 (cit. on pp. 127, 129, 130).