Why so many people? Explaining non-habitual transport overcrowding with internet data

Francisco Pereira, *Member, IEEE,* Filipe Rodrigues, Evgheni Polisciuc, and Moshe Ben-Akiva

Abstract—**Public transport smartcard data can be used to detect large crowds. By comparing smartcard data with statistics on habitual behavior (e.g. average by time of day), one can specifically identify non-habitual crowds, which are often problematic for the transport system. While habitual overcrowding (e.g. during peak hour) is well understood by traffic managers and travelers, non-habitual overcrowding hotspots can be very disruptive given that they are generally unexpected. By quickly understanding and reacting to cases of overcrowding, transport managers can mitigate transport system disruptions.**

We propose a probabilistic data analysis model that breaks each non-habitual overcrowding hotspot into a set of explanatory components. Potential explanatory components are retrieved from social networks and special events websites and then processed through text-analysis techniques. We then use the probabilistic model to estimate each components specific share of total overcrowding counts.

We first validate with synthetic data and then test our model with real data from Singapores public transport system (EZLink), focused on 3 case study areas. We demonstrate that it is able to generate explanations that are intuitively plausible and consistent both locally (correlation coefficient, CC, from 85% to 99% for the 3 areas) and globally (CC from 41.2% to 83.9%).

This model is directly applicable to domains that are sensitive to crowd formation due to large social events (e.g. communications, water, energy, waste).

I. Introduction

Given the quantity and quality of pervasive technologies such as RFID, smartcards and mobile phone communications, we have the ability to detect crowds with minimal risk to privacy in almost real-time. Crowd detection is a valuable measure for safety and security as well as for real-time supply/demand management of transportation, communications, food stock, logistics, water and many other systems that are sensitive to aggregated human behavior. Although these technologies help detect and quantify crowds, they have limited power to explain why crowds happen.

We are less concerned with recurring crowds, such as peak-hour commuting, because we have a better understanding of why these crowds occur. However, we face greater challenges in explaining non-habitual overcrowding scenarios in which we need contextual knowledge in order to discern explanations.

Fortunately, the Internet is a pervasive technology that is rich in local context. Information about public special events (such as sports games, concerts, parades, sales, demonstrations, and festivals), social networks (e.g. Twitter, Facebook) and other platforms that have dynamic context content (e.g. news feeds) are abundant.

In order to assess and treat issues of overcrowding, we must first understand why people are crowding, and where/when they will go next. Then managers can react accordingly by, for example, adding extra buses, trains, or taxis. For example, if we know that an overcrowding hotspot is due because of a concert, we can also estimate the overcrowdings duration (for instance, until shortly after the concert begins) and the next possible overcrowding hotspot (for instance, immediately after the concert ends). If the overcrowding is instead due to a series of small, scattered events, the treatment may be different (e.g. no single ending hotspot). By understanding such impacts on a post-hoc analysis, we can learn from previous events and better prepare for the next time similar events occur.

This paper aims to address the following problems: what are the potential causes of a non-habitual large crowd (an overcrowding hotspot); and how do these potential causes individually contribute to the overall impact? We will particularly focus on public transport overcrowding in special events areas.

Given the importance of these social phenomena, many traffic management centers have teams of people that are responsible for periodically scanning the internet and newspapers in search of special events. The challenge comes when multiple smaller events co-occur in the same area because it is not only harder to find them, but it is also difficult to estimate their aggregated impact.

We identify and measure the overcrowding hotspots by analyzing 4 months of public transport data from Singapore. We define a hotspot as a continuous period where observed demand (e.g. number of arrivals) repeatedly exceeds a high percentile (e.g. 90%). The overcrowding hotspot impact is measured as the total sum of demand above the median line.

During the whole period of the dataset, we collected special events data from 5 websites[1] as well as their Facebook likes and Google hits. While the latter two are numerical in nature, the former include unstructured text descriptions. Hence, we apply an information extraction technique, called topic modeling [1], that transforms such data into a set of

F. Pereira is with the Singapore-MIT Alliance for Research and Technology, SMART, Singapore, e-mail: camara@smart.mit.edu.

F. Rodrigues and E. Polisciuc are with the University of Coimbra, Portugal.

M. Ben-Akiva is with the Massachusetts Institute of Technology (MIT), USA.

[1] These websites were www.eventful.com, upcoming.org, last.fm, timeoutsingapore.com and singaporeexpo.com.sg.

features understandable from a machine learning algorithms perspective.

Since we only have observations of aggregated impacts rather than the individual events, we propose a *Bayesian hierarchical additive model*, where each hotspot is formalized as a sum of the potential explanatory components. We explicitly model uncertainty on the parameters by using the Infer.NET platform [2].

We validate the model in two ways: using synthesized impact data based on real event descriptions; and comparing the sum of estimations from our model with the observed real-world sums. In doing so, we demonstrate that our model approximates the (simulated) ground truth, and that the results in a real world case are feasible.

This methodology is applicable beyond the specific case of public transport overcrowding. For example, during special events, cell-phone, wifi network, energy, waste or catering/logistics systems may be equally disrupted. If data is available to quantify and correlate the impacts, the general procedure remains the same. We provide the source code and the synthesized dataset for interested readers[2].

The main contributions of this paper are a fully implemented model for inferring latent demand contributions in special events scenarios (with code available to the reader, runnable in the Infer.NET platform); the application of a state-of- the-art topic modelling technique (LDA); and a validation technique with synthetic data that follows a realistic methodology and can be used as a benchmark for future work.

In Section II, we present a literature review. In Section III, we explain how we determine the overcrowding hotspots and in Section IV we show how we collect potential explanatory data from the Web. A Bayesian model is explained in Section V and is followed by experimentation and validation in Section VI. We analyze several hotspots in Section VII and end the paper with a discussion and conclusions (Sections VIII and IX, respectively).

II. Literature review

A. Detecting mobility patterns with pervasive data

In June 2014, a search with the keywords *cell phone* and *human mobility* in Google returned approximately 14.4k entries. In Google Scholar, we find over 1000 entries that mention these words explicitly. If we include other types of data such as GPS or smartcard data, these numbers will increase even more dramatically. Therefore, we mention a few papers we consider seminal to the area.

Using a large cell-phone dataset, González and colleagues [3] showed that individual mobility travel patterns generally follow a single spatial probability distribution. This finding indicates that despite their inherent heterogeneity, humans follow simple reproducible patterns. Several other works indicate that human mobility is habitual for the vast majority of the time, for example to estimate disease spreading [4] or vehicular network routing protocols [5].

Despite other studies that stretch the boundaries of that principle and verify that it is widely persistent (e.g. [6], [7]), mobility behavior heterogeneity is recognized to create predictability challenges. This is particularly important when it involves large crowds. As pointed out by Potier et al [8], even for well-known big events (e.g. Olympic games), demand is more difficult to forecast than habitual mobility, particularly in the case of open-gate events. When facing these constraints, authorities tend to rely on trial and error experience (for recurring events), checklists (e.g. [9]) and sometimes invest in a reactive approach rather than planning. This is the case in Germany, with the RTTI (Real-time Traffic and Traveller Information) and its active traffic management [10] and in Netherlands [11]. However, such tools have limited applicability, particularly for smaller and medium events, that are harder to capture and to evaluate.

Calabrese et al [12] use a massive cell-phone dataset to study public home distributions for different types of special events (e.g. sports, concerts, theatre). They identified a strong correlation between public neighborhood distributions and event types. This is a key finding since it implies that such heterogeneous cases are still predictable as long as we have sufficient event information. They did not, however, consider multiple event interactions or deeper explanatory content (e.g. event description text).

B. The role of the Internet

The Internet is the best source for extracting special events information. We can also explore online popularity features, such as Facebook likes or Google trends. In an earlier work [13], we compared an origin/destination (OD) prediction model with and without simple information obtained from the Internet, such as event type or whether the performer/event had a Wikipedia page. We verified that such information could reduce the root mean squared error (RMSE) by more than 50% in each OD. This study was done on a single spatially isolated venue that had one event at a time. When we applied it to more complex locations, we verified that a deeper analysis was needed to handle multiple concurrent events.

The internet is also a valuable source for other aspects of mobility research. For example, Twitter has been used for opinion mining on public bus [14] and inference of home locations [15]; Points of Interest (POIs) from Foursquare, Yahoo! local and others have supported studies on urban region functions [16] and job-related trips [17]; and Flickr has been used to study the geographical distribution of activities (e.g. beach, hiking, sunset) [18] or to recommend touristic routes [19].

C. Topic models

Even if we have all the web pages that announce our events, a considerable amount of the relevant information will be in textual form. To obtain an automated system, we still need to convert such data into a proper representation that a machine can understand. Explicitly including the text, word by word, in a machine learning model would increase its dimensionality much beyond the reasonable. On the other hand, hand coding rules that find certain "relevant" words (e.g. "rock", "pop",

[2] https://dl.dropboxusercontent.com/u/1344277/PereiraEtAl2014.zip

"football", "festival") would incur in plenty of subjective judgment and lack of flexibility. Natural language is rich in synonymy and polysemy, different announcers and locations may use different words, besides it is not always obvious which words are more "relevant" from the perspective of a machine learning model.

The approach of topic modelling research to these questions is to re-represent a text document as a finite set of *topics*. These topics correspond to sets of words that tend to co-occur together rather than a single word associated to a specific topic. For example, a rock festival textual description could have a weight w_1 assigned to topic 1 (e.g. words related to concerts in general), w_2 of topic 2 (e.g. words related to festivals), w_3 of topic 3 (e.g. words related to the venue descriptions) and so on. In particular, we use a specific technique that is called Latent Dirichlet Allocation (LDA). For the readers that are familiar with Principal Components Analysis (PCA), there is a simple analogy: PCA re-represents a signal as a linear combination of its eigenvectors, while LDA re-represents a text as a linear combination of topics. In this way, we reduce the dimensionality from the total number of different words of a text to the number of topics, typically vey low.

In LDA, each document is represented as a distribution over topics, and each topic is a distribution over words. Formally, given each document d defined as a vector $\mathbf{w_d}$ of n words, $\mathbf{w_d} = \{w_{d,1}, \dots w_{d,n}\}$ and the parameter K, representing the number of different topics, LDA assumes the following generative process:

1) Draw a topic β_k from $\beta_k \sim$ Dirichlet(η) for $k = 1 \dots K$
2) For each document d:
    a) Draw topics proportions θ_d such that $\theta_d \sim$ Dirichlet(α)
    b) For each word $w_{d,n}$:
        i) Draw topic assignment $z_{d,n} \sim$ Multinomial(θ_d)
        ii) Draw word $w_{d,n} \sim$ Multinomial($\beta_{z_{d,n}}$)

The parameters α and η are hyperparameters that indicate respectively the priors on per-document topic distribution and per-topic word distribution, respectively. Thus, $w_{d,n}$ are the only observable variables, all the others are latent in this mode. For a set of D documents, given the parameters α and η, the joint distribution of a topic mixture θ, word-topic mixtures β, topics $\mathbf{z}$, and a set of N words is given by:

$$p(\theta, \beta, \mathbf{z}, \mathbf{w}|\alpha, \eta) = \prod_{k=1}^{K} p(\beta_k|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha)$$
$$\prod_{n=1}^{N} \Big(p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_k, k = z_{d,n}) \Big)$$

Broadly speaking, the training task is to find the posterior distribution of the latent variables (the per-document topic proportions θ_d, the per-word topic assignments $z_{d,n}$ and the topics β_k) that maximize this probability. As with most generative models, the exact inference of such values is intractable to compute, therefore approximate inference techniques are used, namely Markov Chain Monte Carlo methods (e.g. Gibbs sampling), or variational inference, or Expectation-Maximization (EM). For further details on this procedure please refer to the original article of David Blei and colleagues [1] and to practical implementation documents (e.g. GenSim [20]).

With a trained LDA topic model, one can apply the same general procedure to assign topics to every new document through euclidian projection on the topics [1], which is generally a very fast procedure.

A final remark relates to the document representation that is typically adopted for LDA and similar techniques, known as the *bag-of-words* representation. Having a dictionary with W different words, this representation translates each document into a vector with dimensionality W, where each element contains the frequency of a dictionary word observed in the document. This technique obviously disregards the original order of words in the text, being based purely on word counts.

D. Hierarchical models

A *hierarchical model* aims to capture effects at two or more levels [21]. The top level represents the most general parameters (e.g. global mean and intercept), and the lower levels introduce effects specific to sub-populations. . In our case, we first break down a hotspots impact into *non-explainable* and *explainable* components. The non-explainable component represents all excessive demand for which we cannot find explanations online. Its existence is more obvious in days without any eventsThis does not correspond to the residual on a regression model since we do not assume it to be normally distributed with 0 mean.. In the second level, the explainable component is expanded into a summation of individual event contributions.

Since this model is a summation of several individual sub-models, it is an *additive model*. We apply the Bayesian framework to estimate its parameters, using the Infer.NET platform [2], hence the title *Bayesian hierarchical additive model*.

III. Identifying overcrowding hotspots

There is no well-defined threshold above which we identify overcrowding. The intuition is that it should happen whenever the supply (e.g. buses) is insufficient to satisfy the demand (e.g. travelers), which leads to heavily loaded vehicles or to denied boarding. The latter is non-observable from our dataset, as are estimates of bus or train loading. Therefore we resort to indirect measurements such as total number of arrivals.

In order to cope with demand fluctuations, transport systems are generally designed with reasonable spare capacity, so we need to define the point above which we consider the system under stress. For any given study area and point in time, we define such points to correspond to the 90% percentile, i.e. whenever the number of arrivals exceeds such threshold, *overcrowding* is occurring. This threshold choice is based on our intuition and experience combined with discussions with local experts. However, our main contribution is methodological and all principles should remain the same if choosing another threshold.

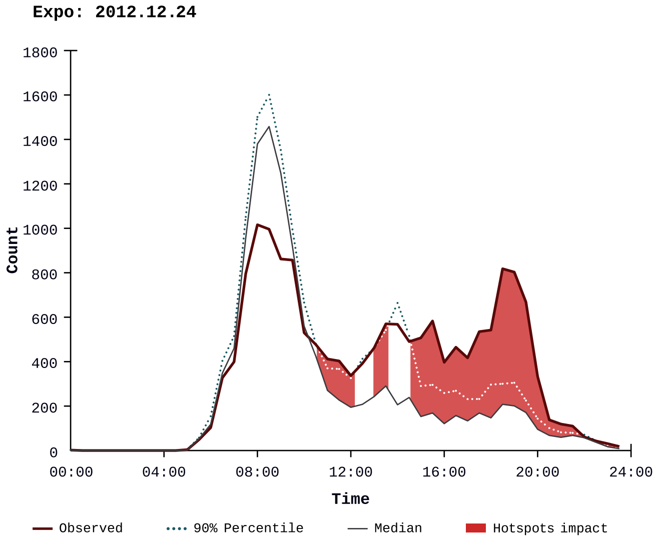


Fig. 1. Overcrowding hotspots detection and measurement.

We quantify the impact by summing up the excess amount of arrivals above the median line in a continuous time frame, discretized by 30 minutes intervals. Figure 1 visualizes this calculation. On 24-12-2012, there were 3 hotspots in this area (Singapore Expo). There were two simultaneous events during several hours (Megatex, related to IT and electronics; and Kawin-kawin makan-makan 2012, an event about Malay food and lifestyle products).

Whenever hotspots are both short in time and have a small relative impact (e.g. below 5% of the mean, only 30 minutes), we remove them as they should not represent a problem for the transportation system.

Our dataset consists of 4 months of smartcard public transport data from Singapores EZLink system. This is a tap-in/tap-out system for both buses and subway (MRT), which means we can infer both departure and arrival locations for any trip. For the purposes of this specific study, we selected trips that start/end in 3 areas that are sensitive to multiple special events: Stadium; Expo; and Esplanade. The Stadium area is dominated by two venues, the Singapore Indoor Stadium and the Kallang theatre. The Expo consists of a single large venue, but it commonly hosts multiple unrelated events simultaneously. The Esplanade has 47 venues and is a lively tourist area near the business district. It has several shopping malls nearby and sits in front of the iconic marina bay of Singapore.

In Table I, we show some descriptive statistics from these areas:

TABLE I. GENERAL STATISTICS: AVERAGES (+-σ) AND TOTALS.

Area	Average daily	Avg. daily	Nr.	Avg. hotspot
	arrivals	events	hspts	impacts
Stadium	4120.4(+-1015.9)	.2(+- .49)	103	382.9(+-680.0)
Expo	14797.5(+-5851.3)	2.5(+- 2.0)	70	2836.7(+-4846.3)
Esp.	4788.7(+-930.5)	17.0(+-6.4)	102	231.6(+-430.8)

IV. RETRIEVING POTENTIAL EXPLANATIONS FROM THE WEB

For each overcrowding hotspot we want to find a set of possible explanations from the web. Two general techniques exist to capture such data automatically: Application Programming Interfaces (APIs) and screen scraping. The choice of technique depends on the website. Some websites provide an exhaustive API that we can use to retrieve the data while with other sites we need to resort to page by page screen scraping. In either method, access may be restricted or prohibited by *terms of service*, therefore we implement individual *event data retrievers* for each website only whenever it is so permitted. We use 5 different websites: eventful.com, upcoming.org, last.fm, timeoutsingapore.com and Singapore Expos website singaporeexpo.com.sg.

For potential duplicate entries that share the same venue/area and day, we use the Jaro-Winkler string distance [22] with a conservative threshold (e.g. ¿ 85% similarity) to identify and merge them. Whenever we find different textual descriptions, we concatenate them.

Each event record contains title, venue, web source, date, start-time, end-time, latitude, longitude, address, url, description, categories, and event price. Unfortunately, this information also contains noise. For example, start and end times are often absent or "default" (e.g. from 00:00 to 23:59), and the same sometimes happens with latitude/longitude (e.g. center of the map). The latter can be corrected by using the venue name, but for the former, we could not determine any particular times. As a consequence, each such event is potentially associated to any impact hotspot of the corresponding day and area.

The description text is run through a latent dirichlet allocation (LDA) process as explained in Section II-C One key parameter for this process is the number of topics. We tested a range of values from 15 to 40 and found that the value of 25 yielded the best model results. We assume this value for the remainder of the paper. The other parameters, the α and η priors, were kept as default (1.0/(number of topics)). To confirm this was a safe choice, we ran several iterations with different initial α and η priors and they generally converged to similar outcomes.

For each event, we capture two online popularity indicators: the number of Facebook likes and the number of hits in Google of the event title query. We retrieve the Facebook page with a semi-automatic procedure: we follow the event URL (which is sometimes a Facebook page) in search of candidate pages. Whenever there is more than one candidate, we manually select the correct one. For Google hits, we search with the event title within and without quotes (yielding two separate features).

In Table II, we summarize general statistics of this dataset.

TABLE II. GENERAL STATISTICS ON DATA FROM THE INTERNET.

Source.	Nr. events	Nr.	Text desc.	Retrieval
	study areas	categories	size (+-σ)	type
Eventful	1221	28	1112.3 (+-1337.1)	API
Expo	58	28	124.9 (+-159.5)	scraper
upcoming	181	13	2423.9 (+-5362.7)	API
last.fm	11	-	901.2 (+-1037.5)	API
timeout	568	49	411.8 (+-866.6)	scraper

We can see that the most comprehensive ones are eventful and timeout, while the one with more detailed descriptions is upcoming. Expo homepage and last.fm have much less, yet very directed information, the former contains all events that happen in Expo (thus a relevant filter in itself) while the latter is only focused on music events.

V. Bayesian hierarchical additive model

The individual event contributions are not observed (i.e. they are *latent*), but we do know they contribute to the global observed impact. We will also assume that individual impacts are mutually exclusive (e.g. no one attends two events), independently distributed and that there will be a parcel that is unexplainable, i.e. some trips will neither be related to the extracted events nor to the usual commuting patterns. Thus, we say that a hotspot impact, h, is given by[3] $h = a + b$, where a is the non-explainable component and b is the explainable one. b is a summation of the k events, e_k. Formally, we define a and b in the following way:

$$a \sim \mathcal{N}(\boldsymbol{\alpha}^T \mathbf{x}_a, \sigma_a) \tag{1}$$

$$b = \sum_{k=1}^{K} e_k, \text{with } e_k \sim \mathcal{N}(\boldsymbol{\beta}^T \mathbf{x}_{e_k}, \sigma_k) \tag{2}$$

where $\mathbf{x}_a$, $\boldsymbol{\alpha}$ and σ_a are respectively the attributes, parameter vectors and variance for the non-explainable component, a. The explainable part, b, is determined by a sum of event contributions, e_k, the second level of our linear model. Each $\mathbf{x}_{e_k}$ corresponds to the individual attributes of event k (e.g. topic-assignments, categories, Facebook likes, etc.), and $\boldsymbol{\beta}$ and σ_k correspond to the event attributes' parameters and the variance associated with that event, respectively. At both levels we assumed a Gaussian distribution for the non-explainable and individual event contributions.

The functional form of the components a and e_k follows a linear model, and we will continue to use this form for this paper. In future work, we intend to extend our work to non-linear models. Note that the general diagram (of Figure 2) will still hold while only functional form of individual components need to be changed.

For the remainder of this section, we apply the Bayesian framework [23], which relies on three concepts: the *likelihood function*, or the probability that a model with a specific set of parameters predicts the observed data; the *prior*, that represents assumptions with respect to model components (e.g. variables, parameters); and the *posterior*, that provides the probability distribution of the model parameters or other variables of interest after observing the data. A major advantage of this framework is that it provides a distribution of values as opposed to a single estimate for each variable of our model. For example, the estimation of a classical linear regression model will result in a set of individual parameter values. Each prediction consists of a single new value, while its Bayesian counter-part results in a probability distribution of values for the parameters and for the predictions themselves.

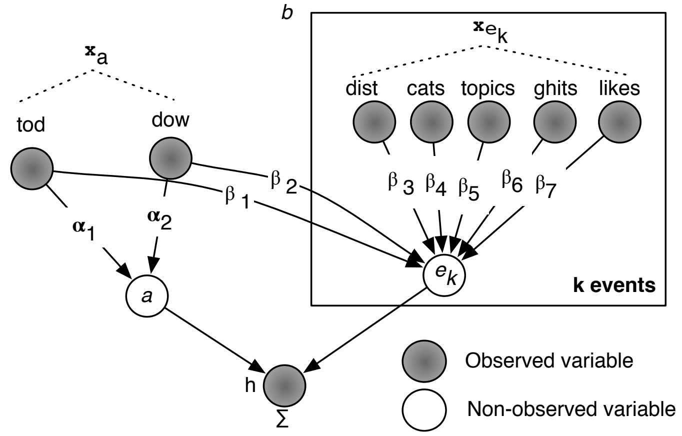


Fig. 2. Plate notation for our model. Legend: tod=time of day; dow=day of week; dist=distance of bus/train stop to venue; cats=categories; topics=lda topics; ghits=Google hits; likes=Facebook likes; a=non-explainable part; e_k=explaining components. w_i and α_j= are model parameters.

While we could simply use the most probable values of each distribution and reduce it to a non-Bayesian model, we risk losing critical information. In our example, by simply choosing the most probable values for the parameters we can obtain predictions for the totals, as in a classic linear regression. However, we can go further by obtaining a distribution of values for totals as opposed to a single point estimate. Since we have these observed totals, we know how well our model is tuned (a good model should provide high probability to the observed value). More importantly, we can use this information to revisit the parameter distributions again to select values that are more consistent with the totals. In practice, for each hotspot the observed totals work together with the parameter distributions to adapt the model to the most likely values. This feedback mechanism is possible with the Bayesian framework, and is embedded in the Infer.NET platform [2].

The advantages and challenges of the Bayesian framework in comparison with other machine learning approaches have been discussed extensively elsewhere and are beyond the scope of this paper. For further information, we recommend the book of Christropher Bishop [23].

In Figure 2 we graphically present our model. Arrows indicate conditional dependence (e.g. a depends on $\mathbf{x}_a$), and nodes correspond to variables. Some are observed (e.g. the sum, h), others are non-observed (e.g. event contributions e_k). Rectangles, or *plates*, are used to group variables that repeat together. This representation is known as *plate notation* [24]. We recall that our main goal is to obtain the values for a and e_k, and that they sum up to h. This relationship can be

[3] For clarity of notation, we will simplify the full notation, $h_{r,j} = a_{r,j} + b_{r,j}$ as $h = a + b$, throughout the article, where r would be the area index, and j the hotspot index.

represented through their joint probability distribution[4]:

$$p(h, a, \mathbf{e}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X}) \\ = p(h|a, \mathbf{e})p(a|\boldsymbol{\alpha}, \mathbf{x}_a) \prod_{k=1}^{K} p(e_k|\boldsymbol{\beta}, \mathbf{x}_{e_k}) \tag{3}$$

where we define $\mathbf{e} = \{e_1, ..., e_K\}$ and $\mathbf{X} = \{\mathbf{x}_a, \mathbf{x}_{e_1}, ..., \mathbf{x}_{e_K}\}$ for compactness. It may be helpful to note the relationship between Figure 2 and the expansion on the right hand side of the equation, where we can see the conditional dependences. The likelihood function for the observed sum h is:

$$p(h|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X}) = \iint p(h, a, \mathbf{e}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X})\, da\, d\mathbf{e} \\ = \iint p(h|a, \mathbf{e})p(a|\boldsymbol{\alpha}, \mathbf{x}_a) \prod_{k=1}^{K} p(e_k|\boldsymbol{\beta}, \mathbf{x}_{e_k})\, da\, d\mathbf{e}$$

By making use of the Bayes rule, we can define the joint posterior distribution of the parameters:

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta}|h, \mathbf{X}) = \frac{p(h|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X})p(\boldsymbol{\alpha})p(\boldsymbol{\beta})}{\iint p(h|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{X})p(\boldsymbol{\alpha})p(\boldsymbol{\beta})d\boldsymbol{\alpha}d\boldsymbol{\beta}} \tag{4}$$

The integral in the denominator is the normalization factor and $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{\beta})$ are the priors, which will follow a standard Gaussian distribution ($\mathcal{N}(0, 1)$).

We can finally estimate the posteriors for a and $\mathbf{e}$ as:

$$p(a|h, \mathbf{X}) = \int p(a|\boldsymbol{\alpha}, \mathbf{X})p(\boldsymbol{\alpha}|h, \mathbf{X})d\boldsymbol{\alpha} \tag{5}$$

and

$$p(\mathbf{e}|h, \mathbf{X}) = \int p(\mathbf{e}|\boldsymbol{\beta}, \mathbf{X})p(\boldsymbol{\beta}|h, \mathbf{X})d\boldsymbol{\beta} \tag{6}$$

where we use equations 1 and 2 for $p(a|\boldsymbol{\alpha}, \mathbf{X})$ and $p(\mathbf{e}|\boldsymbol{\beta}, \mathbf{X})$, respectively, and $p(\boldsymbol{\alpha}|h, \mathbf{X}) = \int p(\boldsymbol{\alpha}, \boldsymbol{\beta}|h, \mathbf{X})d\boldsymbol{\beta}$ and $p(\boldsymbol{\beta}|h, \mathbf{X}) = \int p(\boldsymbol{\alpha}, \boldsymbol{\beta}|h, \mathbf{X})d\boldsymbol{\alpha}$.

We implemented this model in the Infer.NET framework [2], which has the necessary approximate Bayesian inference and Gaussian distribution treatment tools that help make it computationally efficient. We made our code freely available[1].

VI. Model validation

A. Synthesized data experiments

Since we have access to total values but not to the individual contributions, we need to determine how to validate the model. First, we need to test the model as if we had observed individual contributions. We do this by generating simulated data that complies with our assumptions. Afterwards, in the next section, we test how well our model fits with respect to the total (observed) values.

If we cluster the events dataset (from Section IV) using the events characteristics, we end up with sets of events that are somehow related in feature space. We assume that each cluster centroid is manually or randomly assigned its own impact. This value represents the impact of a hypothetical event that does not necessarily exist in the database. Next, we assign impacts to the real events using the distance to their cluster centroid, c. For each event e, its impact is determined by $dist(e, c)^{-1}$.

With this procedure, we are not forcing our model structure into the data (i.e. we are not assigning specific parameter values to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$), instead we use similarity between events to introduce consistency, regardless of area or day.

The individual impacts of simultaneously occurring events are summed and the resulting value is affected by some percentage of noise ($\mathcal{N}(0, 0.1 * b)$). The final result is provided in our model as the observed hotspot impact. The obtained individual impacts are then compared to the ground truth (simulated) values according to three error statistics: the mean absolute error (MAE) provides the absolute magnitude of the error for each impact; the root relative squared error (RRSE) shows the quality of the model relative to a naive predictor based on the average of all observations for that venue; the correlation coefficient (CC) provides insight for how our model results are correlated with the ideal results.

Table III shows the results for the areas of Stadium, Expo and Esplanade.

TABLE III. Synthetic data results

Area	MAE	RRSE	CC
Stadium	410.3	0.21	0.99
Expo	145.0	0.45	0.89
Esplanade	708.1	0.56	0.85

Our model performs differently depending on the area. In Stadium, to the model replicates well the contributions, which is not surprising since this area is more homogeneous than the others (often with only one event in a day). Despite being much more heterogeneous, the Expo and Esplanade models have significant correlation coefficient and considerably outperform the average based predictor.

B. Real data experiments

The observations that we have consist of total hotspot impacts according to Section III. We now want to test our model's capability of recovering such aggregated impacts *without* knowing the individual impacts, it will only count with the known features such as location, day of week, event type, topics, etc. (vectors $\mathbf{x}_a$ and $\mathbf{x}_{e_k}$ as described in Figure 2). We do this by first estimating the parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) with a subset of the observations (training set) then generating the aggregated hotspot impacts with the remaining subset (test set). We apply the 10-fold cross-validation methodology (see e.g. [23]). We use the same error metrics as in the previous section. Table IV shows the summary of the results.

TABLE IV. Real data results

Area	MAE	RRSE	CC
Singapore Indoor Stadium	271.7	0.55	0.68
Singapore Expo	2002.7	0.69	0.84
Marina Promenade	192.6	0.84	0.41

A hotspot can last for many hours, which may lead to large arrival totals, particularly in the Expo and Esplanade areas.

[4] Since $\mathbf{e}$ deterministically dictates b, we replaced b by $\mathbf{e}$ from the beginning.

Thus, the relevance of MAE is difficult to assess. However, the values for RRSE and CC in these areas indicate that the model is able to provide good performance, while the results for the Esplanade are less conclusive.

Notice that this task is not what the model was designed for. The total *observed* sum is not a deterministic constraint anymore, now it becomes an extra unknown. Yet, this exercise serves the validation of our model by allowing us to compare the sum totals (now estimated) with the observed ground truth. Notwithstanding this more complicated task, it is able to approximate the totals well in two of the cases (Stadium and Expo). If our model assumptions were wrong, we would expect the predictions to be considerably off, because the magnitude of the totals varies according to the time duration of the hotspot and because the individual event proportions could be wrong. The specific Esplanade case will be analyzed in the following section.

VII. Explaining hotspots

The ultimate goal of our algorithm is to break down each overcrowding hotspot into a set of explanatory components. In this section, we present the results for our entire dataset. Previously, we validated individual component predictions through a synthetic dataset and the aggregated totals with the observations. Now we do not have observations on individual events, and even if we had access to individual participation data (e.g. through ticket sale statistics), it would not necessarily reveal the correct number of public transport users for that specific event. Thus, our evaluation is now qualitative.

Figures 3, 4 and 5 illustrate some of the results[5]. For each hotspot, we show the global impact (inner circle) and the breakdown (outer circle). The area size of the inner circle is relative to the maximum hotspot impact observed in that location in our dataset. The outer circle contains as many segments as potential explanatory events plus the non-explainable component (in red). For example, on 2012-11-10, Expo had a high impact hotspot (top left diagram in Figure 3) comprised of 8 different events, with roughly the same size. The non-explainable component was small (red segment). Alternatively, on 2012-11-19, the same area had 2 events, one of which explains almost half of a relatively small hotspot as compared to the previous case.

For Stadium and Expo, we can see that the non-explainable component is generally smaller than the explainable one and that the breakdown is not evenly distributed. This happens because the model maximizes consistency across different events. For example, two similar events in two occasions will tend to have similar impacts although the overall totals and sets of concurrent events may be different.

Cases with multiple hotspots in the same day are worth attention. In Figure 3, Expo had 3 hotspots on 2012-11-11, with minor fluctuations in the impacts and individual breakdowns. There were 10 different medium sized events that spanned the course of the day. Conversely, in Stadium (Figure 4) the hotspots for 2013-02- 22 have differing behaviors. There was a fanmeet event with a Korean music and TV celebrity that started at 20:00 (we note that the largest impact is between 17:30 and 21:00). While the algorithm is confident in the first hotspot, it leaves the second hotspot mostly unexplained.

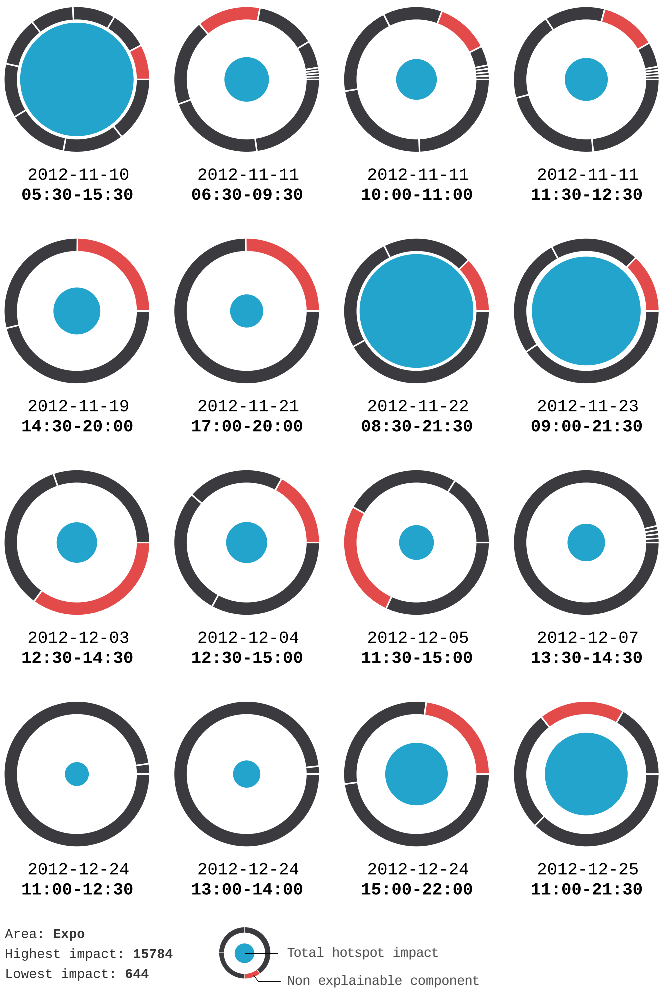


Fig. 3. 12 events from Expo area.

The Esplanade area (Figure 5) shows unclear patterns. A careful examination of the data shows that sometimes there are multiple small events announced for that area, from game watching nights at bars to theatre sessions. Outliers exist (e.g. concerts) but the algorithm would probably need more outlier cases to extract them. Nevertheless, it shows capability of ruling out insignificant events and assigns 0 impact to them.

Let us now analyze a few cases in detail, In Figure 6, we show the hotspot breakdown of Figure 1 according to our model. We note that it was Christmas eve and there were two events: Megatex, an IT and electronics fair; and Kawin-kawin makan-makan, a Malay products event. Our model proposes that the majority of the impacts relate to the electronics event, which is intuitively plausible, particularly on the day before Christmas and knowing that Singapore has a well-known tech-savvy culture.

[5] Full set in https://dl.dropboxusercontent.com/u/1344277/PereiraEtAl2014.zip.

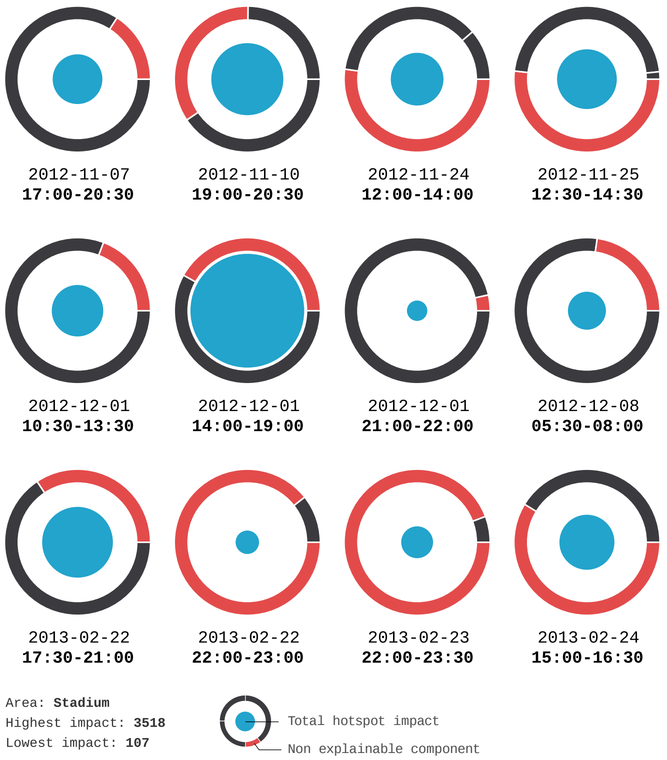


Fig. 4. 12 events from Stadium area.

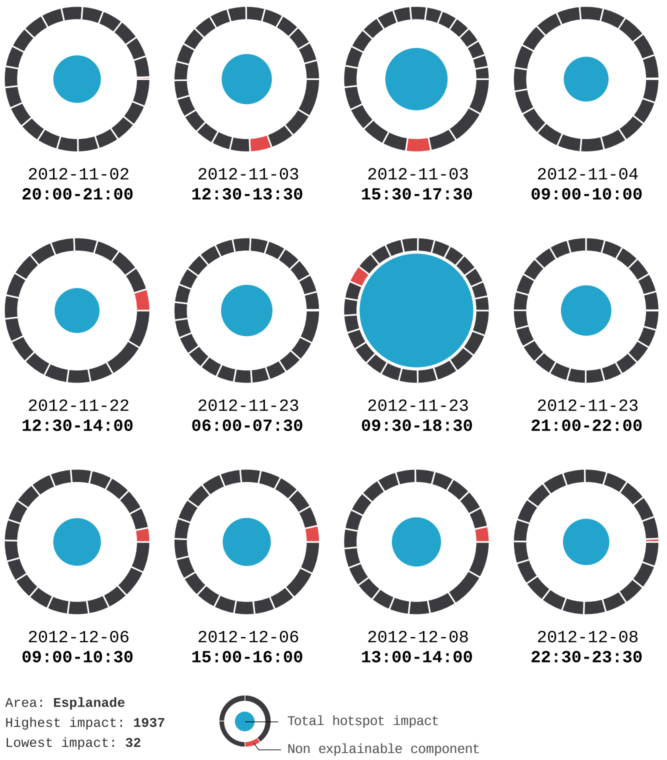


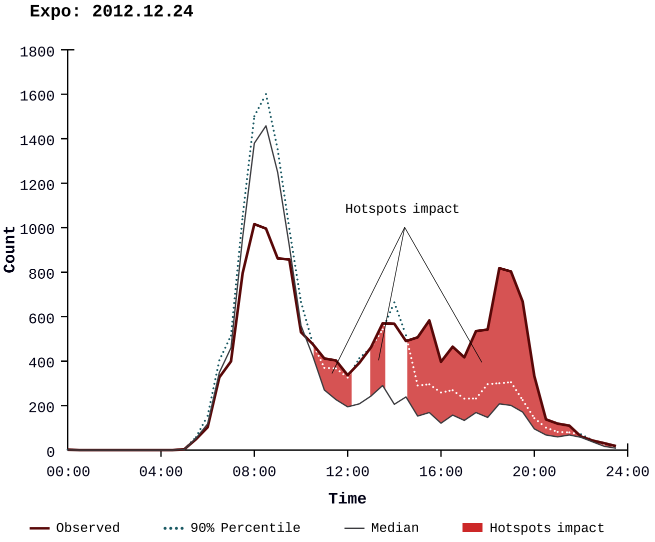
Fig. 5. 12 events from Esplanade area.

In Figure 7, we show the breakdown of a single hotspot, from 12:30 to 14:30 (the other 2 were filtered out due to small impact and duration). This was a tennis event, "Clash of Continents 2012", and most people arrived for the last final matches. The "Dance drama opera warriors" was held at 20:00 at the Kallang theatre. Intuitively, , we may expect an international sports event to attract more people than a classical music event. This is an example where the text description can play a role. If it were a pop concert (also music) and a local basketball game (also sports), the results could be drastically different.

Finally, Figure 8 represents again the most challenging case for our model, the Esplanade. Kalaa Utsavam is an Indian arts festival with multiple events that, when aggregated together, generate the largest impact. Intuitively, this is plausible given Singapores Indian population and culture. However, the results are very clear. For example, "Ten years shooting home" is a photography contest event that may not have elicited as many people as the "International Conference on business management and information systems". Regardless of this subjective analysis, a longer timeline and an improved data cleaning should increase the quality of this model.

VIII. Discussion

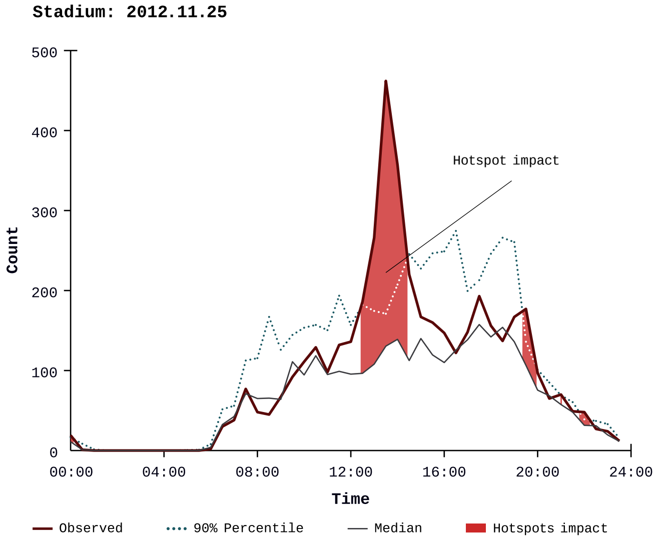
Our model was designed to explain hotspots that were already observed, but it can be used as a demand predictor as could be used in real data experiments. However, in order



Event impacts:

Time	**Megatex**	**kawin-kawin 2012**	**Non-explainable**
11:00—12:30	644	0	0
13:00—14:00	847	14	0
15:00—22:00	2159	1326	1032

Fig. 6. Impact breakdown for Expo 2012-12-24 (same as Figure 1).



Event impacts: 12:30h–14:30h

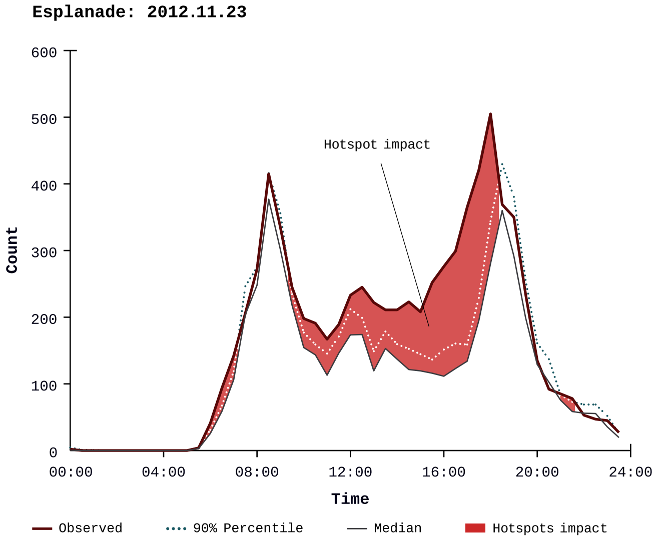
Quantity	Event title
476	Non-explainable
428	Clash of Continents 2012
0	Dance Drama Opera Warriors

Fig. 7. Impact breakdown for Stadium 2012-11-25.

to do so, the model would need to be extended with a few key features: a time series component to cope with seasonality effects; a habitual behavior component to account for baseline demand; and a spatial correlations component to deal with demand variations in other areas (e.g. competing simultaneous events). Each of these extensions deserve particular attention and are more data greedy (e.g. need for larger time window; information about school holidays, weather forecast) and changes to the model itself.

The current model is linear at both levels. It is simple to estimate yet contains all necessary components to prove our concept. However, the problem at hand lends itself to nonlinearities. For example, online popularity will hardly have a linear relationship with real demand (e.g. an artist with millions of likes/Google hits may not attract proportionally more people than another one with thousands). One of our next steps will be to extend the model with a Gaussian Processes component at the second level (individual impacts).

The definition and quantification of hotspots is also a component of our methodology. With negligible changes other than data availability, we can apply it to breakdown influence of events in trips by origin/destination, bus lines, different mode (e.g. taxi), or even go beyond the transport domain (e.g. cell-phone usage, food consumption, credit card usage, water, energy). Generally the model applies to any analysis of large crowds, aggregated both in time and space, under the assumption that these are partially caused by events announced on the web.



Event impacts: 9:30h–18:30h

Quantity	Event title
128	Non-explainable
697	Kalaa Utsavam 2012
247	Kuo Pao Kun: A Life of Practice (theatre)
238	SLO – Mother Daughter Wife and Lover (opera)
145	SRT Holiday Programme – Fantastic Mr Fox (theatre)
128	Travel and Make Money Seminar
123	The Romance of Magno Rubio (theatre)
116	Ten Years of Shooting Home (photography)
115	Int. Conf. on Business Management and Information Systems

Fig. 8. Impact breakdown for Esplanade 2012-11-23.

IX. Conclusions and future work

We presented a machine learning model that classifies aggregated crowd observations into explanatory components. We extract candidate explanations from the Internet and assume that aside from habitual behavior such as commuting, these crowds are often motivated by public events announced on the Web. Since we do not have individual event observations, we treat them as non-observed, or latent, variables.

This model has a two-layer structure, and each one is a sum of components. At the top level, we consider explainable and non-explainable components, and at the lower level, we disaggregate the explainable component into possible explanations retrieved from the Internet.

We tested this hierarchical additive model on a public transport dataset from the city-state of Singapore. We identified *overcrowding hotspots* by comparing the observed people counts (bus or subway arrivals) with a conservative threshold (90% quantile) at 30 minutes intervals. We quantified the hotspots by summing consecutive "excessive" counts. For each hotspot we retrieved the potential explanations from several event announcement websites and extracted relevant available information such as event title, category, venue, and description. We applied Latent Dirichlet Allocation (LDA) [1] to extract topics from the text descriptions.

All these features were organized together in our Bayesian hierarchical additive model, which was implemented on the Infer.NET framework [2]. Results with synthetic data show that the model retrieves the correct results with a correlation coefficient (CC) of at least 85% and a root relative squared error (RRSE) below 56%. Results with real data show that the same model recovers the observed total impacts with a CC between 41.2% and 83.9% and RRSE between 55% and 85%. A qualitative analysis on a case study in Singapore shows that the results of the hotspot impacts breakdowns into different possible explanation are intuitively plausible.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Land Transport Authority of Singapore for providing the data collected from the EZLink system. This work is supported in whole or in part by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Future Urban Mobility (FM).

Special thanks to Steve Robinson, Ajinkya Ghorpade and Youngsung Kim for their helpful reviews and comments. And finally to Katie Rosa, for her very thorough proof reading.

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[2] T. Minka, J. Winn, J. Guiver, and D. Knowles, "Infer.NET 2.5," 2012, microsoft Research Cambridge. http://research.microsoft.com/infernet.

[3] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, Jun. 2008. [Online]. Available: http://dx.doi.org/10.1038/nature06958

[4] B. Adams and D. D. Kapan, "Man bites mosquito: Understanding the contribution of human movement to vector-borne disease dynamics," *PLoS ONE*, vol. 4, no. 8, p. e6763, 08 2009.

[5] G. Xue, Z. Li, H. Zhu, and Y. Liu, "Traffic-known urban vehicular route prediction based on partial mobility patterns," in *Parallel and Distributed Systems (ICPADS), 2009 15th International Conference on*, 2009, pp. 369–375.

[6] C. Song, Z. Qu, N. Blumm, and A.-L. Barabsi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010. [Online]. Available: http://www.sciencemag.org/content/327/5968/1018.abstract

[7] S. Jiang, G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, and M. C. González, "A review of urban computing for mobile phone traces: current methods, challenges and opportunities," in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. ACM, 2013, p. 2.

[8] F. Potier, P. Bovy, and C. Liaudat, "Big events: planning, mobility management," in *European Transport Conference 2003. ETC Proceedings, Strasbourg, France.*, 2003.

[9] FHWA, "Planned special events: Checklists for practitioners," in *U.S. Department of Transportation, Federal Highway Administration, Office of Transportation Management*, 2006.

[10] F. Bolte, "Transport policy objectives: Traffic management as suitable tool," Federal Highway Research Institute (BASt), Bergisch-Gladbach, Germany, Tech. Rep., 2006.

[11] F. Middleham, "Dynamic traffic management," Ministry of Transport, Public Works, and Water Management, Directorate-General of Public Works and Water Management, AVV Transport Research Centre, Rotterdam, Netherlands,, Tech. Rep., 2006.

[12] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, "The geography of taste: analyzing cell-phone mobility and social events," in *Pervasive Computing*, ser. Lecture Notes in Computer Science, P. Floréen, A. Krüger, and M. Spasojevic, Eds., vol. 6030. Berlin, Heidelberg: Springer Berlin / Heidelberg, 2010, pp. 22–37.

[13] F. C. Pereira, F. Rodrigues, and M. Ben-Akiva, "Internet as a sensor: case study with special events," in *Transportation Research Board. Meeting (91st: 2012 : Washington, D.C.).*, 2012.

[14] L. A. Schweitzer, "How are we doing? opinion mining customer sentiment in u.s. transit agencies and airlines via twitter," in *Transportation Research Board. Meeting (91st: 2012 : Washington, D.C.).*, 2012.

[15] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012, pp. 1023–1031.

[16] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 186–194. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339561

[17] F. Rodrigues, A. Alves, E. Polisciuc, S. Jiang, J. Ferreira, and F. Pereira, "Estimating disaggregated employment size from points-of-interest and census data: From mining the web to model implementation and visualization," *International Journal on Advances in Intelligent Systems*, vol. 6, no. 1 and 2, pp. 41–52, 2013.

[18] Z. Yin, L. Cao, J. Han, C. Zhai, and T. Huang, "Geographical topic discovery and comparison," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 247–256.

[19] Y. Sun, H. Fan, M. Bakillah, and A. Zipf, "Road-based travel recommendation using geo-tagged images," *Computers, Environment and Urban Systems*, no. 0, pp. –, 2013.

[20] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta, Malta: ELRA, May 2010, pp. 45–50, http://is.muni.cz/publication/884893/en.

[21] S. W. Raudenbush, *Hierarchical linear models: Applications and data analysis methods*. Sage, 2002, vol. 1.

[22] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage." 1990.

[23] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[24] D. Kollar and N. Friedman, *Probabilistic graphical models: principles and techniques*. The MIT Press, 2009.

Francisco C. Pereira is Senior Research Scientist in Singapore-MIT Alliance for Research and Technology, Future Urban Mobility Integrated Research Group (FM/IRG), where he is working in real-time traffic prediction, behavior modeling, and advanced data collection technologies. In these projects, he applies his research on data mining and pattern recognition applied to unstructured contextual sources (e.g. web, news feeds, etc.) to extract information relevant to mobility. He is also a Professor in the University of Coimbra, Portugal from where he has a PhD degree, and a co-founder of the Ambient Intelligence Lab of that University. He led the development of several projects that apply these context mining principles, some with collaboration other MIT groups such as the Senseable City Lab, where he worked earlier as a postdoctoral fellow.

Filipe Rodrigues is a PhD student in Information Science and Technology at University of Coimbra, Portugal, where he is involved in two research projects whose main goal is to understand the effect of special events in urban mobility.

Currently, he is working on probabilistic models for learning from crowdsourced and noisy data. His research interests include Machine Learning, Probabilistic Graphical Models, Natural Language Processing and Intelligent Transportation Systems.

Evgheni Polisciuc is a PhD student in Information Science and Technology at University of Coimbra, Portugal. He specializes in information visualisation, in particular studies and develops graphical models to revel and explain anomalies in urban mobility and land use.

Currently he is working on estate of the art surveying and, in parallel, on visualisation models that highlight areas with high degree of deviations from the routine in urban mobility.

Moshe Ben-Akiva is the Edmund K. Turner Professor of Civil and Environmental Engineering at the Massachusetts Institute of Technology (MIT), and Director of the MIT Intelligent Transportation Systems (ITS) Lab. He holds a Ph.D. degree in Transportation Systems from MIT and 4 honorary degrees and coauthored the textbook Discrete Choice Analysis.