USING DATA FROM THE WEB TO PREDICT PUBLIC TRANSPORT ARRIVALS UNDER SPECIAL EVENTS SCENARIOS

FRANCISCO C. PEREIRA, FILIPE RODRIGUES, AND MOSHE BEN-AKIVA

Abstract. The Internet has become the preferred resource to announce, search and comment about social events such as concerts, sports games, parades, demonstrations, sales or any other public event that potentially gathers a large group of people. These *planned special events* often carry a potential disruptive impact to the transportation system, because they correspond to non-habitual behavior patterns that are hard to predict and plan for.

Except for very large and mega events (e.g. olympic games, football world cup), operators seldom apply special planning measures for two major reasons: the task of manually tracking which events are happening in large cities is labour-intensive; and, even with a list of events, their impact is hard to estimate, especially when more than one event happens simultaneously.

In this paper, we utilize the Internet as a resource for contextual information about special events and develop a model that predicts public transport arrivals in event areas. In order to demonstrate the feasibility of this solution for practitioners, we apply off-the-shelf techniques both for Internet data collection and for the prediction model development. We demonstrate the results with a case study from the city-state of Singapore using public transport tap-in/tap-out data and local event information obtained from the Internet.

Keywords: Urban computing, public transports, smartcard, web-mining, data mining

1. Introduction

During the past few years, the Internet has become a valuable source for mining information about mobility in the city. Institutional and private websites are complemented by numerous social platforms such as eventful.com, eventbrite.com, upcoming.org, Facebook, Wikipedia,

Key words and phrases. Demand prediction, special events, context mining, machine learning

Twitter, Gowalla, FourSquare, etc. Besides knowing what is happening now and what did happen in the past, we can understand what events will happen in the future. Moreover, through both individual (*crowd-sourced*) contributions and institutional services, the Internet can also provide timely information related to road works, sales, demonstrations, incidents, natural phenomena or the status of the road network.

On the other hand, technologies that generate high quality spatial and temporal data about the city, such as GPS, WiFi, Bluetooth, RFID and NFC, are becoming ubiquitous, to add even more to the already extensive ITS dataset (e.g. (Smith and Venkatanarayana, 2005)). We can find them for example in transit smartcards, toll collection systems, floating car data, fleet management systems, car counters, and personal devices. These mobility traces are increasingly available for practitioners and researchers and represent a reality that can be correlated with online information.

Disruptions due to special events are a well known challenge in transport planning because the transport system is typically designed for habitual demand, and only very large events are given special attention (e.g. olympic games, world cups). The challenge with events that are "only" medium or large is poor information availability: in any large city in the world, there are many such events per day but no systematic way of centralizing relevant information. Even when it exists, little can be done when no estimate of real demand is available.

Currently, the general practice is to rely on formal processes and manual work. For large (e.g. big concerts or football matches) and mega events (e.g. olympic games, football world cup), there are usually procedures in place where event organizers engage with operators and authorities to plan accordingly long time before the event. To capture relevant events where organizers do not inform transport entities, it is common to find, in traffic centers of large cities, teams of people periodically searching through the web and newspapers. This task is labour-intensive and, even with a list of events, their impact is hard to estimate, especially when more than one event happens simultaneously. A timely and accurate notion of demand impact is needed in order to design adequate system changes (e.g. shuttle buses, more buses, road blocking) and to disseminate appropriate information to the public.

In this paper, we focus on the prediction of public transport trips to special events' venue areas. More specifically, we develop a model that is able to predict arrivals with days in advance, thus allowing for enough planning time. Both for operators and for regulators, this can have significant value. Operators can use such information to increase/decrease supply based on expected demand. Regulators can

raise awareness to operators and users on potential non-habitual overcrowding. Furthermore, regulators can also use this research to understand past overcrowding situations (e.g. telling circumstantial from recurrent overcrowding). Notice, however, that we focus on arrivals instead of full Origin/Destination tables. As will be explained later, this is due to data availability constraints. Nevertheless, the overall methodology is extensible for such task.

We introduce a methodology for prediction of public transport arrivals under special events scenarios that extracts events information from the Internet and matches such information with bus and subway tap-in/tap-out data from the city-state of Singapore (the EZLink system). We demonstrate the value of extracted event features for explaining variance under these scenarios and the role of spatial and temporal constraints for such predictive models.

After this introduction, we will review the current literature (Section 2). Then, we describe our datasets and case study (Section 3) and present our model (Section 4). The paper ends with a discussion and conclusions (Sections 5 and 6, respectively).

2. Literature review

From the point of view of transportation planning, recent research in demand modeling for special events, using the traditional 4-step model approach, has been done by Copperman et al (Copperman et al., 2011). They used a questionnaire-based survey at the venue gates and obtained a sample size of nearly 6000 individuals for 20 events in the city of Phoenix, Arizona. The data collected is used to predict, for each event, various indicators: the number of trips by transport mode; trip time-of-day; trip OD; mode; vehicle miles travelled; and transit boardings generated due to the events. This is, to our knowledge, the most comprehensive study of this type. Although they present a very detailed and well-founded model, it is highly dependent on individual participation and it does not explicitly consider events characteristics.

Current transport planning practice gives detailed attention mostly to *mega events* (e.g. olympic games, formula 1, etc.), which involve a large effort from public and private institutions. These events gather huge crowds and flows of people, they are well defined in time and space, and it is often possible to know ahead of time the demand distributions, given for example the accommodation availability (Potier et al., 2003; Coutroubas et al., 2003; Vougioukas et al., 2008). *Large events* are much smaller in terms of audience and duration (e.g. music concerts, sports games, political rallies) but are also more common and

delicate to manage. In practice, the majority of large events does not receive any special treatment or attention in terms of transport planning, which often creates non-recurring congestion and overcrowding because patterns of attendance may not be very clear. It is recognized that such patterns are inevitably more difficult to forecast than those related to daily mobility, particularly in the case of open-gate events (Potier et al., 2003).

In face of these constraints, authorities tend to rely on trial and error experience from recurring events, checklists (FHWA, 2006), and adopt a reactive approach rather than a planned one. For large events, it is clear that the state of the practice is quite advanced in terms of control and actuation techniques. However, demand forecast beyond the immediate vicinity of the event is still very rudimentary, even though it is considered an essential pillar of travel management for special events (FHWA, 2006).

Medium and small events represent the largest portion of special events, but their impact on mobility is extremely hard to measure. Although medium and small events are often negligible, the aggregate effect of simultaneous occurrences of these events can become relevant. Moreover, it is extremely difficult to predict or understand this effect. Information is spread over several sources that are often unrelated, and external effects such as season, holiday, or weather can affect the understanding of the scenario. Consequently, it is not surprising that we find no impact study of medium and small events in current literature, despite its intuitive importance.

An example of innovative approaches for handling special events is the study of Ahas, Kuusik and Tiru (Ahas et al., 2009) about tourism loyalty in Estonia. They show that the sampling and analysis of passive mobile positioning data is a promising resource for tourism research and management because this type of aggregated data is highly correlated with accommodation statistics in urban touristic areas. On the same direction, Calabrese et al (Calabrese et al., 2010) analyze a dataset of over 1 million cellphone users from the Massachusetts area and demonstrated a strong correlation between attendants' home location distribution and event type, thus implying that the *taste of neighborhoods* persists throughout different events of similar types. In a case study in Tawaf during the Hajj, Koshak and Fouda (Koshak and Fouda, 2008) demonstrated that GPS and GIS data could be utilized to perform spatial-temporal analysis of human walking behavior in an architectural or urban open space. Also using GPS data, Guande et al (Qi et al., 2011) analyse the city social profile from a massive dataset

of taxi trips, concluding that the temporal variation of amounts of get-on/off can characterize the social function of a region.

We note that the Internet as a data source for transport modeling is not a novelty. From the simple extraction of public information, such as bus lines, to using Twitter for opinion mining (e.g. (Schweitzer, 2012)), researchers and practitioners have been aware of its value, but its use in transport prediction is poorly explored despite earlier motivating articles, such as (Horvitz et al., 2005; Terpstra et al., 2004), that did little more than raising awareness to the topic. Terpstra et al. (Terpstra et al., 2004) propose a model that includes special events information with weather information and traffic data. They develop a data assimilation technique that applies the principles of extended kalman filters (EKF) and systematically adjusts future predictions by using the difference between current observation and expected current status. However, their events information is limited to location and time.

We believe the Internet can become a tool for transportation modelers and planners at two levels: first, the identification of relevant special events; and second, estimation of demand fluctuations, given discovered events features. Our approach is different because we build a prediction model that includes features easily obtainable via Application Programming Interfaces (APIs) or screen scrapping from online resources. Beyond name, location and time, information such as event category, price, age range, and descriptive text are commonly available. Our model combines a classical machine learning model of transport data with publicly accessible online data and demonstrates an increase in the quality of the predictions, thus uncovering opportunities for transport researchers and practitioners, which often find such an approach as too complex or intimidating.

3. Data description and preparation

The context of this work is the city-state of Singapore, which has a resident population of 5 million people plus roughly 1 million of temporary residents. It has one of the most comprehensive and efficient public transport systems in the world, together with a vibrant environment with many special events, particularly in the center-south and east coast regions. It is also a home to a melting pot of cultures, both from Asia and from the rest of the world, making it an international hub for touristic events, shopping, culture and finance.

Singapore is also an interesting case in terms of Internet usage, with over 82% of the households having Internet access, and over 5.4 million

running 3G subscriptions. There are many events' websites, but only a few have a wide coverage from the very small to the very large events, namely eventful.org, sistic.sg and whatshappening.sg. Other websites with interesting information are eventbrite.com for small and medium alternative events, and yoursingapore.com for very large events.

The time window of the study comprehends 16 days, during February, April and May, 2011. We work with two major datasets: events and public transport. The former is obtained automatically with our own crawling tool that uses available APIs from websites and screen scrapping techniques while the latter is provided by Singapore's Land Transport Authority (LTA) and corresponds to public transport tap in/tap out information from their EZLink system.

3.1. **Events.** For the selection of venues, we considered the following constraints: each venue should have capacity for at least 1000 attendants; any two venues should not share the same bus or subway stops, in order to differentiate demand (otherwise, we remove both venues from the set); they should be served by bus or subway in a radius of at least 800 meters. From an experimental design standpoint, these constraints minimize interactions between events that would turn out difficult to discern. For example, observations from very small events could be affected by other demand attractors, multiple events in the same area could lead to overweighting of individual events (however, there is no reason not to relax these constraints as a next step, provided the availability of a larger dataset). The resulting list has 5 venues:

(1) Singapore Indoor Stadium - Regularly hosts large events, has its own subway station and bus stops;
(2) Singapore Expo - Frequently hosts festivals and fairs, located in the east side of Singapore, has its own subway station;
(3) Singapore Botanic Gardens - Holds regular weekend outdoors events, at a large lawn area, and is a favorite for families and visitors;
(4) Zouk - The largest disco complex in the country, ranked top 10 in the world, a common destination for locals and visitors. Holds special events and performers and is located in the nightlife area of town;
(5) Marina Promenade - Located on the Singapore Grand Prix F1 tracks in the Marina Bay area, near the Singapore Flyer, it is one of the most popular touristic areas of Singapore and hosts many music and art performances.

From the mentioned online resources, eventful.com provides the most comprehensive set of events in Singapore. In Figure 1 and Table 1, we

show an example of an advertised event and its data, as obtained from the API. Notice that it also contains links to other pages, themselves accessible via APIs.

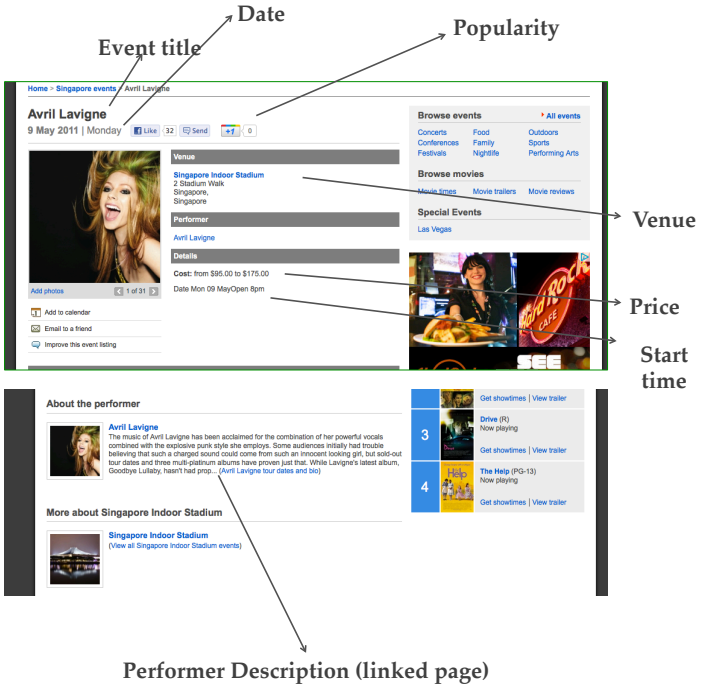

FIGURE 1. Example of an event page in eventful.com.

TABLE 1. Eventful API data example

Field	Content
unique id	E0-001-037869894-7
event title	Avril Lavigne
text description	"Cost: from $95.00 to $175.00. Date Mon 09 May. Open 8pm."
start time	May 9, 2011 08:00 pm — Monday
end time	N/A
location	2 Stadium Walk, Singapore, Singapore (lat=1.30108, lng=103.87411)
venue name	Singapore Indoor Stadium
performer name	Avril Lavigne
eventful link	http://eventful.com/performers/avril-lavigne-/P0-001-000000439-7
ticket price	from $95.00 to $175.00
links to related web content	http://www.timeoutsingapore.com/music/concerts/avril-lavigne
comments	N/A
text tags associated with the event	N/A
list of categories this event is part of	Performing Arts

For the 5 venues, and the 16 days time window available, we extracted 59 events. A few venues have more than one event in the same day and all venues have at least one non-event day.

TABLE 2. Events dataset details

Venue	No. of events	Event days	Non-event days	Nearby buses	Nearby sub-ways	POIs in 1km range	Observations
Singapore Botanic Gardens	6	5	11	2	0	442	Leisure/Touristic, Large space
Singapore Expo	34	15	1	3	1	195	Shopping/Exhibits, isolated
Singapore Indoor Stadium	2	2	14	2	1	96	Leisure, isolated
Marina Promenade	9	2	14	1	1	676	Touristic
Zouk	8	7	9	3	0	1388	Nightlife

Table 2 shows more details on the events dataset. The "POIs in 1km range" column refers to the number of points of interest within 1 kilometer radius, according to yellowpages.com.sg. This is important since all venues share their accessibilities with other nearby attraction points. We selected the bus and subway stops according to the following criteria:

- **Bus**: Consider all bus stops within 250 meters radius of the venue. If no stops are found, extend the radius to 500 meters (this exception was only applied for the Singapore Botanic Gardens).
- **Subway**: Consider all subway stations within 500 meters radius of the venue.

In Figure 2, we present snapshots of each area. The extracted events are assigned to 13 eventful.com categories, as presented in Table 3. Notice that many events are assigned to more than one category.

3.2. **EZLink.** EZLink is the RFID based smartcard system for transit that is implemented in Singapore. It works on a tap-in/tap-out basis, that is passengers tap-in when they board and are discounted the correct fare only when they tap-out and alight. Not taping out implies the payment of the maximum possible cost of the ride. EZLink thus generates full boarding and alighting data for all public transport system in Singapore. It comprises bus, light rail (LRT) and subway (MRT), with 4581 bus stops serving 318 bus routes, 33 LRT stations for 4 lines, and 73 MRT stations for 4 lines with a total extension of 130 Km. If a passenger alights a vehicle and boards another in a different mode or

FIGURE 2. Maps of the 5 venues studied

line within 45 minutes, these subsequent legs are counted as a single trip, thus paying a lower fare. Up to 5 such transfers can count as a single trip.

Our dataset corresponds to the following days: 22 February; 11-17, 23 and 28 of April; and 9, 17, 18 and 26-28 of May. For these days,

TABLE 3. Number of events per category

Category	Nr. events
Business & Networking	3
Concerts & Tour Dates	4
Conferences & Tradeshows	8
Festivals	3
Food & Wine	2
Neighborhood	5
Nightlife & Singles	1
Other & Miscellaneous	16
Outdoors & Recreation	2
Performing Arts	2
Sales & Retail	15
Science	1
Technology	2

we only have records for trips started after 2 p.m. The total number of trip legs recorded is over 63 million while the number of trips is 47.5 million.

For each venue, we count the trip arrivals (i.e. the end of the final leg) that occur at bus or subway stops selected as explained in section 3.1. We aggregated such arrivals into 30 minutes intervals. The choice of an half-hour over 5, 15 and 60 minutes was based on three considerations: the peak travel periods identified according to LTA's Household Interview Travel Survey (HITS 2008) have durations of either 30 min or 1 hour; according to the same survey, the mean travel time across all modes in Singapore is approximately 30 minutes; the planned bus headway scheme consists of 30 minutes or less between two subsequent buses of the same line.

4. Arrivals prediction model

Our prediction model should demonstrate the *hypothesis* that contextual information is significant for the purposes of public transport arrivals prediction in the vicinity of special event venues. Generalizing to other cases, and upon available transport data, this should be valid to any area in which demand can be somehow associated with available contextual information, as for example school areas and information about school holidays, shopping areas and large sales, and main governmental buildings and public demonstrations. Occasionally, and in small numbers, such events can be dealt with manually by practitioners, but in large numbers and high frequency, automated methodologies become necessary.

4.1. **Preliminary analysis.** To have a first insight on our dataset, we compared venues in terms of descriptive statistics, specifically mean and variance of arrivals. In areas with regular high number of arrivals, only events with high participation originate significant impact, whereas in areas with high variance, event information may add insufficient explanatory power. For each venue and each half-hour, we calculated statistics on maximum number of arrivals, mean standard deviation, maximum standard deviation, and the signal-to-noise ratio (SNR). The latter corresponds to the mean divided by the standard deviation. Typically, values below 5 indicate that the signal has high variance (i.e. high noise) although this heuristic, called the *Rose criterion*, is mostly used in image recognition. Table 4 shows the results obtained.

TABLE 4. Preliminary Analysis Results

Venue	Max. No. Of Arrivals	Mean Std. Dev.	Max. Std. Dev.	Mean SNR
Singapore Botanic Gardens	91.88	16.45	35.65	2.51
Singapore Expo	396.38	278.86	561.38	2.88
Singapore Indoor Stadium	247.13	45.11	216.33	1.05
Marina Promenade	314.50	68.45	134.75	2.89
Zouk	85.56	17.97	35.75	3.33

Notice that we are only counting the final destinations, so the actual number of people using public transport in these areas is higher if we also count with transfers. Also, isolated areas, such as the Singapore Indoor Stadium and the Singapore Expo have high variance, probably because events do cause impact to an otherwise quiet area. This is confirmed by their low SNR values. Notice that, for the Expo, a venue with events in 15 out of 16 days, the SNR is still low. In fact, from our sample data we cannot infer if it is quiet in non-event days, but different event days originate different impacts, so detailed features such as event categories may help explain such variance.

4.2. **Model selection.** For the model selection and implementation, we explored the common portfolio of already well-established algorithms available in the Weka platform (Hall et al., 2009), which can both demonstrate our hypothesis and be easily replicable by practitioners. With average Java or Python programming skills and common

expertise in web service and open source APIs, our methodology can be implemented by others in a short time.

Our model aims to predict the number of arrivals to a given venue area, occurring at a given half-hour interval of a given day. In other words, for each venue, such model will generate a half hour resolution time series. In order to better understand the value of Internet information, we built two models (Table 5): a *basic* one, with no online information; and a model with *all features* extracted from the Internet.

TABLE 5. Model configurations

Model configuration	features
basic	venue, day of week, half hour and arrivals count
all features	*time to next event* features + event categories

More specifically, the basic model has only spatial (venue) and temporal (half-hour, day of week) information; the *all features* model adds information on category frequencies for each venue as well as the time to next event. Regarding category frequencies, for example, if there are two events in the same area, one with "Performing Arts" and "Literary and books", the other with "Performing Arts" and "Concerts and tour dates", the "Performance Arts" category would have value of 2 while the others would have 1. Regarding the time to next event, we created 3 distinct features: one nominal, called stage, with a set of values determined by us (*NoEvent*, *PreEvent* - 2 hours before the event, *EventStarting* - 30 minutes before and after start time, *AllDay* - when it's a whole day event with no explicit start time); an integer, indicating the number of half-hours before the event starts; and a real value, corresponding to the inverse of the latter. We also added a dummy variable that identifies whether there was or not on that day at all. We provide these apparently redundant features to let the models pick the representation that is the best predictor.

We have the following additional considerations:

- Although the dataset is very large and comprehensive in terms of country and population coverage, it is not extensive in terms of number of days;
- The total number of vectors per venue is 320 (20 half-hours $\times$ 16 days);
- The feature sets to consider are heterogeneous (with nominal, binary and numeric data types);
- The task to solve is essentially of regression (although it is feasible to discretize the arrival counts into bins for classification);

- To exploit the variety of events and venues and maximize cross-learning between them, we build a single model for all venues.

Following these considerations, our experimental design uses the classical n-fold cross-validation methodology, where we split data into n folds. We repeatedly leave one fold out for testing and use the remaining $n-1$ folds for training. We repeat such procedure n times and the final evaluation takes into account the whole performance of the n folds. There is, however, an important aspect to notice. Our training samples are not purely independent and identically distributed (i.i.d.). In fact, on a within-day basis, they form a time series with the associated autocorrelation parameters. For this reason, we split the n *folds* such that no sample falls within the same day/event as any other sample from another fold. In practice this becomes a 16-fold cross validation where each fold corresponds to an entire day of data. In this way, we can replicate the realistic case where the system is predicting for some event in advance and has a set of past events in the database. Furthermore, this methodology maximizes the use of the training set, which is important given the small size of the dataset.

We seek predictions that can be made far ahead in time (several days as opposed to only a few hours before), so typical time series analyses methods such as ARIMA would add very little value since we can not count with the trends during the event day. Furthermore, our dataset is too limited in size to capture seasonality patterns.

We did a preliminary test with several algorithms, namely neural networks, K-nearest neighbour, Gaussian processes, radial basis functions, linear regression, regression trees, and support vector regression. We tested using 16-fold cross-validation and Weka default parameters for each model. In table 6, we summarize the results. We compared the different algorithms using Pearson's correlation coefficient (CC), mean absolute error (MAE) and root mean squared error (RMSE), computed as follows:

$$\text{CC}(\hat{x}) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(\hat{x}_i - \bar{\hat{x}})^2}} \tag{4.1}$$

$$\text{MAE}(\hat{x}) = \frac{1}{N}\sum_{i=1}^{N}|x_i - \hat{x}_i| \tag{4.2}$$

$$\text{RMSE}(\hat{x}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \tag{4.3}$$

(4.4)

where N denotes the number of instances in the dataset, $\hat{x}_i$ is the predicted arrivals count for the i^{th} instance, x_i is the corresponding true arrivals count, and $\bar{\hat{x}}$ and $\bar{x}$ are the means of predicted and observed counts, respectively. Table 7 shows the results.

TABLE 6. Comparing several machine learning models.

	All Features			**Basic**		
	CC	MAE	RMSE	CC	MAE	RMSE
RBF	0.69	53.88	143.85	0.34	69.73	193.92
K-Nearest Neighbour	0.55	53.82	173.88	0.34	67.78	194.88
Neural Network	**0.85**	**45.9**	**105.42**	**0.38**	**71.24**	**185.6**
Gaussian Processes	0.62	75.61	161.91	0.35	83.13	184.96
Support Vector Regression	0.54	74.77	189.59	0.35	63.93	188.5
Regression trees	0.56	58.58	174.21	0.38	72.54	181.71
Linear Regression	0.61	77.06	167.62	0.34	83.62	185.27

A classical multi-layered perceptron neural network consistently outperformed the others for this dataset, so we will henceforth assume this technique in the paper.

4.3. **Model architecture.** Our neural network uses a typical architecture, as shown in Figure 3: one hidden layer, sigmoid activation functions, back-propagation learning, and regression output. The input layer has the following neuron set: 5 binary neurons for venues; 20 binary for half-hour identifiers (from 2pm to midnight); 1 binary for weekend/weekday; 13 numeric for event categories; 1 numeric for number of half-hours to next event (*halfhours_to_next_event*); 1 numeric for inversed half-hours to next

event (*inv_halfhours=1/halfhours_to_next_event*); and 5 binary identifiers that indicate time to next event (*NoEvent*, *PreEvent*, *EventStarting*, *EventStarted* and *AllDay*). This configuration will change depending on the model, for example the basic model has only the first three sets of neurons.

Regarding the hidden (sigmoid) layer, we determined the number of neurons by adding the number of attributes to number of classes and dividing by two. In a regression model, this number is heuristically determined by the discretization of outputs into intervals, in which each relevant sized interval is considered a "class" for this estimate. Therefore, in our experiments, this consisted of 16 and 33 neurons, for the above mentioned models, respectively. The output layer consists of one neuron that transmits the weighted linear sum of hidden layer activations.

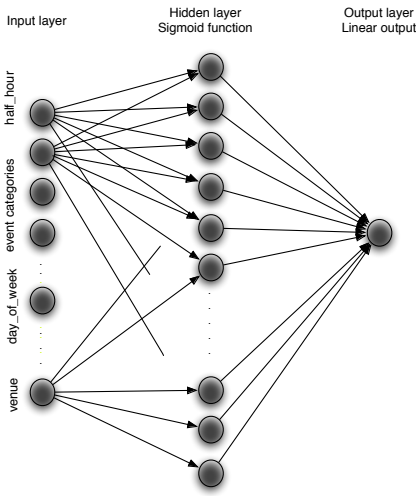

FIGURE 3. Artificial neural network architecture used.

4.4. **Results.** We now show the results for the neural network model globally and per venue, according to the two configurations mentioned in the previous section. We use four indicators: the correlation coefficient of predicted and actual arrivals; the mean absolute error (MAE); the root mean squared error (RMSE) of the arrivals; and the root mean squared error normalized (RMSN). The latter is computed as follows:

$$\text{RMSN}(\hat{x}) = \frac{\text{RMSE}(\hat{x})}{(\sum_{i=1}^{N} x_i)/N} \tag{4.5}$$

where N denotes the number of instances in the dataset, $\hat{x}_i$ is the predicted arrivals count for the i^{th} instance, and x_i is the corresponding true arrivals count. Table 7 shows the results.

TABLE 7. Results by venue.

	All Features				**Basic**			
	CC	MAE	RMSE	RMSN	CC	MAE	RMSE	RMSN
Singapore Expo	**0.84**	**119.47**	**210.59**	**0.09**	0.15	198.464	391.92	0.17
Marina Promenade	0.80	51.91	88.59	0.18	**0.81**	**49.32**	**77.09**	**0.16**
Zouk	**0.57**	**16.79**	**22.73**	**0.18**	0.20	36.78	44.99	0.36
Singapore Indoor Stadium	**0.90**	**26.47**	**48.61**	**0.06**	0.37	48.84	97.91	0.12
Singapore Botanic Gardens	**0.73**	**14.82**	**22.16**	**0.13**	0.42	22.79	37.79	0.19

In Figures 4 and 5, we also show the variation of MAE and RMSN through time for the five venues. For each case, we calculate MAE and RMSN obtained by half-hour for the entire dataset. We let the Y scale differ in each plot to clarify the performance of the model relative to the habitual number of arrivals of the area. For transport planning, the same value of error in Zouk, a less busy area in terms of arrivals, would raise more concerns than in Singapore Expo, which is a busier area.

With respect to the mean absolute error (MAE), we can see a clear advantage of the model with Internet information in 4 out of the 5 venues. The exception is the Marina Promenade, which by itself should be sufficiently isolated to allow the model to learn well. However, its bus and MRT stops can also serve other areas, namely other tourist attractions (e.g. Esplanade, Singapore Flyer, river walk) and a busy shopping area, the Suntec City. Interestingly, it leads to a case where less information allows for better generalization, which is a clear sign of over fitting of our model in this area. Besides training with a more balanced dataset, one could also add a regularization term or a bayesian prior to the model parameters. This would demand a different type of model, not directly available in Weka and falls outside the scope of this document.

We can also find a few patterns in the observed errors. First, at the end of the day, the error considerably decreases for all cases, which is not surprising because the variance will decay considerably as the day ends (and less trips happen). Another aspect are the spikes that can be seen roughly around the starting time of the events. The *all features* model behaves more robustly but it over-reacts in Marina Promenade, possibly induced by other similar cases (e.g. Singapore Indoor Stadium) with much more public. This indicates that some missing

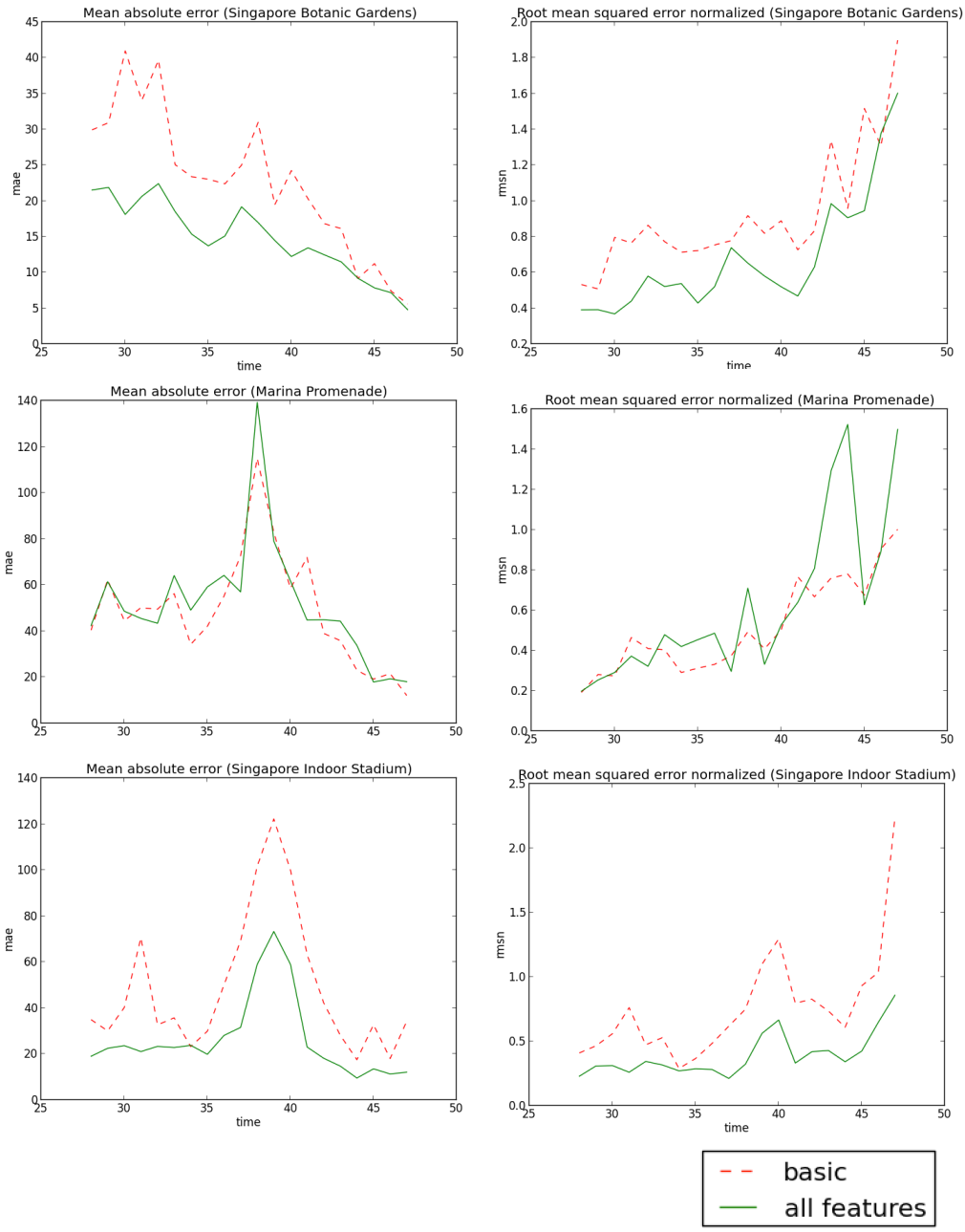

FIGURE 4. Mean Absolute Error (MAE) and Root mean square error normalized (RMSN) by half-hour from 2pm to midnight.

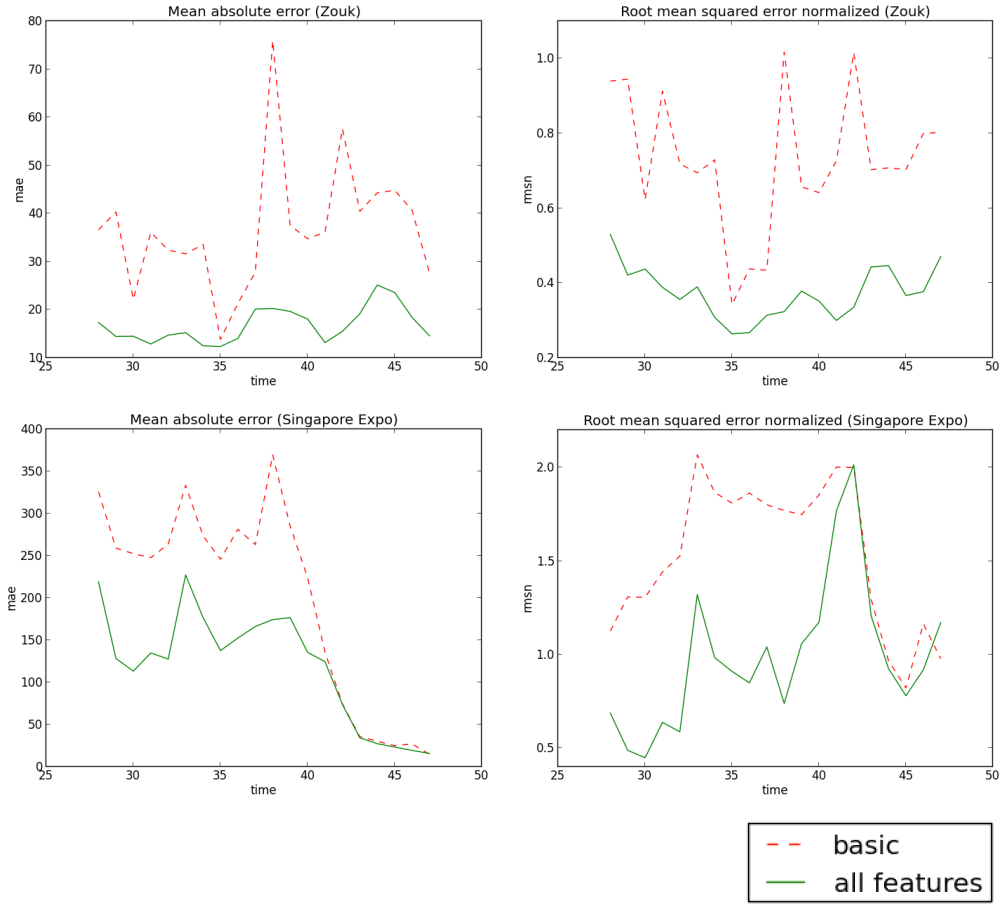

FIGURE 5. Mean Absolute Error (MAE) and Root mean square error normalized (RMSN) by half-hour from 2pm to midnight (cont.).

information, not always easily retrievable from the Web, such as venue size and price might have an important influence here.

Another persistent spike in the errors happens during peak-hour period, affecting both models, although the *all features* behaves more robustly in this regard.

The RMSN provides a different perspective since it normalizes the error with respect to the average of the observed values. Here too, the model with *all features* generally outperforms the *basic* one. In some cases, however, one could say that neither model is totally reliable,

particularly whenever the RMSN reaches values as high as 0.8 for example, as in Marina Promenade and Singapore Expo. We remind that the latter area has very high variance, as shown in Table 4. A longer time series and more events should be sufficient to improve the RMSN.

To further compare the two models, we also plot the *observed vs predicted* values in Figure 6. The closer the points are to the 45 degree line, the more accurate the model is. For easier comparison, we also added the linear centerline for each model. As this line approaches 45 degrees, the model's average error approaches zero. We can visually confirm that the correlation coefficient (observed before in Table 6) is much higher for the model with Internet information.

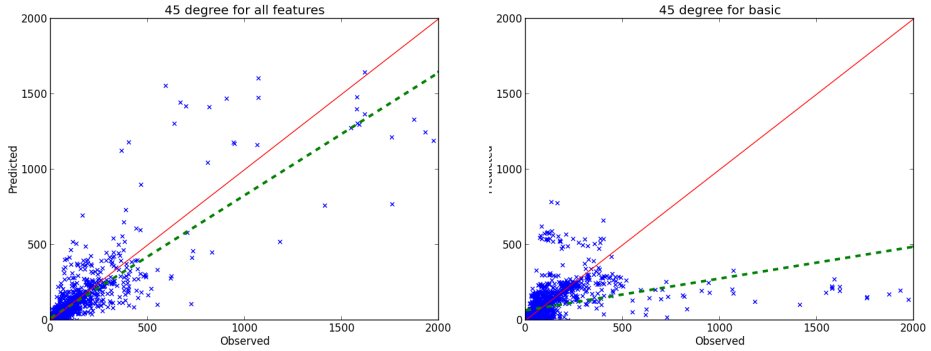

FIGURE 6. Plotting *observed vs predicted* in a 45 degree plot. Points that fall in the 45 degree line are fully accurate predictions. The dashed line represents the linear fit.

Figure 7 complements this analysis by showing the comparison of the ground truth (blue line indicating the true number of arrivals) with the models with basic and all features for 6 events. We notice that the basic model learns the arrival values by half-hour resolution per venue. The model with context is also affected by the same up and down trends acquired by the basic one, but it adds corrections learned from the context.

As expected, the model with *all features* outperforms the other, but there is still potential for improvement. The error ranges from about 0 to 800 arrivals in 30 minutes in the worst case, and, in terms of public transport management, this difference can represent a relevant overcrowding problem, depending on the supply capacity. On the other

hand, the basic model performance is much worse in these cases, with errors above 2000 arrivals.

Finally, Figure 8 presents a global comparison of both models with ground truth, averaged by half-hour for all venues and all days. -

5. Discussion

Our results show that there are spatial and temporal characteristics, associated with the venues, that play a crucial role in our model. When the venue is isolated, even with a small dataset, it is possible to capture the role of events up to some extent. Notice that we have two very isolated venues: Singapore Expo and Singapore Indoor Stadium. The former has events in all days except one, while the latter is the opposite, with only two events in the whole period. However, their models show comparable performance.

On the other hand, the model seems to be sensitive to venues that are not isolated, when other activities potentially occur simultaneously, and where other facilities also attract transport demand. Although this is a limitation of the model proposed here, there is no reason not to overcome it with a model that assumes multiple attractors. An obvious answer would be a mixture model where the arrivals at certain moment would correspond to the *sum* of the individual events. To calibrate it, one would need a dataset with enough such cases where different events interact in neighbor areas.

To gain further insight into our model, we ran a linear regression over the entire dataset. By doing this, we obtain the coefficients of each feature as well as their p-value. The coefficients provide a notion of the relative importance of each feature with respect to predicting the target variable, the number of arrivals. The p-value is an indicator of statistical significance of the coefficients found. Table 8 and Figure 9 present the results. For each feature, we also briefly present its descriptive statistics before normalization (min, max, mean and standard deviation). The overall correlation coefficient for this linear regression model was 0.8, while the MAE was 64.1 and RMSE 115.9. We can further check the goodness of fit for linear regression with R^2, which is 0.65. Notice that this is a fit to the *entire* dataset (as opposed to a test set evaluation) and nevertheless our best model effectively outperforms this one. Unfortunately, the weights of a neural network are much harder to interpret than those from a linear regression.

After normalizing the input vector, the algorithm used applies a greedy feature selection mechanism that eliminates collinear features, which explains why some features (e.g. weekend=FALSE,

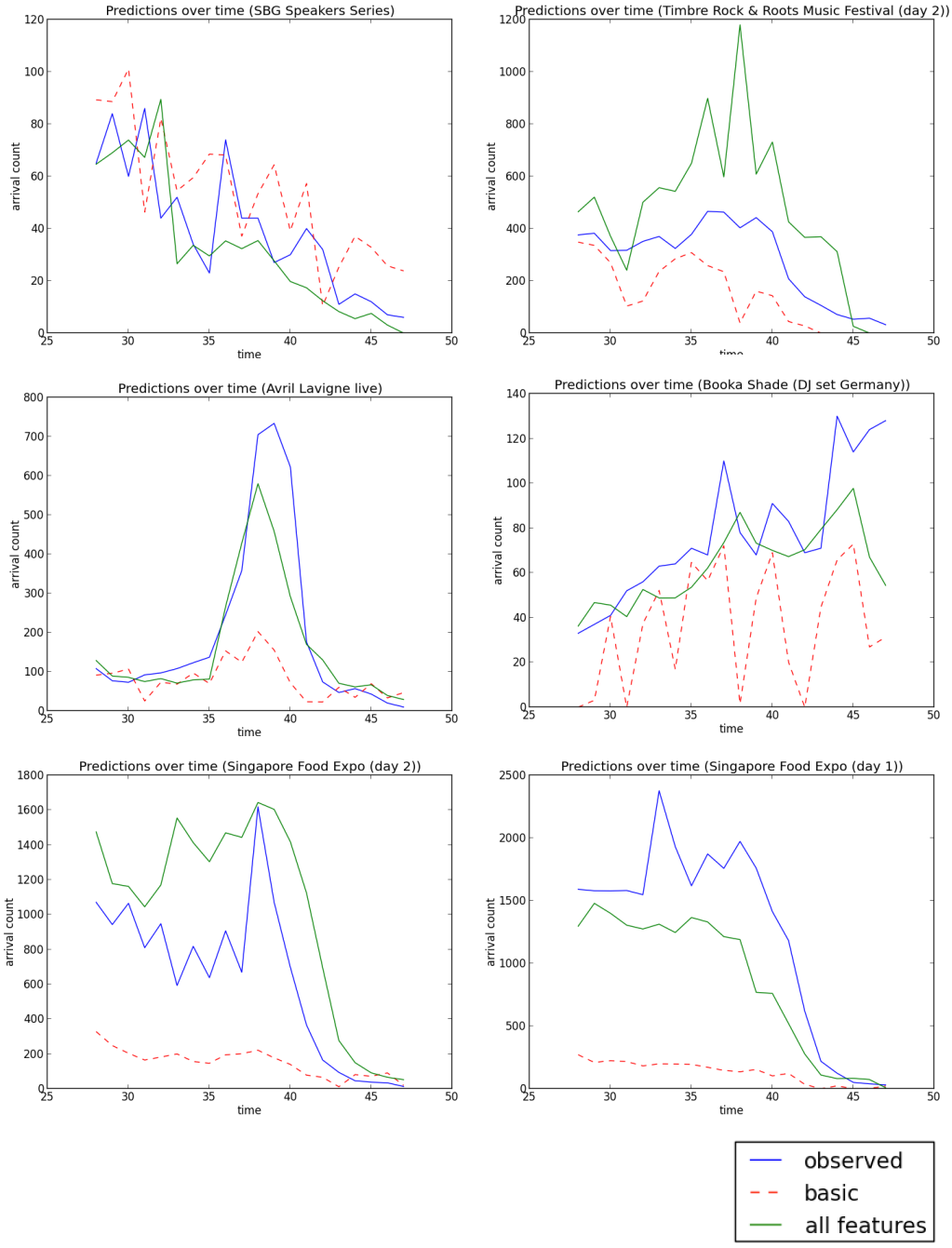

FIGURE 7. Comparison with the ground truth for 6 different event/days.

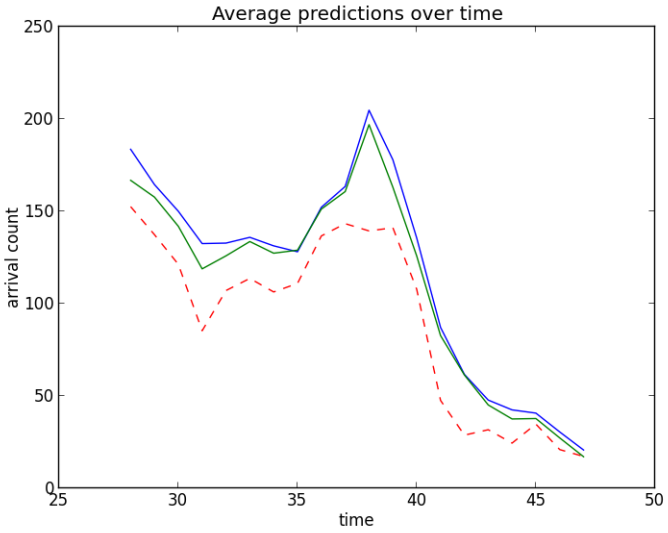

FIGURE 8. Averages comparison of ground truth vs basic vs all features.

venue=Singapore Indoor Stadium, time_tag) are absent in Table 8. They were not found to be relevant for the model.

We note that the simple location and time attributes (e.g. venue and half-hour) have individually less influence than those based on event information (e.g. event categories and inv_halfhour). It is also interesting to notice the extreme values. *Business/networking* events seem to drive much less attendance than any other, while with *food/wine* the opposite happens. A hypothetical explanation might be that the public is very different in mode share (attendants of business events use less public transport) and dimension (food/wine events may be directed to a larger audience). Our *inv_halfhour* feature seems to be more relevant than number of *half-hours_to_next_event*. We remind that the latter is simply the inverse of the former. The temporal stage feature is only relevant for distinguishing days with and without events, and other values (e.g. EventStarting) seem not to help much.

Another limitation in this work is lack of ground truth in terms of bus stop location choice by the public. Our distance-based rules mentioned in section 3.1 may need to be replaced by a case-by-case selection, with the help of experts from transport authorities.

In terms of further explaining the variance, one can also explore other data, available on the web. A particularly appealing concept is *online popularity*, which could be based on number of Google hits of

TABLE 8. Linear regression coefficients, descriptive statistics and p-values (codes=*$p < 0.1$; **$p < 0.05$; ***$p < 0.001$; ****$p << 0.001$). Sample size=1600 vectors; $R^2 = 0.65$. Definition of features: **venue** is the event location; **weekend** is a dummy variable to distinguish weekday/weekend; **halfour** is the time tag for the vector, within the day; **event** is a binary feature, with the values {NoEvent, Event}; **Performance Arts**...**Nightlife/Singles** are the category counters (nr. events tagged by category in a half-hour); **halfhours_to_next_event** is the number of half-hours till the next event and **inv_halfhour** is its inverse.

attribute	type	mean (std)	coeff.	std.error	p-value	obs
venue = Singapore Botanic Gardens	Binary	0.2(0.4)	-29.17	8.91	0.001	***
venue = Marina Promenade	Binary	0.2(0.4)	75.24	9.33	0.000	****
venue = Zouk	Binary	0.2(0.4)	-32.49	9.69	0.001	****
weekend = TRUE	Binary	0.25(0.4)	78.11	7.78	0.000	****
halfhour = h28	Binary	0.05(0.2)	48.81	13.60	0.000	****
halfhour = h29	Binary	0.05(0.2)	30.10	13.60	0.031	**
halfhour = h37	Binary	0.05(0.2)	23.36	13.56	0.104	
halfhour = h38	Binary	0.05(0.2)	65.93	13.53	0.000	****
halfhour = h39	Binary	0.05(0.2)	40.77	13.53	0.003	***
halfhour = h41	Binary	0.05(0.2)	-55.77	13.52	0.000	****
halfhour = h42	Binary	0.05(0.2)	-80.78	13.51	0.000	****
halfhour = h43	Binary	0.05(0.2)	-72.06	13.70	0.000	****
halfhour = h44	Binary	0.05(0.2)	-83.32	13.93	0.000	****
halfhour = h45	Binary	0.05(0.2)	-77.61	13.73	0.000	****
halfhour = h46	Binary	0.05(0.2)	-87.51	13.71	0.000	****
halfhour = h47	Binary	0.05(0.2)	-91.60	13.78	0.000	****
stage = NoEvent	Binary	0.79(0.4)	-60.50	16.13	0.000	****
stage = AllDay	Binary	0.14(0.3)	68.22	19.40	0.000	****
Performing Arts	Integer	0.02(0.1)	106.59	20.61	0.000	****
Other/Miscellaneous	Integer	0.2(0.4)	-125.12	23.40	0.000	****
Food/Wine	Integer	0.04(0.2)	762.25	21.38	0.000	****
Concerts/Tour Dates	Integer	0.15(0.7)	40.73	27.46	0.177	
Business/Networking	Integer	0.01(0.1)	-747.08	38.97	0.000	****
Neighborhood	Integer	0.06(0.2)	43.25	14.97	0.004	***
Nightlife/Singles	Integer	0.01(0.1)	50.52	27.39	0.078	*
inv_halfhour	[0,1]	0.06(0.1)	156.93	35.55	0.000	****
halfhours_to_next_event	{0..48}	38.5(14.7)	88.73	17.09	0.000	****
(Intercept)			57.56	21.90	0.010	***

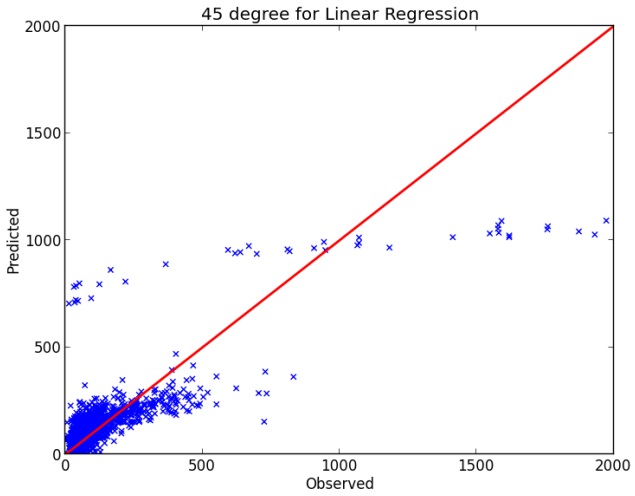

FIGURE 9. Linear Regression model 45 degree plot.

an event (or performer), Facebook likes, twitter mentions, news texts mentions, etc. For these metrics, however, the experimental design is challenging because one needs to track them in time to simulate the "prediction" moment properly. For example, to predict for the Avril Lavigne concert (2011-05-09) one week ahead, one would need to obtain those exact quantities as of 2011-05-02. Thus, we need to pay close attention to concepts such as event tracking (Lin et al., 2010; Tirado et al., 2011) or popularity evolution in time (Couronn et al., 2010).

Other potentially good data sources are the event textual descriptions themselves, news feeds, and micro-blogs that often hold detailed explanatory power. For example, two soccer matches with the same general event information but with different importance (e.g. friendly match vs national league) might originate completely different patterns. Information extraction techniques (e.g. Named Entity Recognition (Krishnan and Manning, 2006)) will be necessary to extract such distinctive features from the event description. Other contextual information is also commonly available such as weather reports/forecasts, seasonal information (e.g. school holidays, moving holidays, strikes).

6. Conclusions and future work

We demonstrated that, using online information, we can improve the quality of transport prediction under special events scenarios. The opportunity to collect data from the Internet and use it to understand

real-world phenomena is not negligible. While the size and complexity of online information alerts us for the need to apply sophisticated techniques such as information retrieval and extraction, we can already obtain relevant results with simple procedures such as using available APIs with events data.

We combined information extracted from the web with public transport data to build a predictive model of arrivals to special event venue areas. This is typically a difficult case for transport planning since special events originate high variance in demand. It is common to find, in traffic centers of large cities, teams of people manually scanning through the web and newspapers to find events with potential high impact. Our approach not only helps with automatic scanning for such sources, but also gives a better sense of what such impact means and what event features are the most important.

Our mobility dataset has limited size in total duration, but it has full coverage of Singapore, a small country with a massive public transportation system. This allowed our model to generalize across different events in different venues but with similar characteristics (e.g. same event categories). More than the actual prediction results, our main contribution is the demonstration of value and usefulness of this approach and what we can learn from the types of features discussed here.

There are a few important future steps to pursue, namely the disaggregation of the model into the level of Origin/Destination trips prediction and the extension to other transport modes. By having Origin/Destination models, the bus operator can optimize its fleet, such as reallocating vehicles to specific lines. The other future direction is the implementation of such models for private vehicles and taxi fleets. The former will be based on road sensors such as inductive loop counters, RFID toll readers and video-camera counters, while for the latter we will work with GPS floating car data from taxis.

The ultimate goal of this project is to provide scenario-based multimodal demand predictions for a dynamic traffic assignment (DTA) real-time traffic prediction model, DynaMIT (Ben-Akiva et al., 2012). Although DynaMIT is able to self-calibrate demand parameters in real-time using incoming sensor data, this complex optimization process is sensitive to initial conditions. Under special events scenarios, and in fact any situation that is better identifiable with Web data, our methodology can help DynaMIT's online calibration process. This principle can obviously be extended to other, non-DTA, traffic prediction systems (e.g. (Zhang and Ye, 2008; Puzis et al., 2013)).

We end with a final note on the generalization potential of this methodology. By its nature, any demand prediction model for social events needs to be adapted to local context. We do not expect that the same model, applied to a different city, or even a different area of the same city, would have a similar performance. Different neighborhoods, social groups and their lifestyles will lead to different behaviors in terms of event participation, mode share and relationship with event location. This should be more relevant for the smaller events because they often focus on audience niches and specific neighborhood interests. Furthermore, regarding Internet data, different web sites would have to be crawled, and adaptations in terms of language and style might be needed. Summarizing, although we see that the methodology itself is easily extendable to other problems (e.g. predicting Origin/Destination, departures, event attendance, mode share), it will need a careful fine-tuning for each individual local context.

Acknowledgments

The authors would like to thank the reviewers for their insightful comments that strongly contributed to the improvement of this article.

The authors also 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), and also by Fundação para a Ciência e Tecnologia (FCT), reference PTDC/EIA-EIA/108785/2008.

References

Ahas, R., A, A. K., and Tiru, M. (2009). Spatial and temporal variability of tourism loyalty in estonia: Mobile positioning perspective. In *Proceedings of the Nordic Geographers Meeting (NGM09)*. Department of Geography, University of Turku, Finland.

Ben-Akiva, M. E., Gao, S., Wei, Z., and Wen", Y. (2012). A dynamic traffic assignment model for highly congested urban networks. *Transportation Research Part C: Emerging Technologies*, 24(0):62 – 82.

Calabrese, F., Pereira, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2010). The geography of taste: analyzing cell-phone mobility and social events. In Floréen, P., Krüger, A., and Spasojevic, M., editors, *Pervasive Computing*, volume 6030 of *Lecture Notes in Computer Science*, pages 22–37, Berlin, Heidelberg. Springer Berlin / Heidelberg.

Copperman, R., Kuppam, A., Rossi, T., Livshits, V., and Vallabhaneni, L. (2011). Development of a regional special events model and forecasting special events LRT ridership. In *Proceedings of the 13th TRB National Transportation Planning Applications Conference*. Washington, DC: Transportation Research Board, National Research Council.

Couronn, T., Stoica, A., and Beuscart, J.-S. (2010). Online social network popularity evolution: An additive mixture model. In Memon, N. and Alhajj, R., editors, *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 346–350. IEEE Computer Society.

Coutroubas, F., Karabalasis, G., and Voukas, Y. (2003). Public transport planning for the greatest event - the 2004 olympic games. In *Proceedings of the European Transport Conference 2003*. ETC Proceedings, Strasbourg, France.

FHWA, editor (2006). *Planned Special Events: Checklists for Practitioners*. U.S. Department of Transportation, Federal Highway Administration, Office of Transportation Management.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Horvitz, E., Apacible, J., Sarin, R., and Liao, L. (2005). Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *Twenty-First Conference on Uncertainty in Artificial Intelligence*.

Koshak, N. and Fouda, A. (2008). Analyzing pedestrian movement in mataf using gps and gis to support space redesign. In *Proceedings of the the 9th International Conference on Design and Decision Support Systems in Architecture and Urban Planning*.

Krishnan, V. and Manning, C. D. (2006). An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 1121–1128, Morristown, NJ, USA. Association for Computational Linguistics.

Lin, C. X., Zhao, B., Mei, Q., and Han, J. (2010). Pet: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 929–938, New York, NY, USA. ACM.

Potier, F., Bovy, P., and Liaudat, C. (2003). Big events: planning, mobility management. In *Proceedings of the European Transport*

Conference 2003. ETC Proceedings, Strasbourg, France.

Puzis, R., Altshuler, Y., Elovici, Y., Bekhor, S., Shiftan, Y., and Pentland, A. S. (2013). Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks. *Journal of Intelligent Transportation Systems*, 17(1):91–105.

Qi, G., Li, X., Li, S., Pan, G., Wang, Z., and Zhang, D. (2011). Measuring social functions of city regions from large-scale taxi behaviors. In *Proceedings of the 2011 Pervasive Computing and Communications Workshops (IEEE PERCOM 2011 Workshops)*, pages 384–388.

Schweitzer, L. A. (2012). How are we doing? opinion mining customer sentiment in u.s. transit agencies and airlines via twitter. In *Proceedings of 91st Transportation Research Board Meeting (TRB 2012)*. Washington, DC: Transportation Research Board, National Research Council.

Smith, B. L. and Venkatanarayana, R. (2005). Realizing the promise of intelligent transportation systems (ITS) data archives. *Journal of Intelligent Transportation Systems*, 9(4):175–185.

Terpstra, F., Meijer, G., and Visser, A. (2004). Intelligent adaptive traffic forecasting system using data assimilation for use in travel information systems. In *The Symposium on Professional Practice in AI a stream within the First IFIP Conference on Artificial Intelligence Applications and Innovations AIAI-2004, Toulouse France*.

Tirado, J. M., Higuero, D., Isaila, F., and Carretero, J. (2011). Analyzing the impact of events in an online music community. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, New York, NY, USA. ACM.

Vougioukas, M., Divane, S., and Thymiakou, G. (2008). Transport and tourism investments for hosting big events: the case of the 2013 Mediterranean games in Volos, Greece. In *Proceedings of the European Transport Conference 2008*. ETC Proceedings, Leeuwenhorst, Netherlands.

Zhang, Y. and Ye, Z. (2008). Short-term traffic flow forecasting using fuzzy logic system methods. *Journal of Intelligent Transportation Systems*, 12(3):102–112.

Singapore-MIT Alliance for Research and Technology (SMART), Singapore
E-mail address: camara@smart.mit.edu

Department of Informatics Engineering, CISUC, Coimbra, Portugal
E-mail address: fmpr@dei.uc.pt

Department of Civil and Environmental Engineering, MIT, Cambridge, USA
E-mail address: mba@mit.edu