Predicting Traffic Congestion in Presence of Planned Special Events

Simon Kwoczek  
Group Research  
Volkswagen AG  
simon.kwoczek@volkswagen.de

Sergio Di Martino  
Group Research  
Volkswagen AG  
sergio.di.martino@volkswagen.de

Wolfgang Nejdl  
L3S Research Center  
University of Hanover  
extdl@l3s.de

Abstract

The recent availability of datasets on transportation networks with high spatial and temporal resolution is enabling new research activities in the fields of Territorial Intelligence and Smart Cities. Within these domains, in this paper we focus on the problem of predicting traffic congestion in urban environments caused by attendees leaving a Planned Special Events (PSE), such as a soccer game or a concert. The proposed approach consists of two steps. In the first one, we use the K-Nearest Neighbor algorithm to predict congestions within the vicinity of the venue (e.g. a Stadium) based on the knowledge from past observed events. In the second step, we identify the road segments that are likely to show congestion due to PSEs and map our prediction to these road segments. To visualize the traffic trends and congestion behavior we learned and to allow Domain Experts to evaluate the situation we also provide a Google Earth-based GUI. The proposed solution has been experimentally proven to outperform current state of the art solutions by about 35% and thus it can successfully serve to reliably predict congestions due to PSEs.

Keywords: traffic prediction, planned special event, event analysis

1. Introduction

In times of ongoing urbanization and steady growth of mobility demands, traffic congestions cost billions of dollars to the society every year [12]. Within the last years, thanks to the advances in sensor technologies (like Smartphones, GPS handhelds, etc.) and storage capabilities, datasets about traffic with high spatial and temporal resolution have become available. This has led to advanced investigations on the impact of different influencing factors, as traffic lights, daily rush hour, construction zones, etc., on traffic congestions (i.e. [2, 7, 9]). Most of these factors are either recurring on a regular base (i.e. rush hour), exist only once for limited time (i.e. construction zones) or their occurrences are unpredictable (i.e. accidents). Current state of the art commercial solutions are able to work well with recurring traffic situations as it’s behavior can easily be learned from historical data [18]. On the other hand, also non-periodic events with an expected large attendance (known also as Planned Special Events, or PSE as introduced in [6]), such as concerts, soccer games, etc., play a major role for delays in everyday transportation [9]. As example, the concert of Rihanna in Johannesburg (South Africa) in October 2013 caused people to sit in traffic for as long as five hours, trying to reach the stadium. Similarly, the concert of Robbie Williams in London, in 2003, caused tailbacks up to 10 miles on the highway A1 towards the stadium. An interesting aspect is that the traffic due to PSEs has a quite typical behavior, having two subsequent waves of congestion [10]. The first one is caused by people going to the event, while the second one is due to people leaving the venue, and may be even bigger that the first wave. Very few research attention has been devoted to predict the congestion due to a PSE. At the same time, even the most advanced available commercial systems are incapable of predicting this kind of non-recurring traffic ahead of time.

To address this open issue, in this paper we describe a solution we developed to predict the spatio-temporal impact of the second wave of traffic due to a PSE around its venue. In particular, the proposed approach is meant to be executed while the event is happening, and takes as input the category of the PSE, like Concert, Entertainment, etc., and the information on the first wave of traffic, coming from traditional traffic providers. Then, using an adaptation of the K-Nearest Neighbors (K-NN) algorithm, we look for the most similar past PSEs (cases) among historic observations, in order to derive a prediction of the impact of the second wave of traffic. Such a prediction is done in terms of average delay over the road segments around the venue that have been found to be highly correlated with the congestions due to PSEs. To graphically visualize the results, we provide a Google Earth-based tool 1, recalling the one de-

1http://earth.google.com

357
veloped in [4] or in [13]. Thanks to this tool, it is possible for Decision Makers to understand how the traffic situation is influenced by a PSE.

To assess the proposed approach, we used data about traffic and events from June to December 2013 in the inner city of Cologne, Germany. The traffic data has been provided by one of the most prominent traffic providers in Europe. It comes mainly from Floating Car Data, i.e. data collected from GPS sensors in vehicles, and it covers most of the streets within the inner city. As for the PSEs, we considered all the 29 events hosted in the Cologne LANXESS arena, in the same temporal span. By performing a leave-one-out cross validation on the event dataset, we compared our proposal with current state of the art, intended as real time traffic informations, and with a baseline consisting simply of replicating the impact of the first wave as predicted second wave. Results show that our proposals outperforms the alternative solutions, providing prediction that are better up to 41%.

The remainder of the paper is structured as follows: section 2 presents related work within the field of traffic predictions and PSEs and some background terminology. In section 3 we present the approach to predict the congestions caused by outbound traffic after an event, plus the GUI to visualize the results. In section 4 we describe the Research Questions and the evaluation protocol to assess the proposed approach. In section 5 the results of the empirical assessment are presented and discussed. Finally in section 6 conclusions are outlined, together with some future research directions.

2. Background and Definitions

Since years, traffic congestion predictions have been widely studied within the research communities of ITS, Smart Cities and Territorial Intelligence, leading to a rich body of literature on these topics. Indeed, by knowing in advance the traffic patterns it is possible to optimize mobility. For instance, route calculation engines can compute more energy efficient routes, able also to save time for the drivers.

In general, predicting traffic congestion in urban environments is a highly complex task. Early approaches for traffic predictions used simulations and theoretical modeling (e.g. [2,3]). Nowadays thanks to the availability of new massive datasets on traffic, several different statistic and data driven approaches have been presented to the community. Examples include generalized linear regression ([21]), nonlinear time series ([8]), Kalman filters ([11]), support vector regression ([20]), and various neural network models ([11, 17, 19]). A combination of some of the latter kind of approaches is used also by current commercial navigation solutions, able to predict recurring congestions by identifying characteristic traffic flow patterns for street segments from historical data. On top of that, these commercial systems can also optimize the route planning based on the real-time traffic situation [18]. In general, traffic congestion can be divided into recurring congestions, usually caused due to a mobility demand that exceeds the capacity of the road network (e.g. due to rush hour), and congestions that are non-recurring (e.g. due to incidents or special events) [9]. The effects of nonrecurring traffic congestions and their prediction is a widely investigated topic within the research community (e.g. [14–16]). Although these approaches showed a significant improvement in prediction they use data from stationary loop sensors that are not always capable of reflecting the traffic state in a granularity required for urban scenarios. In addition, their focus lies on one-directional street segments as highways whereby usually in the inner cities the impact is a multidimensional problem, evolving in a 2D, more complex route network.

Previous researches have highlighted that also PSEs are a possible influencing factors [7,9], since they may lead thousands (or even hundreds of thousands) of people to travel towards the same destination in a very limited time interval, and then to leave the venue again in a very short time span. To the best of our knowledge, the only work available that has its focus on the influence of PSEs on traffic is presented in [10]. The authors report a generic overview about the influence of PSEs on the road network, derived from an event classification defined by the Chinese State Council. They also introduce management plans for the different types of events, but there is no quantifiable solution for the prediction of the traffic.

2.1. Definitions

A PSE may have an impact on the traffic behavior in a specific region over time. Such an impact region can be defined as the list of congested segments of the road network. A congested segment is a piece of road network where the difference between expected and actual traffic speed is bigger than a certain threshold. The expected traffic velocity on a given segment can be obtained from a number of historical observations, coming from special sensors in the infrastructure and/or Floating Car Data. In this way it is possible to learn the typical traffic behavior for a given road segment on a certain day of the week, at a given time. If a segment is congested it takes more time to pass by it. This additional time is defined as delay time and is measured in seconds. As an example, Figure 1 shows the summed up delay time in a circular impact region, with a radius of 500 meters around the LANXESS Arena in Cologne, Germany for all Tuesdays in our considered dataset, where no events happened. The red line presents the model that is derived from historical observations as the average delay time on all Tues-
days. From this trend line it is possible to detect the rush hour behavior in the morning and in the evening. As stated above, the graph only contains Tuesdays without events and the generated trend line can therefore be seen as the regular pattern around the venue on Tuesdays. As described in the Introduction, a PSE leads to two waves of traffic. An example can be seen in figure 2. It shows the traffic behavior around the LANXESS Arena on Tuesday the 2nd of July 2013, when Mark Knopfler gave a concert in the arena. The figure shows the difference between the observed delay time at that day and the trend line for that day of the week. From the figure we can clearly see the two peaks of congestion, between 18:00 and 20:00, and between 22:00 and 23:00. Since the concert started at 20:00 and ended around 22:00, the assumption arises that the two peaks are caused by people going to the event (inbound traffic) and leaving after its end (outbound traffic). To quantify the impact of PSEs on traffic, we focused on the traffic data happening before and after the events. From the 29 events we analyzed, and for this specific location, we found out that the inbound traffic is always contained in a time frame within two hours before the scheduled begin of the event. As for the outbound traffic, in our dataset the time frame starts two hour after the begin of the event and lasts for two hours. These time frames ensure that the entire event-caused traffic is captured. For both time frames, we define a normalized timescale, using the difference to the planned begin of the event as scale. In this way, we can compare traffic situations happening in the same relative time corridor, for events that start at different times. For both time frames we observe the average delay time, defined as \( \text{avgdelay} \) (measured in seconds) as measure for the impact of the PSE on traffic.

3. The Proposed Approach

In this section we define two different approaches to predict the outbound traffic, namely the Category-Based Modeling Approach (CBMA) and the Category Specific Inbound-Based Prediction Approach (CIPA). Afterwards, we introduce a tool to visualize the results of the predictions.

3.1. Impact of event category

The LANXESS arena hosts different events of different categories. During the time span of our study, we observed 18 concerts, 6 entertainment events, 3 comedy events and 2 sporting events. The naming of the categories is taken from the official arena schedule. Although most of the considered events started roughly at the same time (between 19:00 and 21:00) the different categories show significant differences in their impact on traffic. In total, 12 out of the 18 concerts showed significant inbound congestions and 11 of these events showed also strong outbound congestions. While comedy events did not account for any significant raise in congestions, entertainment events caused congestions in 2 out of 6 times for the inbound and 1 time in the outbound traffic period. The only events that differed significantly in their start time were the sporting events (in the time span of our research the sporting events were two handball games). Both of them showed a comparable congestion behaviour for in- and outbound traffic. To further illustrate the correlation between event category and congestion behaviour, we report in figure 3 the behavior for each category using the \( \text{avgdelay} \) observed as congestion measure. The boxplot highlights the different behavior of the different event categories.

These differences motivated us to develop a first approach that focuses on the typical congestion behavior for different categories that we named Category-Based Modeling Approach (CBMA).

3.2. Category-Based Modeling Approach (CBMA)

The CBMA is aimed at detecting the specific congestion patterns of each observed category and uses these patterns for predictions. We learn the behavior of each category by accumulating the congestion behavior during the outbound traffic for all events of that category in our training set. From that, we consider the average of all observations and
use the generated trend line for predicting the outbound traffic congestion behavior for all events in the test set. Since the only information CMB uses for predictions is the category of the events, it can be used without any knowledge about the traffic situation on the specific day of the targeted event. Thus, it can be applied early ahead of time.

3.3. Impact of Inbound Traffic Congestion

While each category of events shows a different pattern, some events of the same category also vary a lot in their congestion behavior. For example, the Rihanna concert on the 26th of June 2013 and the concert of Barbra Streisand on the 12th of June 2013 are both of the same category (concert) and start almost at the same time. Thus, one would expect that they have a similar impact on traffic. However while the outbound congestion after the Rihanna concert shows an avgdelay of around 4600 seconds, the outbound congestion after the Barbra Streisand concert only shows an avgdelay of 2100 seconds. A likely explanation for the phenomenon lies in the different amount of people going to these two concerts. Although the actual number of attendees is not available, we were willing to explore if the inbound congestion of an event is a good indicator of how severe the outbound congestion is going to be. In other words, we expect the congestion during the inbound traffic to describe the "hype" about the event by serving as an indicator about how many people are going to the event. This motivated us to develop a second approach, named Category Specific Inbound-Based Prediction Approach (CIPA) that includes this information into the prediction model.

3.4. Category Specific Inbound-Based Prediction Approach (CIPA)

The CIPA is based on the idea that both category and inbound traffic are explanatory variables for the expected outbound traffic. One possible way to incorporate these information into the prediction process lies in applying Machine Learning techniques on an adequate dataset of historic data to automatically learn the correlations between these variables. For CIPA, the dataset contains information about the category of an event and the observed avgdelay during the inbound traffic. We use these features as predictors of the avgdelay during the outbound traffic. In our case, we applied the K-Nearest Neighbor regression [1] Machine Learning technique. In order to do a prediction, this technique aggregates the values the k "closest” examples in the training set, where k is an input parameter to the algorithm. To compute the distance among observations we used the Euclidean distance function, while as aggregation formula, we used the mean of the observed values. Further details on K-NN can be found in [1]. A drawback of this approach compared to CBMA lies in the fact, that it can only be used for predictions on the event day, after the inbound traffic has already been observed.

3.5. Mapping Delay on the Road Network

In order to use the generated predictions in Advanced Driver Information Systems, the resulting delay time needs to be mapped onto the street network. We identify road segments that are likely of being congested due to outbound traffic from historic data by selecting all segments that were affected in at least 1/3rd of all PSE caused congestions. Then we spread the predicted outbound delay time for an PSE over the segments, assuming a normal distribution. From the results, we calculate the delay per meter DM index (in seconds) for that specific event, which gives an overview about the severity of congestion in the area. The index is also used in the GUI we describe in the next section, to allow Domain Experts to get an overview about the expected impact of the PSE.

3.6. The proposed interactive GUI

Once the results are produced by the one of the previous prediction engines, they can feed a visualization layer. To this aim, we have developed a Graphical User Interface, shown in figure 4, that embeds Google Earth to render in 3D the predicted spatio-temporal impact area, over a geo-referenced satellite image, optionally also enhanced by additional informative layers. The goal is to facilitate Domain Experts and Decision Makers to visually understand spatial relationships among the datasets and their spatial context. The proposed GUI contains a main frame that holds the current representation of the predicted traffic severity level in Google Earth. On the left side, the user can pick a location from a list of available data. After selecting a location, the set of scheduled PSEs for the venue is shown in the left panel. If the information is available, the inbound traffic of
the event will be represented in a graphic way at the bottom of the left panel. By selecting one of the developed algorithms (CBMA or CIPA) the prediction process starts and the user gets a visual representation of the predicted congestion area as a new layer on the map. The severity level can be manually adjusted by the settings tab, allowing Domain Experts to easily customize the representation of the results.

4. Experiments

This section describes the setup of the experiments conducted to assess the validity of our proposals.

4.1. Dataset Description

The traffic information used in this research is collected by one of the most prominent traffic providers all over Europe. It covers the main road network within the inner city of Cologne. The information is generated from a combination of various different sources. These include Floating Car Data (originated from millions of vehicles equipped with GPS sensors), GSM probe data and data from stationary sensor (e.g. loop detectors, camera sensors) obtained from local traffic management centers. More information about the different sources and their aggregation can be found in [18]. From these combined sources, an accurate description of detected traffic congestions can be derived. Specifically, for each detected congestion, the data contains the resulting delay time on a segment of the road network (in seconds) and the congestion level, on a scale from 1 to 5, where 1 means no relevant traffic congestion, 4 is the highest level and 5 means that the congestion level is unknown. We received data from June the 1st 2013 until December the 31th, 2013. The dataset we used in this research accounts for about 50 Gigabytes which leads to a favorable coverage of the area around the arena in Cologne. All measurements presented in this section show the accumulated values of all road segments within an area of 500 meters around the stadium. As for the events, we collected all the PSEs happening in the LANXESS arena in the same time span covered by the above described traffic information. In these seven months, 29 events were scheduled. Regarding these observed events, 16 showed congestion due to inbound traffic. For these events, the observed an accumulated avgdelay on the considered road segments varied from very minor with 420 seconds (Andreas Gabalier, 17th of October 2013) to severe traffic congestion with accumulated 9422 seconds (Rihanna, 27th of June 2013), that is a delay of almost 3 hours in the considered area. The outbound traffic caused congestion in 17 of the 29 events whereas the avgdelay varied from 124 seconds (Cirque du Soleil, 25th of October 2013) up to 4632 seconds (Rihanna, 26th of June 2013).

4.2. Baseline Approaches

As stated before, to the best of our knowledge, there is no research that has studied the impact of PSEs on traffic in a similar manner before. Therefore, we introduce two approaches that serve as baseline for our proposed CBMA and CIPA, named the Zero approach and the Start approach.

Zero approach: In the Zero approach, we simply assume that there is no traffic in the area during outbound traffic. While this approach seems fairly simplistic, it actually reflects the current state of the art in navigation solutions, and is therefore suitable for comparing newly developed algorithms against it. As matter of fact, by relying just on historic data, current navigation solutions are unable to predict congestions due to nonrecurring events.

Start approach: While the Zero approach simply assumes no additional traffic at all above the historic model, the Start approach goes one step further. It is based on the idea that the amount of people going to the event is the same as the people leaving the event. Consequently, the outbound traffic could be estimated as the same as the inbound traffic. Thus, we could use the exact information about the inbound avgdelay as a prediction for the outbound avgdelay.

4.3. Research Questions (RQ)

Given the above described approaches, the research questions underlying this paper can be formulated as follows:

1. Are the predictions of expected delay due to outbound traffic using CBMA better than those obtained with the Zero and the Start approaches?

2. Are the predictions of expected delay due to outbound traffic using CIPA better than those obtained with the Zero and the Start approaches?

3. Are the predictions of expected delay due to outbound traffic using CIPA better than those obtained with CBMA?

For these research questions, the performance of the different approaches is measured using the root mean square error (RMSE) defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$  \hspace{1cm} (1)

where $y_i$ and $\hat{y}_i$ represent the actual and predicted delay time respectively and $N$ is the number of predictions.
5. Results

To evaluate the validity of the proposed approaches, and thus to answer the Research Questions, we applied the four solutions described in the previous section on the dataset of 29 events by means of a cross-validation, that is splitting the data set into training and validation sets. Training sets are used to build models, while validation sets are used to validate the obtained prediction models. In particular, we exploited a leave-one-out cross validation, which means that the original data set is divided into \( n=29 \) different subsets of training and validation sets, where each validation set has just one event.

As for the use of the K-Nearest Neighbors algorithm, one of the issue is the optimal choice of the parameter \( k \), that is the number of closest training examples to be considered in the feature space. This choice usually depends upon the data. It is outside the scope of this research to investigate hyperparameter optimization techniques (e.g. as in [5]). In our case we investigated the use of \( k=3 \) and \( k=5 \), obtaining very similar results. Consequently, in the following we report only on the use of \( k=5 \).

Figure 5 shows the RSMEs of the predicted outbound traffic \( \text{avgdelay} \) for each approach separately. The figure shows, for the entire dataset of PSEs, the prediction performance for each method in terms of the RMSE values obtained from the cross validation. The data shows, that the baseline which uses the \( \text{avgdelay} \) from the inbound traffic as predictor has the highest RMSE of 2095 seconds. The Zero approach, which has no a-priori information about the PSEs or the traffic information at all, shows an RMSE of 1918 and therefore slightly outperforms the Start approach.

To answer the RQ1, we can see that the RMSE for the CBMA is 1364, outperforming both the Zero- and the Start approach by far. Indeed, the proposed approach has a predicted delay error that is 28% lower than current state of the art, and 35% lower than the Start approach. Consequently, we can positively answer the RQ1, since CMBA outperforms the two baselines.

As for RQ2, the RMSE for the CIPA, which incorporates the most information about the situation, is 1230. This reflects a reduction of the prediction error of around 41% percent compared to the Start approach and still 35% percent compared to the Zero approach. Consequently, we can positively answer the RQ2, since CIPA outperforms the two baselines.

As for the RQ3, we can see from 5 that CBMA and CIPA provide quite similar results. This motivated us to further investigate the performances, splitting the results for each
category of events. The results are shown in figure 6. We found that for very homogeneous categories (as for Sport and Comedy in our examples), the CIPA was outperformed by the CBMA. For instance, CBMA showed an RMSE of 80 seconds for the Sport category while the RMSA for CIPA is 418 seconds. At the same time, for highly fluctuating categories (as for the categories Concert and Entertainment in our examples), the CIPA shows much better results of for instance an RMSE of 1441 seconds for the concert category, compared to 1621 using CBMA. Consequently we cannot provide a definitive answer to RQ3, since CIPA and CBMA have comparable performances.

5.1. Mapping results on the Road Network

We identified a set of 52 street segments that showed high correlation of the congestion with PSEs, which account for a total distance of 2105 meters. For these links, calculating the delay per meter index $DM$ shows the different severity levels of the events. As an example, the CIPA predicted a delay time of 2129 seconds during the Mark Knopfler concert, that, given the total length of the considered road segments, leads to an severity index of around 1s/m. In a future integration with a route calculation engine, this means that the travel time of each route passing in the vicinity of the arena, in the considered timeframe, takes about 100 additional seconds per 100 meters of the route. Consequently, the route calculation engine has now new travel times for each segment, and can reroute the driver on the most efficient path.

6. Conclusion and Future Work

Improving mobility is a paramount issue due to the huge environmental and social impact of vehicular traffic. Thanks to the recent availability of high quality spatio-temporal datasets coming from Floating Car Data and other data sources, new and unprecedented research opportunities are arising in the direction of smarter solutions for mobility. Within this topic, to the best of our knowledge, a very limited attention has been dedicated to the prediction of the impact of Planned Special Events, such as concerts or sport games, on urban traffic, even if some researches highlight that they are a significant fraction of the total traffic. PSEs have a very specific impact on mobility, inducing two waves of traffic. The first one is due to the incoming spectators while the second one is due to the people leaving the event. In this paper we have proposed an approach to predict the impact of the second wave of traffic due a PSE, given information on it’s category and on the first wave of traffic. To this aim, we have defined a two-step solution, where at first we created models able to predict the out-bound congestion based on historic data about past events and, in the second step, mapped these congestion prediction on to the route network by identifying the road segments that are most likely of being congested due to PSEs. The approach is complemented by a GUI that allows Decision Makers to visualize on an interactive 3D map the dynamic extension of the impact area and the severity of the expected congestions. An empirical assessment has been done using floating car data covering 7 months of mobility information in the city of Cologne, Germany, and using all the PSEs hosted in the LANXESS arena in the same period. The results show that our proposal can positively identify the influence of the PSE on the traffic, being up to 35% better than actual state of the art solutions. Consequently it can be a viable module to be integrate in any navigation device, in order to reroute drivers away from the location of the event and optimize urban mobility. As future work, the next obvious step is to predict also the first wave of traffic, but our experiments showed that relying only on the time schedule is not sufficient, since there is no way to predict the amount of people that will attend the event. The idea we are starting to investigate is to use social networks, such as Facebook or Twitter, to understand what is the popularity of a specific event or artist, considering also the geographical and age distribution of the supporters.

7. Acknowledgment

The research leading to these results has been partly funded by the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement n 610990 - Project COMPANION. The authors wish also to thank TomTom NV for providing the traffic data used in this research.

References


