Text analysis in incident duration prediction

Francisco Pereira, Filipe Rodrigues, and Moshe Ben-Akiva

1. Singapore-MIT Alliance for Research and Technology (SMART),
e-mail: camara@smart.mit.edu.
2. University of Coimbra, Portugal
3. Massachusetts Institute of Technology (MIT), USA.

Abstract

Due to the heterogeneous case-by-case nature of traffic incidents, plenty of relevant information is recorded in free flow text fields instead of constrained value fields. As a result, such text components enclose considerable richness that is invaluable for incident analysis, modeling and prediction. However, the difficulty to formally interpret such data has led to minimal consideration in previous work.

In this paper, we focus on the task of incident duration prediction, more specifically on predicting *clearance time*, the period between incident reporting and road clearance. An accurate prediction will help traffic operators implement appropriate mitigation measures and better inform drivers about expected road blockage time.

The key contribution is the introduction of *topic modeling*, a text analysis technique, as a tool for extracting information from incident reports in real time. We analyze a dataset of 2 years of accident cases and develop a machine learning based duration prediction model that integrates textual with non-textual features. To demonstrate the value of the approach, we compare predictions with and without text analysis using several different prediction models. Models using textual features consistently outperform the others in nearly all circumstances, presenting errors up to 28% lower than models without such information.

Keywords: Incident duration prediction, text analysis, topic modeling, regression models

1. Introduction

On a daily basis, traffic incidents demand quick reaction and adaptability from both operators and motorists. In order to provide accurate and timely travel advisory, a key information is the expected duration of the incident [29, 2, 30], as defined by the time period that spans from incident occurrence to road clearance. During this period, the traffic operators need to efficiently execute a response strategy, which in turn depends on a variety of factors, some

objective and measurable such as number of lanes affected, location or traffic conditions, others subjective or difficult to assess such as capacity reduction, drivers behavior or potential for generating secondary accidents. Traffic operators also need to provide guidance information for drivers, and it is crucial that this guidance is consistently trustworthy and accurate [3].

To support a timely response, traffic management centers establish workflows that consist of collecting information, analyzing it and executing the chosen strategy, continuously using updated information to control traffic, disseminate information and manage incident response resources (e.g. [26, 17, 2]). Each of these steps can be supported by automatic tools, but the nature of the problem demands systematic presence of humans in the loop. Since the exception is the norm, the range of different problem configurations is so wide that constrained-value interfaces (e.g. buttons, check boxes, pull-down list) become insufficient and the use of written or verbal messages is needed to communicate between different actors.

The advantage of human language is its virtual lack of limits in terms of transmitting the subtle details that allow others to assess the situation, while the disadvantage is the difficulty that machines find in interpreting it. With a few notable exceptions (e.g. [27]), earlier works on incident analysis and prediction have completely ignored this dimension or based themselves on crude simplifications. In this article, we propose the integration of *topic modeling*, a state of the art text analysis technique, with a machine learning model, as an effective means to continuously predict incident duration (or, equivalently, remaining time to clearance) taking into account the latest report updates. It turns out that topic modeling is particularly well suited for the domain of incident messages since these have a relatively homogeneous lexicon and each message aims to objectively convey clear and synthetic information to help emergency response.

Our main contribution is thus an incident duration prediction model that is able to consider text messages as well as numeric and nominal information. As with earlier literature (e.g. [25, 11, 22]), we define our duration variable as the period from reporting to clearance, thus excluding the unknown period from occurrence till reporting. We focus on accident duration prediction thus leaving aside other kinds of incidents such as vehicle breakdowns, road works and others. We will demonstrate the value of the approach by comparing with models without text analysis and by observing the performance of the model on a realistic setting, whereby information becomes sequentially available over time.

The remainder of this article is organized as follows. The next section will motivate this work within the framework of traffic incident management systems. The subsequent two sections shall provide the reader with the necessary background, first with a general literature review on incident duration prediction, then with an overview of topic modeling. Sections 5 and 6 describe the available data and the prediction models, respectively. The article will end with a short discussion and conclusions (section 7).

2. Motivation: Traffic Incident Management

Traffic incidents have been identified as one of the major contributors to increased congestion, causing about one-quarter of the congestion on U.S. roadways [1]. They are estimated to cause more than 50% of delay experienced by motorists in total for all urban areas. Furthermore, for every minute that the primary incident remains a hazard, the likelihood of a secondary crash increases by 2.8 percent [20].

In order to mitigate such consequences, traffic incident management (TIM) procedures take advantage of surveillance information and communication systems to improve response via the coordination afforded by a Traffic Operations center, as well as the real-time sharing of information among the affected agencies. Responders need to estimate its magnitude, expected duration, as well as the expected vehicle queue length, and then should set up the appropriate temporary traffic controls for these estimates [2]. Calculation of magnitude can follow earlier incident rating systems (e.g. [6]), but determining expected duration and queue length is a more complex task that needs to take into account characteristics of the incident, historical records, availability of other response systems (e.g. ambulance), weather and traffic status.

Together with managing incident area clearance, TIM needs to provide predictive guidance information to motorists (e.g. via VMS, radio) to help reduce considerably the average travel time for most vehicles traversing the incident, therefore reducing the total delay in the network [19]. However, real-time traffic prediction systems that provide consistent prediction under incident scenarios, such as DynaMIT [8], also need incident information in the form of capacity reduction and expected duration. Again, while capacity reduction can be approximated by directly observable characteristics, such as number of lanes blocked and weather status [23], incident duration demands consideration of many other factors. Testing with seven different case-studies, Lopes [25] demonstrates that an incident information feed (with predicted duration and capacity reduction) can improve DynaMIT's normalized root mean squared error (NRMSE) performance by up to 15%, 44% and 53% for predicting speeds/travel times for 10, 20 and 30 minutes ahead, respectively.

Finally, beyond real-time incident response, post-hoc analysis [5, 30] is a crucial tool for performance assessment and TIM process redesign. The typical measures are incident duration, delay, secondary incidents. Other variables, such as economic cost/loss are calculated from those measures. The key is to contrast the observed values with reasonable benchmarks, often set up by regulators [5]. As suggested by Feyen and Eseonu [13], such benchmarks should reflect statistical evidence and help expose "root causes of incident response performance". The task is thus to study how factors that characterize the incident (e.g. location, severity, time of day, weather) correlate with the target measures and design statistical tools that determine an "expectable performance".

Our research is motivated by the framework of traffic incident management, both contributing for duration prediction in real-time, as new information arrives, as well as for off-line post-hoc analysis. On an real-time basis, it ulti-

mately aims to provide supply parameter updates to dynamic traffic prediction systems, such as DynaMIT. On an off-line basis, the goal is become a tool for performance assessment for transport agencies such as the Land Transport Authority of Singapore (LTA).

For practical reasons, we will focus exclusively on data from incident reports. Traffic and weather data are not available for the same period of time. Methodologically, this will allow us understand the individual contribution of text-analysis to the problem of incident duration prediction. In future work, we will enrich our dataset with other information such as traffic or weather status and will evolve the prediction model itself considering more sophisticated options (e.g. survival analysis [28]).

3. Literature review

Task definition. According to the highway capacity manual [4], traffic incident time is organized into 4 periods: *reporting*, the time between occurrence of the incident and reporting to traffic operators; *response*, the period between reporting and arrival of response team; *clearance*, the time needed for response teams to assist involved parties and clear the road; and *recovery*, the period from clearance until the restoring of normal traffic conditions. As with earlier works (e.g. [25]), we define our incident duration as the sum of response and clearance periods, since they are objectively measurable, differently to the other components.

Sequential and one time models. There are two general types of models, one that continuously predicts duration (or remaining time) as new information becomes available; another that assumes that all information is available at reporting time and makes a "prediction" at that time. The former, called *sequential model* (e.g. [22, 25, 11, 34]) aims for deployment in real-time settings while the latter, the *one time* model, is generally motivated by a post-hoc incident analysis perspective rather than real-time prediction (e.g. [18, 15, 14]).

Modeling techniques. The set of modeling techniques is wide, from statistics based to non-parametric machine learning models. Valenti et al [33] present a comparative analysis of several machine learning techniques that are typical in the literature, including artificial neural networks (ANN), decision trees, support vector regression/relevance vector machines (SVR/RVM), and linear regression (LR). They applied the one time model perspective and concluded that SVR/RVM outperformed the others for long durations and ANN for short durations. ANN were also used by Lopes [25] in a sequential model with 4 neural networks with incremental inputs. With the maximum information model, he achieved the lowest Root Mean Squared Error (RMSE) so far in literature of 12.45 for major accidents and 7.31 for minor ones. It should be noted, however, that his scenario is highly homogeneous (a single highway corridor), information rich, and such results are often achieved at later stages of the incident response process.

From the statistical realm, Nam and Mannering [28] applied survival analysis to re-estimate duration conditioned on elapsed time and information updates

at three stages, namely incident occurrence, incident notification and rescuer arrival. According to their results, factors such as month, rain conditions, geographic information, incident characteristics (e.g. existence of injuries), and response agency are important variables in forecasting duration. Also taking into account elapsed time, Khattak et al. [21] designed a sequential model with a series of truncated regression models.

Feature set. The key component of all these models is indeed the set of features used, sometimes specific to the dataset or location, other times extremely difficult to compute accurately. A non-exhaustive compilation from the existing literature (e.g. [22, 18, 14, 35, 7]) includes number of affected lanes, number of vehicles, time, location, incident type (e.g. vehicle breakdown, accident), type of vehicles involved, life/equipment impacts (e.g. with/without injuries, damaged infrastructure), response agency (e.g. police, fire department), weather status, existence of secondary incidents, geographic information, and traffic conditions. As mentioned earlier, some of these details are sometimes available in text form.

Synthesis. Historically, research in incident duration prediction had an initial phase of problem understanding and analysis, mostly centered on statistical approaches, with crucial works from Golob et al [15], Khattak et al [21] and Nam and Mannering [28]. More recently, the emergence of powerful computing capabilities together with massive datasets led to a predominance of machine learning approaches (e.g. [33, 27, 11, 25]). In this set, one work, from Miller and Gupta [27], is the closest to ours in the sense that they use text based features. They apply string matching rules to extract certain features (e.g. type of vehicle) from the text. This approach depends essentially on manual work and on subjective assumptions on what is and what is not relevant from the text, which implies important limitations in terms of scalability and flexibility.

4. Topic modeling

Generally, a topic model describes a collection of documents as a statistical distribution of abstract "topics" or "classes". The emergence of these statistical natural language processing (NLP) approaches started in the 1990's when the field reached limited success with strictly rule-based, deterministic approaches (e.g. grammars and description logics). In 1990, Deerwester et al [12] proposed a document indexing model that used singular-value decomposition (SVD), in which a large term by document matrix is decomposed into a set of about 100 orthogonal factors from which the original matrix can be approximated by linear combination. This general approach was later coined as Latent Semantics Indexing (LSI). With documents and search queries represented by such reduced dimensionality, the process to retrieve documents that maximize cosine values becomes more efficient and noise resistant.

Later, Papadimitriou et al [31] and then Hofmann [16] laid a more comprehensive theoretical foundation for LSI and introduced probabilistic Latent Semantics Indexing (pLSI). In contrast to LSI with SVD, the probabilistic variant defines a generative data model in which each document is probabilistically assigned a latent topic z, each topic being defined as a distribution over words.

Latent Dirichlet allocation (LDA) is a generalization of pLSI developed by Blei et al [10], that allows documents to be mixtures of topics. In LDA, each document is represented as a distribution over topics, and each topic is a distribution over words. In contrast to other approaches, LDA has the practical advantage of being more interpretable. In general, topics can be intuitively associated to a specific higher-level meaning. For example in our case, some topics can be associated to "seriousness" indicators, such as whether there are injuries or oil spillage, while other topics are related to progress monitoring, such as if the police arrived, if participants exchanged information, photos were taken, etc.

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

The parameters α and η are hyperparameters that indicate respectively the priors on per-document topic distribution and per-topic word distribution. Thus, $w_{d,n}$ are the only observable variables, all the others are latent in this model[1]. 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 thus 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 Expectation-Maximization (EM). For further details on this procedure please refer to the original article of David Blei and colleagues [10].

[1]For the interested reader, we present the plate notation of our LDA model - specifically called "smoothed LDA" - in Figure 9 of the Appendix, with a brief explanation

With a trained LDA topic model, one can apply the same general procedure to assign topics to every new document [10]. In this case, the problem is considerably simpler since all parameters except for $z_{d,n}$ and θ_d are already defined, so the assignment for small documents (e.g. under 1000 words) is practically instantaneous.

The key point is that, for each document, LDA assigns a vector of topic weights in a similar fashion as Principal Components Analysis (PCA) does with numeric data. In that sense, LDA is also a dimensionality reduction technique for unstructured text whereby a "signal" of N words is reduced to a vector of K topics and, in general, $K << N$.

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. In most circumstances, this is not a relevant limitation. In our case, it is only necessary to preserve sentence order according to timestamps, so we keep the chronology of report messages as a sequence of different documents.

5. Data

5.1. Data description

The data available for this work consists of 2 years of records from all accidents that occurred in the expressways of Singapore, from January 2010 to December 2011. Each record is internally created at the traffic management center and can be originated by external information (driver's call, police, local traffic agents) and via traffic camera observation. The time delay between actual occurrence and the initial reporting is estimated to be below 5 minutes, according to local agents. The incident response is then continuously adapted to the situation as new information becomes available. Initially, the operator keys in the affected lanes, location data (road name, coordinates, distance to on/off ramps, zone ID), direction, traffic status (congested/non congested), and number of vehicles affected. Then, as new information comes in, it is gradually inserted in textual form, together with a time tag. When the accident is fully cleared, the report is closed.

Below is an example of an incident record[2].

```
id: 473586
zone ID: 2
Location (X, Y): 26266.6, 34916.9
Road name: AYE
```

[2]Due to data non-disclosure compromises, this case does not correspond to a real occurrence although it strictly exemplifies the content of original data.

```
Location type: 3
Lane blockage: Lane 1, Shoulder blocked
Down point: 20.32
Congestion status: 0
Queue length: 500m
Start time: 2010-08-20 22:50:01
End time: 2010-08-20 23:31:45
Number of vehicles: 2
2250hrs - TP Joe X spots an accident. car and
          bike involved.
2255hrs - Passers-by shift the bike to the
          shoulder.
2300hrs - Ambulance arrives at location. LTM
          arrives at location.
2309hrs - Ambulance conveys rider to National
          University Hospital.
2310hrs - TP arrives at location.
2311hrs - Notify by LTM the rider is seriously
          injured. The accident involves a car
          and bike.
2331hrs - TP requests RC and LTM to resume
          patrolling. All other vehicles move
          off. Shoulder clear.
```

From literature review ([25, 22, 33]), discussions with traffic authorities and data inspection, we established an upper limit of 180 minutes, leaving out nearly 100 cases, which are either particular outliers or in fact reports wrongly left open. After some further data cleaning, where we eliminated cases without any text or explicitly wrong (e.g. wrongly created, duplicated, zero duration), we obtained a total of 10139 accident cases for the expressways, each one having up to 16 sequential messages, unevenly spaced in time. Figure 1 presents the resulting distribution. One interesting observation is that, quite differently from some earlier literature that mentions distribution fittings to lognormal (e.g. [18, 15, 14]), our data indicates a strong fit to an exponential distribution ($\lambda = 0.029, p < 0.01$).

Differently to other earlier studies, we did not have access to traffic data for the same entire period, therefore our model will not consider such type of information. The question of how much the model would improve with that additional information will be left open for future work.

5.2. Data preparation

Our feature set comprises two types of data: the basic set of values originally created by the operator (location, lanes blocked, etc.) together with values computed from them; and topic distributions assigned by our LDA algorithm. Regarding the basic set, we started by converting the nominal attributes (e.g. zone id, road name) into dummy binary variables, and normalizing the numeric ones. Then, we computed a few other features that should contribute to the prediction. For each record we calculated the number of accidents that occur at several distances (100m, 1000m, 5000m, same road) during a time window before the current accident. This follows the recommendation from Khattak et al. [22], on the interactions between different accidents that are close in time

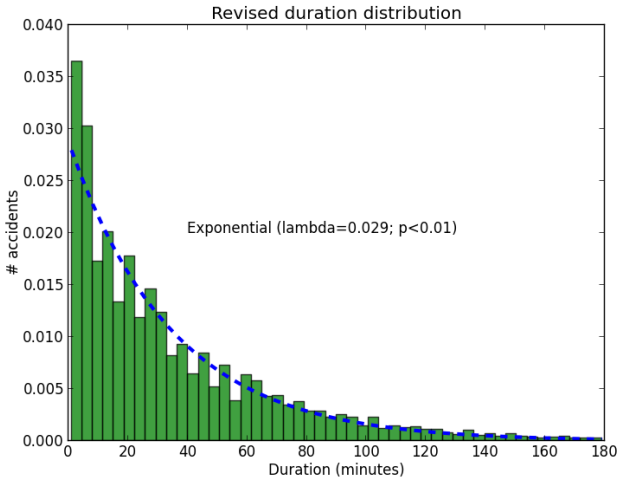

Figure 1: Distribution of revised durations.

and space. We also identified the type of day of week and time of day (peak, non-peak, morning, night, etc.), and calculated a simplified capacity reduction value by dividing the number of affected lanes by the total number of lanes in the affected area.

The second data type, the topic assignments from the text, is the subject of the next section. For now, we explain its data preparation. LDA receives as input a set of documents in the bag-of-words representation mentioned above. Thus, now we explain the task pipeline that generates, for each incident text description, its corresponding bag-of-words vector:

- timestamp matching: The timestamp is actually inserted by hand by the operator, varying its form in many ways, sometimes not exactly in the beginning of the line. For example using colon (":") or not. We apply regular expressions to match these patterns;

- abbreviation recognition: Given the nature of the messages involved, we can easily match some common abbreviations and acronyms. For example, the word for ambulance can either be "amb", "ab", "ambul" or "ambulance", the word "m/cycle" means "motorcycle" and so on. We identified less than 20 of such sets of synonyms. This task is important because a more uniform set of words will increase LDA quality;

- treatment of negation: Treating negation is a typical problem in natural language texts. In this case, fortunately the texts are objective enough such that a simple concatenation will generate the relevant word. For example, "no injuries" becomes "noinjuries";

- stop word removal: It is also important to remove all words that have no semantic relevance for the problem, as for example articles, prepositions, pronouns, or common verbs such as *to have, to be* or *to do*;
- stemming: Stemming is the process of reducing each word to its word family root. For example "seriously injured" and "serious injuries" both reduce to "serious injur";
- dictionary construction: The above processes aim to reduce a wide set of words into a smaller set of relevant words to be catalogued in a dictionary. Each word in this dictionary will thus correspond to an index in the bag-of-words representation.

Except for the first and last parts, these tasks are all typical of NLP and there are plenty of available open source tools to use such as NLTK [9].

To illustrate this process, we now transform the text example given above into the set of words generated, with corresponding timestamps:

```
[0: 2250] tp joe x spot acc veh bike involv
[5: 2255] passer shift bike sh
[10: 2300]  ab arr ltm arr
[19: 2309] ab convei rider nation univers hosp
[20: 2310]  tp arr
[21: 2311]  notifi ltm rider serious injur acc
            involv veh bike
[41: 2331]  tp request rc ltm resum patrol veh
            move off sh clear
```

The last step of data preparation for LDA is then to generate the bag-of-words vector. In our example, we only have 26 different words, and our dictionary has 3545 words, which leads to very sparse vectors. For this reason, its representation is generally in compressed format as a list of pairs <word id: count>.

5.3. LDA training and topic assignment

The training of the LDA model requires defining three parameters, namely K, α and η. The latter two are in fact vectors of dimension K and W, respectively, the general practice being to assume the same value for each element, often $1/K$ for both cases. Regarding K, it represents the set of different topics that we expect to be present in a message. The training and assignment procedure is then an iterative process using the variational EM algorithm (see [10] for more details and [32] for a stable python package that applies this method).

To get a small illustrative example, we trained an LDA model on our dataset, with $K = 6$, $\alpha = 1/6$ and $\eta = 1/6$. Below, we show the list of all 6 topics obtained. Due to space limitations, we restrict each topic to its top 6 words.

tp #0: 0.15*noinjur + 0.09*veh + 0.08*noinfst + 0.07*came + 0.06*across + 0.05*selfdriven

tp #1: 0.11*open + 0.09*spill + 0.09*oil + 0.09*scdf + 0.08*close + 0.06*2ln

tp #2: 0.06*polic + 0.06*drive + 0.04*crew + 0.04*drink + 0.03*div + 0.03*driver

tp #3: 0.07*ab + 0.07*tow + 0.06*convei + 0.06*rider + 0.06*hosp + 0.05*bike
tp #4: 0.08*damag + 0.07*call + 0.04*tow + 0.04*veh + 0.03*tp + 0.03*vig
tp #5: 0.06*tp + 0.05*convei + 0.05*ab + 0.04*tow + 0.04*hosp + 0.04*veh

We can recognize that topic 0 represents minor accidents without injuries in which vehicles are self driven out of the area; topic 1 relates to oil spillages; topic 2 is about drink driving cases; topics 3 and 5 involve accidents with injuries ("ab" is ambulance) with bike and car, respectively; and topic 4 relates to infrastructure damages. Applying the assignment procedure mentioned in Section 4, our earlier incident example would obtain the following topic assignment $(0.01, 0.02, 0.02, 0.32, 0.13, 0.50)$, meaning that our document can approximately be re-represented as $0.5 \times topic5 + 0.32 \times topic3 + \ldots$

To choose appropriate values for K, α and η, we trained a set of linear regression models for incident duration prediction. The features were defined in the previous section, and each model would only differ from the others in the combination of K, α and η. Since the model estimation is computationally costly, we randomly sampled from combinations of these values. For the best results of correlation coefficient and mean average error, we further explored neighbor values. The best configuration was with $K = 25$, $\alpha = 0.5$ and $\eta = 0.75$ (see table 3 in Appendix for more details). Unless otherwise stated, these values will be assumed throughout the rest of this article. We should note, however, that we only verified significant sensitivity to different values of K, while variations in α and η lead to minimal changes in the results. We also ran experiments with values much larger than 1 and did not find meaningful difference. This is not surprising since α and η are simply parameters of prior distributions for θ and β, respectively, and these are adjusted during the training phase. The fact that they converge to very similar values independently of their prior indicates that the dataset is sufficiently large with regards to robustness of the LDA process.

6. Prediction models

6.1. One time prediction

The one time prediction model assumes that all information is available when the incident is reported. This assumption was followed implicitly or explicitly by some of the earlier studies (e.g. [18, 33]), mostly to understand the predictive power of individual features. We will follow the same reasoning and leave the more realistic scenario of sequential prediction for the next section.

To maximize consistency with the state of the art, we use the same models as Valenti et al [33], which is a comprehensive set that is quite popular for the subject: linear regression (LR), support vector regression (SVR/RVM), artificial neural networks (ANN), and decision/regression trees (DT). We also add radial basis functions (RBF). Also following earlier literature, performance is assessed in terms of correlation coefficient between observed and predicted (CC) and Mean Average Error (MAE). Similarly to Boyles et al [11], we also use the median error (ME), which is less sensitive to outliers.

Given that we have an exponentially distributed target variable, we applied a logarithmic transformation to approximate it to a normal distribution during the training process. Of course, for the purposes of evaluation we transformed the results back to the original values, and recalculated all measurements on this scale.

We have four sets of feature configurations: *basic set* (which includes original and computed attributes); *initial message* (basic set+first message); *full message* (basic set+complete report); *only text* (no data from basic set). We use 10-fold cross validation in all experiments. Tables 1 and 2 show the results.

For illustrative purposes, we also included RBF*, where we test the RBF model against the training set, which should represent in practice the best possible values given the data.

Table 1: Comparison of different one time prediction models (Correlation Coefficients, all $p < 0.05$).

Model	basic	initial msg	full msg	only text
LR	0.45	0.61	0.74	0.63
SVR	0.39	0.44	0.49	0.5
ANN	0.28	0.44	0.54	0.64
DT	0.46	**0.62**	0.74	0.71
RBF	**0.48**	**0.62**	**0.75**	**0.73**
RBF*			0.96	

Table 2: Comparison of different one time prediction models. MAE (ME in parenthesis) are in minutes.

Model	basic	initial msg	full msg	only text
LR	22.2 (17.3)	18.4 (13.0)	16 (11)	22.3 (17.1)
SVR	22.1 (**14.2**)	20.9 (**12.6**)	21.2 (12.8)	20.4 (11.8)
ANN	28.3 (21.9)	23.6 (15.4)	21.9 (13.1)	19.3 (14.5)
DT	22.2 (17.2)	18.3 (12.7)	15.9 (10.7)	16.4 (10.9)
RBF	**21.9** (16.6)	**18.2 (12.6)**	**15.5 (10.5)**	**15.9 (10.4)**
RBF*			4.4 (2.6)	

From tables 1 and 2, we can see that the radial basis function outperforms the others in almost all cases, being closely followed by the linear regression and regression tree models. Notice also that, although the SVR model shows a very low correlation coefficient, it is competitive in terms of mean and median errors particularly in the models with less text information, which would be coherent with the findings from Valenti et al [33]. However, with a closer look at the results, we verified that it is highly biased towards durations between 0 and 50 minutes while completely ignoring higher ones.

The major performance increment is between the basic (best CC=0.48; MAE=21.9; ME=14.2) and the initial message model (best CC=0.62; MAE=18.2; ME=12.6), which means that the topics extracted from the initial text alone provide significant information, even though it is often very incomplete at that early stage.

Another interesting point is that, if we are left only with text information, our model seems to be lightly affected. From the point of view of the sequential model, this becomes a big advantage in that it eliminates the dependency on having a minimal set of features from the beginning. It could rely solely on a textual message feed. We will explore this aspect in the next section.

In order to understand the error relative to observed duration, a common measure has been the Mean Absolute Percentage Error (MAPE), defined as:

$$\text{MAPE} = \frac{1}{|M(t)|} \sum_{i \in M(t)} \left| \frac{\hat{x}_i - x_i}{x_i} \right| \times 100\%$$

where $M(t)$ corresponds to the set of measurements that occurred in time t and x_i and $\hat{x}_i$ the observed and predicted values, respectively. Figure 2 shows the MAPE value by duration. We highlight the upper thresholds proposed by Lewis [24], of 50%, 20% and 10% to qualitatively assess prediction performance, as "reasonable", "good", and "highly accurate", respectively. Although such type of thresholds are always very arbitrary, we use them to better compare the different models, as with earlier literature (e.g.[34]). According to these criteria, our model has "reasonable" performance for incidents between 20 and 110 minutes in total duration, and it outperforms the model without text analysis except for incidents that have a duration between 15 and 30 minutes.

From these experiments, we can conclude that the models that use text analysis are significantly more correlated with the data than those with no text. To provide further insight on the comparative role of information, we analysed the parameter coefficients from the linear regression model as well as the word distributions in topics. We can identify topics that strongly suggest longer duration (related to injuries, oil spillage, need for towing vehicle) while others go on the opposite direction, such as when the driver is able to self-drive the car off the road. The topics, regression coefficients and other statistics are in Tables 4 and 5 of the Appendix, respectively.

Despite the better results of our text-analysis approach, the amount of error involved is still considerable which is not surprising given the high variance of the duration throughout the entire dataset. Figures 3 and 4 illustrate this discussion and let us visualize the difference between the basic and the more complete model. We also point out its performance in terms of 10 and 20 minutes interval accuracy. For example, the full message model predicts duration with error less than 10 minutes in 54.5% of the times.

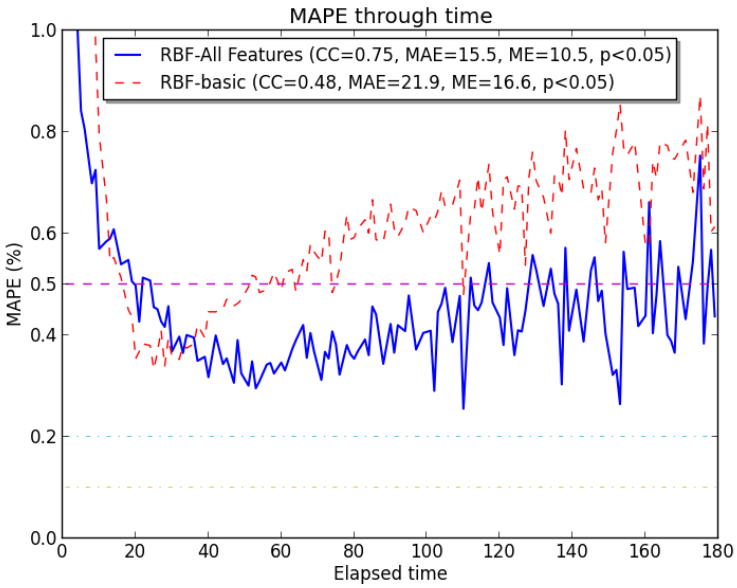

Figure 2: MAPE performance of models with and without text analysis features.

6.2. Sequential prediction

In reality, only little information is available when a report is created, and as time progresses it becomes continuously updated. As mentioned in section 5.2, we can extract message update timestamps that we can use to simulate information flow in time. Regarding the other features (the *basic* and *computed sets*), except for those objectively related to location and time, we cannot exactly determine their recording times from the reports since those fields might have been overwritten during the incident resolution. We will explore two assumptions, one where all information is available at the beginning (all features model), and the other where such information is not available at all (conservative model). We reckon that the latter is utterly pessimistic, since all incidents will have some intermediate moment where information such as number of blocked lanes or number of vehicles involved is available.

Since temporal sequence is now important, the experimental design differs considerably from the previous section. We split the dataset into two parts: a training set comprising one year and nearly 2 months (66% of the overall dataset) and a test set with the last 10 months (33% of the dataset). Each case is now split into a sequence of vectors, each one tagged with its elapsed time and topic assignment from total text available at that moment. Thus, each incident will have a topic assignment that will "evolve" as time advances.

We maintain the same algorithm configuration parameters: $K = 25; \alpha =$

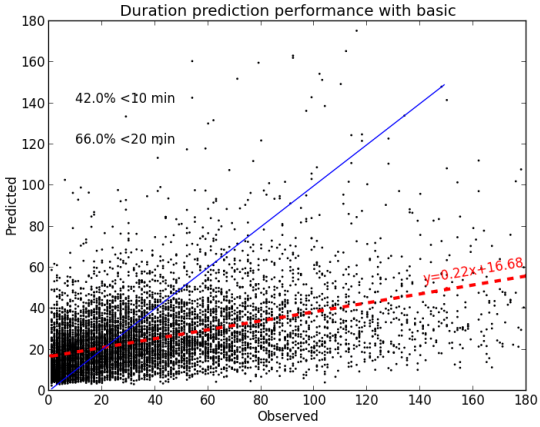

Figure 3: Best model with basic features (Radial Basis Function). Blue 45 degree line indicates ideal performance. Red dashed line corresponds to linear fit to observations.

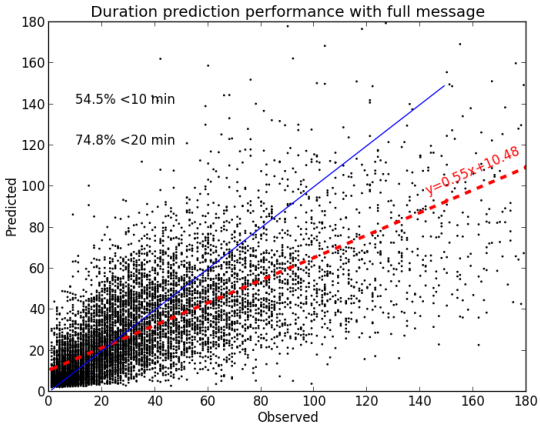

Figure 4: Best model with full message (Radial Basis Function).

$0.5; \eta = 0.75$; RBF algorithm; and logarithm of duration as target. We will test three model types: the *basic set* model, which also includes basic and computed attributes; the *all features* model, which has all available features, including the topic assignments; and the *conservative* model, which only has text information (topic assignments) as well as incident location and reported start time.

As time progresses, these models generate prediction updates whenever new text information is received, also taking into account the elapsed time. For comparison purposes, at each of these moments, we also obtain prediction updates from the basic set model.

Intuitively, the quality of predictions should gradually improve as new information becomes available, particularly when also considering the elapsed time. In Figure 5, we present the MAPE diagram through time. The difficulty in predicting for lower than 15 minute durations together with the lack of information in the beginning turn all models effectively useless during that initial period. Throughout the entire period, the model with complete information outperforms the one without text features and it is notable that after minute 120. It is also interesting to observe that the conservative model, which has only textual features and their time stamps, has good performance during the first hour after which it has an erratic behavior. This result is not trivial to interpret but it suggests that, after that period, a combination with other factors not obtainable from textual topics (e.g. capacity reduction, time of day) is necessary to discriminate between different incident duration expectations.

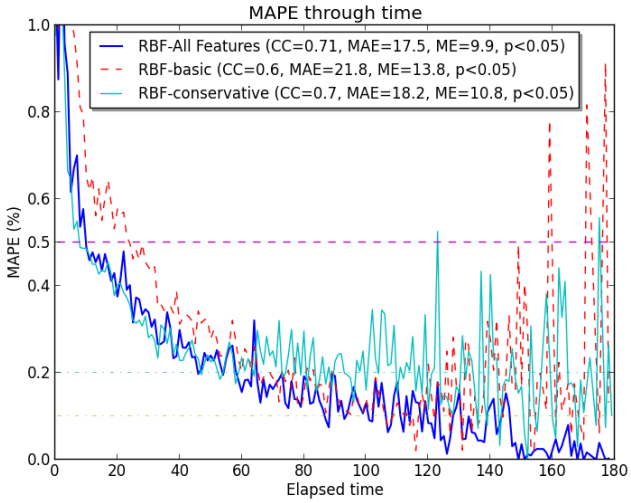

Figure 5: Mean Absolute Percentage Error through time.

To illustrate the performance improvements as a function of the number of messages received, we plot different MAPE results in Figure 6. Now we show

the incremental progress of both models that use text-based features, while receiving messages 1 to 16. With arrival of each new information, the model is able to improve gradually. Notice that there is not necessarily a direct mapping between number of messages and elapsed time, which complicates a connection with the previous figure. Sometimes, during the first five minutes we observe many messages, and in other cases message updates arrive at a much lower pace.

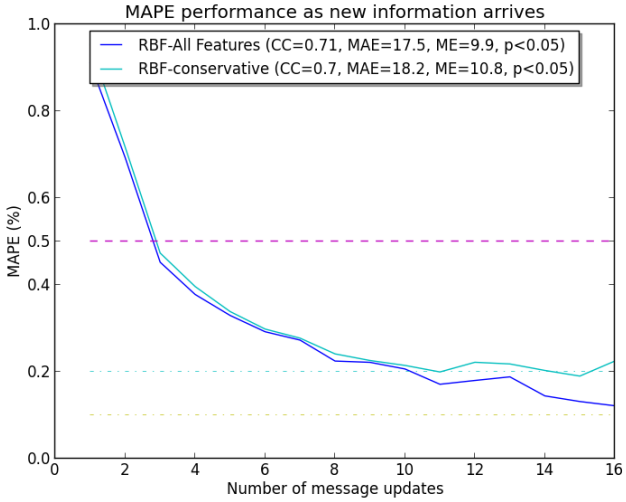

Figure 6: Mean Absolute Percentage Error with number of messages.

An interesting change of perspective on these models is given by counting down the time to clearance. This provides a plot where the timeline is reversed and the first prediction is placed at $t = duration$ since it will correspond to the total remaining time to clearance. Hence, this lets us see the performance with regards to a key indicator: remaining time. In Figure 7, we compare four models, *basic set*, *all features*, *conservative* and *full message*. The latter, shown as a dashed purple line, will serve as the best possible baseline where the full text report is available at initial reporting time. In fact, this corresponds to the best one time prediction model from Section 6.1, spread along the reverse timeline. The goal is to enhance the importance of timeliness.

We can consider the basic and full message models as the upper and lower error bounds, respectively, with respect to timeliness of message updates. For example, if we systematically delay the message updates, the blue and cyan lines (all features and conservative, respectively) will come closer to the basic model performance. The intuition is simple: if the message arrives too late, text features will not add relevant information in comparison to the elapsed duration itself. On the other hand, if we obtain the messages earlier in time, we improve

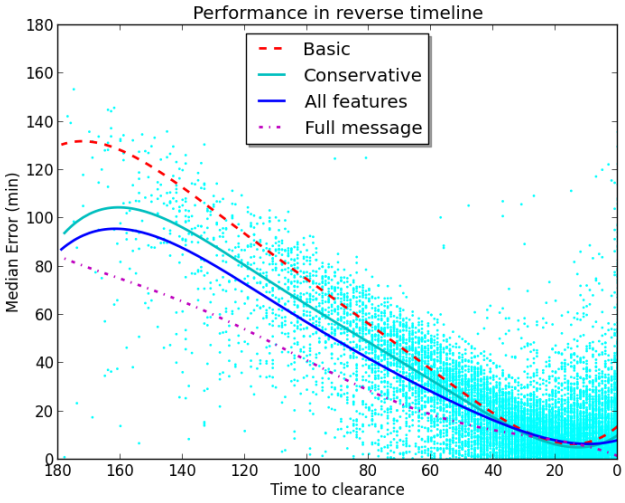

Figure 7: Performance of 4 models with respect to remaining time to clearance. Light colored dots represent prediction errors for the *conservative* model to illustrate the spread of the errors, which is visually similar to all models. The four lines are 5 degree polynomial fits to respective data.

the predictions, in the limit being as good as the *full message* model, which has all information at time zero. This lets us visualize the practical benefits of a more efficient workflow of the reporting process at the traffic management center.

Another observation is that the last 15 minutes before clearance present a general decline in the median/mean performance of the models (except for the model with full message). Again, the problem lies in the large set of incidents that last *less than* 15 minutes, that now concentrate on the far right of the graph. A possible solution to this problem would be a two-tier model where a binary classifier first distinguishes between incidents with more or less than 15 minutes duration, and then two regression models estimate the duration.

Finally, Figure 8 illustrates the performance through time of the model with all features in terms of absolute error. Each of the 16018 individual points indicates a prediction error obtained at a specific moment in time. Each incident thus generates as many predictions as message updates, including the last one (often a clearance or closing notification). The error is highly skewed to the higher durations due to exceptional outliers, thus the use of the median line. Interestingly, many such outliers correspond to repeated prediction failures for the same incident.

Relative to the model without topics (*basic*), the overall median error of the *all features* model decreases by 28% (9.9 vs 13.8). Except for the very long durations, we notice that there is always some residual error, of at least 5 minutes.

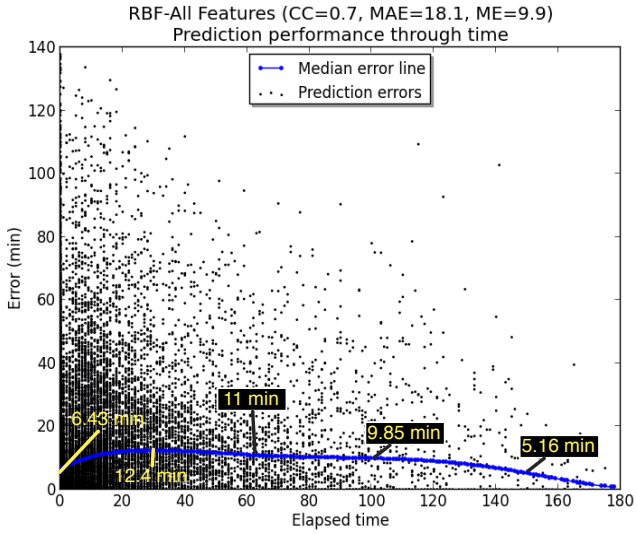

Figure 8: Performance through time.

Although this may be mostly related with the inherent statistical heterogeneity of this problem, our intuition is that there is potential for improvement, namely by adding new information, such as traffic conditions, distance to emergency response bases or weather status.

7. Conclusions and future work

This paper presented a machine learning framework that uses topic modeling, a text analysis technique, in addition to typical features (number of vehicles involves, lanes blocked, etc.) to generate predictions of accident duration. The model continuously generates predictions as new information becomes available in form of textual updates to the incident report.

We evaluated extensively the range of possible parameters for our topic modeling technique, called *latent dirichlet allocation* (LDA) [10], and concluded that our report messages can be re-represented as combinations of 25 different topics. Following earlier literature [33], we evaluated several machine learning models, and for our context the Radial Basis Function revealed the best results. We compared predictions with and without text analysis using several different models. For a single prediction with full information, our model could achieve a correlation coefficient of 0.75 and a median error of 10.5 minutes, as opposed to 0.48 and 16.6 minutes for the best model without text analysis, representing a decrease in the error of more than 35% and a relevant improvement in terms of the overall behaviour of the model. For the sequential prediction, it is able

to generate reliable predictions after 15 minutes and consistently improves the estimates as new information arrives, gradually reducing the median error from 12.5 to less than 5 minutes after 150 minutes of elapsed time. Comparing the sequential models with and without topics, the overall median error decreases by 28%.

The model here presented did not have access to traffic sensing information, which means that it didn't take into account the traffic conditions. However, it is arguable that, together with text-based evaluation of accident severity, such information should help the model increase further its accuracy. For example, if there exist injuries or oil spillage, the traffic conditions on the roads that connect to the emergency support services will influence the clearance time. Thus, a natural next stage of this research is to include a real-time information feed that reveals the traffic conditions in the relevant areas.

This methodology is applicable to other areas within transportation beyond incident scenarios. In fact, it applies to any other context where relevant information is available in natural language form. For example, in the Internet, we may find traffic information in text form (e.g. from Twitter or Waze) as well as special events announcements (e.g. Facebook, Eventful). Verbal communications, such as radio stations, often carry last-minute information about the transport system before becoming available to authorities or detected by sensors. With appropriate analysis tools, one can extract uniquely rich information to feed applications in transport operations, planning and management.

Acknowledgments

The authors gratefully acknowledge Land Transport Authority of Singapore for providing the dataset. This research was supported by the National Research Foundation Singapore through the Singapore MIT Alliance for Research and Technology's FM IRG research programme and also by Fundacão para a Ciência e Tecnologia (FCT), reference PTDC/EIA-EIA/108785/2008.

We'd also like to thank Constantinos Antoniou, Yaniv Altshuler and Oren Lederman for their careful reviews.

References

[1] Benefits of Traffic Incident Management. National Traffic Incident Management Coalition (NTIMC), 2006.

[2] Unified Response Manual for Roadway Traffic Incidents. FHWA, USDOT, 2006.

[3] Manual on Uniform Traffic Control Devices. U.S. Department of Transportation, FHWA, 2009.

[4] Highway Capacity Manual. Transportation Research Board (TRB), Washington, DC: National Research Council, 2010.

[5] Measuring and Improving Performance in Incident Management. UC Berkeley Transportation Library, 2010.

[6] Traffic Incident Management Handbook. U.S. Department of Transportation, FHWA, 2010.

[7] Ahmed, M., Abdel-Aty, M.. A data fusion framework for real-time risk assessment on freeways. Transportation Research Part C: Emerging Technologies 2013;26:203–213.

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

[9] Bird, S.. Nltk: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive presentation sessions. Stroudsburg, PA, USA: Association for Computational Linguistics; COLING-ACL '06; 2006. p. 69–72.

[10] Blei, D.M., Ng, A.Y., Jordan, M.I.. Latent dirichlet allocation. Journal of Machine Learning Research 2003;3:993–1022.

[11] Boyles, S., Fajardo, D., Waller, T.. Naive bayesian classifier for incident duration prediction. In: Transportation Research Board. 86th Meeting: 2007, Washington, D.C. 2007. .

[12] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.. Indexing by latent semantic analysis. Journal of the American Society for Information Science 1990;41(6):391–407.

[13] Feyen, R.G., Eseonu, C.I.. Identifying Methods and Metrics for Evaluating Interagency Coordination in Traffic Incident Management. Intelligent Transportation Systems Institute Research and Innovative Technology Administration, 2009.

[14] Giuliano, G.. Incident characteristics, frequency, and duration on a high volume urban freeway. Transportation Research Part A 1989;23(5):387396.

[15] Golob, T.F., Recker, W.W., Leonard, J.D.. An analysis of the severity and incident duration of truck-involved freeway accidents. Accident Analysis and Prevention 1987;19(5):375395.

[16] Hofmann, T.. Probabilistic latent semantic indexing. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM; SIGIR '99; 1999. p. 50–57.

[17] Hu, J., Chan, Y.. Stochastic incident-management of asymmetrical network-workloads. Transportation Research Part C: Emerging Technologies 2013;27(0):140 – 158.

[18] Jones, B., Janssen, L., Mannering, F.. Analysis of the frequency and duration of freeway accidents in seattle. Accident Analysis and Prevention 1991;23(2):239–255.

[19] Kamga, C.N., Mouskos, K.C., Paaswell, R.E.. A methodology to estimate travel time using dynamic traffic assignment (dta) under incident conditions. Transportation Research Part C: Emerging Technologies 2011;19(6):1215 – 1224.

[20] Karlaftis, M.G., Latoski, S.P., Richards, N.J., Sinha, K.C.. Its impacts on safety and traffic management: An investigation of secondary crash causes. Intelligent Transportation Systems Journal 1999;5:39–52.

[21] Khattak, A., Schofer, J.L., Wang, M.H.. A simple time sequential procedure for predicting freeway incident duration. IVHS J 1995;2(2):113138.

[22] Khattak, A., Wang, X., Zhang, H.. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays. IET Intelligent Transport Systems 2012;6(2):204214.

[23] Knoop, V.L., Hoogendoorn, S.P., van Zuylen, H.J.. Capacity reduction at incidents: Empirical data collected from a helicopter. Transportation Research Record: Journal of the Transportation Research Board 2008;(2071):19–25.

[24] Lewis, C.D.. Industrial and Business Forecasting Method. Butterworth Scientific, London, 1982.

[25] Lopes, J.A.. Traffic prediction for unplanned events on highways. Ph.D. thesis; Instituto Superior Tecnico, IST; 2012.

[26] Lou, Y., Yin, Y., Lawphongpanich, S.. Freeway service patrol deployment planning for incident management and congestion mitigation. Transportation Research Part C: Emerging Technologies 2011;19(2):283 – 295.

[27] Miller, M., Gupta, C.. Mining traffic incidents to forecast impact. In: Proceedings of the ACM SIGKDD International Workshop on Urban Computing. New York, NY, USA: ACM; UrbComp '12; 2012. p. 33–40.

[28] Nam, D., Mannering, F.. An exploratory hazard-based analysis of highway incident duration. Transportation Research Part A: Policy and Practice 2000;34(2):85102.

[29] Noland, R.B., Polak, J.W.. Travel time variability: A review of theoretical and empirical issues. Transport Reviews 2002;22(1):39–54.

[30] Ozbay, K.M., Xiao, W., Jaiswal, G., Bartin, B., Kachroo, P., Baykal-Gursoy, M.. Evaluation of incident management strategies and technologies using an integrated traffic/incident management simulation. World Review of Intermodal Transportation Research 2009;2(2/3):155–186.

[31] Papadimitriou, C., Raghavan, P., Tamaki, H., Vempala, S.. Latent semantic indexing: A probabilistic analysis. In: Proceedings of ACM PODS. 1998. .

[32] Řehůřek, R., Sojka, P.. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. p. 45–50.

[33] Valenti, G., Lelli, M., Cucina, D.. A comparative study of models for the incident duration prediction. European Transport Research Review 2010;2:103–111. 10.1007/s12544-010-0031-4.

[34] Wei, C.H., Lee, Y.. Sequential forecast of incident duration using artificial neural network models. Accident Analysis and Prevention 2007;39(5):944 – 954.

[35] Yang, S.. On feature selection for traffic congestion prediction. Transportation Research Part C: Emerging Technologies 2013;26:160–169.

APPENDIX

7.1. LDA plate notation

Figure 9 depicts the LDA model in plate notation. This is a common graphical model representation, particularly useful for generative models. It allows us to see the dependencies and groupings between different variables. The dark shaded nodes represent observed variables. The boxes (the "plates") represent multiple iterations of the same sub-graph. For example, there exist a set of K samples of β_k, drawn from a (dirichlet) distribution parameterized by η.

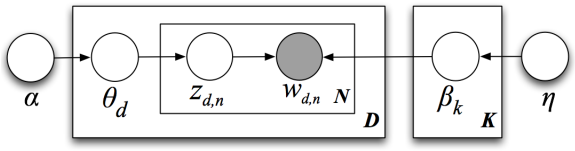

Figure 9: Plate representation for the LDA model.

D denotes the number of documents, N the number of words in each document, β_k represents the word distribution of a topic, θ_d represents the topic distribution of a document, $z_{d,n}$ the topic assigned to word n of document d, and $w_{d,n}$ the word itself, which is determined by the topic $k = z_{d,n}$ using a multinomial distribution with parameter β_k.

7.2. LDA parameter analysis

To select an appropriate set of values for K, α and η, we sampled sequential prediction runs with $K \in \{5, 10, 15, 18, 25, 30\}$, and $\alpha, \eta \in \{0.01, 0.25, 0.5, 0.75, 1\}$. We chose sets of fixed values to easily compare pairs of runs with only one differing parameter. We defined a limited number of different values to constrain the search space. In this way, we have $6 \times 5 \times 5 = 150$ possible runs. For practical reasons, we did not run all of them, and we used linear regression instead of slower algorithms such as Radial Basis Functions.

For the best case ($\alpha = 0.5, \eta = 0.75$), we further explored the neighborhood values. We chose $K = 25$ instead of $K = 30$ since the latter brings a negligible gain (of 0.1 minutes) at the expense of more 5 variables. Notice that the resulting values are inferior to the final ones reported in the paper (obtained with the RBF), which would be expectable. To certify the overall consistency of the results, we also re-ran some of the sampled combinations, confirming similar performance relative to the chosen set of parameters ($K = 25$, $\alpha = 0.5, \eta = 0.75$).

7.3. One time model linear regression

Tables 4 and 5 present the coefficients of the LR model and the LDA topic list, respectively. The LR algorithm used also applies feature selection, using the M5's method, which steps through the attributes removing the one with

the smallest standardised coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion (AIC).

We can identify topics that are relevant, namely 1, 4, 6 and 14 suggest longer duration (related to injuries, oil spillage, need for towing vehicle) while others suggest shorter times such as when the driver is able to self-drive the car off the road. We remind the reader that the target variable is in logarithm form so a strong negative value (as in topic 17) simply forces to predict zero duration. We can also note that some topics are essentially noise that effectively was dropped out by the linear regression model, for example topics 3, 19 and 20.

Table 3: Selection of hyper parameters

K=5 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	0.57		0.57		0.57
0.25					
0.5	0.57		0.56	0.57	
0.8	0.57				
1	0.56	0.57			

K=5 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	19.1		19.0		19.0
0.25					
0.5	18.9		19.1	19.0	
0.8	18.9				
1	19.0	18.9			

K=10 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	0.56				
0.25					
0.5				0.56	
0.8	0.57				
1					

K=10 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	19.1				
0.25					
0.5				19.2	
0.8	19				
1					

K=15 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01				0.58	
0.25					
0.5				0.57	
0.8	0.57				
1			0.56		

K=15 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01				18.9	
0.25					
0.5				19	
0.8	18.8				
1			19.2		

K=18 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	0.58			0.58	0.56
0.25				0.6	
0.5	0.57	0.58	0.58	0.6	0.6
0.8	0.57			0.56	

K=18 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	18.9			18.9	19.1
0.25				18.5	
0.5	18.9	18.9	19	18.4	18.4
0.8	19			19.2	

K=25 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	0.58	0.58			
0.25				0.61	
0.5		0.58	0.59	0.61	0.61
0.8		0.58		0.59	
1				0.57	

K=25 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01	18.7	18.7			
0.25				18.3	
0.5		18.8	18.4	18.2	18.2
0.8		18.8		18.6	
1				19	

K=30 CC

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01		0.58		0.58	
0.25	0.58				
0.5			0.61	0.61	0.61
0.8					
1		0.58			

K=30 MAE

alpha \ Eta	0.01	0.25	0.5	0.75	1
0.01		18.6		18.9	
0.25	18.6				
0.5			18.1	18.1	18.2
0.8					
1		18.8			

Table 4: Linear regression results for $K = 25$, $\alpha = 0.5$ and $\eta = 0.75$ (codes=*$p < 0.1$; **$p < 0.05$; ***$p < 0.001$; ****$p << 0.001$)

Attribute	Coefficient	Std. Error	Std. Coeff.	Tolerance	t-stat	p-value	Code
time_of_day = night	0.264	0.034	0.598	0.943	7.87	0.0	****
time_of_day = morn_peak	-0.051	0.033	-0.109	1.0	-1.565	0.148	
time_of_day = aftrn_npeak	-0.057	0.035	-0.162	0.995	-1.621	0.131	
time_of_day = aftrn_peak	-0.069	0.032	-0.146	0.994	-2.129	0.038	**
time_of_day = dinner	0.024	0.035	0.064	0.999	0.685	0.499	
time_of_day = morn_npeak	-0.049	0.033	-0.112	0.999	-1.466	0.184	
CONGT_STATUS	0.382	0.03	0.624	0.983	12.713	0.0	****
NUM_BLOCK_LANES	0.056	0.035	0.044	0.994	1.593	0.139	
capacity_reduction	0.396	0.063	0.366	0.988	6.295	0.0	****
block_shoulder	0.258	0.038	0.36	1.0	6.73	0.0	****
qlength	0.014	0.01	0.029	0.96	1.387	0.217	
y_coord	-0.003	0.001	-0.001	0.999	-2.998	0.003	***
topic1	2.333	0.092	2.713	0.713	25.293	0.0	****
topic2	-0.836	0.185	-0.997	0.993	-4.509	0.0	****
topic4	0.999	0.266	0.973	0.979	3.748	0.0	****
topic6	1.623	0.122	1.538	1.0	13.283	0.0	****
topic7	-1.482	0.221	-1.669	0.986	-6.722	0.0	****
topic8	-4.45	1.023	-1.901	0.808	-4.35	0.0	****
topic11	-0.878	0.232	-0.772	0.987	-3.792	0.0	****
topic14	2.834	0.157	3.262	0.904	18.065	0.0	****
topic15	-2.978	0.325	-2.614	0.95	-9.171	0.0	****
topic17	-17.566	1.397	-5.702	0.649	-12.569	0.0	****
topic18	-1.744	0.276	-1.598	0.981	-6.328	0.0	****
topic21	-2.261	0.104	-1.824	0.598	-21.812	0.0	****
topic22	-0.49	0.181	-0.629	0.996	-2.704	0.008	***
msg_size	0.001	0.0	0.001	0.644	20.076	0.0	****
same_road	0.054	0.012	0.111	0.973	4.55	0.0	****
less100	0.065	0.027	0.256	0.994	2.425	0.017	**
less5000	0.086	0.008	0.124	0.96	10.129	0.0	****
(Intercept)	2.779	0.066			42.209	0.0	****

Table 5: List of 25 topics inferred by our LDA model with $K = 25, \alpha = 0.5$ and $\eta = 0.75$ (we show the top 7 words)

Topic	word distribution
topic #1	0.085*tp + 0.071*convei + 0.070*ab + 0.064*tow + 0.062*hosp + 0.056*rider + 0.048*bike
topic #2	0.054*ir + 0.049*confirm + 0.045*congest + 0.034*traffic + 0.027*case + 0.023*soe + 0.018*heavi
topic #3	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #4	0.064*exit + 0.030*road + 0.025*hit + 0.022*wait + 0.020*io + 0.019*bu + 0.014*case
topic #5	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #6	0.138*tow + 0.122*noinjur + 0.092*veh + 0.080*present + 0.066*clear + 0.061*owner + 0.047*nodamag
topic #7	0.074*scene + 0.053*imt + 0.019*rd + 0.018*check + 0.016*op + 0.015*inf + 0.014*inv
topic #8	0.024*taken + 0.012*photo + 0.010*drove + 0.010*nodetail + 0.008*told + 0.006*two + 0.006*avail
topic #9	0.288*vr + 0.113*ltm + 0.089*rta + 0.061*awai + 0.029*disp + 0.025*incid + 0.019*ln
topic #10	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #11	0.101*assist + 0.056*taxi + 0.045*notifi + 0.032*requir + 0.029*came + 0.029*p3 + 0.027*passeng
topic #12	0.076*polic + 0.041*otm + 0.033*2nd + 0.033*last + 0.023*div + 0.022*tm1 + 0.016*1st
topic #13	0.077*damag + 0.071*call + 0.043*skid + 0.034*tp + 0.029*near + 0.028*self + 0.027*vig
topic #14	0.072*left + 0.057*activ + 0.050*tp + 0.045*2ln + 0.040*scdf + 0.035*open + 0.029*spill
topic #15	0.160*minor + 0.049*exchang + 0.047*particular + 0.032*advis + 0.029*refus + 0.025*detail + 0.013*involv2
topic #16	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #17	0.011*sd + 0.001*bin + 0.001*wish + 0.001*skip + 0.001*veh + 0.001*tel + 0.001*design
topic #18	0.112*report + 0.093*rc + 0.063*tm + 0.041*btw + 0.036*lorri + 0.034*bef + 0.032*actv
topic #19	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #20	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #21	0.112*veh + 0.107*off + 0.086*move + 0.077*selfdriven + 0.071*1ln + 0.064*clear + 0.057*rc
topic #22	0.037*ado + 0.033*unabl + 0.033*resourc + 0.016*send + 0.016*spot + 0.015*ema + 0.014*activ
topic #23	0.001*veh + 0.001*nch + 0.001*spot + 0.000*rc + 0.000*clear + 0.000*ltm + 0.000*1ln
topic #24	0.031*tid + 0.017*messag + 0.015*msg + 0.010*manual + 0.010*test + 0.009*implement + 0.008*depatch
topic #25	0.077*alreadi + 0.049*selfskid + 0.035*a1 + 0.031*invl + 0.028*e2 + 0.026*em1 + 0.026*c1