Minining Traffic Incidents to Forecast Impact

Mahalia Miller
Dept. of Civil Engineering, Stanford University
& Hewlett Packard Labs
mahalia@stanford.edu

Chetan Gupta
Hewlett Packard Labs
chetan.gupta@hp.com

ABSTRACT
Using sensor data from fixed highway traffic detectors, as well as data from highway patrol logs and local weather stations, we aim to answer the domain problem: “A traffic incident just occurred. How severe will its impact be?” In this paper we show a practical system for predicting the cost and impact of highway incidents using classification models trained on sensor data and police reports. Our models are built on an understanding of the spatial and temporal patterns of the expected state of traffic at different times of day and locations and past incidents. With high accuracy, our model can predict false reports of incidents that are made to the highway patrol and classify the duration of the incident-induced delays and the magnitude of the incident impact, measured as a function of vehicles delayed, the spatial and temporal extent of the incident. Equipped with our predictions of traffic incident costs and relative impacts, highway operators and first responders will be able to more effectively respond to reports of highway incidents, ultimately improving drivers’ welfare and reducing urban congestion.

Keywords
Geomining, event analysis, traffic prediction, cyber-physical

1. INTRODUCTION
Traffic congestion causes business losses due to increased travel time, requires increases in public infrastructure investment, and due to extra emissions, threatens urban air quality. While work zones, weather, fluctuations in normal traffic, special events, traffic control, and physical bottlenecks can all lead to congestion, a key contributor to congestion is traffic incidents—“events that disrupt the normal flow of traffic, usually by physical impedance in the travel lanes” [2].

The wealth of data from transportation sensor networks offers the opportunity to understand traffic incidents in order to make urban travel more efficient. Incident detection was among the first incident management problems to be addressed. Researchers have proposed machine learning and rule-based approaches for detecting incidents or outliers at individual locations, e.g. [8, 11]. A recent extension is to detect and identify causal relations between anomalous events in traffic data streams from GPS traces [12]. This work offers insights into understanding the spatio-temporal impact of traffic events. Another approach uses data from fixed sensors to build a model of recurrent traffic congestion and retrospectively identify the spatio-temporal impact of an incident [10]. This work leaves open some specific problems: (1) the definition of the impact region is based on a fixed threshold instead of values tuned to expected conditions at different locations and times of day and (2) the definition of the impact region is limited to only the highway on which the incident originated.

A second area in incident management is forecasting the severity of an incident and its impact. Analytical formulae based on queuing or shockwave models have been used, often demonstrated on simulator data without validation from sensor recordings, e.g. [3]. Khattak et al. [9] use sensor data to build regression models for incident duration, but forecast delay using analytical formulae. In contrast, Garib et al. [5] have applied regression on historical data for predicting incident delay. However, they use incident duration as a predictor variable for incident delay, which means that delay can be predicted only after the incident is cleared. This work suggests the challenge of forecasting the incident impact promptly after an incident is detected to enable the best possible emergency response. Possible responses include variable message signs, reducing flow of traffic on up-stream onramps, and prioritizing emergency responders.

In this paper, we propose a practical system for predicting the cost and impact of highway incidents using classification models trained on sensor data and police reports. We do this by applying machine learning models to the incoming stream of recordings from traffic sensors embedded in many highway systems. Our system allows for the rapid prediction of how severe a newly-detected incident will be by mining historical traffic sensor recordings as well as semi-structured police reports, weather data, and day of week to build a classification model of incident impact. The following are the paper’s key contributions:

- Using historical traffic data from thousands of sensors, we build a baseline model that captures the expected conditions at each location and time of day over the study area.
- We predict the impact of an incident that has just started by using a classification model trained on his-
torical events. We show that combining traffic sensor recordings, initial police reports, weather data, and time of day we are able to reliably predict absolute and relative impact of a highway incident. We measure the impact by the duration of incident-induced delays and by economic losses from cumulative travel time delay.

- We demonstrate the high predictive power of our model on two real datasets of reported highway incidents from different regions, each over a two-month period. We also show the high level of transfer learning possible by training our model on one region and testing the model on incidents from a different year and region.

Although in this work our framework is specifically applied to traffic, our approach points a way toward studying the impact of events in other operational domains in which we have sensor data and external incident (event) logs.

2. INCIDENT IMPACT COMPUTATION

We first introduce relevant definitions and then outline an algorithm to identify the impact region for a particular incident considering the true road topology.

2.1 Preliminaries

In this work, we analyze contiguous sections of freeway. We use data from the California Performance Measurement System where tens of thousands of loop detectors fixed along major highways statewide measure quantities such as velocity, density, and flow every 30 seconds. For incidents, we use data from the California Highway Patrol (CHP). When an incident is reported, CHP notes details such as when and where it occurred. We provide further details in Section 4.1.

We model the regional highway network as a spatial network graph, \( G = (U, E) \). Each node represents a detector, \( u_i \) (located at postmile \( x_i \)) and an edge \( e_{i,j} \) exists for every pair of detectors such that there is a single road segment between two sensors where the road segment is part of a highway with PeMS detectors. The edge \( e_{i,j} \) is directed such that \( j \) is upstream of \( i \), where upstream is defined as in the opposite direction of normal traffic flow. In practice, at detector \( u_i \), the vehicle count data is aggregated over all lanes to compute the total recorded flow in a given direction during a five-minute time window. This quantity is known as flow.

The average occupancy of the highway over all lanes in the given direction is denoted by \( \rho(i,t) \), with \( 0 \leq \rho(i,t) \leq 1 \), where 0 indicates an empty highway and 1 indicates the theoretical limit of full coverage of the road by vehicles. The flow-weighted mean velocity \( v(i,t) \) at location \( i \) and time \( t \) is the chosen measure for the vehicle speed at a given location \( i \) and time \( t \):

\[
v(i,t) = \frac{\sum_{k=1}^{N_i} v_k(i,t) q_k(i,t)}{\sum_{k=1}^{N_i} q_k(i,t)}
\]

where the speed \( v_k(i,t) \) for each lane \( k \) and at detector \( i \) and at time \( t \) is computed according to the algorithm described by Chen.

Critical for determining abnormal traffic behavior is the identification of recurrent traffic conditions. Kwon use a nearest-neighbor approach to find \( k \) days that had traffic conditions similar to the traffic conditions just before the start of an incident of interest. The average of the conditions at \( k \)-days is then taken to be the recurrent (or normal) behavior. For small \( k \), the method leads to the possibility of choosing a date in the past with an incident—the opposite of trying to find recurrent non-incident conditions. Instead of using nearest-neighbors, we calculate the median of the speeds at each sensor location \( i \) and each point in time \( t \) and call it the recurrent velocity, \( v^*(i,t) \). The median is robust to outliers, i.e. to incidents.

2.2 Understanding Cost of Delay

In this work, the economic loss due to lost productivity associated with an incident is directly proportional to the cumulative incident delay value according to the cost-benefit framework used in the State of California. Thus, we will now detail the computation of the cumulative delay for all drivers in an incident impact region. The delay on each road segment per unit of time is defined as the additional vehicle-hours traveled driving below free-flow speed, \( v_{ref} \), per unit of time. In this study \( v_{ref} \) is 60 mph as is standard with other researchers using similar data. This analysis further assumes that the flow, occupancy, and speed remain constant during every five-minute interval per road segment. We begin with the standard domain definition of delay \( d(i,t) \) in the road segment of length \( l_i \) starting at detector \( i \) at time \( t \). The delay essentially measures the vehicle-hours lost due to traveling below reference velocity:

\[
d(i,t) = l_i \times q(i,t) \times \max \left( \frac{1}{v(i,t)} - \frac{1}{v_{ref}}, 0 \right)
\]

where \( v(i,t) \) refers to velocity at location \( i \) and time \( t \) as discussed in 2.1. The flow at location \( i \) and time \( t \) and \( v_{ref} \) is considered to be constant for all time \( t \) and location \( i \) with a value equal to 60 mph.

We are interested in the cumulative delay for all affected drivers, not simply the per segment per unit time delay, \( d(i,t) \). Thus integrating \( d(i,t) \) over a region \( A \) and time interval \( T \) the total delay over the spatial and temporal region defined by \( A \) and \( T \) would be given as

\[
D_{tot} = \int_A \int_T d(i,t) dudt.
\]

The total delay over region \( A \) and time \( T \) can have many causes. An approach as described by Kwon and Varaiya is to separate total delay as follows:

\[
D_{tot} = D_{inc} + D_{rec} + D_{rem}
\]
where $D_{tot}$ is total delay calculated from flow and speed, $D_{inc}$ is the delay caused by non-recurring incidents such as collisions ($D_{nc}$ is referred to as $D_{col}$ in [10]), $D_{rec}$ is the recurrent delay and $D_{rem}$ is the remaining delay. Of our interest is $D_{nc}$, which is computed over a contiguous region called the impact region of an incident. The work of previous researchers to divide delay into components offers a general framework, but there is no work to date that mathematically derives expressions that capture the different terms. We now provide a derivation of the various components of delay.

We begin by defining an impact region, which we understand to be a contiguous region (contiguous in both space and time) beginning at the time and location of the incident, and where the traffic conditions deviate from the recurrent traffic patterns. This contiguous region goes forward in time and upstream in space, since after an incident, traffic backs up (Figure 1 illustrates the spread of the impact region with time for an example incident). Precisely, let the closest upstream sensor to incident $I_a$ denote by $u_a$ and the time of incident $I_a$ be $t_a$. Then:

**Definition 1.** The impact region is collection of connected subgraphs $A_t$ induced from $G$. $A_t$ is constructed for each time $t$ starting at $t = t_a$, such that for all nodes $i \in A_t$, either $v(i, t-1) < v^*(i,t-1)$ or $v(P(i,t-1)) < v^*(P(i,t-1))$ where $P(i)$ is the parent of the detector $i$. At $t_a$, $A_{t_a}$ is $u_a$.

Now we can derive some related quantities. Substituting Equation 2 into Equation 3 and replacing the integral with summation (since measurements are discrete), Equation 3 can be rewritten as

$$D_{tot} = \sum_{t} \sum_i l_i \times q(i,t) \times \frac{v^*(i,t) - v(i,t)}{v(i,t)} + \sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v(i,t)}.$$  

Equation 5 shows that the delay is trivially 0 when the current speed is greater than the free-flow speed, i.e., $v(i,t) \geq v_{ref}$, and 5 allows a rewriting of the total delay into two non-zero delay cases, as defined below:

1. Current speed is below the threshold speed and the free-flow speed, $v(i,t) < v^*(i,t)$ \& $v(i,t) \leq v_{ref}$.

$$D_{tot} = \sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v^*(i,t)} - \frac{1}{v(i,t)} + \sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v^*(i,t) - v_{ref}}.$$  

(6)

2. Current speed is greater than the threshold speed but less than the free-flow speed, $v(i,t) \leq v(i,t) \leq v_{ref}$.

$$D_{tot} = \sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v(i,t) - v_{ref}}.$$  

(7)

For the sake of simplicity, it is assumed that each spatiotemporal region of delay is caused by a unique delay-causing incident. So, when computing the impact region of an incident, we do not include regions corresponding to the start of a subsequent incident. Now, let $A_t$ be the spatial extent of a particular incident’s impact at time $t \in T'$ where $T'$ is its temporal extent. Then,

$$\sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v^*(i,t)} - \frac{1}{v(i,t)} =$$

$$\sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v(i,t) - v_{ref}} + \sum_{t} \sum_i l_i \times q(i,t) \times \frac{1}{v(i,t) - v_{ref}}.$$  

(8)

Combining Equation 4, which states that $D_{tot} = D_{inc} + D_{rec} + D_{rem}$, with Equations 6, 7 and 8 yields the following definitions for $D_{inc}, D_{rem}$ and $D_{tot}$:

**Definition 2.**

If $v(i,t) < v^*(i,t)$,

$$D_{inc} = \sum_{t} \sum_i l_i \times q(i,t) \times \left( \frac{1}{v^*(i,t)} - \frac{1}{v(i,t)} \right)$$

$$D_{rem} = \sum_{t} \sum_i l_i \times q(i,t) \times \left( \frac{1}{v^*(i,t)} - \frac{1}{v_{ref}} \right)$$

$$D_{rec} = \sum_{t} \sum_i l_i \times q(i,t) \times \left( \frac{1}{v^*(i,t)} - \frac{1}{v_{ref}} \right)$$

If $v(i,t) \geq v^*(i,t)$,

$$D_{inc} = D_{rem} = 0$$

$$D_{rec} = \sum_{t} \sum_i l_i \times q(i,t) \times \left( \frac{1}{v^*(i,t)} - \frac{1}{v_{ref}} \right) \times \left( \frac{1}{v^*(i,t)} - \frac{1}{v_{ref}} \right).$$  

(9)

These equations precisely define quantities that can be understood as calculating the cumulative extra travel time experienced by all drivers during a certain time period and spatial region when traffic is not freely flowing ($v < v_{ref}$).

The recurrent delay, $D_{rec}$, is the component of the currently experienced delay that drivers might always expect traveling through a region during a certain time of day. The incident delay, $D_{inc}$, is the cumulative extra time that all drivers spend in traffic associated with a given incident. The equations essentially codify our intuition that: if the current velocity $v(i,t)$ is below the expected velocity $v^*(i,t)$, then one component is the recurrent delay (from $v_{ref}$ to $v^*(i,t)$) and the remaining component (from $v^*(i,t)$ to $v(i,t)$) is either incident delay if we are in an incident region or remaining delay otherwise. If the current velocity is above the expected velocity, all of the delay is the recurrent delay. The cost of an incident is then simply a constant dollar amount times $D_{nc}$ (We take it to be $11.20 for California [1].)

### 2.3 Computing Duration and Cost of Delay

Now that we have defined the impact region and cost of delay, we will briefly outline our algorithm for computing the delay associated with an incident.

We define the set of detectors of congested segments at time $t = 0$ as a group of detectors in which the speed at each detector is below the variable threshold speed, $v^*(i,0)$, as defined in 2.1. Note that in published algorithms to date, e.g., [10], a fixed threshold speed such as 50mph has been used to determine membership in the preliminary impact region.

Intuitively, for each incident, $a \in A$, we begin at $t_a$ with the closest upstream sensor $u_a$. Our algorithm continues in the next time steps $t > 0$ and a node is appended to the set, $S_t$, of nodes at time $t$, if the node is congested, i.e., if the flow-averaged speed is less than $(1-c)v^*(i,t)$, (where $c$ is a constant), and the sensor was congested in the previous time step or is directly upstream of a congested sensor at time $t$. If the speed is greater than $(1-c)v^*(i,t)$ and less than $(1+c)v^*(i,t)$ and if the speed at the immediate downstream sensor was not between $(1-c)v^*(i,t)$ and $(1+c)v^*(i,t)$, the sensor is appended to the set of congested sensors at this time step. When both stop conditions, not congested and no upstream sensors at the previous time step, are met, time is incremented and the search upstream from $u_a$ is repeated. The total process of identifying the impact region for a given incident $a$ ends when $S_t = \emptyset$ at the end of
a timestep iteration. This allows us to compute the impact region, \( A_t \) at any time \( t \) (In our experiments, we found \( c = 0.05 \) to be the most suitable). We skip the details for lack of space.

In contrast to the work to date on impact regions \[10\] in which the algorithm identifies congestion on one highway in isolation, our algorithm recursively searches upstream, thus identifying possible incident impacts to connecting highways. Our algorithm considers the true freeway topology. For example, Figure 1 shows an example from our Los Angeles dataset that illustrates the automatic incident impact region detection applied to real data. The incident starts in the southbound direction of the highway at the lower right of the black line shown in the figures. The incident then spreads upstream and the impact region spills over onto a connecting highway temporarily. This figure also illustrates how our algorithm considers the true road topology.

We are also interested in predicting how long the incident impact lasts. We define the duration, \( t_{\text{tot}} \), as the duration of the contiguous incident impact region characterized by \( A' \) and \( T' \). In contrast, the PeMS traffic database system used in California (URL mentioned before) computes the “duration” of an incident by subtracting the last time stamp of a police report from the first report. Because this method does not account for effects of an incident after police log reports, we instead computed the “duration” directly from the sensor recordings. We compared the durations we computed directly from the impact region with those of the police incident logs and found that impact regions tend to last longer than the PeMS “duration”. In the data, we see examples such as one incident with one report of a crash and a second and final report of an officer dispatched. In this example, our algorithm will compute a longer duration because we measure until traffic conditions return to recurrent conditions.

3. PREDICTING INCIDENT IMPACT

3.1 Problem Definition

The impact of an incident can be characterized in multiple ways. One possibility might be how much the slowdown will be at a particular location or another might be predicting a metric over the entire impact region, as previously defined. As mentioned early, we predict the latter—the macro level impact caused by an individual incident. The problem is: “Given an incident just occurred, what will its impact be?”

We define impact in two ways: (i) The monetary productivity costs of the cumulative non-recurrent delay in Equation 9 that will be caused by a particular incident (ii) The duration which can understood to be the temporal extent of the impact region. Both metrics are moderately correlated and are different measures of impact. The cost of delay is a constant factor accounting for the value of lost time times the integration of the delay magnitude over space and time, which could cause incidents of different duration, but different magnitudes, to have the same delay value.

Ideally, we would aim to predict the precise numerical values for the two quantities of interest. We tried several regression techniques and models such as regression trees. We found that predicting the precise values (for either cost or duration) is a difficult problem. For example, the least root relative square error obtained with regression trees for duration was close to 100% in many cases, meaning that our error would be the same as if were to make the trivial prediction of the overall mean as the predicted value. To overcome this, we mapped the problem to predicting the impact as a class variable. This makes the problem more tractable and from the domain expert level no less useful. Traffic operators are very often satisfied with knowing the general range of the impact, namely “Will the impact be negligible, moderate, or severe?” Or, “Given two incidents, which one will have a greater impact?” Answering these questions is the first step to solving the domain-specific need of identifying how to prioritize limited resources to mitigate the effects of incidents reported to emergency dispatchers.

3.2 Building the Feature Vector

A big challenge in predicting impact is building a feature vector that combines data from disparate structured and unstructured data sources. The first step is to collect sensor recordings near the incident location. We map a reported incident to the closest upstream sensor on the directed graph \( G(U, E) \). Note that the impact of an incident typically spreads upstream, i.e. there is back-up behind an incident (as depicted in Figure 1). At this closest upstream sensor as well as one directly upstream and one farther upstream, we collect the current conditions including speed \( v(i,t) \), expected speed \( v'(i,t) \), and road occupancy \( \rho(i,t) \). Figure 2 depicts the spatial relationship.

![Diagram of sensor locations upstream (“b”, “c”) and downstream (“a”) of the incident.](image)

\( \text{City costs of the cumulative non-recurrent delay} \)

\( \text{A timestep iteration. This allows us to compute the impact region, } A_t \text{ at any time } t \text{ (In our experiments, we found } c = 0.05 \text{ to be the most suitable). We skip the details for lack of space.} \)

\( \text{In contrast to the work to date on impact regions }[10]\text{ in which the algorithm identifies congestion on one highway in isolation, our algorithm recursively searches upstream, thus identifying possible incident impacts to connecting highways. Our algorithm considers the true freeway topology. For example, Figure 1 shows an example from our Los Angeles dataset that illustrates the automatic incident impact region detection applied to real data. The incident starts in the southbound direction of the highway at the lower right of the black line shown in the figures. The incident then spreads upstream and the impact region spills over onto a connecting highway temporarily. This figure also illustrates how our algorithm considers the true road topology.} \)

\( \text{We are also interested in predicting how long the incident impact lasts. We define the duration, } t_{\text{tot}}, \text{ as the duration of the contiguous incident impact region characterized by } A' \text{ and } T'. \text{ In contrast, the PeMS traffic database system used in California (URL mentioned before) computes the “duration” of an incident by subtracting the last time stamp of a police report from the first report. Because this method does not account for effects of an incident after police log reports, we instead computed the “duration” directly from the sensor recordings. We compared the durations we computed directly from the impact region with those of the police incident logs and found that impact regions tend to last longer than the PeMS “duration”. In the data, we see examples such as one incident with one report of a crash and a second and final report of an officer dispatched. In this example, our algorithm will compute a longer duration because we measure until traffic conditions return to recurrent conditions.} \)

\( \text{3. PREDICTING INCIDENT IMPACT} \)

\( \text{3.1 Problem Definition} \)

\( \text{The impact of an incident can be characterized in multiple ways. One possibility might be how much the slowdown will be at a particular location or another might be predicting a metric over the entire impact region, as previously defined. As mentioned early, we predict the latter—the macro level impact caused by an individual incident. The problem is: “Given an incident just occurred, what will its impact be?” We define impact in two ways: (i) The monetary productivity costs of the cumulative non-recurrent delay in Equation 9 that will be caused by a particular incident (ii) The duration which can understood to be the temporal extent of the impact region. Both metrics are moderately correlated and are different measures of impact. The cost of delay is a constant factor accounting for the value of lost time times the integration of the delay magnitude over space and time, which could cause incidents of different duration, but different magnitudes, to have the same delay value. Ideally, we would aim to predict the precise numerical values for the two quantities of interest. We tried several regression techniques and models such as regression trees. We found that predicting the precise values (for either cost or duration) is a difficult problem. For example, the least root relative square error obtained with regression trees for duration was close to 100% in many cases, meaning that our error would be the same as if were to make the trivial prediction of the overall mean as the predicted value. To overcome this, we mapped the problem to predicting the impact as a class variable. This makes the problem more tractable and from the domain expert level no less useful. Traffic operators are very often satisfied with knowing the general range of the impact, namely “Will the impact be negligible, moderate, or severe?” Or, “Given two incidents, which one will have a greater impact?” Answering these questions is the first step to solving the domain-specific need of identifying how to prioritize limited resources to mitigate the effects of incidents reported to emergency dispatchers.} \)

\( \text{3.2 Building the Feature Vector} \)

\( \text{A big challenge in predicting impact is building a feature vector that combines data from disparate structured and unstructured data sources. The first step is to collect sensor recordings near the incident location. We map a reported incident to the closest upstream sensor on the directed graph } G(U, E). \text{ Note that the impact of an incident typically spreads upstream, i.e. there is back-up behind an incident (as depicted in Figure 1). At this closest upstream sensor as well as one directly upstream and one farther upstream, we collect the current conditions including speed } v(i,t), \text{ expected speed } v'(i,t), \text{ and road occupancy } \rho(i,t). \text{ Figure 2 depicts the spatial relationship.} \)

\( \text{Figure 2: Diagram of sensor locations upstream (“b”, “c”) and downstream (“a”) of the incident.} \)
way Patrol (CHP) reports since initial testing indicated an improvement in accuracy by including these extra features. In addition to basic information including reported start time of day, the structured part of the CHP records offers a type class (a total of 36 types), which we mapped to 9 types: Traffic hazard, Collision+no/minor injuries, Collision+major injuries/ambulance, Natural weather hazard, Lane closure, Fire, Collision-no details, Hit and run, Other. We also include features parsed from unstructured police logs in the first 2 minutes after the reported start of an incident.

Another feature we used is whether within one hour of an incident it was raining and if so, by how much.

In addition we considered features pertaining to the topology of the highways. As of now, we consider the number of lanes as a feature. In future work, we aim to add more features such as the proximity to an exit.

In addition to building feature vectors based on data for each incident, we construct feature vectors for possible pairwise combinations of incidents. This dataset is used for predicting which incident of a pair will have relatively higher impact.

### 3.3 Bin Selection

For every incident of interest, we compute the incident impact both in terms of cost of delay and duration. This is done using the algorithm and definitions introduced in 2. Once we have obtained these numerical quantities, we map the two prediction variables to classes or bins.

There are many ways to construct bins [7], such as equi-width, equi-height and clustering. We experimented with many of these approaches and found that either the bins lead to a very skewed distribution of elements across classes (such that it makes sense to just learn a trivial learner) or have no semantic meaning for domain experts. The latter is important since this an emerging application to be implemented in practice by internal company teams and external agencies. Finally, as reported in our results, we chose to do manual binning. This allows us to divide the range of values into sub-ranges, so that they have a semantic meaning. For example, for a range of duration $d$ from 0—140 minutes, the range of $d \leq 5$ indicates negligible impact, $5 < d \leq 30$ indicates moderate impact, $30 < d \leq 60$ indicates high impact and $d > 60$ indicates severe impact.

### 3.4 Building Prediction Models

Once we have constructed the feature vector and mapped the continuous prediction variables to discrete classes, we are ready to build prediction models. We can use classification models for prediction, where each bin becomes a class.

Before building classification models, we do feature selection. We remove those features that provide no or very little information gain. For prediction, we have constructed a suite of classification algorithms, including: (1) a classification tree model based on C4.5 trees with various settings, (2) an ensemble of short trees constructed using Adaboost, (3) a k-NN classifier (with various ks and distance weights) (4) a multiclass classification algorithm built with 5 models by varying C4.5 tree model parameters (5) a multi-layer perceptron and (6) radial-basis function network. For the decision trees we varied the confidence threshold (up to 0.2), minimum number of nodes at leaf nodes (1-5), and whether or not to force binary splits for nominal variables. For the nearest neighbor we varied the number of neighbors (from 1-7) and how they were distance weighted. The intuition was to select techniques from various classes of classifiers (tree-based, nearest neighbor, and function construction). We also put a minimum threshold for recall, precision for each class (typically set to 0.25). Then given a dataset, we choose the classifier that gives us the lowest error given that the recall and precision for each class is greater than the threshold.

A further refinement would be to introduce a cost matrix once the relative trade-offs between different types of prediction error are quantified for a particular transportation district.

### 4. EXPERIMENTAL RESULTS

#### 4.1 Dataset Description

As mentioned earlier, we extracted our traffic sensor datasets from PeMS. We chose two different regions and seasons to help test the versatility of our framework. The first dataset is two months of sensor recordings in the Los Angeles area (Caltrans District 7), namely January and February 2009. The dataset includes 28,819,432 recordings from 1696 mainline (open freeway) detectors. Note that the full dataset includes detectors on ramps and freeway to freeway connectors, but we consider only a subset (1696 detectors here) that are embedded the main highway lanes. The second dataset is July and August 2011 for the San Francisco Bay Area (Caltrans District 4) and is 32,587,200 recordings from 1825 mainline detectors. Figure 3 shows the spatial distribution of sensors and freeways in the two regions.

![Figure 3: Map of the freeway sensors (black dots) and major highways (in grey) used for predicting the impact of incidents.](https://example.com/figure3.png)

(a) San Francisco Bay Area  (b) Los Angeles Area

We focused on incidents starting on I-5 in eastern Los Angeles and US Route 101 in Silicon Valley in the San Francisco Bay Area, because they provide samples with high traffic volumes and high potential impact on productivity. For the first dataset, we specifically investigated incidents on southbound interstate I-5 between postmiles 116.9 and 130 and their effect on northbound and southbound I-5 traffic as well as on connecting highways. There were 173 incidents in this subset of the corridor during January and February 2009. This second dataset focuses on incidents on southbound US Route 101 between postmiles 400 and 410, which numbered 244 during the study period, and their effect on US Route 101 and connecting highways. We investigated weekday incidents only and used recurrent speeds calculated for weekdays only to be consistent.

For incident details, we used two types datasets from the California Highway Patrol (CHP) as collected by PeMS and
also freely available in near-real-time from CHP. First, there were structured logs with incident start times and locations. Second, the incident details were semi-structured with free text police logs. After parsing the data with aid from the CHP dictionary of shorthand (can be found at the URL mentioned above) and including some common misspellings, we extracted features such as the number of officers on the scene, if a truck was involved, if a tow truck was mentioned, and the number of vehicles reported. For example, one line in the police log for a particular incident is: “11779132.01/23/2009 00:00:00.RP IN A GRY MERZ BENZ VS WHI BIG RIG L/UP94757, ADD”. This police log was made available one minute and 2 seconds after the reported start of the incident and says that the reporting party (RP) was in a gray Mercedes Benz and in a crash with a white semi-truck (BIG RIG). From this line of raw text, our natural language processor identifies that two vehicles are involved. Additionally, it notes that a truck is present. These clues are then added to the feature vector.

In addition, we extracted hourly rainfall data from the California Department of Water Resources databases by mapping each sensor to the closest weather station.

The distribution of both prediction variables is quite skewed and in Figure 4(a) we plot the distribution of incident delay (cost divided by a constant factor) in log-log scale from the LA dataset and observe the high concentration of low delay events. We observe that the distributions follow an exponential distribution. Second, in Figure 4(b) we see a similar trend for the incident duration identified by our method.

![Figure 4: Probability of an incident having an impact greater than x, where x is incident delay in (a) and incident duration in (b) (Complementary CDF, in log-log scale). Note the high number of low magnitude events.](image)

### 4.2 Results

We evaluate our technique for its accuracy in predicting the impact class, measured by either incident cost or duration. The results are obtained with 10-fold cross validation. We do 10 runs of 10-fold cross validation and present results averaged over the 10 runs. For each run we choose the maximum cross validation accuracy (recall that in our model every run had a suite of classifiers) obtained in the run and the overall accuracy result is the average over these runs. For each maximum accuracy, we note the precision, recall and f-measure for each class and present the average of these values. We first present results for cost and then for duration. We present results rounded to the second decimal place.

5California Highway Patrol (CHP) Traffic Incident Information Page, [http://cad.chp.ca.gov](http://cad.chp.ca.gov)
6California Department of Water Resources California Data Exchange Center, [http://cedc.water.ca.gov/intro.html](http://cedc.water.ca.gov/intro.html)

### 4.2.1 Predicting Cost

In Table 1 we provide the accuracy results for predicting the cost of delay on the San Francisco Bay Area (SF) dataset. For each experiment, we describe the number of classes, the bounds for the classes, the counts in each class and the overall accuracy. For example, experiment SF-3 has three classes, with buckets being, cost ≤ $10, $10 < cost ≤ $250 and cost > $250. We have chosen semantically meaningful buckets and not buckets that might give us higher accuracy. A cost of less that $10 can be thought to be insignificant, a cost of $10 to $250 can be thought to be moderate, $250 to $1000 as high and a cost in excess of $1000 as exorbitant. We also provide the precision, recall and f-measure for each class in Table 1. This is useful since for a skewed distribution we need to make sure that we have reasonable recall and precision for classes with fewer examples. We obtain good results for two classes both in terms of accuracy and precision/recall for each of the individual classes. In particular, we can decide with high degree of confidence if the cost of an incident is going to be low (less than $10). As expected, as the number of classes increases, our accuracy goes down, but even for four classes (experiment SF-4) we have reasonable results. Table 1 also details the results from the Los Angeles region (LA) dataset. It is interesting to note the cost of incidents are in general higher and the bin boundaries are shifted accordingly (The range of $250 - $1000 is removed because we have very few training examples). Since traffic in LA is notoriously bad, these results are expected.

### 4.2.2 Predicting Duration

In Table 2 we show the results for the SF and LA datasets for predicting the temporal duration. The bin boundaries are in minutes, so the buckets for the three experiments are (1) duration ≤ 5 min and duration > 5 min, (2) duration ≤ 30 min and duration > 30 min, and (3) duration ≤ 5 min, 5 min < duration ≤ 30 min and duration > 30 min. For emergency responders, a major decision support need is identifying a “false alarm” or near 0 duration event. In this 2 class paradigm, our model is over 90% accurate for LA, and 87% for SF.

### 4.2.3 Comparing Incidents

Finally, from a relative magnitude perspective, we are interested in knowing which of two incidents will have a higher impact, whether measured by incident cost of delay or temporal duration. Table 3 shows that our model can predict the relative magnitude of a pair of incidents with high accuracy.

### 4.2.4 Extensions: Transfer Learning and Multi-tiered Classification Model

To investigate the versatility of our proposed framework we asked the following question: to what extent will a model trained on one region transfer to another region? Table 4 show the accuracy for a model trained on SF incidents and directly applied to predict the impact class of incidents in LA. The results suggest that we do transfer learning from region to another and demonstrates the versatility of our framework. In comparison to the LA results when trained on LA data, the LA results when trained on SF data generally show only marginal losses in accuracy. This result is encouraging and suggests that for practical deployment,
4.3 Discussion

The overall classification results are encouraging. In addition, we discovered models that brought new insights to the domain experts. For example, the pruned C4.5 tree depicted in Figure 5, which was discovered for predicting false alarms in the Los Angeles dataset (70/30 train/test) has over 90% accuracy. The root node of the classification tree makes a decision based on how different the speed at a sensor (two sensors upstream from the incident) is from the recurrent speed. Then, if the speed is significantly below the recurrent speed for that location and time of day, the model predicts that an accident is occurring. If not, the model checks how densely packed the road is. If the road is relatively empty, then the model predicts the report as a false alarm. If not, the model checks to see if any police report mentioned vehicles involved within the first two minutes. If so, the model predicts that an accident is occurring. If not, the model classifies the report as a "false alarm" (meaning that impact would be negligible). This automatically creates and pruned tree is consistent with intuition and offers insights, such as that being only a relative magnitude of incident cost and duration. Our model is able to reliably predict the relative magnitude of the impact.

4.3 Discussion

Table 1: Results for predicting cost. We showed high overall accuracy and high precision and recall for each class.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Classes</th>
<th>Bounds ($)</th>
<th>Counts</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-1</td>
<td>2</td>
<td>10</td>
<td>114, 124</td>
<td>95.59</td>
</tr>
<tr>
<td>SF-2</td>
<td>2</td>
<td>100</td>
<td>157, 81</td>
<td>87.86</td>
</tr>
<tr>
<td>SF-3</td>
<td>3</td>
<td>10, 250</td>
<td>114,54,70</td>
<td>81.09</td>
</tr>
<tr>
<td>SF-4</td>
<td>4</td>
<td>10, 250, 1000</td>
<td>114,54,34, 36</td>
<td>73.73</td>
</tr>
<tr>
<td>LA-1</td>
<td>2</td>
<td>10</td>
<td>75, 97</td>
<td>91.40</td>
</tr>
<tr>
<td>LA-2</td>
<td>2</td>
<td>250</td>
<td>94,78</td>
<td>88.84</td>
</tr>
<tr>
<td>LA-3</td>
<td>2</td>
<td>1000</td>
<td>120, 52</td>
<td>82.33</td>
</tr>
<tr>
<td>LA-4</td>
<td>3</td>
<td>10, 1000</td>
<td>75, 45, 52</td>
<td>75.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>f-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-1</td>
<td>1</td>
<td>0.95</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>SF-2</td>
<td>1</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>SF-3</td>
<td>1</td>
<td>0.92</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>SF-4</td>
<td>1</td>
<td>0.94</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>LA-1</td>
<td>1</td>
<td>0.90</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>LA-2</td>
<td>1</td>
<td>0.91</td>
<td>0.88</td>
<td>0.90</td>
</tr>
<tr>
<td>LA-3</td>
<td>1</td>
<td>0.87</td>
<td>0.88</td>
<td>0.87</td>
</tr>
<tr>
<td>LA-4</td>
<td>1</td>
<td>0.87</td>
<td>0.92</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Table 2: Results for predicting duration of an incident. We show high overall accuracies and good precision and recall for most experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Class</th>
<th>Precision</th>
<th>Recall</th>
<th>f-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>SF-1</td>
<td>1</td>
<td>0.89</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>SF-2</td>
<td>1</td>
<td>0.83</td>
<td>0.86</td>
<td>0.83</td>
</tr>
<tr>
<td>SF-3</td>
<td>1</td>
<td>0.92</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>LA-1</td>
<td>1</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>LA-2</td>
<td>1</td>
<td>0.95</td>
<td>0.74</td>
<td>0.83</td>
</tr>
<tr>
<td>LA-3</td>
<td>1</td>
<td>0.91</td>
<td>0.86</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Table 3: Prediction accuracy (%) for pairwise relative greater magnitude of incident cost and duration. Our model is able to reliably predict the relative magnitude of the impact.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Region</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>San Francisco</td>
<td>96.61</td>
</tr>
<tr>
<td>Cost</td>
<td>Los Angeles</td>
<td>94.89</td>
</tr>
<tr>
<td>Duration</td>
<td>San Francisco</td>
<td>96.3</td>
</tr>
<tr>
<td>Duration</td>
<td>Los Angeles</td>
<td>92.21</td>
</tr>
</tbody>
</table>
Table 4: Prediction accuracy (%) by each bin selection choice for \( k \) classes of incident delay trained on the SF dataset and tested directly on LA: the model results suggest good transfer learning.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Bounds</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>$10, $100</td>
<td>86.05</td>
</tr>
<tr>
<td>Duration</td>
<td>5, 30 Minutes</td>
<td>91.28</td>
</tr>
<tr>
<td></td>
<td>10, $100</td>
<td>90.07</td>
</tr>
</tbody>
</table>

Figure 5: Classification tree for predicting “false alarms” for LA, which predicts with >90% accuracy and is consistent with intuition of domain experts.

often good predictors. This might point to a way for constructing models for physical systems—namely that we look at some equations from domain experts and capture some of those relationships in our feature vectors. Furthermore, in this work we collected features of different types describing: the incident, changes in the state variable (such as deviation from recurrent velocity), the environment (such as rain), and the topology (refinement is ongoing). We have shown that this exercise also improves the accuracy of our models.

Furthermore, for the predicted variable it is important to first choose a physically useful quantity to predict (such as the ranges for cost or duration) even at the expense of accuracy in models. Making such a choice often causes a problem with class distribution and makes an already challenging problem even more so. Nevertheless this is needed to get useful models that domain experts want to use.

Our results show that given a cyber-physical system, machine learning techniques can help predict the impact of an incident. We believe that this way of obtaining details of incidents from some event database, computing impact using sensor data and building a model that correlates the two can be extended to other domains.

5. CONCLUSIONS

In this paper we have proposed a system for rapid prediction of the cost and impact class of highway incidents that demonstrates the real-world applicability of a predictive model based on classification models trained on available urban data. The feature vector is built from structured data from sensor networks on highways as well as semi-structured text collected at different points in time. With experiments on real-world data, we have demonstrated that our models are good predictors of incident impact. Thus, our work supports a decision-response to a highway incident—that until now has relied on human expertise. In addition to advancing the use of machine learning for highway operations in practice, these results open the door to further studies on applying statistical methods to traffic management and related applications. We are discussing these results with HP Services and a government agency and expect to deploy the underlying techniques to real-time operations.

6. REFERENCES