Predicting Emergency Incidents in San Diego

CS229 Final Project Report

Tyler Romero (tromero1@stanford.edu)
Zachary Barnes (zbarnes@stanford.edu)
Frank Cipollone (fcipollo@stanford.edu)

Abstract—**By predicting details of future emergency incidents such as type and location, we can provide a way for emergency responders to better allocate their resources and save more lives.**

Keywords—machine learning; decision trees; clustering; neural nets; prospective modelling; background; trigger; emergency events;

I. Background

A. Motivation

Each year, emergency responders assist in millions of critical events across the country, costing billions of dollars. For example, In 2015, 1.3M of these events were fires, resulting in over 15,700 civilian injuries and $14.3B in estimated property damage [1]. As emergency events, the time it takes for first responders to arrive on scene is critical, with minutes often making the difference between life and death. Because of these factors, standard staffing and resource use is very high to make sure enough responders are available at any given time. These factors make emergency response an important potential application for optimization based on predictions. Thus, a model that can learn and make predictions on the location, frequency, and type of these events would be extremely useful to government and department management in making staffing and resource allocation decisions.

B. Related Work

A similar application of machine learning to help emergency responders was done by Bayes Impact. They analyzed Seattle police report data in order to determine ways in which Seattle could better deploy officers with the goal of minimizing serious and violent crime [4].

II. Approach

A. Goals

Our goal is to use historic emergency incidents in a specific geographic region to predict where future emergencies might occur, and of what type. We will frame this application as a supervised learning problem where training examples will be drawn from historic data on emergencies for the region as well as relevant weather, geographical, structural, and demographic features. This will allow our model to learn the incident likelihood over our region of interest which can then be subsequently turned into a prediction.

B. Data Sources

Our source of historic emergency incidents was obtained directly from the San Diego Fire Department which provides details for every emergency incident responded to in the last year. This dataset is comprised of the type, location, date, time, response time, and category of severity for approximately 1,000,000 incidents.

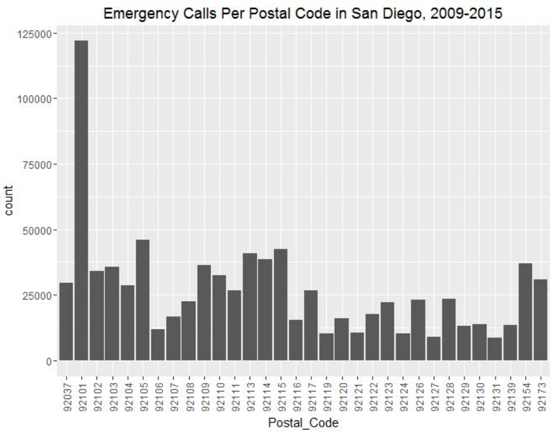
C. Data Preprocessing

We performed two significant transformations to our data in order to make it more suitable to our needs. First, we used Google and MapQuest APIs to transform each street addresses into latitude and longitude coordinates. Second, we split the PhonePickUp timestamp into year, month, and day of week columns.

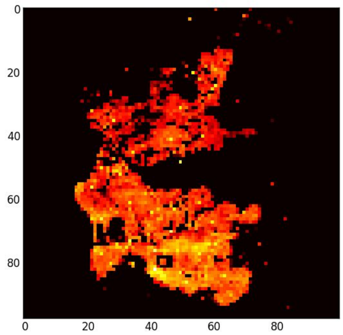
D. Data Visualization

Before beginning to design a model, we wished to get a better understanding of the data that our model will be built on. We proceeded to generate several different charts to summarize our data:

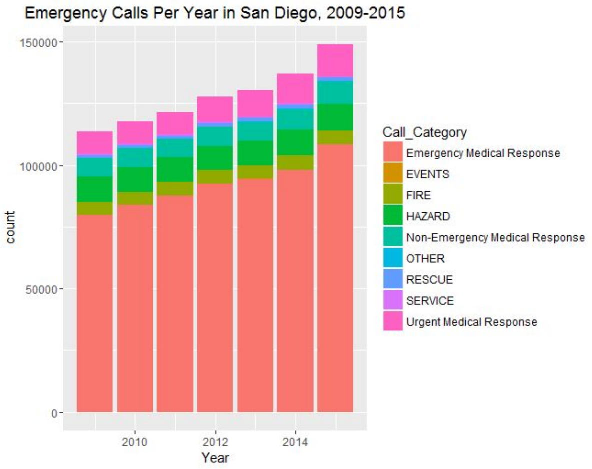
First of all, since we will be attempting to make predictions based on locations within San Diego, we plotted the number of emergency incidents per postal code.



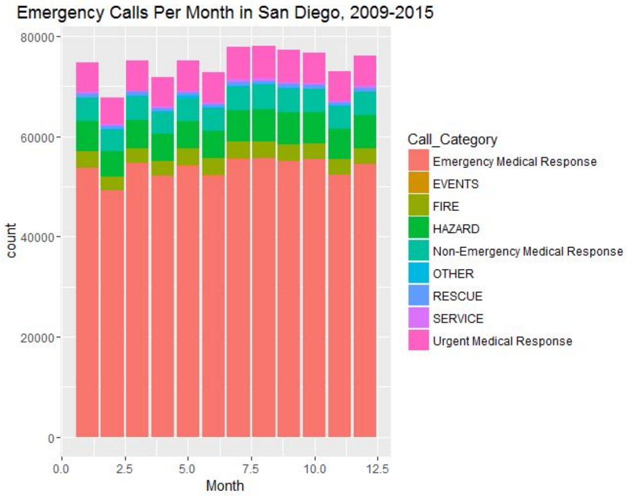
Clearly emergencies occur more frequently in some postal codes than others. This particular chart could be misleading due to the different sizes of various postal codes. Once we were able to get the exact latitude/longitude values for our incidents, we were able to create a heatmap to display incident frequency.



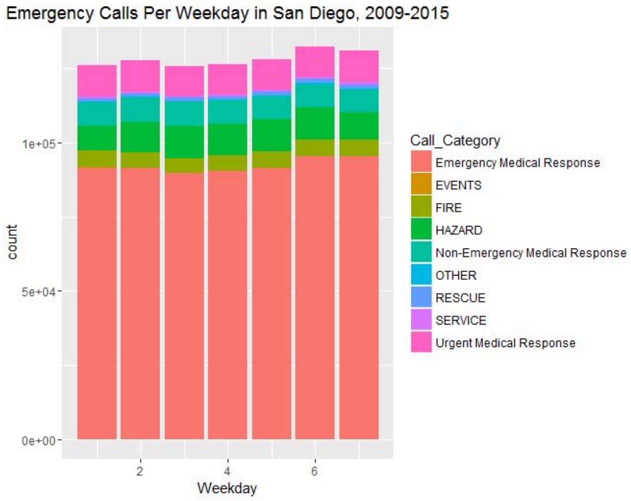
We then wished to visualize the data over time. It is evident that the number of emergency incidents per year is increasing with time, most likely due to the increasing population of San Diego.



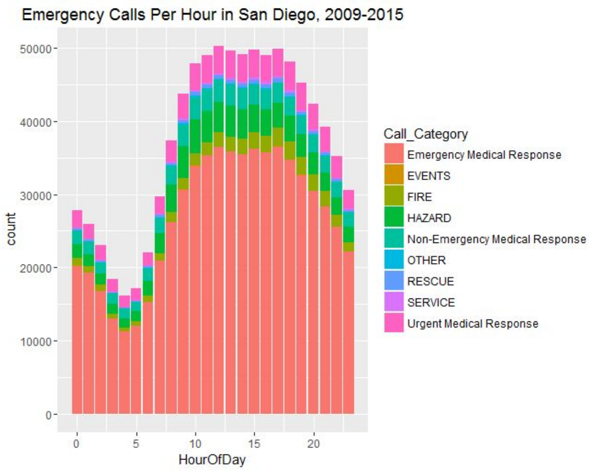
In contrast, it seems that the number of incidents per month is relatively constant throughout the year.



The same holds for days of the week, although it does seem that Saturdays tend to have the most incidents.



Finally, we plotted number of incidents per hour of the day:



It is evident that there is a strong relationship between time of day and number of incidents.

2. Experiments and Results

A. Decision Tree Regression

Due to the above noted relationship between time and incidents per hour as well as the slight monthly dependence, a test on the predictive power of these features was needed. We constructed a decision tree using hour, day, month, and zipcode as covariates, in order to predict the number of incidents per hour at a given location. We chose a decision tree due to the highly categorical nature of the features and with the goal of understanding the different effects that features based on location have on the predicted number of incidents.

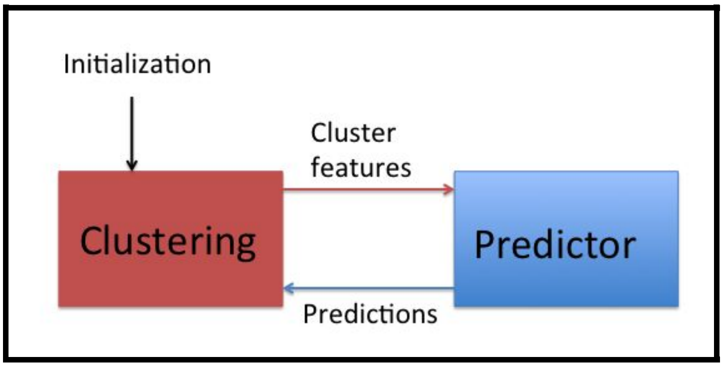
A 2009 dataset on incidents in the San Diego region containing 150,000 training elements was partitioned and used in 10-fold cross-validation. We trained several different decision tree models varying the max allowable depth. These models were then tested against the actual number of incidents per hour, and for each the mean squared error calculated. The table below summarizes these results in the case of unbounded depth and a max depth of 10.

Experimental Model: (postal code sized regions)	**RMSE:** (incidents/hour)
Decision Tree Regression (DTR)	0.7010
DTR with Max Depth 10	0.4827

The decision tree with a max depth of 10 performs better than the unbounded decision tree when using 10-fold cross-validation, because limiting depth helps to reduce overfitting to the training data.

B. Iterative Clustering and Prediction with Neural Nets

While the decision tree was a good first thing to try, it did not take into account the correlations between nearby grid points, instead making a decision for each grid location. The fact that the grid locations are correlated with each other in some unknown way is a good reason to try unsupervised learning. The model we arrived at is one that combines unsupervised learning, in order to take into account the correlation between nearby grid points, and supervised learning, in order to retain the predictive power. Our model alternates between using k-means clustering and a neural network predictor, giving the neural network predictor features from the output of the clustering algorithm until convergence.



The cluster features that we used were the following:

1. Cluster size
2. Average number of incidents in cluster
3. Maximum incidents in gridpoint in cluster
4. Minimum incidents in gridpoint in cluster

By training the neural network with information on each cluster we are able allowing to learn more about specific spatial groups of grid locations.

Experimental Model: (20 x 20 sized regions)	**RMSE:** (incidents/hour)
Neural Nets (no clustering)	3.161
Max Depth DTR	2.013
Iterative Clustering with Neural Nets	1.977

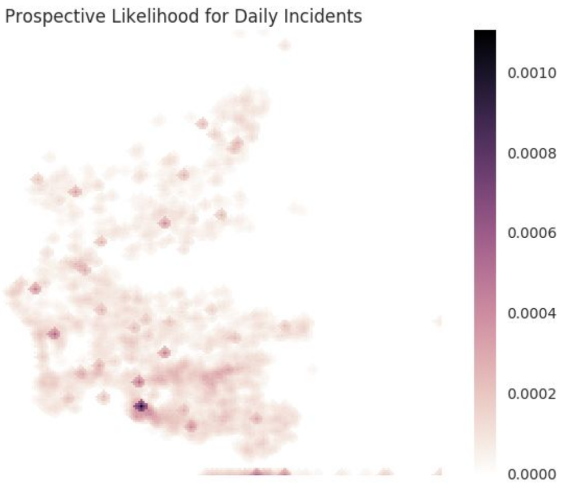
B. Spatial-Temporal Prospective Prediction Models

While the above methods serve as a reasonably strong method of predicting the number of incidents in a region over a time window, we have come to realize the need for a stronger prediction that leads to more actionable insight to the stakeholder (fire departments). We seek to achieve this goal by narrowing the operating region for fire departments by predicting a set of very likely locations, 400m x 400m grid locations, for incidents on any given day. We approach this goal by trying to model future daily incidents with a probabilistic likelihood based on past observations that are both temporally and spatially relevant.

In order to do this, we frame the problem similar to a contagion model where likelihood of future events decay inversely over time and space from witnessed events. Thus, for any given day of interest, the likelihood of witnessing events at column location x* and row location y* is determined by the space time kernel K_{st}.

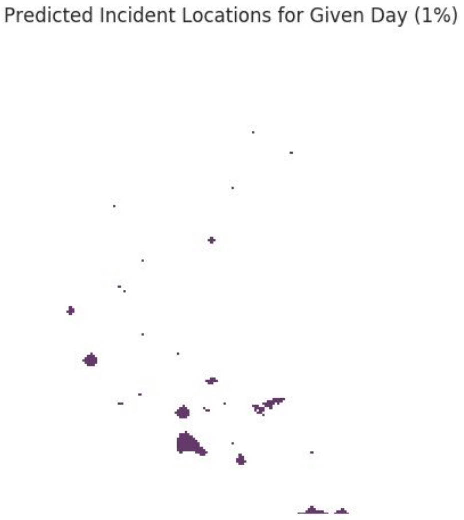
$$p(x^*, y^*, t \mid \theta) = \frac{1}{m} \sum_{t' \in D_{t'<t}} K_{st}([x^* \ y^*], [x' y'], t - t' ; \theta)$$

Where, $K_{st}\,(s,\ s',\ t\ ;\ C,\ q,\ p)\ = \frac{C}{(\|s - s'\|^2 + 1)^q * (t+1)^p}$ determines the spread of likelihood in both space and time, where the probability is the kernel density estimation. The relevant kernel parameters are learned via validation through a grid search.



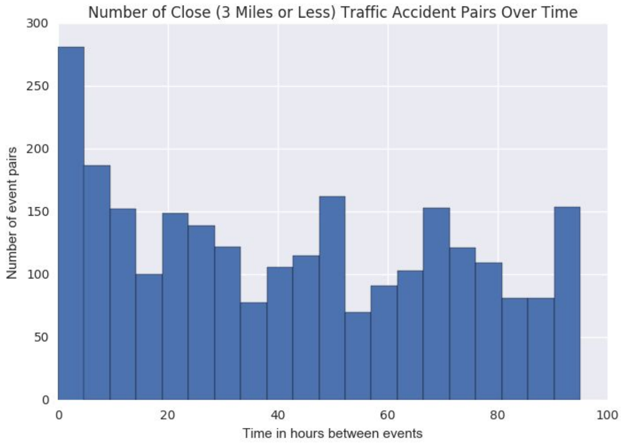
Thus, for any given day we have a prospective likelihood of incidents occurring in every grid region (400m x 400m). From these likelihoods, we take the top feasible number (1%) and select them as predicted locations.

As is shown below, even though the number of predicted locations is 238, the map coverage of the locations is relatively small, meaning many contiguous regions could be dealt with by the same emergency response team.



Experimental Model: (400m x 400m grid sizes)	**% Events Captured:** (for 1% of locations)
Spatial-Temporal Prospective	33.4%

One way we looked at to improve this model is to account events being triggered by events of the same type. This is similar to how earthquakes are responsible for triggering aftershocks. The following graph displays this phenomenon for traffic accidents.

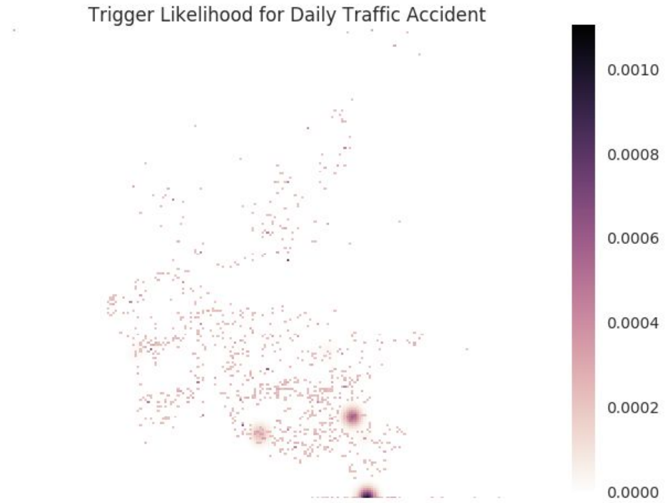


With this information, we frame the problem of the likelihood of a traffic accident occurring in any given day to be the probability of triggering a traffic accident or the probability of one randomly occurring.

$$p(x^*, y^*, t \mid \theta) = p_b(x^*, y^*) * \ b(x^*, y^*)$$
$$+ \ p_t(x^*, y^*) * \frac{1}{m} \sum_{t' \in D_{t'<t}} K_t(x^* - x', y^* - \ y', t^* - t' \ ; \ \theta)$$

Where b is the background likelihood of traffic accidents, p_b is the likelihood of the point being a background emission, and p_t is the likelihood of the point being a trigger emission. Here $K_t(x, y, t) = e^{-\sigma_t * t} * e^{-\frac{-x^2}{\sigma x^2}} * e^{-\frac{-y^2}{y^2}}$ represents the fact that triggered events are highly temporally and spatially correlated and thus can be seen as the joint distribution of log-normal over time and gaussian in both x and y directions.

In generating the parameters for the above model, we make the naive assumption that incident pairs that are close temporally and spatially are triggered and any others are background emissions. This allows us to easily estimate the above model using Monte Carlo methods. However, a more appropriate further approach would be to have triggered or background be unknown and use an EM approach to learn inside of that unknown.



Experimental Model: (400m x 400m grid sizes)	**% Events Captured:** (for 1% of locations)
Spatial-Temporal Prospective	35.8%
Trigger-Background Model	37.5%

3. Conclusion

A. Discussion

Initially, our goal was to predict the number of incidents that would happen in over the course of a day in a specific location. In order to do this, we created two models, using decision tree regression and iterative clustering with a neural network predictor. We found that both of these methods achieve similar performance. We believe this is due to the fact that hourly incidents have low variation and both models they both effectively learn by segmenting up different locations into groups and learning on those groups.

After consulting with a career firefighter about what sort of predictions would be most helpful when it comes to real-life emergency response, we decided to pivot and create models that instead attempt to isolate the areas most susceptible to emergency incidents within the next day, based on previous days. We created two models in order to make these predictions: Spacial-Temporal Prospective Likelihood and Trigger-Background Likelihood. Trigger-Background performs marginally better because it is able to use the distinctions between types of events to its advantage.

In general, we found this to be a difficult problem due to the semi-random nature of emergency incidents, and the large area we needed to cover with our predictions. We believe that similar models could be useful for ride-sharing companies: predicting the top n% of most likely positions that pick-up requests will come from could help lower response time.

B. Next Steps

In the future, we would like to improve upon our current models by acquiring more demographic and census data regarding emergency incidents. In addition, we would like to see if our results hold up in other major cities. In our models, we used the top 1% of locations as a method of comparison between models, but we would also like to learn the optimum threshold for percent of locations to isolate, given the available resources of a fire-department.

REFERENCES

[1] Haynes, Hylton JG. “Fire loss in the United States during 2015.” National Fire Protection Association. Fire Analysis and Research Division, 2016.

[2] *Weather Underground - Weather Forecast & Reports*. N.p., n.d. Web. 21 Oct. 2016.

[3] "San Diego Demographic Data." *Census.gov*. N.p., n.d. Web. 21 Oct. 2016.

[4] Wong, Jeff. "Walking the Beat: Mining Seattle’s Police Report Data." *Bayes Impact*. N.p., n.d. Web. 21 Oct. 2016.