Predicting Emergency Incidents in San Diego

CS229 Project Milestone

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; 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. Historic emergency data will be supplemented with weather [2] and demographic data [3] corresponding to the region of interest.

C. Data Preprocessing

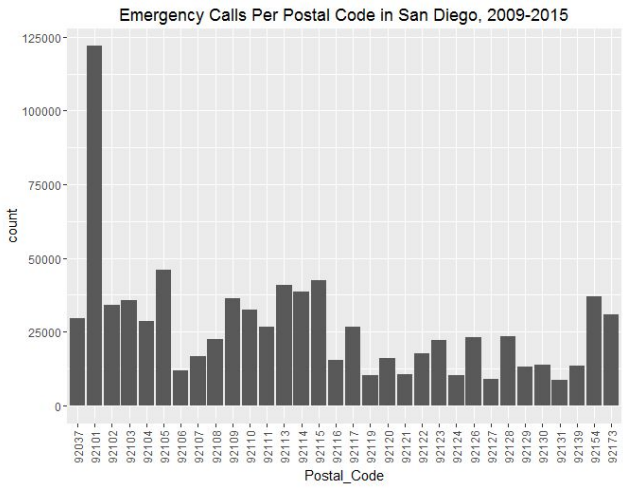
We performed two significant transformations to our data in order to make it more suitable to our needs. First, we used Google and MapQuest APIs to transform each street addresses into latitude and longitude coordinates. Second, we split the PhonePickUp timestamp into year, month, and day of week columns.

D. Data Visualization

Before beginning to design a model, we wished to get a better understanding of the data that our

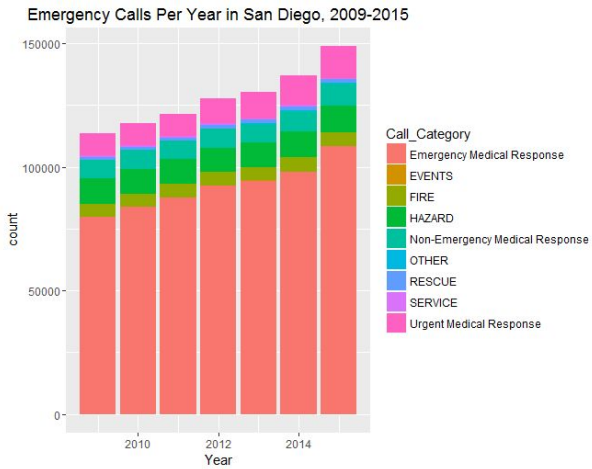
model will be built on. We proceeded to generate several different charts to summarize our data:

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.

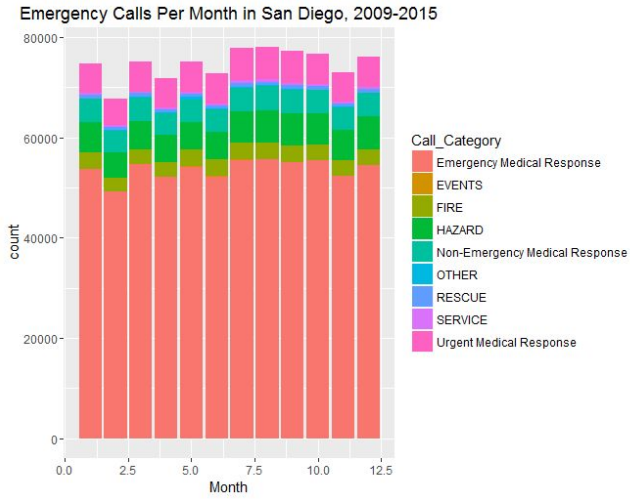


Clearly emergencies occur more frequently in some postal codes than others. This particular chart could be misleading due to the different sizes of various postal codes. Our preprocessing of latitude and longitude is still in process, but once that is complete, we will divide up our data into equal area cells.

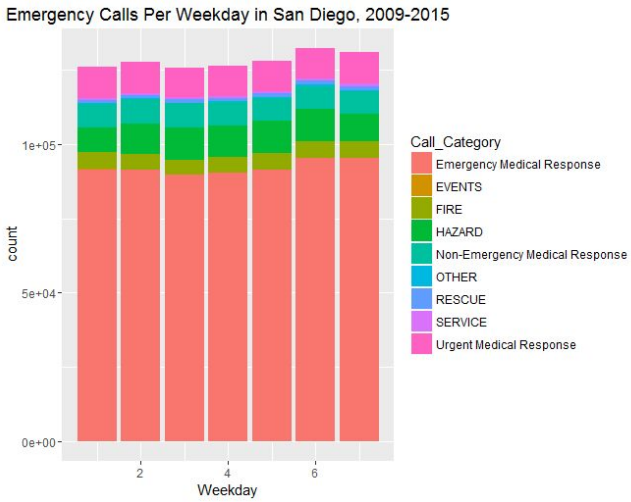
Following our analysis based on postal codes, we wished to visualize the data over time. It is evident that the number of emergency incidents per year is increasing with time, most likely due to the increasing population of San Diego.



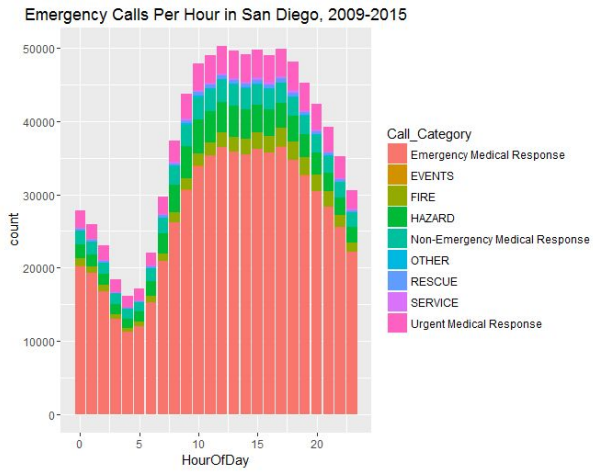
In contrast, it seems that the number of incidents per month is relatively constant throughout the year.



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.

III. 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:	**RMSE:**
Decision Tree Regression	0.7010 (incidents/hour)
Decision Tree Regression with Max Depth 10	0.4827 (incidents/hour)

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.

IV. Conclusion

A. Discussion

The above results suggest that large regional stratifications (such as zipcodes) are able to relatively easily determine the resulting distribution of the number of incidents per hour (when overfitting is curtailed). However, the large areas covered by postal codes, and the lack of prediction of the type of emergency incident that might occur leave this simple model lacking for real world application. Unpacking the decision tree and the distribution of events per hour suggest a heavy relationship between location and event type, and a potential log-normal distribution over number of event occurrences. Moving forward, more statistical sharing must be achieved both spatially and temporally to allow for predictions and learning with finer grain resolution and applicability.

B. Next Steps

We plan to try several additional models in order to reduce testing error as much as possible and achieve the above goals.

One possible approach could be clustering based on location and incident. Using unsupervised learning could allow us to learn key relationships between types of incidents allowing for greater predictive power in a given day. Additionally, clustering could also allow us to learn regions that behave in similar ways, allowing us to learn better location based models.

A different approach to tackling this problem is to learn a temporal heatmap distribution over the region that focuses on the time-spatial differences between events. This heatmap model could then be leveraged to help guide predictions for future events.

In general, we believe that we need to leverage a more finely grained spatial indicator (such as latitude longitude) in order to better relate similar instances, and provide more actionable predictions. We also plan to incorporate additional data sources, such as weather and demographics, in order to gain additional relevant covariates.

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.