Fire Emergencies in Seattle

Correlations with Human Activity and Rain Patterns

S. C. Vargas

Contents

1 Problem

Can we establish trends in 911 fire calls in Seattle to predict and find patterns, or correlate them with factors such as rain patterns?

2 Client(s)

Governmental agencies might be interested in this study to refine strategies that pin point influential factors in the manifestation of fires in Seattle. Insurance agencies might also find this relevant. Ultimately, the objective is to reduce the significant human and financial cost generally associated with fires. Reducing the number of fires will ultimately allow the police, fire department and other dependencies to divert their resources in other issues faced by the city and its inhabitants.

3 Sources and Data Sets

Seattle Observed Monthly Rain Gauge Accumulations [1]

These monthly data goes from October 2002 to May 2017, containing measurements of 17 rain gauges located throughout Seattle city limits. There is no information on the units describing the amounts of rain. The locations of the rain gauges are given indirectly in a image (see Figure 1).

Freely available for download, modification and distribution, under the license CC0 1.0.

Seattle Fire 911 Calls [2]

Fire 911 calls in Seattle, from 2010 to 2011. It contains latitude and longitude of the location of the caller, in addition to date, time and type of call. This version corresponds to the September 2, 2018 update.

Freely available for download, modification and distribution, under the license CC0 1.0.

Holidays in Washington State in 2010 and 2011 [3, 4]

Federal and state holidays in 2010 and 2011 in the state of Washington. These were retrieved in October 10, 2018.

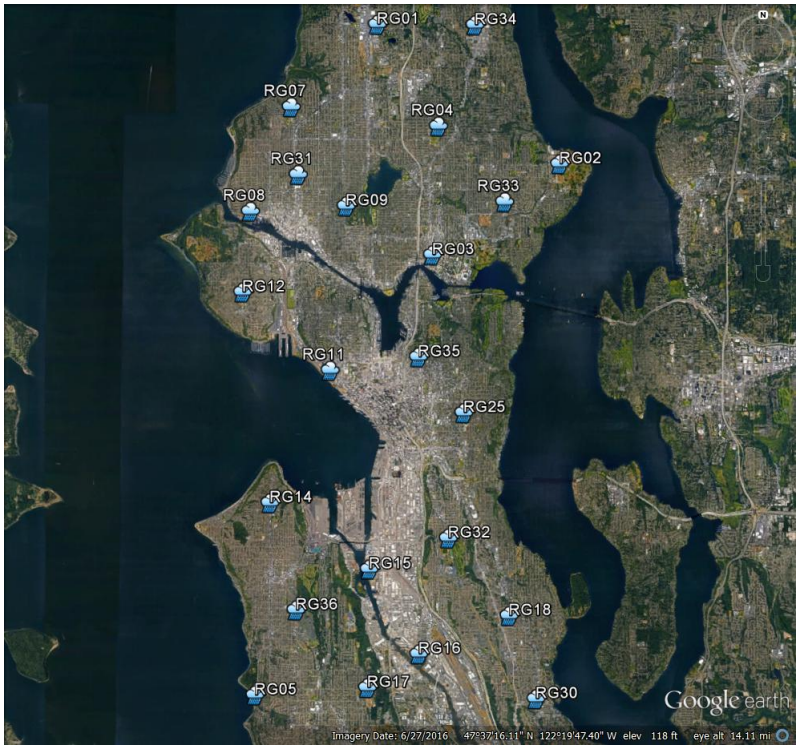

Figure 1: *Locations of rain gauges in Seattle. In reality, measurements for only 17 of these gauges are reported. This image is part of the data set in [1].*

4 Data Wrangling and Exploratory Data Analysis

4.1 Rain Data Set

With the help of Figure 1, we can deduce an approximate location for the rain gauges (latitude and longitude) which we store in a dictionary. As we mentioned, only 17 rain gauges actually appear in the data set.

We proceed to import the `csv` file and read it as a data frame. The data has a date column, and a column for each 17 rain gauges. The date column shows that measurements were taken monthly, while the other columns indicate values for rain accumulation (in unknown units). These cover the period from November, 2002 to May, 2017.

We first proceed to search for null values or outliers. There are no null or non-existent values in the set and a direct plotting of all the data shows it to be fairly reasonable between the expectations of seasonal behavior. For instance, we can see this in the case of the gauge RG01, as shown in Figure 2.

We are interested in the correlation between these weather data and the frequency of fire emergency calls. The fire 911 calls set covers only the period between 2010-07 and 2011-02, so we select the rain data corresponding to this interval in a separate data frame which we label `df_rain_10_11_F`.

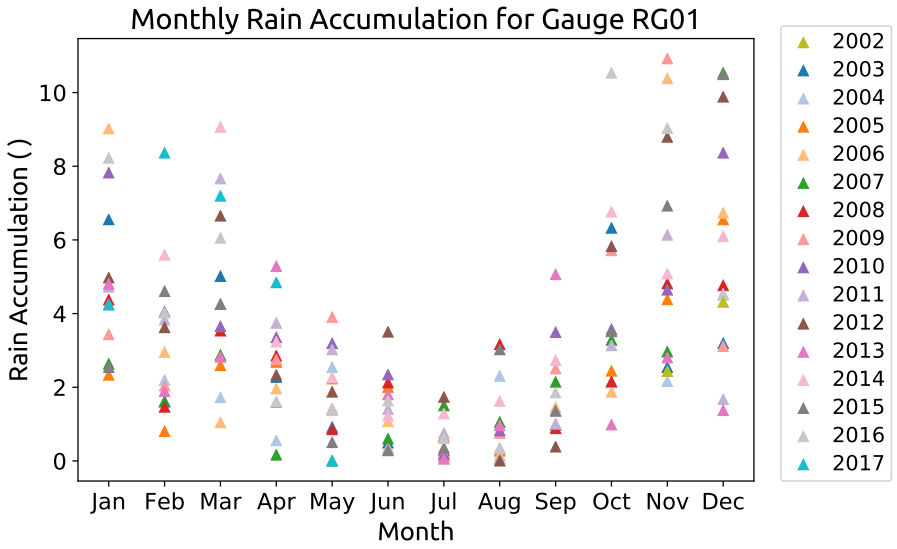

Figure 2: *Rain accumulations measured by the gauge RG01 for all years.*

4.2 Fire 911 Calls

Let us consider the 911 fire calls set. We can import it directly from the Seattle data website and read it as a data frame. It has 4 columns: `Type` of the call, `Datetime`, `Latitude` and `Longitde` of the caller. It has no null values.

In order to see if the fire 911 data is correlated to the rain patterns, we have to associate their location with the location of the rain gauges. In order to do this, we write two functions:

```
great_circle_dist(lat1,lon1,lat2,lon2)
```

which computes the distance in kilometers between two points, given their latitude and longitude, and

```
closest_g(lat,lon)
```

which picks a point and finds the closest rain gauge. By using the latter, we can find the nearest gauge to each 911 caller. Effectively, this defines rain gauge regions as partitions or sectors that cover the city of Seattle. In a way, we are using a nearest neighbor approach indirectly.

We add a new column to the fire 911 calls data frame, `C_gauge`, which gives the nearest rain gauge. In addition, by taking a look of the `Type` column, we see that not all calls are explicitly related to fires. We select here the calls that have `'Fire'` in the `Type` description.

We build a new data frame where we keep only the `Datetime` and the closest gauge of these Fire calls, `C_gauge`. We may also consider in this set only month precision for the `Datetime` column, as we did with the rain data.

This 2-column data frame, `df_fire_10_11_S` already presents some interesting information. We can get a first visualization of the fire data, by displaying the totality of 911 fire calls for each gauge region, as shown in Figure 3.

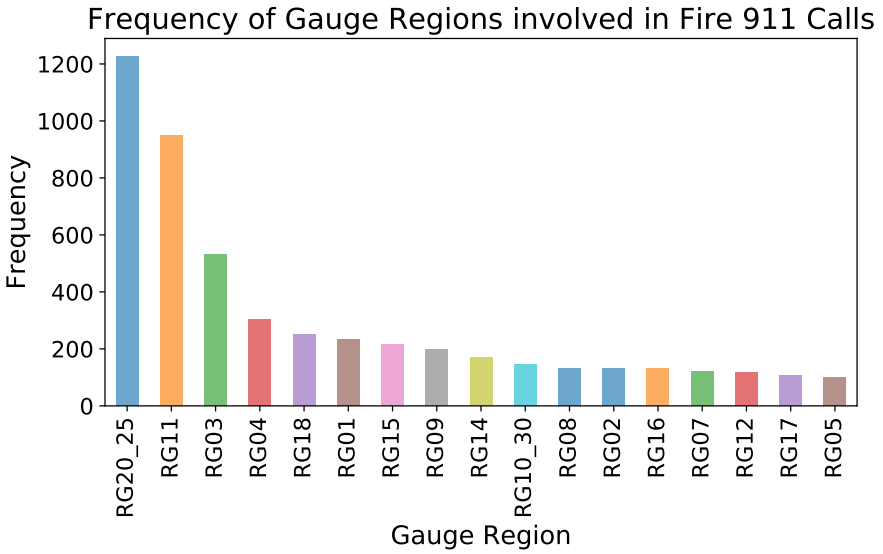

Figure 3: *Totality of fire 911 calls in each gauge region.*

As we see, there are some gauge regions that seem to posses an abnormal number of fire 911 calls. In order to be sure that the functions we built to assign the nearest gauge are working, let us make a graphical double check. We can first grab the totality of the original data, plot their frequency using a 2D histogram which follows latitude and longitude as axes. Then we can add the locations of the gauges and directly observe that the anomalies are real. This corresponds to Figure 4.

There are indeed accumulations of 911 calls around central Seattle, which explains the excessive counting of calls for specific rain gauges. It is interesting to see how the number of calls is enough to reproduce the coastal line in central Seattle.

Since we are interested in the correlation between calls and rain accumulations, we create a pivot table which aggregates the number of calls for each gauge region and each month. Looking at the set, we find that the months corresponding to 2010-06 and 2011-03 have incomplete data, and therefore we remove them. We can do some basic visualization of this information.

We can directly plot the number of 911 fire calls per gauge region per month, as seen in Figure 5. We again find that the gauge regions that we identified previously with abnormal number of calls display this behavior consistently through time.

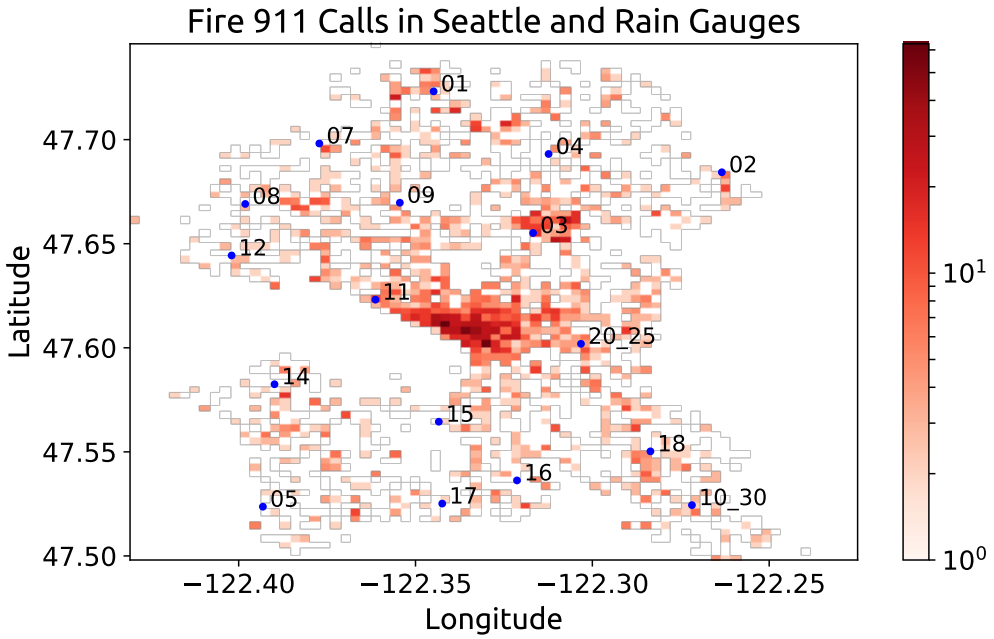

Figure 4: *Distribution of fire 911 calls in the city of Seattle. In blue we see the locations of the rain gauges of the weather data set.*

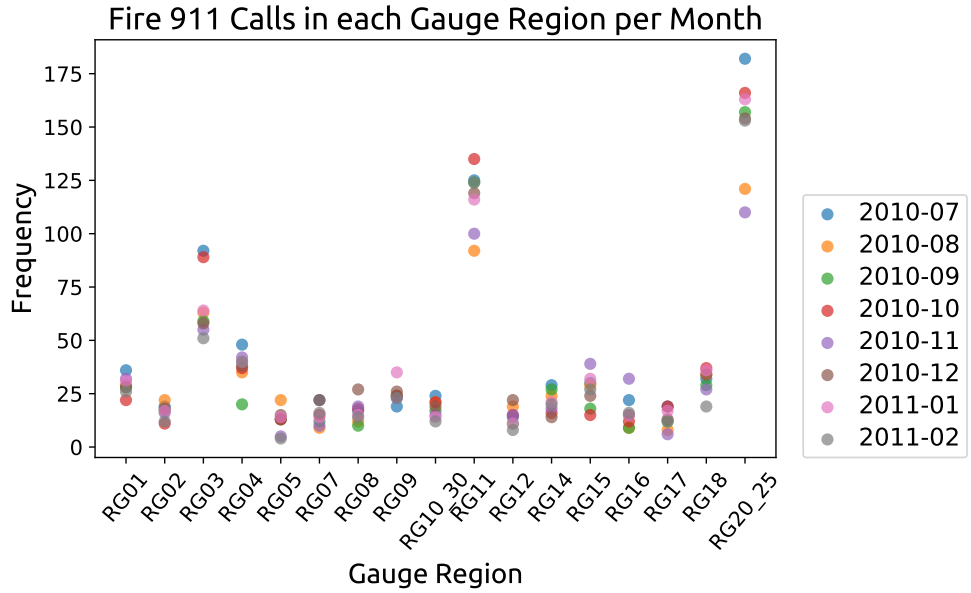

Figure 5: *Number of fire emergency calls per gauge region, per month. We see again that some sectors present an abnormal number of calls.*

4.3 Off days

With the help of the information in [3, 4] on holidays in the state of Washington in 2010 and 2011, we can easily produce a list containing these dates. We store them as `Timestamp` objects that we will be able to relate to the dates in the

fire calls data set. We add the weekends to this list in order to complete a list called `off_days`. Saturdays and sundays in a specific year can be produced with `datetime` functions such as `date` and `timedelta`. With these we write a function `all_x_days(x_day,year)`, which provides all the `weekdays = x_day` (e.g. all mondays) in a specific year.

We have in total 231 off days in 2010 and 2011 together, which represents 31.6% of this period of time. Together with the rain accumulations information, we build a data frame called `df_fire_10_11_off` which contains the data on 911 fire calls together with the rain gauge region and a column that indicates whether the call happened in an off day.

5 Findings and Analysis

We explore a couple of potential effects in the manifestation of fire emergencies.

5.1 Rain and Fire

Let us consider the possibility of a correlation between rain patterns and number of 911 fire calls. Here we present several pieces of evidence against this hypothesis.

As we found in the EDA process, the gauge regions RG20_25 and RG05 display the opposite behaviors when it comes to number of fire 911 calls: the former has more than 12 times the number of calls than the latter. If rain was a dominant factor, we would expect a noticeable difference between the rain patterns for these 2 gauges in the dates of relevance. As we see in Figure 6, this does not seem to be case.

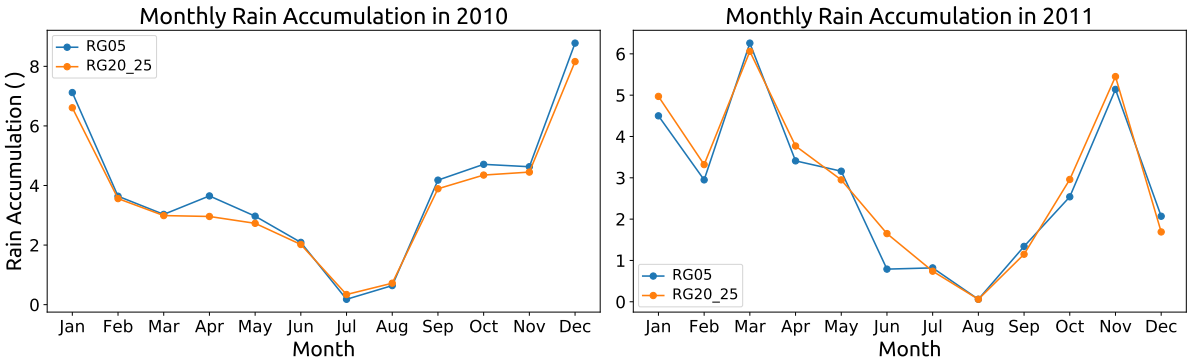

Figure 6: *Rain accumulations in gauges RG20_25 and RG05 for the years 2010 (left) and 2011 (right).*

We can certainly find more compelling evidence. We have already organized our data in such a way that we can compare directly number of 911 fire calls and rain accumulation per gauge region per month. We must pair this information which lays in two distinct data frames. Once this is done, we can make a scatter plot of these points, which we see in Figure 7.

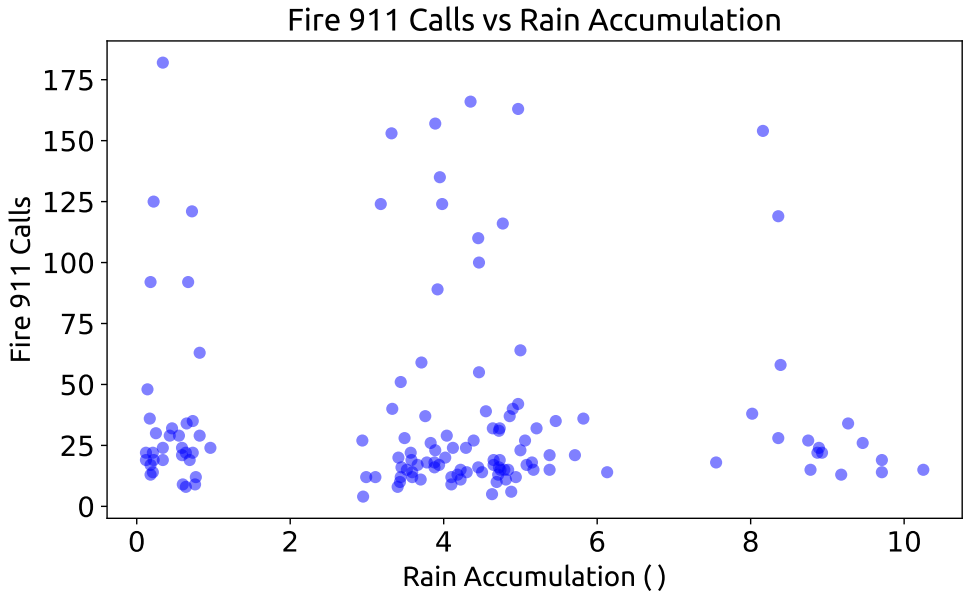

Figure 7: *Number of fire 911 calls paired with rain accumulations for all the months and rain gauges covered in the data sets.*

There is little indication of a clear correlation. We also notice that there is a differentiation of rain accumulation in three distinct subgroups. Indeed, the correlation coefficient is -0.036, indicating that there is about 0.1% of probability of a linear correlation. Even more, the p-value indicates that there is a 67.8% in favor of a null hypothesis for absence of a correlation. This is well beyond any reasonable significance value. There is abundant evidence against the rain patterns being relevant in determining the number of 911 fire calls.

Since we found that the amount of rain accumulation per gauge seems to organize in 3 subgroups, we may be interested in finding the median for the number of 911 calls for each of these 3 subgroups. We may call them points with *reduced*, *medium* and *large* rain accumulations, according to the ranges observed in the x axis of Figure 7. The medians are respectively, 24, 19 and 24. Hence, even when slicing the data in these subsets, we can not find a particular behavior that correlates these two variables.

5.2 Time of the day and Fire

The previous exploration makes us think that instead of weather conditions, human activity might be the most relevant at exploring 911 fire calls. While we did

not found significant difference between month to month data, in these sets we might be able to see whether time of the day plays an important role in the number of 911 calls.

Let us first do a basic counting of *day time* and *night time* calls, by which we will simply mean calls between 6 am and 6 pm for day time and the rest for the night. A first analysis shows a significant abundance of calls in day time relative to the amount of calls in night time. We explore this more clearly by slicing the data hourly.

When we do this, we find an abnormal number of calls in the range between 7 and 8 am. In addition, we found no calls made between 12 pm and 1 am. Due to the difference in orders of magnitude, we do not expect these to be taken as reliable points. For now, we will remove these hours out of our analysis. It is unclear if there is a justification for this odd behavior, which may be due, for instance, to the procedure by which calls are reported or stored at these particular times.

With these corrections taken into account, we find that there are in total 1383 fire emergency calls in day time and 1971 in night time. These numbers show that it is the night time that has more number 911 fire calls after all.

Nevertheless, we should explore the possibility that this is an artifice of the data sample. To study this alternative, we can test this proportion against a null hypothesis of a 50% amount of night 911 fire calls. To do this we resort to a cumulative binomial distribution and compute the probability of obtaining the observed proportion or higher, assuming the null hypothesis. This p-value is of the order of 10^{-24}, so it is vanishing in practice. This is well below any reasonable level of significance, which suggests we should reject the null hypothesis. This indicates that there are certainly more 911 fire calls in the night. We can see this behavior more clearly in Figure 8.

5.3 Off days and Fire

Let us explore the relation between off days and number of emergency calls. We can take a look of the raw number of 911 calls that happened in off days and in regular work days. We find that about 32.1% of calls happen in off days. This is extremely close to the fraction of off-days to total number of days in 2010 and 2011 that we mentioned before (31.6%).

It is then reasonable to explore the possibility that the calls in off days simply follow the same proportion that one observes for off days versus the totality of days in 2010 and 2011. We can use this as a null hypothesis for a one-sample proportion test. If the null hypothesis can not be rejected, we would have an indication that whether the days are work days or not does not substantially influence the total number of emergency calls.

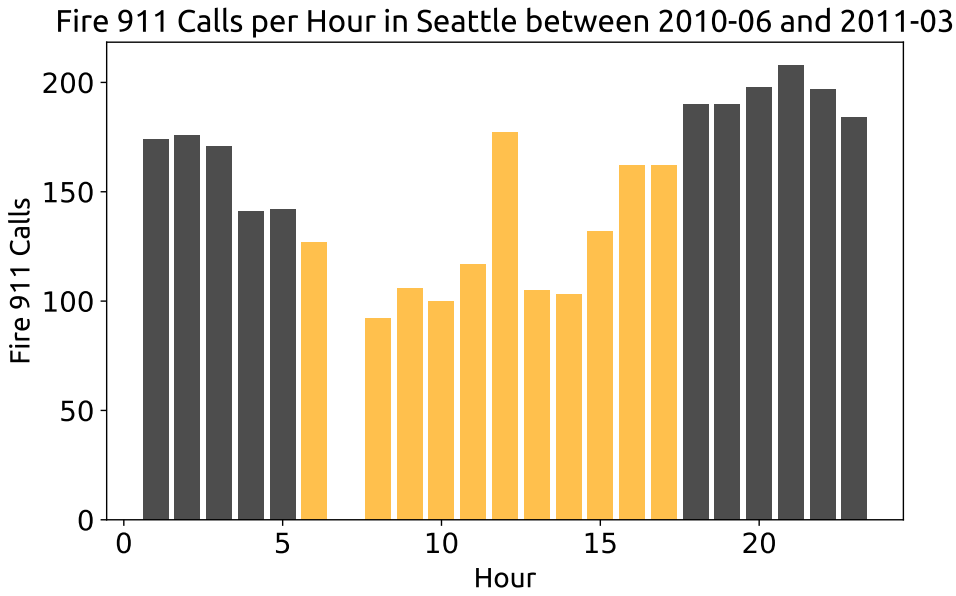

Figure 8: *Number of fire 911 calls grouped by the hour of the call. There is spurious data for the periods between 7 and 8 am and between 12 pm and 1 am, which are hence missing in the plot.*

By using a cumulative binomial distribution, we find that the probability of the proportion of off day-calls being as extreme or larger than the one we observed under the assumption of the null hypothesis is $p = 0.236$. This is well above any reasonable level of significance, which certainly does not allow us to reject the null hypothesis. This is an indicator that, if the calls are taken as a whole and not split into any categories, there seems to be no influence of the day being a work day or an off day in the number of 911 fire calls.

6 Machine Learning Analysis

Let us take a look of these data sets from the point of view of machine learning techniques. To do this, we perform some additional data wrangling to set it ready for a clustering process. We focus fundamentally on the categories `Longitude`, `Latitude`, `Off-Day` (which indicates whether the call happened in an off day) and a set of 27 columns which span the categorical information in the `Type` column.

6.1 Clustering

We explore different approaches to the clustering of this data set. These are:

1) Clustering only `Longitude` and `Latitude` of the emergency calls. We do this with `MinMaxScaler` and `KMeans` with distinct number of clusters k. We compute the most

common parameters to decide the ideal k, such as inertia, silhouette, gap statistic and the gap criterium of [5].

2) We then tackle the whole data set with a `PCA` dimensional reduction. By computing variance we try to establish the ideal dimension. This can be done with and without previously scaling with `StandardScaler`. We find that in this case the model responds better without previous scaling.

Looking at the behavior of the cumulative explained variance, one sees that 2D PCA represents accurately about 69% of the data, and the elbow rule would suggest that the ideal dimension is likely 3, representing about 76% of the data. We perform clustering with `MinMaxScaler` and `KMeans` of the projected 2D and 3D data, using the same parameters we mentioned in 1) to establish the ideal k.

3) Finally, we also experiment with clustering all data without using PCA. Here we do this with `MinMaxScaler` and `KMeans`, testing, as before, different possible k's.

An in depth analysis of these distinct approaches and related plots can be found in Appendix A.

6.2 A Clustering Story

With the help of the routines we implemented, we can establish some interesting points regarding these data. For presentation purposes, we leave the graphical study of this story to the Appendix B, from which we are deriving the following observations.

Let us start with the clustering location only, in which we intentionally limit ourselves to a subset of the data. By doing so, we have the opportunity to present a data-based method to establish a sectorization of the city, based on 911 calls. This in itself is useful as one may be interested in distributing resources to deal with these fire emergencies based solely on the raw amount of calls. Unsurprisingly, the distributions of latitudes and longitudes are well localized for this type of clustering.

The proportion of off day calls in each cluster is relatively uniform throughout the whole city, oscillating between 30% and 35%. This goes along the expectations of the exploratory data analysis we made for the totality of calls. When we go to the data on type of calls, we verify that central Seattle (mostly covered by cluster 3) is a priority in most of the categories. Automatic fire alarms are the biggest component of emergency calls, and both buildings and cars are important locations in fire production. Together with central Seattle, a significant percentage of emergency calls is found to come from the north in all categories.

We then moved to analyzing the totality of data with dimensional reduction. We first attempted to scale the data before performing a PCA fit, but we found that it

required a really high dimension to produce an effective description of the data. Instead, we decided to use PCA without previously scaling, and the performance was much better. Looking at the behavior of the cumulative explained variance, one sees that 2D PCA represents accurately about 69% of the data, and the elbow rule would suggest that the ideal dimension is likely 3, representing about 76% of the data. We then performed clustering with the projected 2D and 3D data.

Let us start with the 4 clusters we found with 2D PCA. The first notable observation is that the clusters align clearly in terms of the off day category; each cluster possesses a clear preference for work or off days. In addition, the second priority of these clustering seems to be the type of call category: the two most frequent types are directly covered by pairs of clusters. More specifically, cluster 0 are the off day calls that fill the most popular category, while cluster 1 are the weekday calls that fill the same category. We find the same structure for the second most popular category and clusters 2 and 3. Even more, clusters 0 and 1 fill all the remaining types of calls, again splitting the off and work day calls. These clustering is then very useful when it comes to studying the off day condition. Hence, when we go to the geographical distribution of the 4 clusters, we have a picture of the distribution of off/work day fire emergency calls.

In the case of 3D PCA, we picked a 6-clusters model. These are also well aligned according to the off-day category. In addition, the first 4 clusters fill the two most popular call types as in the 2D PCA case. In contrast to the 2D case, clusters 0 and 1 fill all the remaining types of calls except for one: this one type is filled by the last clusters 4 and 5. The direct comparison between these two dimensions shows quite clearly some features of these clustering methods. PCA is clearly putting the priority on the features with a bigger ratio of `True` (or 1) values, as is expected. This is why the less popular categories are the less aligned with respect to the clusters. Then, once we allow a higher PCA dimension or a bigger number of clusters, the next-to-most popular categories are aligned with the clusters. Consequently, these methods automatically and systematically select the most relevant features of a data set and allow us to study their rich intersections.

Lastly, we approached the case of clustering without PCA. We used a 8-cluster model in this case. The first 4 clusters are following the same structure of the 4 clusters in the 2D PCA case. The remaining clusters then align more to the type of call category and mix week and off days. Thanks to that, we get to explore more clearly less popular type of calls, such as building and rubbish fires.

7 Concluding Remarks

We found no indication that rain accumulations and the number of fire 911 calls are correlated. Fire seems to be more directly related to human activity, as we found that about 53.5% of the fire emergencies happen in central Seattle. It would

be interesting to try to study a correlation of 911 fire calls with population density, if the data were available.

Most of fire emergency calls happen during night time, which seems to reinforce the relevance of human factors over weather. Nevertheless, there is a small peak of activity in the midday, which could match a peak of solar activity. In addition, the proportion of total 911 fire calls is not affected by the day being a work day or not.

With machine learning we obtained a data-based approach to the sectorization of the city in accordance to the number of fire emergency calls. This can be seen in the bottom-left plot of Figure 9. We found that automatic fire alarms are the most significant cause of 911 responses, even after removing the false alarm calls. Car fires contribute substantially in these events and Central followed by North Seattle are a priority in the amount of emergencies. Coastal activity is secondary in the manifestation of 911 emergencies. The amount of rubbish fires is unfortunately notable, a subset of emergencies that can be specially treatable with a better observance of waste disposal.

Policies for awareness and prevention of fires are a good approach to the reduction of fire related emergencies. These can be focused in fires related to urban activity, specially in highly populated regions of the city, and for events that unfold typically during night time.

8 Bibliography

[1] N. Daniels, "Seattle observed monthly rain gauge accumulations," July, 2018. https://www.kaggle.com/city-of-seattle/seattle-observed-monthly-rain-gauge-accumulations/version/16. Version 16. Retrieved August 20, 2018.

[2] C. of Seattle Fire Department Management Information Systems, "Seattle fire 911 calls from 3/1/2010 to 3/1/2011," Sept., 2018. https://data.seattle.gov/Public-Safety/Seattle-Fire-911-Calls-from-3-1-2010-to-3-1-2011/d9j6-s59d. Version: September 2, 2018 update. Retrieved September 2, 2018.

[3] OfficeHolidays, "Public holidays in washington in 2010," Oct., 2018. https://www.officeholidays.com/countries/usa/regional.php?list_year=2010&list_region=washington. Retrieved October 10, 2018.

[4] OfficeHolidays, "Public holidays in washington in 2011," Oct., 2018. https://www.officeholidays.com/countries/usa/regional.php?list_year=2011&list_region=washington. Retrieved October 10, 2018.

[5] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** no. 2, 411–423, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00293. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00293.

A Clustering Analysis

We start with a clustering of only the location of the 911 emergency calls. This is registered in the columns `Longitude` and `Latitude`. Fisrt we pick store this location data in an array and scale it with `MinMaxScaler`; this sets the range of data to [0,1]. We follow with the application of `KMeans`. As we will do repeatedly, we use a series of parameters to determine which number of clusters is ideal. This includes inertia, average and individual silhouettes, gap and the gap criterium.

1) Inertia corresponds to the sum of squared distances of samples to their closest cluster center and can be computed with the attribute `inertia_`.

2) `silhouette_samples` takes into account both the distance of each point to points in its own cluster and to points in others to give a measure of how solid the structure found in a model really is. `silhouette_score` computes an average over all points.

3) The gap statistic $\mathrm{Gap}(k)$ compares the inertia of the model and the inertia of models built from random arrays of identical size and range. Ideally, high values of this parameter are preferred. In addition, we use the criterium which suggests that the ideal k is the lowest one that satisfies [5]

$$\mathrm{Gap}(k) - \mathrm{Gap}(k+1) + s_{k+1} \geq 0\,, \tag{1}$$

where s_k is a (weighted) standard deviation of the random sample clusterings. We store these parameters and we plot them for all the clusterings we performed, as we show in the following subsections.

Let us take a look of the plots in Figure 9. The elbow rule in the inertias plot seems to make a case in favor of the values $k = 5, 6$. These also have high average silhouettes, together with $k = 9$, although it should be mentioned that all values are below 0.5. $k = 1, 2, 6, 10$ have high $\mathrm{Gap}(k)$, but when we take a look of the gap criterium (1), we see that $k = 2$ is the lowest number of clusters that have a positive score. It should also be noticed that $k = 3, 6, 7$ also have a positive gap-criterium score. All factors being taken together, we decide to pick $k = 6$ in this study, in the spirit of following fundamentally the silhouette and $\mathrm{Gap}(k)$ score. We can also see in Figure 9 a silhouette plot for $k = 6$ that provides a graphical representation of individual scores.

We then proceed to perform dimensional reduction via PCA. First, we consider the case of scaled data via `StandardScaler`. After doing a PCA fit of the data, we study the behavior of explained variance for different number of dimensions. By looking at the top two plots in Figure 10, we find that this is not an ideal approach. It seems to suggest that only high dimensionality can explain a significant percentage of the data behavior. On the other hand, we can instead make a PCA analysis without performing a scaling. The results can be seen in the middle two plots of Figure 10, where we find a more promising scenario. With 2 dimensions we can account for about 69% of the data and with 3 we reach almost 76%. The elbow rule indicates that $D = 3$ is a good choice for dimensional reduction. We then proceed to avoid scaling in further studies of this data set.

Interestingly, we have very well delimited clusters after performing these 2D and 3D reductions, as can be seen in the projected data in the bottom two plots of Figure 10. We then proceed to use a `MinMaxScaler` + `KMeans` clustering on the reduced data and we follow the same steps we used in previous searches to pick the ideal k.

In the 2D case, the gap criterium seems unreliable, as seen in Figure 11. We conjecture that one possible reason behind this odd behavior for this particular data set could be the radical change in scale for the inertias as k reaches the elbow point. On the other hand, $\mathrm{Gap}(k)$ is monotonically increasing. The average silhouette and the elbow rule (see Figure 11) both seem to point to the $k = 4$ expected result, that we could have predicted from the 2D projected data. The silhouette and labeled data plots for $k = 4$ show the desired clustering as shown in the bottom plots of Figure 11.

Now we proceed to perform PCA reduction to $D = 3$. As in the prvious case, the gap criterium does not seem to be a useful parameter and $\mathrm{Gap}(k)$ is a monotonically increasing functionm as seen in Figure 12. $k = 6$ seems like a good choice from the point of view of the elbow rule in the inertias plot and the behavior of the silhouette score. This also seems reasonable from the point of view of the projected data scatter plot.

Finally, we attempt to perform a clustering of the complete data set without performing a dimensional reduction. We follow the same `MinMaxScaler` + `KMeans` recipe and perform the same steps to explore the ideal k. As we see in Figure 13, while the gap behavior is again not ideal, the choices $k = 6, 8, 9$ seem promising by looking at the inertias and the silhouettes. We pick $k = 8$ and plot silhouettes for this value in Figure 13.

A.1 Clustering Location only

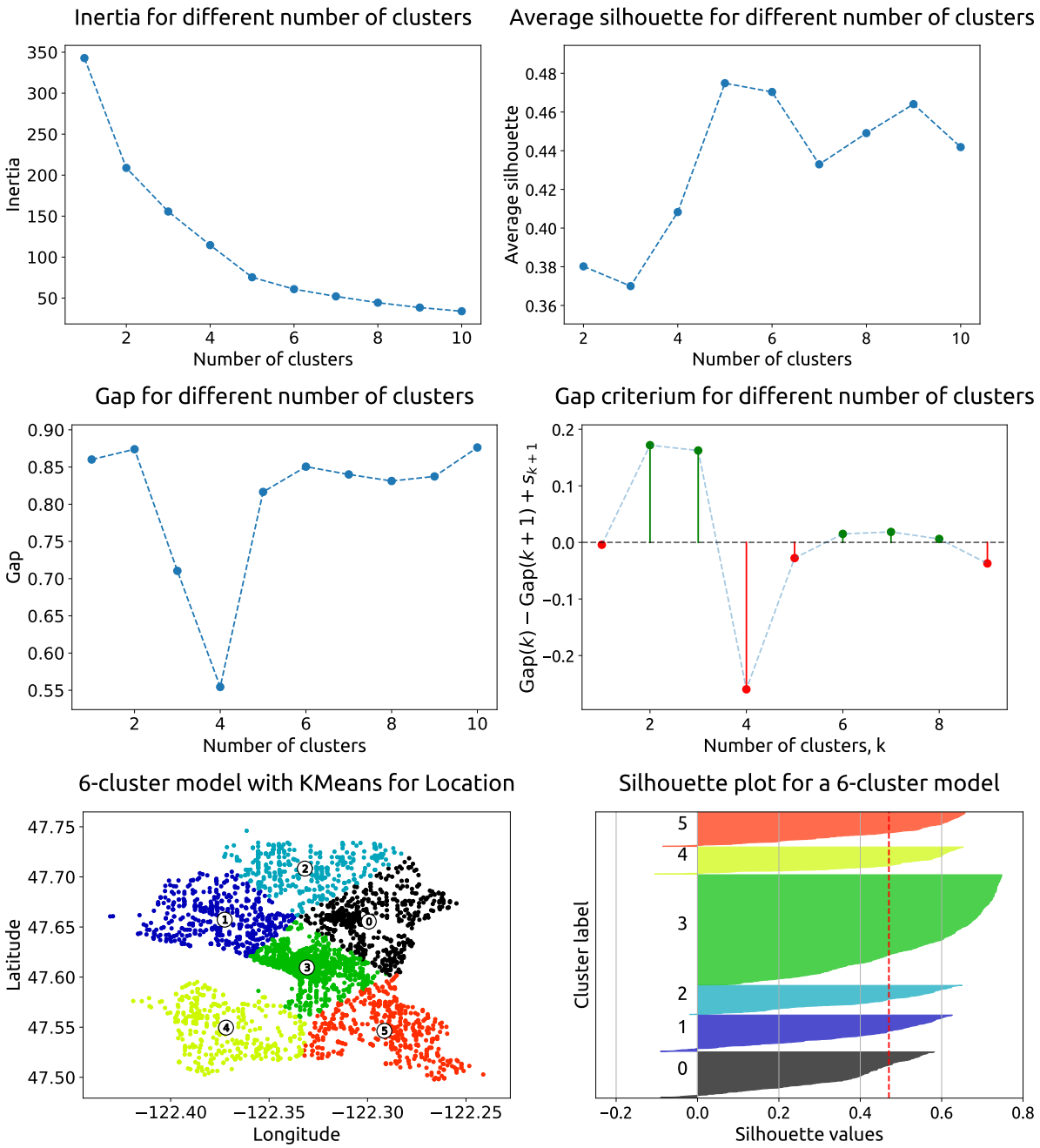

Figure 9: *Top 4 plots show the explored parameters in the clustering of location as functions of the number of clusters k. At the bottom, graphical distribution of clusters and silhouette plot for the chosen number of clusters, $k = 6$.*

A.2 Clustering with PCA with and without previous Scaling

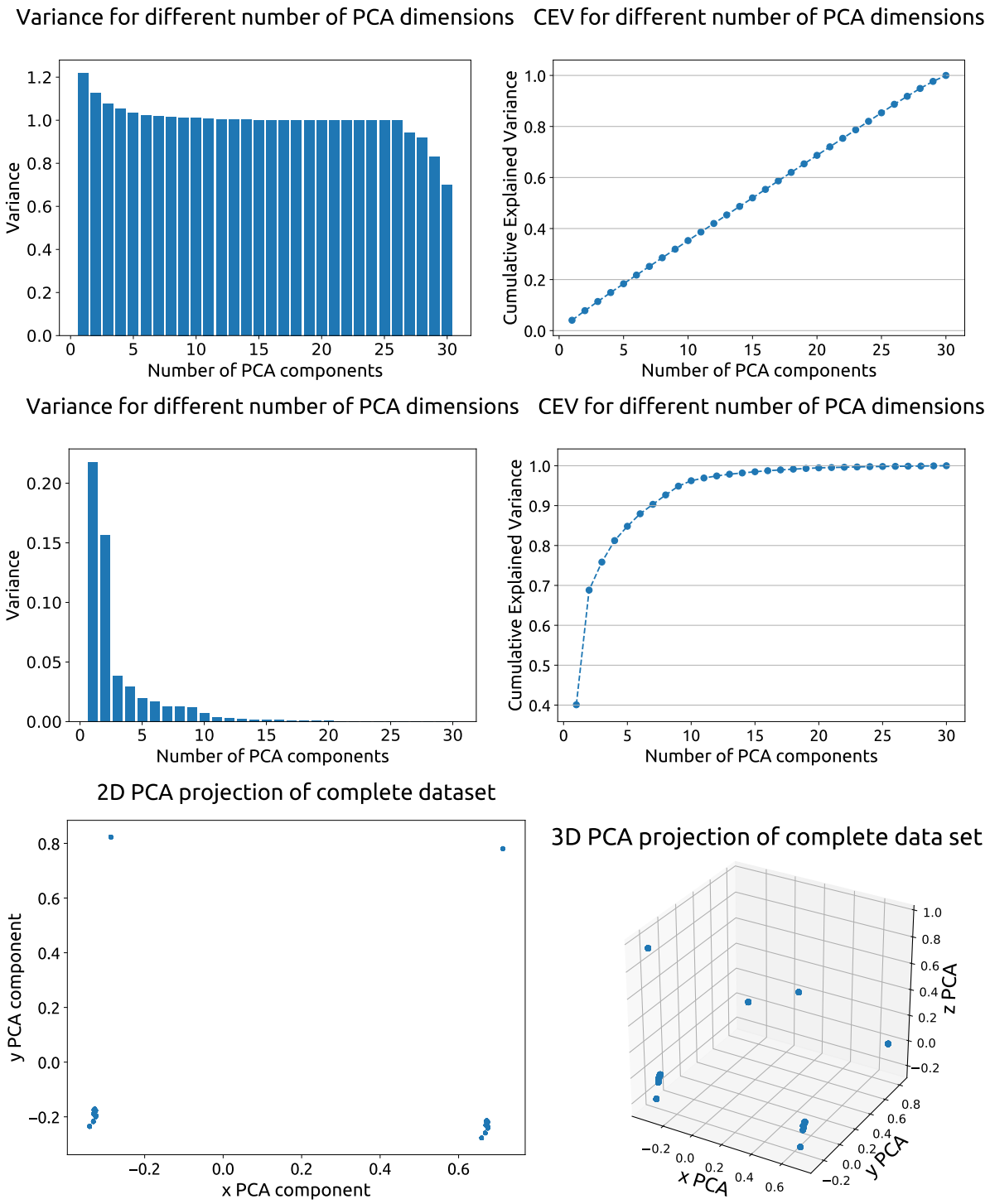

Figure 10: *Top four plots show the explained variance for a PCA fit with different number of dimensions: top two obtained* **with** *previous scaling while middle two obtained* **without** *previous scaling. The bottom two plots show the data after doing PCA* **without** *scaling when one picks 2D (left) and 3D (right) projections.*

A.3 Clustering 2D PCA

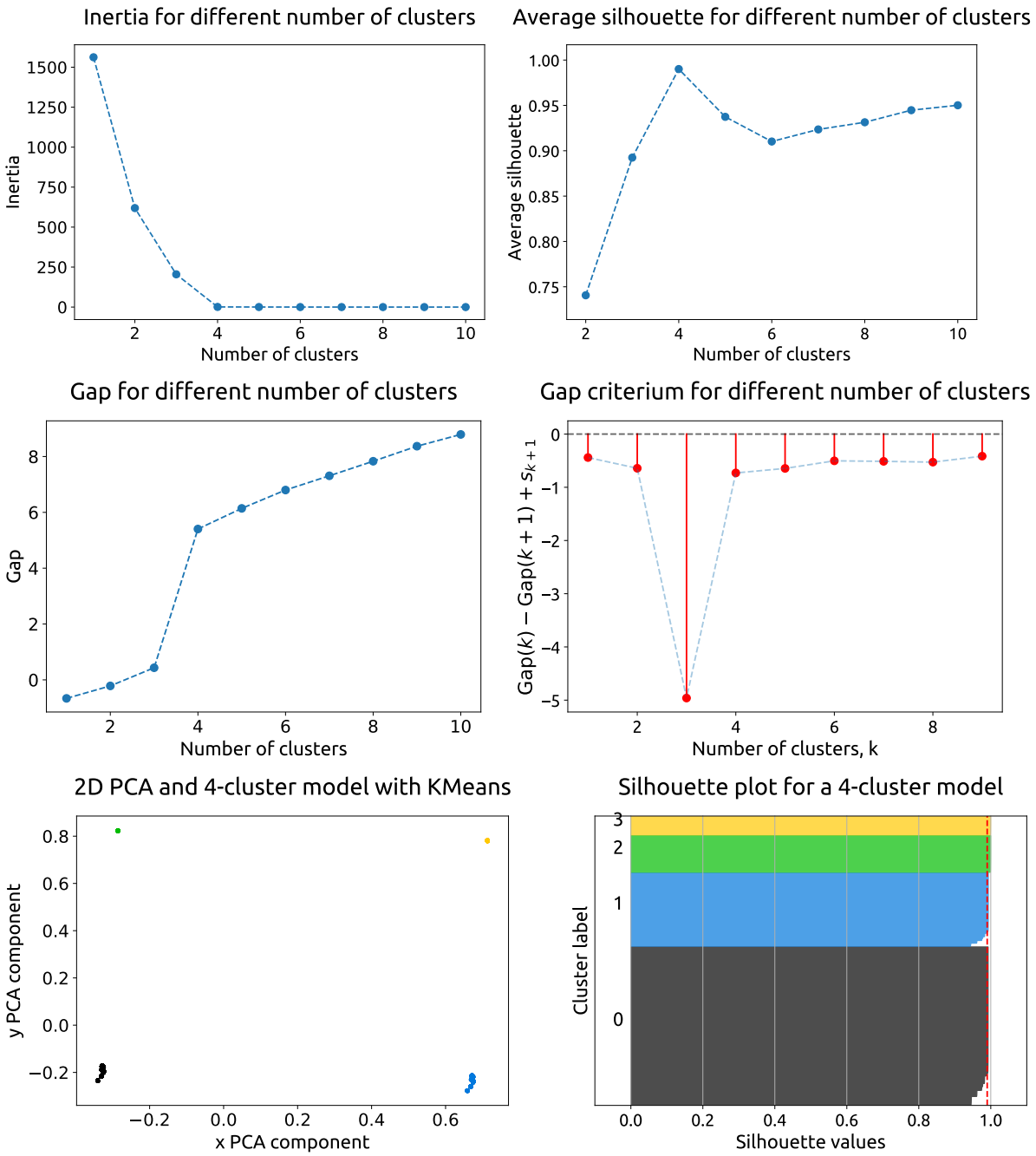

Figure 11: *Top 4 plots show the explored parameters in the clustering of 2D PCA as functions of the number of clusters k. At the bottom, distribution of clusters in the PCA subspace and silhouette plot for the chosen number of clusters, $k = 4$.*

A.4 Clustering 3D PCA

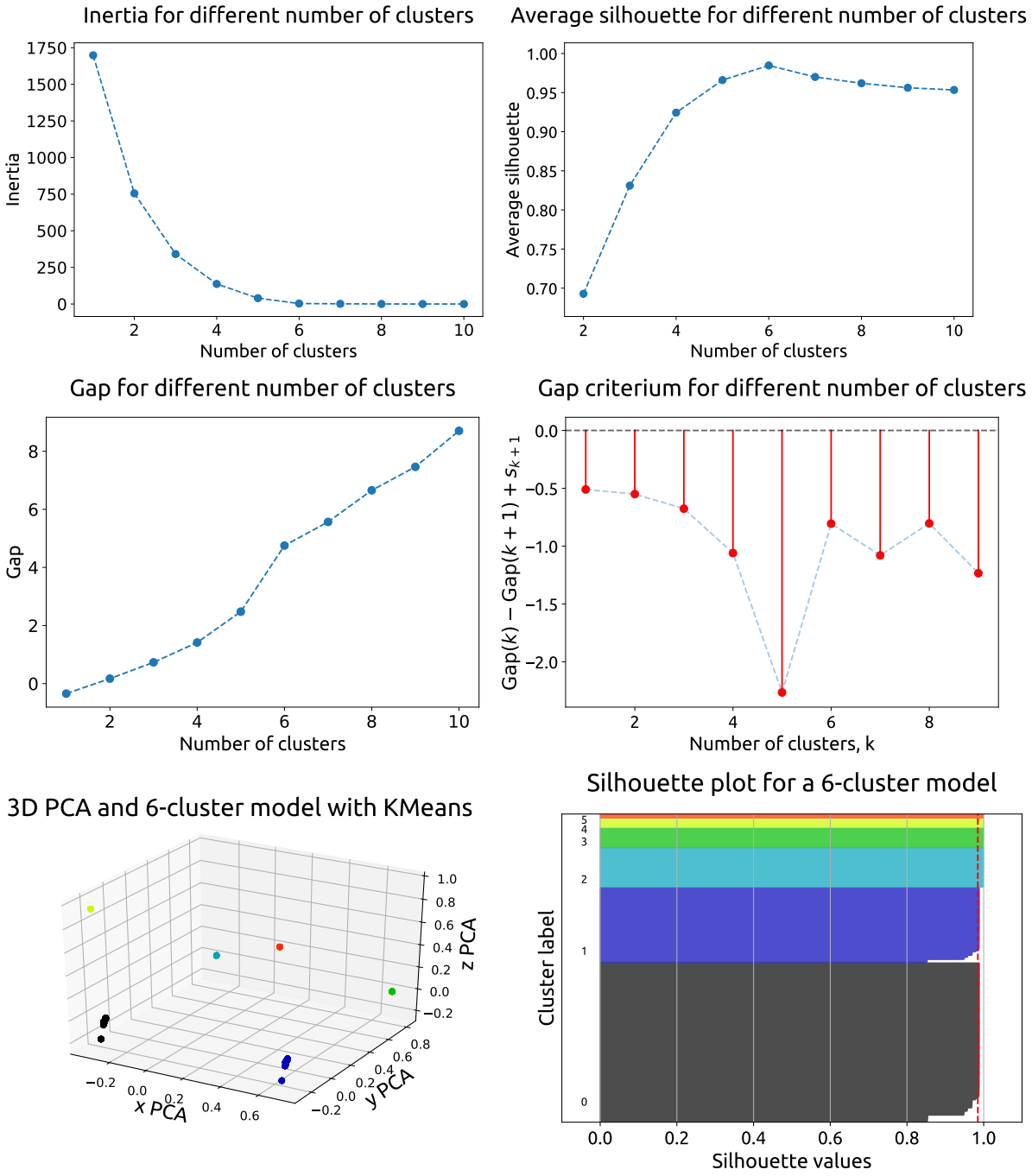

Figure 12: *Top 4 plots show the explored parameters in the clustering of 3D PCA as functions of the number of clusters k. At the bottom, distribution of clusters in the PCA subspace and silhouette plot for the chosen number of clusters, $k = 6$.*

A.5 Clustering without PCA

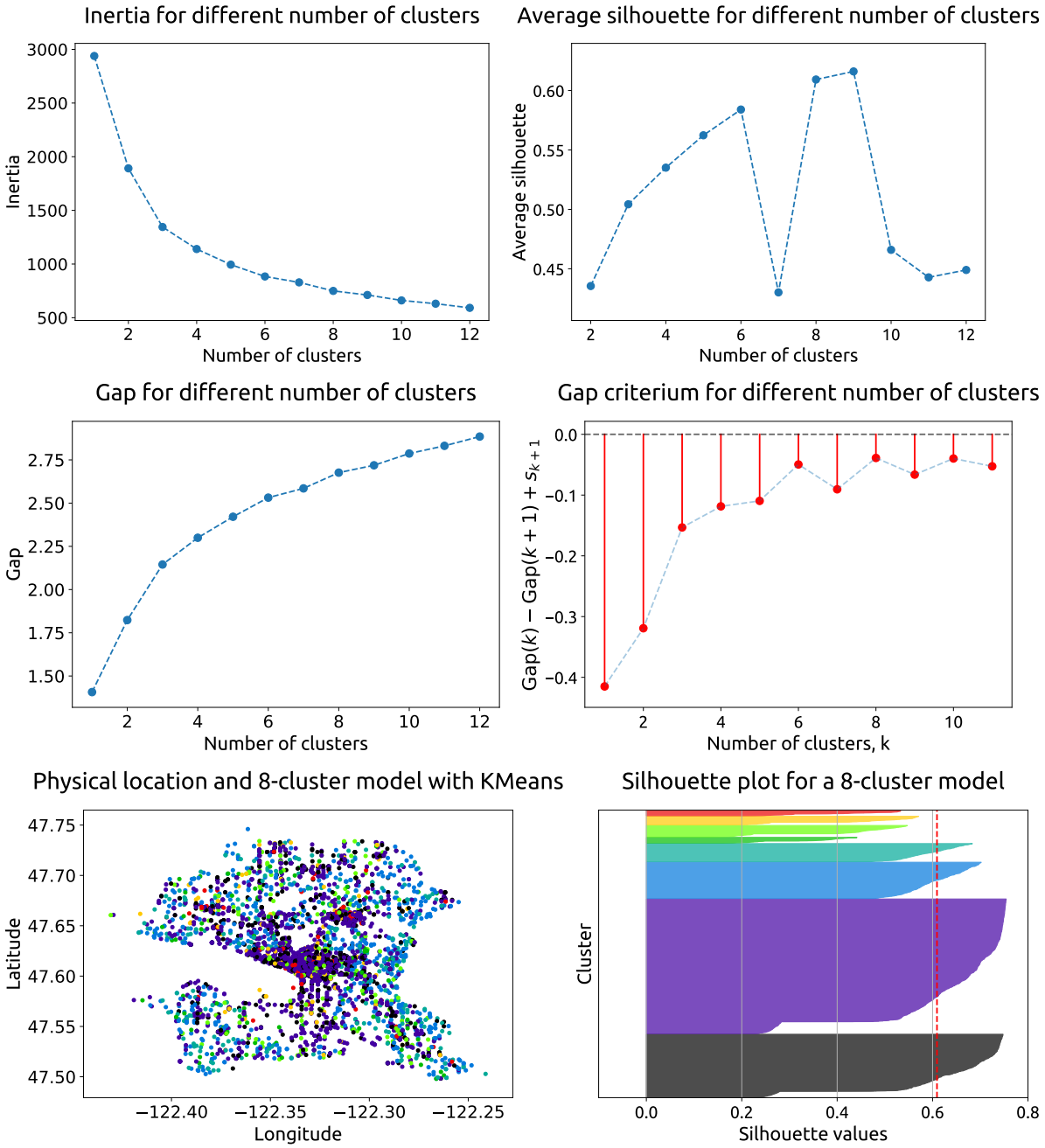

Figure 13: *Top 4 plots show the explored parameters in the clustering without PCA as functions of the number of clusters k. At the bottom, graphical distribution of the clusters and silhouette plot for the chosen number of clusters, $k = 8$.*

B Clustering Feature Plots

Here we explore how every clustering we made covers the distinct features of the data set. We present 3 types of plots for each clustering:

1) location plot with the cluster labels as colors,

2) distribution of off day and work day calls in each cluster,

and

3) distribution of clusters in the categories of the `Type` column.

B.1 Clustering Location only

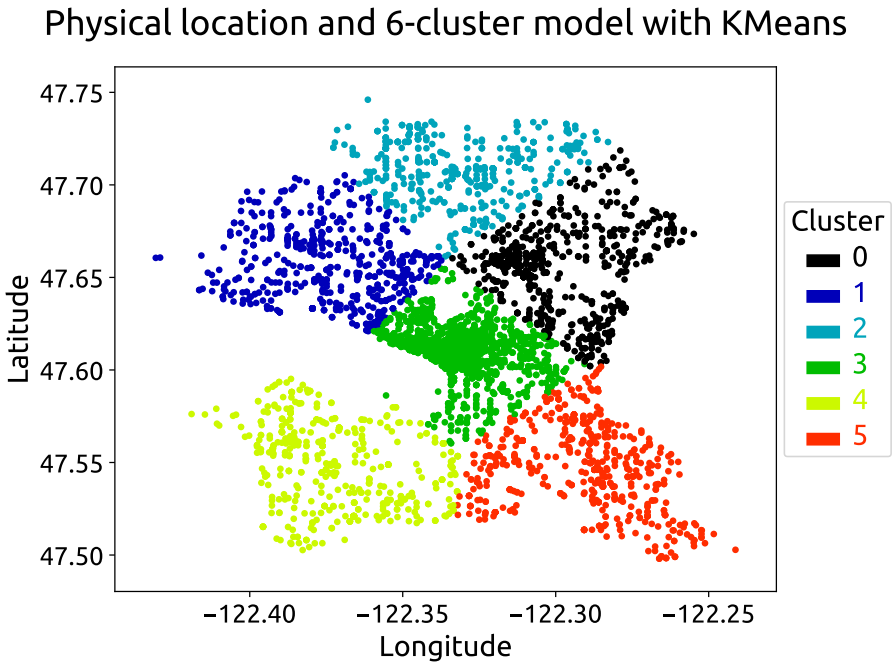

Figure 14: *Geographical distribution of clusters when clustering location only.*

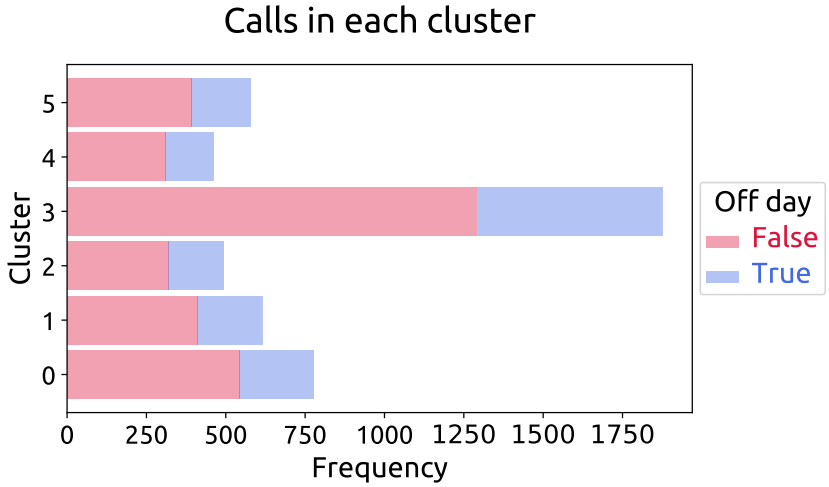

Figure 15: *Distribution of off day and work day calls when clustering location only.*

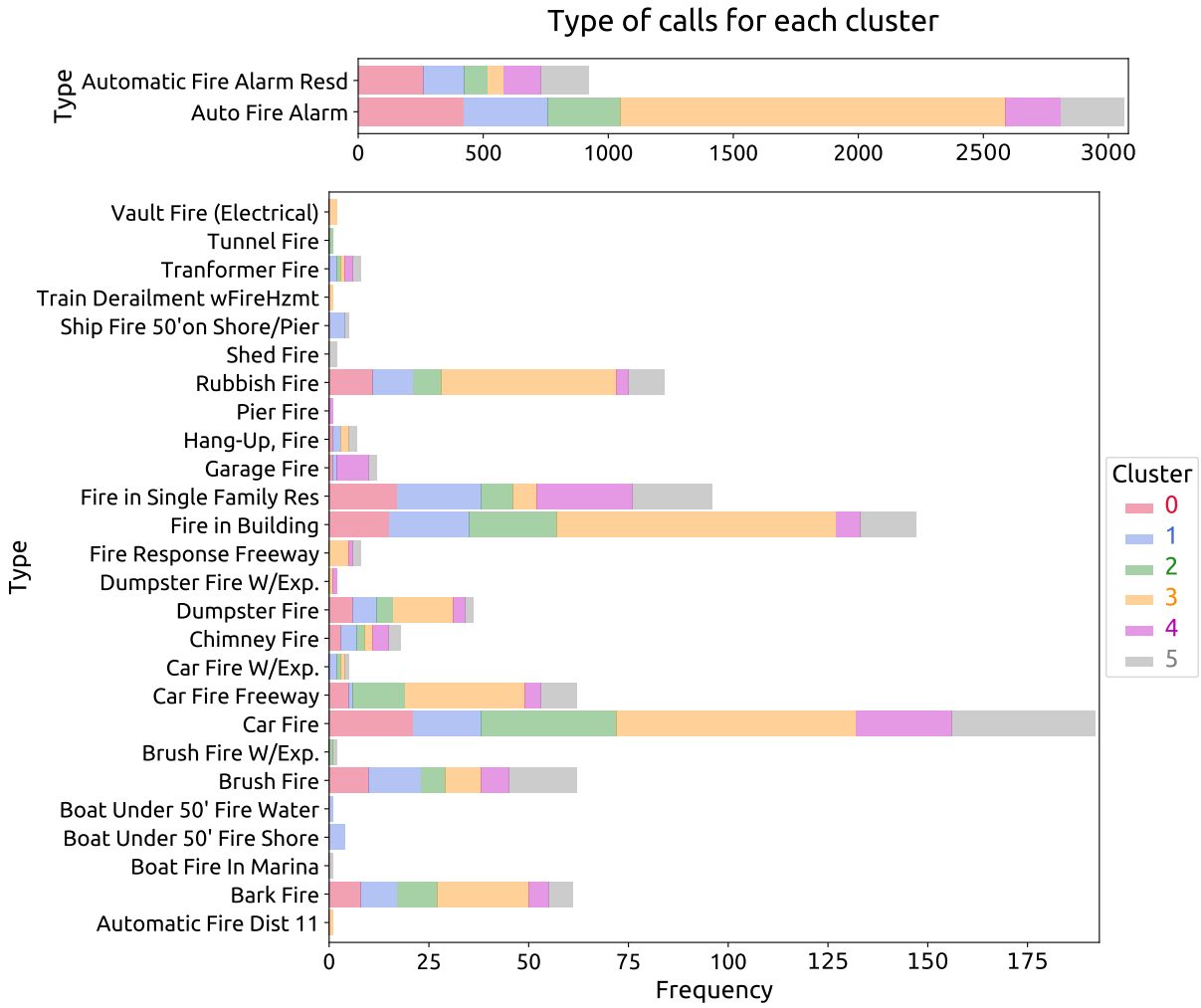

Figure 16: *Distribution of clusters in the type of calls categories when clustering location only.*

B.2 Clustering 2D PCA

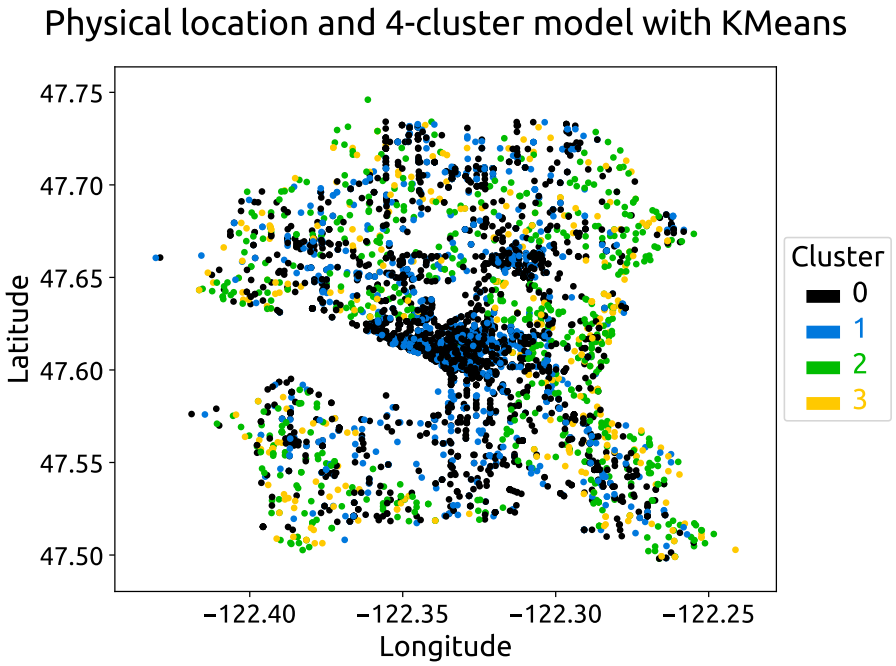

Figure 17: *Geographical distribution of clusters when clustering 2D PCA.*

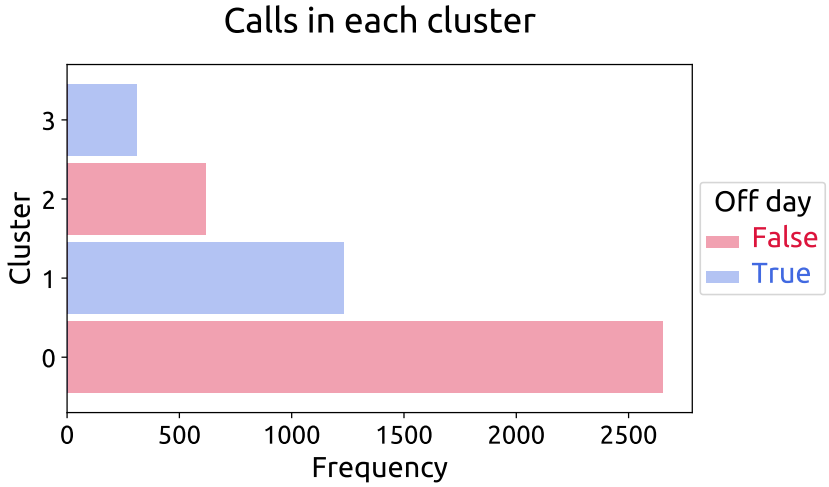

Figure 18: *Distribution of off day and work day calls when clustering 2D PCA.*

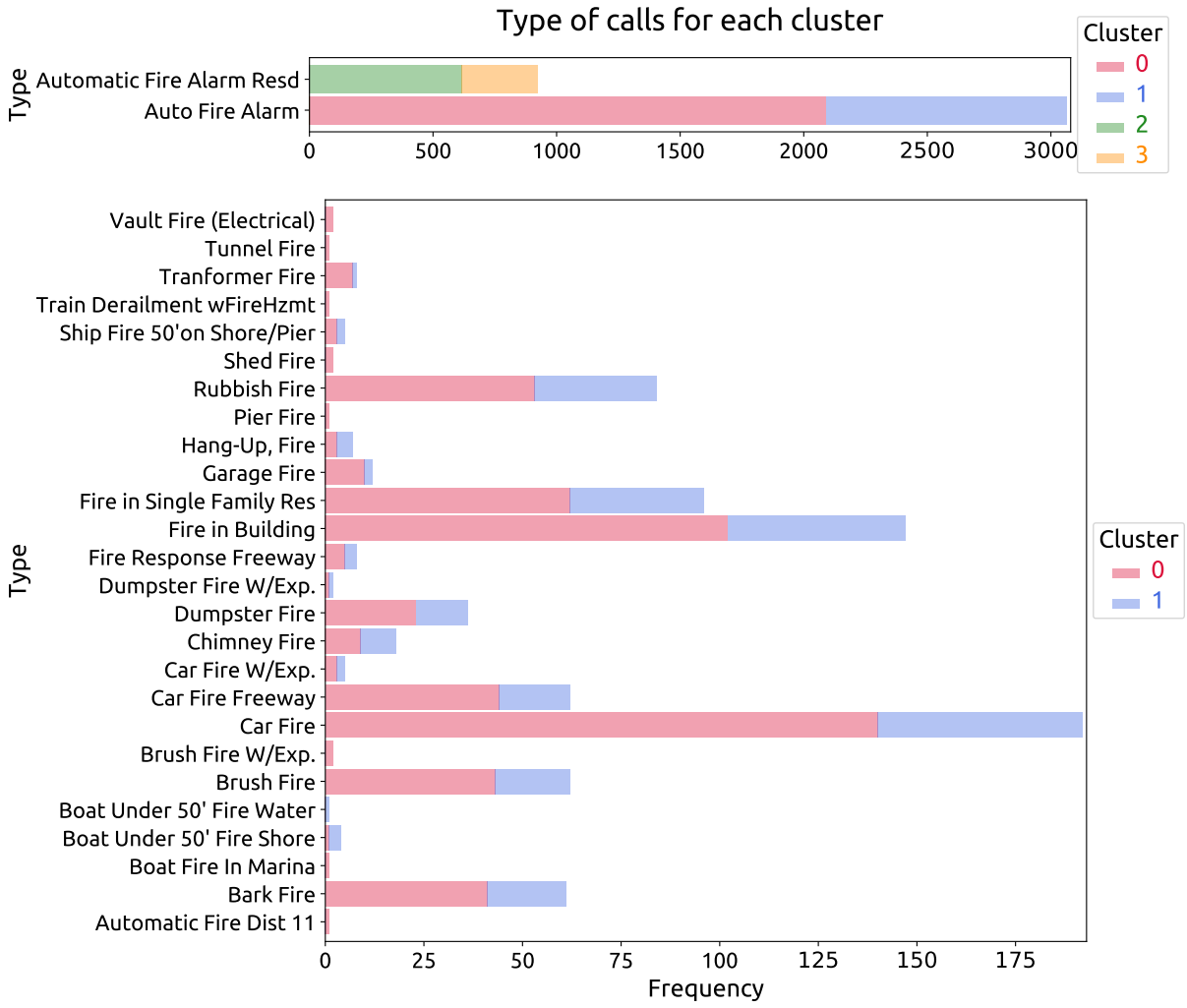

Figure 19: *Distribution of clusters in the type of calls categories when clustering 2D PCA.*

B.3 Clustering 3D PCA

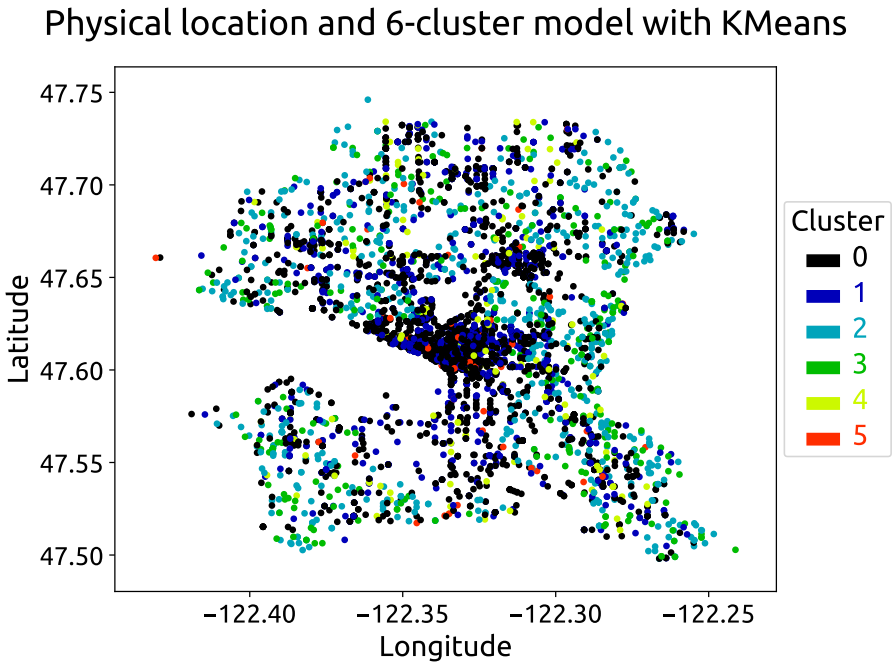

Figure 20: *Geographical distribution of clusters when clustering 3D PCA.*

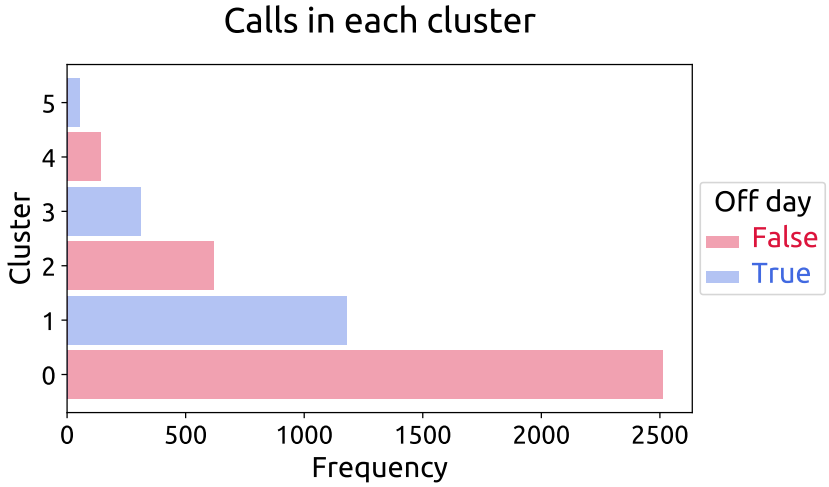

Figure 21: *Distribution of off day and work day calls when clustering 3D PCA.*

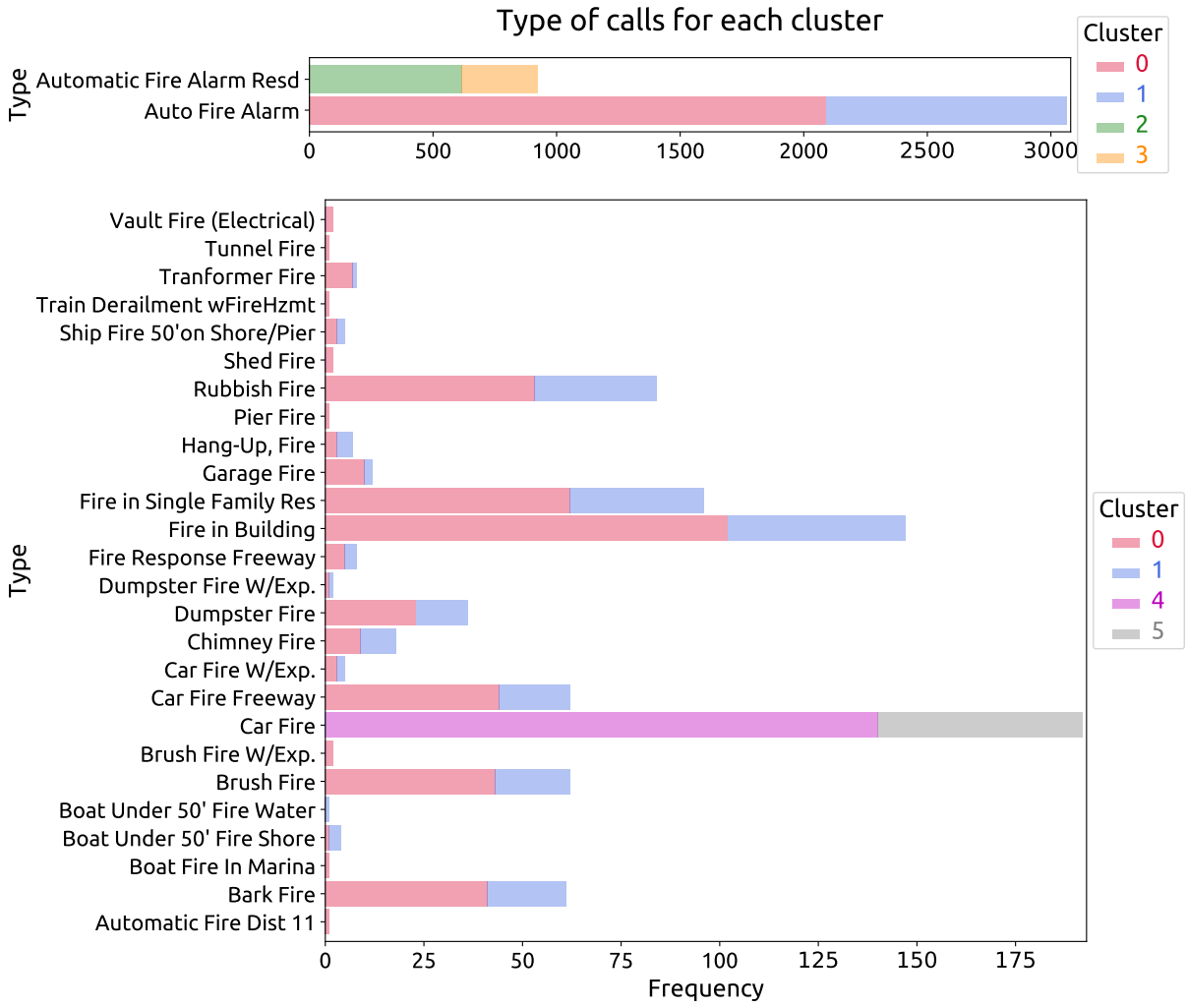

Figure 22: *Distribution of clusters in the type of calls categories when clustering 3D PCA.*

B.4 Clustering without PCA

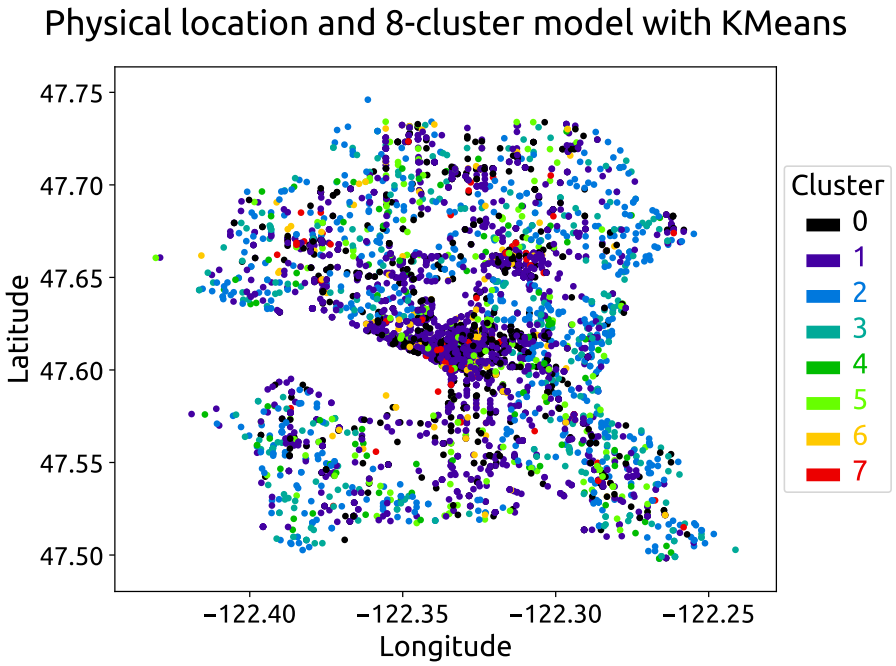

Figure 23: *Geographical distribution of clusters when clustering without PCA.*

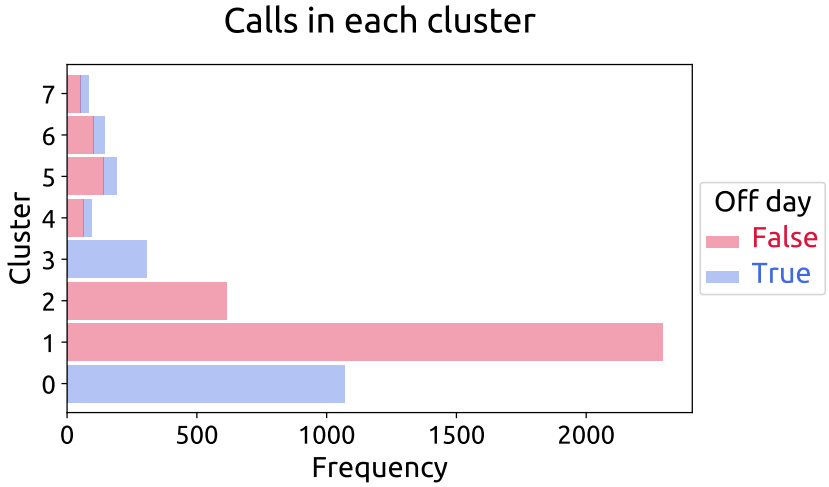

Figure 24: *Distribution of off day and work day calls when clustering without PCA.*

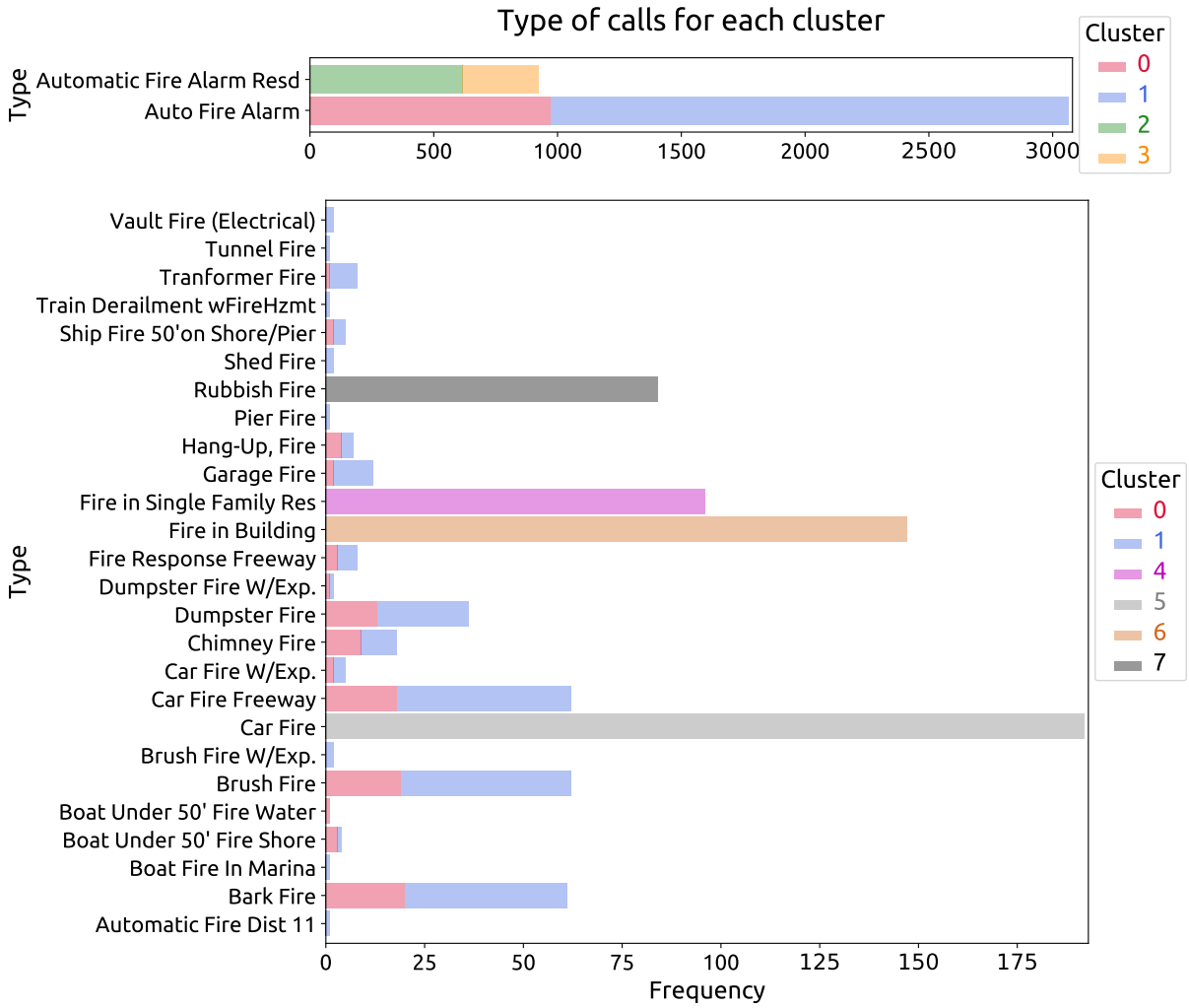

Figure 25: *Distribution of clusters in the type of calls categories when clustering without PCA.*