Insights from Time and Spatial Analysis Using Behavioral Health Calls in Detroit

Client : Anna Bauer
Investigator : Geonhyeok Jeong

30 November, 2023

1. Abstract

The purpose of this statistical consultation is to identify suicide 'hot spots' rather than individual risks, utilizing a population-level approach. To discern patterns and conduct a more thorough analysis, our focus centers on behavioral health data in Detroit, which often indicates mental health problems that could potentially lead to suicides. Several analyses and hypothesis tests are conducted with time and spatial analysis.

Before constructing a predictive model for suicide, we scrutinize the data and identify significant variables to ensure the development of a robust model. We have a general idea of what might drive someone to attempt suicide and acknowledge its highly time and context-dependent nature. Our analysis supports these assumptions:

- According to the Chi-square test, the results demonstrate a significant temporal dependency of call volume, varying hourly, daily, and monthly.
- Through simulation and ANOVA tests, we discover that the frequency of calls is not proportional to the population. In fact, regions with smaller populations proportionally have more calls relayed to 911.

These analyses reveal the influence of both time and population size on call counts, highlighting the need for the consideration of additional variables when constructing an effective predictive model.

2. Exploratory Data Analysis

The dataset, provided by Anna Bauer, encompasses 911 calls in Detroit in 2017, and is compromised of both original and revised data for analysis. The original dataset has around 190000 observations, reflecting a rich collection of information.

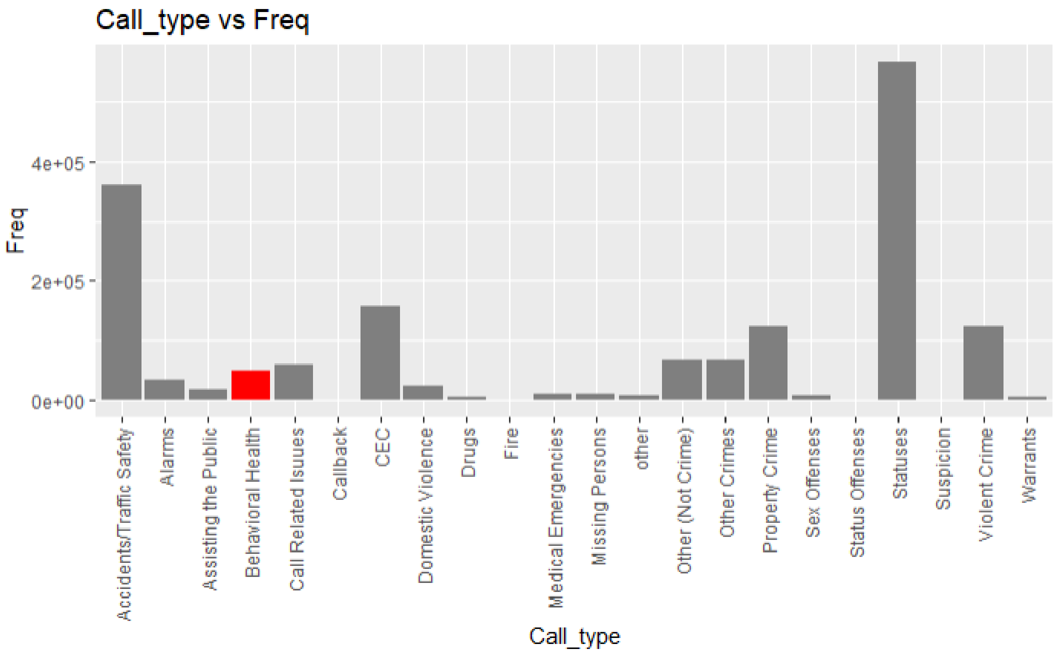


Figure 1: Bar plot depicting various call types, with the red bar specifically representing Behavioral Health Calls

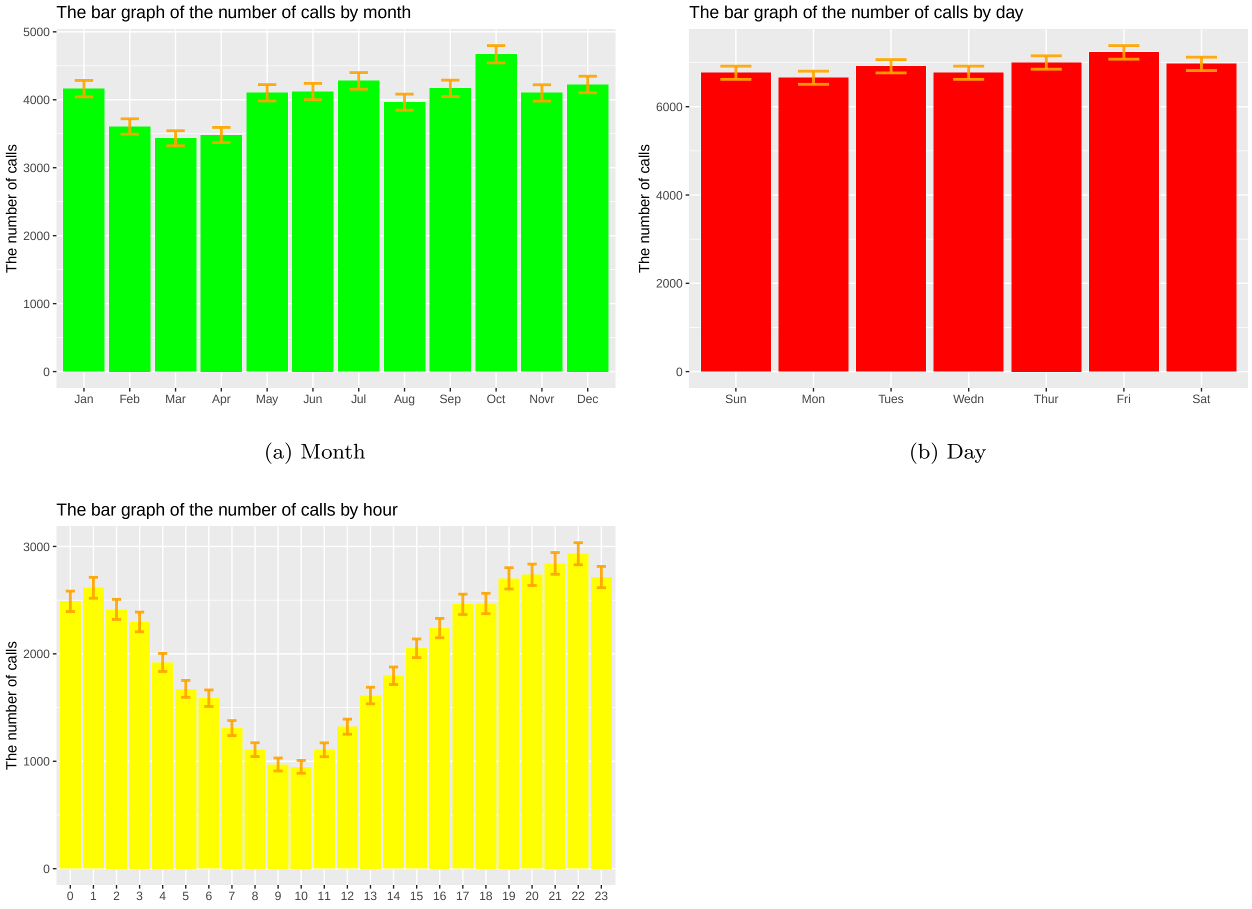
Figure 1 illustrates the diversity of call types present in the dataset. Our primary focus is on 'Behavioral Health' calls, which often indicate mental health problems that could potentially lead to suicides. After filtering for 'Behavioral Health' calls, we narrowed down our dataset to approximately 50,000 observations. This focused subset enables a more targeted analysis of mental health-related emergency calls. We examined the data from a spatial perspective, identifying 218 unique locations associated with Behavioral Health calls. Unfortunately, three locations lack population information, compelling us to exclude them from further analysis, as they represent only 200 observations out of the total 50,000.

3. Methodology

3.1 Time Analysis

3.1.1 Distribution of ‘Behavioral Health’ Calls Over Time

Our initial focus is to examine the distribution of calls over time, exploring potential differences in call counts based on hours, days, and months. For this analysis, we generate a bar plot illustrating the counts of calls, accompanied by a 95% confidence interval for each bar.



(a) Month

(b) Day

(c) Hour

Figure 2: Bar Plot of the number of calls for each of e Month, Day, and Hour

Based on Figure 2, variations in call counts between different times are noticeable. Monthly and hourly bar plots reveal significant differences, as indicated by non-overlapping confidence intervals between time periods. However, the day bar plot does not exhibit this distinction clearly, except for Monday and Friday, where non-overlapping intervals are observed. To rigorously assess these differences, we conduct a Chi-square hypothesis test. This test assumes that all groups should have a similar number of calls if there is no influence from times. The ‘p-value’ helps us determine if there is a real difference between groups of data. If the p-value is less than 0.05, it indicates a strong likelihood that time influences the number of calls.

While the distinction in the plot for days may not be visually apparent due to the large number of observations, Table 1 reveals that all p-values associated with the time variables (hour, day, and month) are much

Table 1: Chi-Square result

	X-squared	df	p-value
Months	348.52	11	$< 2.2e-16$
Days	31.597	6	1.95e-05
Hours	4800	23	$< 2.2e-16$

less than the threshold of 0.05. This suggests that time is a statistically significant factor in predicting the number of calls.

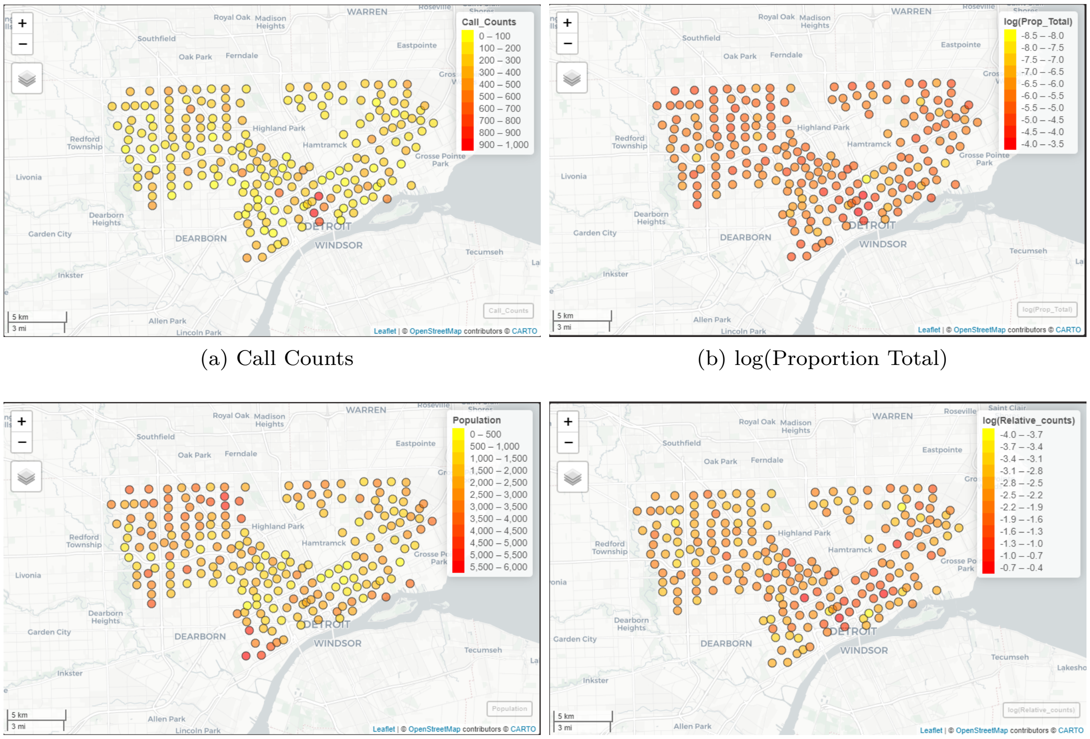
Additionally, it's worth noting that, although the p-value for days is less than 0.05, when compared to the other p-values, it is relatively larger. Consequently, the apparent difference may not be visually evident in the bar plot. The results of our hypothesis tests underscore the importance of including time variables—hour, day, and month—in building a predictive model for future analyses.

4.2 Spatial Analysis

4.2.1 Distribution of 'Behavioral Health' Calls Over Population Density

Secondly, through spatial analysis, our goal is to uncover patterns, trends, and relationships between location and call density that might not be apparent. This analysis contributes to understanding the geographical aspects of 'Behavioral Health' calls, providing a comprehensive view beyond temporal considerations.

To gain a general overview, several heat maps are generated based on calls, proportion of total calls, population, and relative calls (calls/population).



(a) Call Counts

(b) log(Proportion Total)

(c) Population

(d) log(Relative Calls)

Figure 3: Heat map illustrating the relationships among Call Counts, log(Propportaion Total), Population, and log(Relative calls), progressing diagonally from the upper left to the lower right

We generate four plots: Call Counts, log(Prop_Total), Population, and log(Relative Calls). The intensity of red indicates higher frequencies compared to other regions. The decision to express Proportion of Calls with Total Calls and Relative Calls in logarithmic scale is due to their small values, making differences challenging to discern with original data. Upon constructing the plots, these features become evident.

A quick look at the Call Counts heatmap suggests increased activity in the central area of Detroit. However, without considering population density, this observation may be misleading, as areas with higher populations naturally tend to have more calls. To address this, the Population and Relative Calls heatmaps offer a clearer perspective. These plots unveil that the central area's population is not particularly high, and the Relative Calls heatmap emphasizes regions with proportionally more calls.

To rigorously assess the proportionality of Call Counts to the population, we utilize the Kolmogorov-Smirnov (KS) test with simulation. Assuming that Call Counts are proportional to the population, we generate 50,000 sample data points with a distribution based on the probability of Prop_Pop (population divided by the total population). The goal is to compare the empirical cumulative distribution functions (ECDFs) of the simulated data with the actual Relative Calls. By visualizing the ECDF plots, we can evaluate the similarity in the distribution of the data. If the plots align, we can reasonably infer that the number of calls is proportional to the population, and the Kolmogorov-Smirnov test should not reject this assumption.

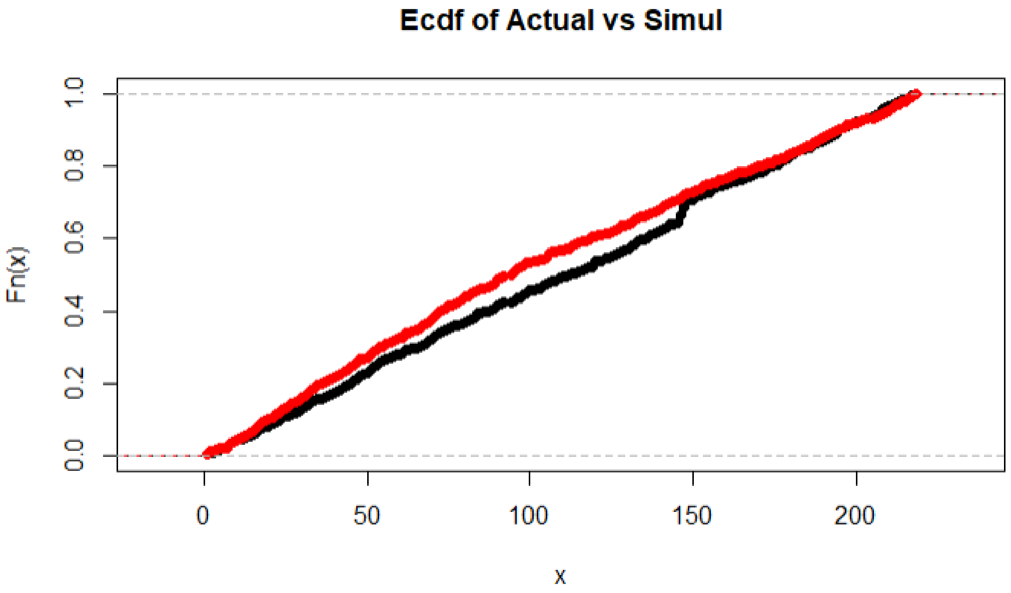


Figure 4: Empirical Cumulative Distribution Functions: The red line represents simulated data, while the black line represents actual data.

Upon examining Figure 4, it becomes apparent that the red line representing the ECDF of simulated data differs from the black line depicting the ECDF of actual data. This discrepancy is further supported by the results of the Kolmogorov-Smirnov test.

Table 2: KS test result

D	p-value
0.082585	< 2.2e-16

As depicted in Table 2, the test results reveal significant differences between the ECDFs, where D represents the value of the statistics and the p-value is less than the threshold of 0.05. This suggests that Call counts in each region are not proportional to the population.

However, according to the 2017 CDC report[1], individuals in urban areas are less likely to commit suicide than those in rural areas. This report implies that population density might inversely affect suicide rates.

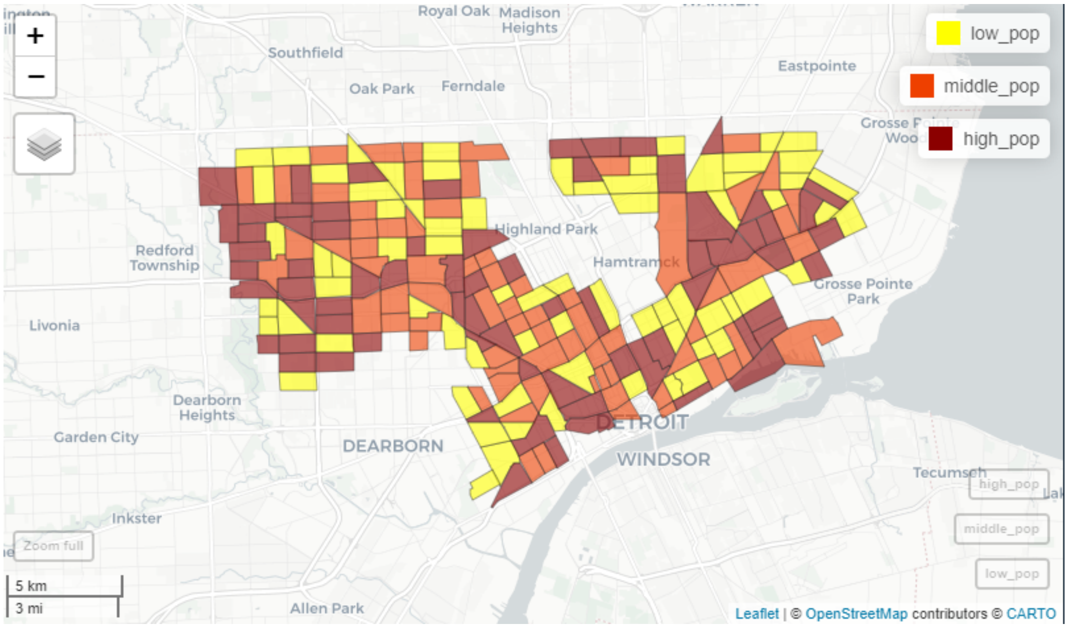


Figure 5: Detroit map divided into three regions based on population: low population group, middle, and high

To explore this potential relationship, we conduct an Analysis of Variance (ANOVA) test. Initially, we categorize regions into three groups based on population: low_population, middle_population, and high_population, as illustrated in Figure 5. The figure depicts a random distribution of population across these groups. The ANOVA test, designed to assess average differences of relative calls between multiple groups, is then employed for a more in-depth analysis of the relationship.

Table 3: ANOVA test result for 3 groups

	Df	Sum Sq	Mean Sq	F value	Pr(>F))
Group	2	10.34	5.171	19.94	1.16e-08
Residuals	212	54.98	0.259	-	-

Table 4: Relative counts of each groups

Group	Population	Relative Counts
1	Low	0.12756975
2	Middle	0.09926116
3	High	0.07452479

In the analysis presented in Table 3, the result of the ANOVA test indicates a statistically significant difference between the groups, with a p-value much less than 0.05. The F-value represents the statistics used to obtain the p-value. Other columns, such as DF (Degree of Freedom), Sum sq (Sum of Squares), and Mean sq (Mean of Squares), are explained in detail in this link.[2]

Focusing on Table 4, we observe that Group 1, representing areas with low population density, exhibits a higher relative calls rate compared to the other groups. To enhance the robustness of our findings, we

[1] https://www.cdc.gov/suicide/facts/disparities-in-suicide.html#:~:

[2] https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/#:~:

conduct an additional ANOVA test, expanding the scope to six groups. We choose six as we determine that the number of groups would be large enough to highlight the distinctive features of each group.

Table 5: ANOVA test result for 6 groups

	Df	Sum Sq	Mean Sq	F value	Pr(>F))
Group	5	13.32	2.6643	10.71	3.55e-09
Residuals	209	52	0.2488	-	-

Table 6: Relative counts of each groups

Group	Population	Relative Counts
1	Lowest	0.15646221
2	Low	0.10516675
3	Lower_Middle	0.10267224
4	Upper_Middle	0.09613883
5	High	0.07696163
6	Highest	0.07247128

The analysis with six groups reveals clearer distinctions, as the hypothesis is more strongly rejected, with a p-value significantly lower than that of the three-group analysis. Additionally, the differences in relative counts become more pronounced than in the previous analysis. This suggests a significant influence of population density on the number of counts, indicating an inverse proportionality.

4.2.2 Distribution of 'Behavioral Health' Calls Over Location

Our attempt to examine the relationship between emergency calls and geographic locations encountered challenges. The data collection method, organized based on regions, did not reveal any discernible patterns or informative insights. One statistical analysis method, the Ripley K function, which is used to identify sparsity or clustering, was applied, but the data does not show clear sparsity or clustering. Further exploration and consideration of the spatial aspects of the data would be needed.