Project Report
Sam Whittman
DSCI 410

Discussion:

How does CAHOOTS call volumes change with weather events such as air temperature, precipitation, solar radiation, and other environmental events?

I chose this research question because I believe it is crucial to understanding what CAHOOTS does, and when they do it. Specifically understanding when resources are needed and when they are not. Understanding when the demand for CAHOOTS services is highest allows for a more strategic allocation of crisis workers, improving response times and benefiting the community. Conversely, identifying periods of low demand can enable CAHOOTS to adjust shift schedules, optimizing their limited resources and addressing their funding challenges more effectively. With a comprehensive analysis on call volume with environmental factors, CAHOOTS should be able to prepare for the day, week or month with greater confidence.

Data:

I will be using data from CAHOOTS. The CAHOOTS data has an observation for every CAHOOTS call CAHOOTS was dispatched to. Each observation has date, time of call, age, gender, race, language, city (Eugene or Springfield), and reason for dispatch. This data has quite a few missing values for some of these variables such as age, gender and race. However, for my analysis these variables seem unimportant.

I will also be using publicly available data through NOAA, https://www.ncei.noaa.gov/pub/data/uscrn/products/hourly02/?C=D;O=A. The weather station is in Corvallis, OR, as it is the closest NOAA station from Eugene. For this data, there is an observation at the beginning of every hour. There are 38 different variables included in each observation, which are posted in the README.

I have 2 CAHOOTS datasets, one for 2023 and one for 2021-2022. I will create a Date Time column for both of these datasets. I then aggregate these 2 datasets.

I then get rid of columns I do not need, such as race, gender, age and language, as they do not pertain to the questions I want to answer. With the CAHOOTS dataset having a Date Time variable, and reason for dispatch, I can move onto the weather data.

For my weather data, I have 3 datasets. One for each year 2021- 2023. My weather dataset has hourly data. I first need to create a Date Time column for all three of these datasets.

I can then concat weather for 2021 to 2023 into one dataframe, using the date time column.

Figure out which weather variables I want, so I can then drop the rest. My variables will be date and time, average, min and max air temperature, total precipitation in the hour, average, max, and min global solar radiation, average, max, and min infrared surface temperature, as well as average soil moisture and temperature 5 cm below the surface. How the average was taken depends on the variable, see README.

My variables do have some missing data points, but because we have such a large dataset I am just going to replace these numbers with NaN values and drop NAN rows.

I can now aggregate my weather and CAHOOTS data with a Date Time index. I am using merge_asof which allows me to merge the data based on the closest time. For example, because the weather data is at the start of every hour, but our CAHOOTS data takes place at random times, the CAHOOTS data is paired with the weather data which happened closest to the time of CAHOOTS observation.

Methods:

I then plan to make plots for each weather variable I have. I will do daily calls on the y axis, with the weather variable on the x axis. I create a column which has the count of calls per day, and a column which averages each weather variable for each day. These are my x and y inputs. My output is a 4x3 of 12 graphs, with Daily Calls on the y axis, and the weather variable on the x axis.

I will make a few more plots which help me explore the data. This includes a plot of daily call volumes. With the already made column of Daily Calls, I use the date as my x axis, and count of calls as my y axis. Output is a graph of call volumes per date.

(Support): At this point in the project I plan to Meet with David, who is addressing the same research question. This should be greatly beneficial. I plan to check our data, the similarities and differences between the processes we took in the cleaning, and address any mistakes we may have. With this we can look at our plots, which should again help to identify any mistakes either of us have.

From this I will make a few linear regressions. First, I am going to start with one variable at a time. This means I am going to make 12 different linear regressions to begin with. The inputs of these linear regressions are a single weather variable as X, and my daily calls mentioned earlier, as my Y. The output will be the coefficient, intercept, MSE, and R^2 to identify the change in call volume from each weather variable.

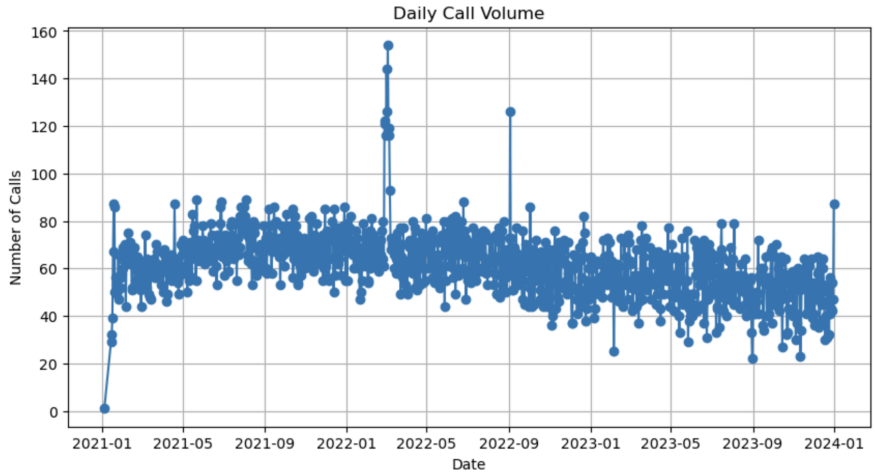
(Support): At this point I plan to meet with David once again, or Rori if I feel necessary. Once again looking for more mistakes, similarities and differences between my and David’s results.

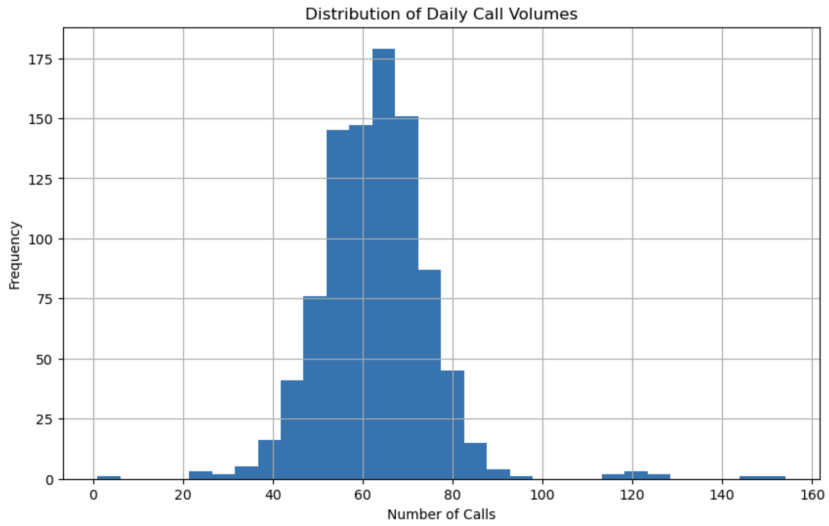
I will then create some multiple linear regressions, which I hope will capture more of the variance of call volume. I will do this using the highest R^2 values and coefficients.

I will be using Pandas, Matplotlib and seaborn for graphical representation, and scikit-learn and statsmodels for regression analysis.

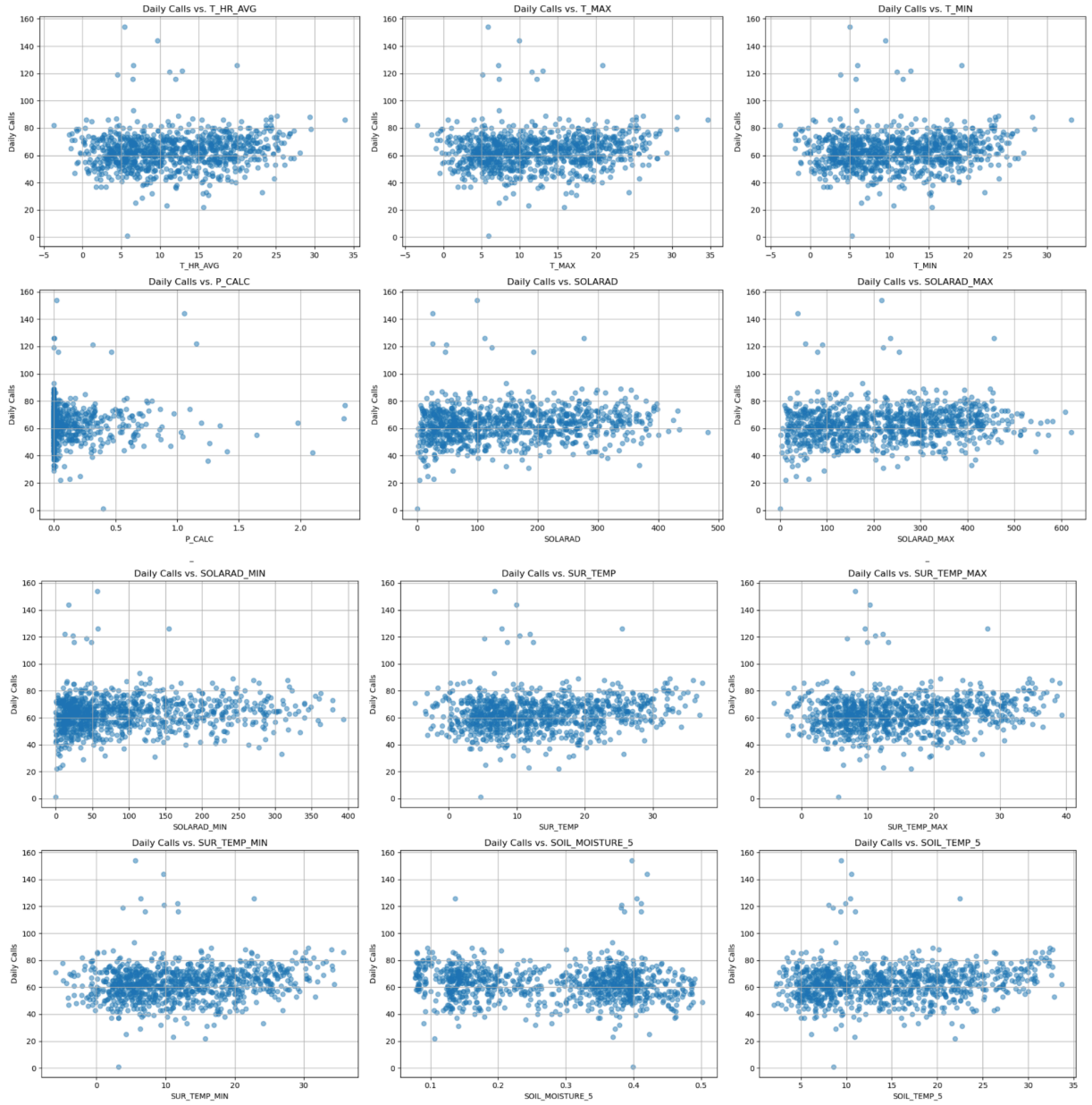
Git Hub link: https://github.com/swhittman/CAHOOTS-WEATHER.git

Results:
First, I started with my exploratory graphs. Looking at the distribution of call volume, and call volume through time. This is important to explore the data, in order to both understand where to go next, but also to understand my results from those upcoming steps.





Next, I wanted to plot each weather variable I was looking at, with daily calls on the y axis.



Next, I created a regression for each plot above. Again, these are 12 separate regressions, with call volume as the dependent variable, and the weather variable being the independent variable. The first 12 regressions are daily calls, while the second 12 regressions are hourly calls.

```
T_HR_AVG: Coef = 0.186, Intercept = 60.768, P-Value = 0.016, MSE = 138.585, R2 = 0.025
T_MAX: Coef = 0.182, Intercept = 60.686, P-Value = 0.016, MSE = 138.539, R2 = 0.025
T_MIN: Coef = 0.189, Intercept = 60.854, P-Value = 0.017, MSE = 138.625, R2 = 0.025
P_CALC: Coef = 0.794, Intercept = 62.835, P-Value = 0.685, MSE = 142.176, R2 = -0.000
SOLARAD: Coef = 0.015, Intercept = 60.588, P-Value = 0.002, MSE = 138.364, R2 = 0.026
SOLARAD_MAX: Coef = 0.012, Intercept = 60.172, P-Value = 0.001, MSE = 138.626, R2 = 0.025
SOLARAD_MIN: Coef = 0.014, Intercept = 61.533, P-Value = 0.016, MSE = 139.122, R2 = 0.021
SUR_TEMP: Coef = 0.191, Intercept = 60.352, P-Value = 0.001, MSE = 137.131, R2 = 0.035
SUR_TEMP_MAX: Coef = 0.180, Intercept = 60.235, P-Value = 0.001, MSE = 137.319, R2 = 0.034
SUR_TEMP_MIN: Coef = 0.202, Intercept = 60.504, P-Value = 0.001, MSE = 137.106, R2 = 0.035
SOIL_MOISTURE_5: Coef = -5.322, Intercept = 64.393, P-Value = 0.199, MSE = 139.980, R2 = 0.015
SOIL_TEMP_5: Coef = 0.229, Intercept = 59.558, P-Value = 0.001, MSE = 137.416, R2 = 0.033
```

```
T_HR_AVG: Coef = 0.023, Intercept = 2.619, P-Value = 0.000, MSE = 2.711, R2 = 0.016
T_MAX: Coef = 0.023, Intercept = 2.607, P-Value = 0.000, MSE = 2.710, R2 = 0.016
T_MIN: Coef = 0.023, Intercept = 2.629, P-Value = 0.000, MSE = 2.711, R2 = 0.015
P_CALC: Coef = 0.017, Intercept = 2.867, P-Value = 0.568, MSE = 2.755, R2 = -0.000
SOLARAD: Coef = 0.000, Intercept = 2.857, P-Value = 0.179, MSE = 2.755, R2 = -0.001
SOLARAD_MAX: Coef = 0.000, Intercept = 2.861, P-Value = 0.392, MSE = 2.755, R2 = -0.001
SOLARAD_MIN: Coef = 0.000, Intercept = 2.856, P-Value = 0.073, MSE = 2.755, R2 = -0.001
SUR_TEMP: Coef = 0.010, Intercept = 2.742, P-Value = 0.000, MSE = 2.737, R2 = 0.006
SUR_TEMP_MAX: Coef = 0.009, Intercept = 2.740, P-Value = 0.000, MSE = 2.738, R2 = 0.006
SUR_TEMP_MIN: Coef = 0.011, Intercept = 2.748, P-Value = 0.000, MSE = 2.736, R2 = 0.006
SOIL_MOISTURE_5: Coef = -0.264, Intercept = 2.944, P-Value = 0.020, MSE = 2.753, R2 = 0.000
SOIL_TEMP_5: Coef = 0.020, Intercept = 2.586, P-Value = 0.000, MSE = 2.722, R2 = 0.012
```

I then created two multiple linear regressions, I chose these by using the greatest magnitude R^2 and coefficient. The first regression are the 4 weather variables with the highest R^2. The second regression are the 4 weather variables which have the highest magnitude coefficients. In these tables you will notice the coefficient for each x variable, as well as standard error, t stat, p value, and confidence interval.

OLS Regression Results

Dep. Variable:	Call Volume	R-squared:	0.032
Model:	OLS	Adj. R-squared:	0.026
Method:	Least Squares	F-statistic:	5.246
Date:	Thu, 06 Jun 2024	Prob (F-statistic):	0.000366
Time:	18:54:52	Log-Likelihood:	-2567.4
No. Observations:	648	AIC:	5145.
Df Residuals:	643	BIC:	5167.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	61.2911	1.502	40.817	0.000	58.342	64.240
T_HR_AVG	-0.9591	0.338	-2.836	0.005	-1.623	-0.295
SOLARAD	-0.0003	0.008	-0.030	0.976	-0.017	0.016
SUR_TEMP	0.7356	0.398	1.847	0.065	-0.047	1.518
SOIL_TEMP_5	0.1976	0.267	0.741	0.459	-0.326	0.721

OLS Regression Results

Dep. Variable:	Call Volume	R-squared:	0.036
Model:	OLS	Adj. R-squared:	0.030
Method:	Least Squares	F-statistic:	6.046
Date:	Thu, 06 Jun 2024	Prob (F-statistic):	8.86e-05
Time:	18:54:52	Log-Likelihood:	-2565.8
No. Observations:	648	AIC:	5142.
Df Residuals:	643	BIC:	5164.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	57.7437	2.891	19.973	0.000	52.067	63.421
T_HR_AVG	-0.9523	0.315	-3.027	0.003	-1.570	-0.335
P_CALC	2.0709	2.013	1.029	0.304	-1.883	6.024
SUR_TEMP	1.0067	0.245	4.115	0.000	0.526	1.487
SOIL_MOISTURE_5	8.9429	6.382	1.401	0.162	-3.588	21.474

Discussion:

From our exploratory plots, we can clearly see the mean call volume per day is about 60 calls. We see maximum calls per day at about 150 calls, and possibly a slight decrease in calls over time. Looking at our 12 graphs of weather variables and call times, we can see very little indication of any trends or correlations.

Looking at our 24 regressions, we are going to focus solely on the first 12, the daily calls. This is because our R^2 are generally higher, but I also believe the first 12 regressions are very similar

to the second 12, with the main difference being the magnitude of the statistics in which I believe is proportional to using daily vs hourly calls. We see our highest R^2 is 0.035 for surface temperature. This means surface temperature makes up for 3.5% of the variance of daily call volumes, and all other weather variables we are looking at make up for less than 3.5% variance each. We can also see the Max's and Min's of each variable typically mimicking the average (Air temp, surface temp, solar radiation). Because of this, we begin to ignore the variables which are max and min.

Now, we look at our multiple variable OLS models. We first see our R^2 are 0.032 and 0.036, respectively. Again, we see these regressions are not making up much variance for call volumes. We see our constants for both of these regressions are statistically significant, along with average air temperature. However, in both of these regressions, the coefficients for average air temperature are negative. This is telling us, as air temperature increases by 1 degree celsius, call volumes decrease by 0.95 calls, for both of these regressions. In our second regression, we see surface temperature is also statistically significant with a coefficient of 1. This is telling us as surface temperature increases by 1 degree celsius, our call volume increases by 1 call per day.

This is where our analysis gets a little tricky. As we see in our first regression, we are looking at air temperature, solar radiation, surface temperature, and soil temperature. However, air temperature is affected by solar radiation, which both affect surface temperature and soil temperature. All of these variables are related to each other. In linear regressions multicollinearity occurs when independent variables are correlated with each other. Multicollinearity creates less reliable results. If we assume these weather variables are not just related, but highly correlated, then we have to be extremely weary of our analysis.

Now with this being said, we can really dive in. In both of our multiple variable OLS, we see call volume increasing due to a decrease in air temperature. However, in our second MV OLS, we see call volume increasing due to an increase in surface temperature. We can clearly see, these results contradict each other. Because of the multicollinearity in our independent variables, I believe our regressions are superfluous. With all 26 linear regressions we made, we see a maximum variation of call volumes explained by a single regression being 3.6%.

I believe weather, such as surface and air temperature, certainly affect call volumes. However, from our results, I would conclude it has a fairly minimal effect, while call volume mostly varies from other factors outside of weather, which we see in our constant variables in our multiple variable OLS. Due to the multicollinearity of weather variables, along with possible confounding variables, our regression statistics should be taken as a grain of salt.