Analysis of Response Time to Pothole Repair Requests in Chicago

Melanie Hanna

1. INTRODUCTION

Potholes—along with bone-chilling winters, jam-packed expressways, and suspected municipal corruption—are a favorite gripe of longtime Chicagoans. Formed when soil beneath the pavement weakens or shifts, potholes are worsened through freeze/thaw cycles and heavy traffic—factors that would seem to be universal to all areas of the city. But any Chicago driver would tell you that potholes seem more prevalent in some parts of the city versus others. Is this true? And if so, what factors are correlated with increased pothole creation?

Using data from the City of Chicago and the Chicago Metropolitan Agency for Planning (CMAP), we'll explore how location within the city affects the number of potholes as well as the response time for the City's Department of Transportation (DoT) to patch the road. Potential clients of this analysis include Chicago residents, who can better understand how the DoT prioritizes patching requests in their neighborhood and based on what factors. City residents can also use this analysis to determine if a route through certain areas of the city is more likely to have potholes than another.

1.1 Data Sets

The City of Chicago has compiled a dataset of 311 service requests, ranging from 2011 to the present, for pothole repair, which can be downloaded from the site below.

https://data.cityofchicago.org/Service-Requests/311-Service-Requests-Pot-Holes-Reported/7as2-ds3y

The dataset consists of more than 477,000 pothole reports and includes several columns on location (latitude/longitude, zip code, ward number, police district, street address, community area). The number of potholes found at that location for the entire block is also provided, as well as the date each request is created and completed. If two pothole patching requests within a buffer of four addresses are added to the system before the request is closed, they are listed as duplicates ("-Dup") and closed simultaneously when the crew arrives to patch the block.

There are some limitations with this data set, specifically surrounding what action the DoT has taken on the request. More than half of the entries list "Pothole Patched" but many cite several other outcomes, such as "Not Within CDOT Jurisdiction" or "Street Resurfaced". As we are only investigating 311 service requests in total and the response time (no matter the outcome), we can ignore this column for those analyses.

I also used CMAP's Community Snapshot data set, which includes data by community area from the U.S. Census Bureau's 2010-14 American Community Survey, Longitudinal Employment-Household Dynamics data for 2014, and 2014/2015 data from the Illinois Department of Employment Security and the Illinois Department of Revenue. This data set consists of 155 metrics (median income, percent vacant housing, etc.) for each of the city's 77 community areas.

https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data

1.2 Data importation

I imported the pothole dataset as a JSON file, which needed to be parsed before converted into a database. The data was downloaded as a nested JSON string, which included the data rows, along with metadata, which included information about the dataset itself such as the number of downloads as well as the column names. I needed to break down the meta dictionary and extract the column headings and compile these into a list.

Then, I cleaned the data by removing several initial columns on each row that referenced unnecessary internal ID's. I also removed the 'LOCATION' column at the end since this data only contained the longitude and latitude pairings, and the latitude and longitude both already had their own columns. Once the rows were cleaned, I could convert the data into a panda dataframe and add column headings from the 'meta' data.

The CMAP data was imported using a .csv file into a panda data frame and did not require any data wrangling.

1.3 Data cleaning

After the data was imported into a panda data frame, I cleaned the Completion Date and Creation Date columns which were added as strings. Each date string was converted to a timedate object (while leaving any 'None' data entries alone) and placed into a new column in the data frame. I also calculated a "Time Passed" vector by subtracting the creation date from the completion date to test operations on the new data types.

I then converted the column containing the number of potholes filled on the block to an integer from strings. Any None values were converted to blank spaces. Latitude, longitude, and the x & y coordinates were converted to floats.

The CMAP data was merged with the pothole data set by first creating a data frame "key" that linked the name of each community area (column 'GEOG' in the CMAP data) to the community area's number (included in the pothole data set) and adding an extra column specifying which region the community area belonged to (Far North, Southwest Side, etc.).

Outliers

Some completion dates were much later than their creation dates. The maximum time elapsed was 1194 days or about 3.27 years. The next four largest values for time elapsed were 1009, 997, 720 and 688 days. In all, there are 4,795 rows that show a wait time of more than a year so it seems unlikely that these were key errors during data entry. Therefore, as I have no reason to believe that these data entries are incorrect, I chose to keep them in the dataset.

Similarly, the maximum number of potholes filled on a block in the dataset was 320. As the average block length (on the long side) in the city is 660 feet, this works out to one pothole every two feet! The next 30 largest values for number of potholes filled per block are all 300, and the number of entries above 100 potholes is 2,161 so again it seems very unlikely that these were key errors. Therefore, these entries remain in the dataset.

Mysterious Ward 0

While sorting by ward, I found 1,365 rows associated with Ward 0, which is not a legitimate ward number in the city of Chicago. After theorizing that perhaps these addresses were found on ward borders, I looked up various addresses associated with Ward 0 but could not find any apparent pattern on the city map. The City simply does not have a ward (or a community area) associated with those addresses as found at the following site:

https://www.cityofchicago.org/city/en/depts/mayor/iframe/lookup_ward_and_alderman.html

All rows associated with Ward 0 and Community Area 0 were removed from the data set.

2. DATA EXPLORATION

I initially explored preliminary relationships in the pothole data by ward but the City redraws ward boundaries every few years with the latest iteration occurring in 2015. Consequently, Ward 2 in 2012 in the pothole data set is not the same Ward 2 in 2016. Organizing the data by community area allows a more stable analysis and better alignment with the CMAP data.

Unlike Chicago's wards, the community area boundaries were established in the 1920's according to neighborhood borders and have not changed since then. A 76th community area was added to the original 75 in 1953 to include O'Hare Airport, and the final 77th area was created when Edgewater split from Uptown in 1980. These community areas were defined to track population and economic data over time, and census tracts generally lie neatly within the area boundaries to allow easy data aggregation. Please see the Appendix for a map of the community areas.

2.1 Validating the DoT's Response Time Claims

The City made two claims when introducing the pothole data set. The first was that "pothole repairs are generally completed within 7 days from the first report of a pothole to 311." I found this to be true as 49.1% of all pothole repair requests were serviced in 7 days or fewer.

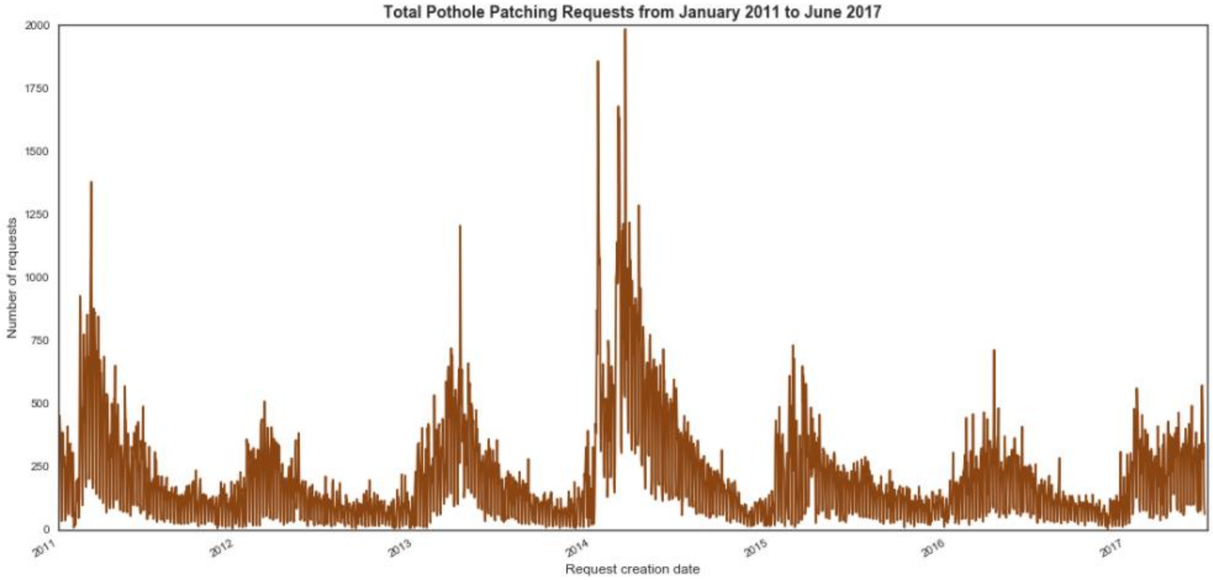
Next, I evaluated the statement that "Weather conditions, particularly frigid temps and precipitation, influence how long a repair takes." I segregated each request by the season in which it was submitted (all requests made in March, April, and May were classified as spring; June, July, August as summer; September, October, November as fall; December, January, February as winter) and looked at response time in each season.

The average number of days between request and response was 43.3 (almost a month and a half) in the summer compared to only 14.0 in the winter. A Welch's t-test confirms that this difference in means is statistically significant and further invalidates the City's claim that pothole patching is delayed by frigid and inclement weather.

2.2 Service Requests over Time

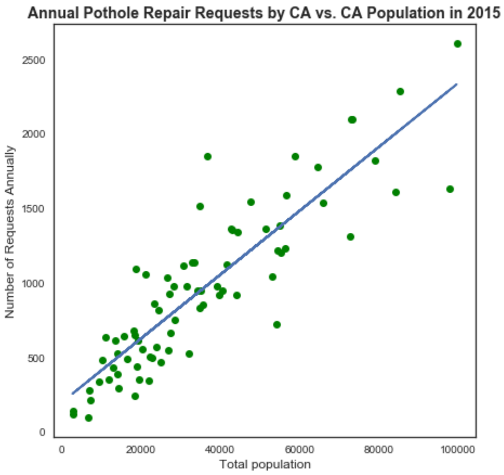
Analyzing service requests by season shows that almost half (43.5%) of all requests are made in the spring months. After a winter of multiple freeze/thaw cycles, this result is in line with expectations.

Next, we look at how pothole repair requests have varied over the timespan of our data set (January 1, 2011 to June 1, 2017). From the graph below, we see a major spike in the early months of 2014, which coincides with the infamous "polar vortex" that plunged temperatures in Chicago to record lows. That winter also had the third-highest snowfall in Chicago history, leading to frequent snow plowing and further contributing to pothole creation.



2.2.1 Service Requests by Community Area

After creating a bar plot of annual service requests by community area (CA) and another bar plot of total population in each community area, I noticed that these graphs had similar shapes, leading me to plot population against annual total service requests directly.

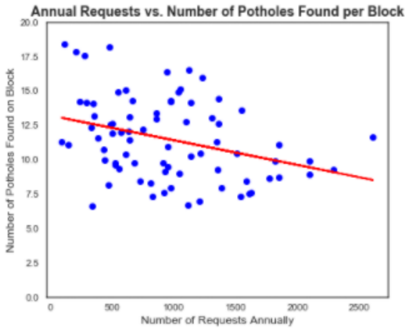


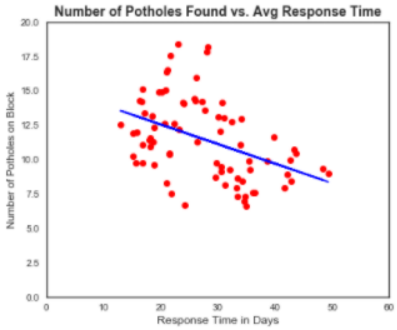
This plot has a clear linear relationship, and a linear regression line overlaid onto the graph has an r-value of 0.88 and a p-value of 4.69 x 10^{-26}. We can conclude that Chicago residents submit pothole repair requests at roughly the same rate (2.1 requests per 100 residents per year, based on the slope of the regression line) regardless of the community area in which they live.

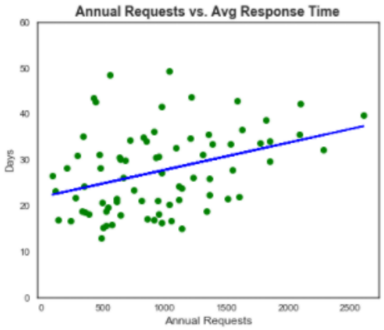
2.3 Comparing Pothole Metrics

I identified three key metrics from the pothole dataset for evaluation against the CMAP data: number of potholes found per block, number of annual requests, and average response time. These, of course, can be correlated with each other. Perhaps when fewer requests are filed with the DoT, the patching crew finds more potholes to fill due to less frequent service. Or when the patching crew arrives more quickly after a request is filed, they find fewer potholes as less time has passed for more potholes to appear.

Scatterplots comparing these are shown below. There does appear to be a correlation among all three—a hypothesis confirmed by the correlation matrix also shown below with r-values of 0.33 or greater.







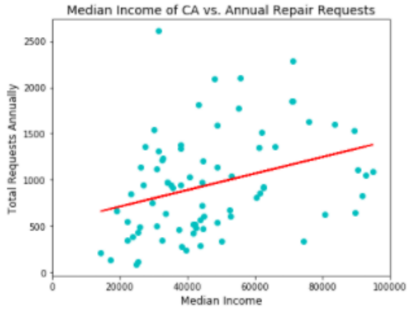
	Response time (days)	Annual requests	Average number of potholes found
Response time (days)	1	0.36570821	-0.4282473
Annual requests	0.36570821	1	-0.33370708
Average number of potholes found	-0.4282473	-0.33370708	1

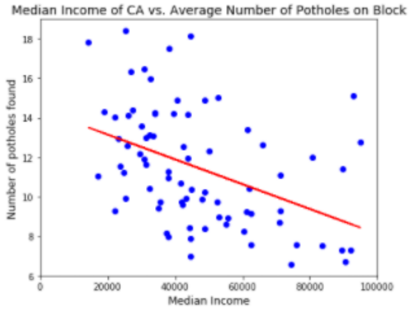
The number of potholes found per block is negatively correlated with both response time and annual requests, validating our theory that fewer potholes are discovered with more frequent and prompt service. Response time is positively correlated with annual requests, suggesting the DoT does not respond as quickly to CA's filing more repair requests or that the community areas file more requests when the DoT responds slowly.

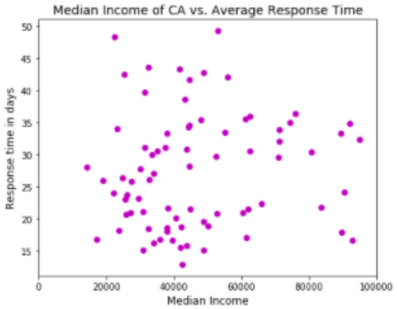
2.3.1 Effect of Median Income on Pothole Metrics

Using CMAP's median income data for each community area, we can determine how the income of a certain area affects response time, the number of potholes found, and the number of requests filed. A decent linear relationship was discovered between the median income and total annual requests (r-value = 0.33, p-value = 0.003), with a slope suggesting an additional nine requests are submitted annually for pothole repair for every extra $1000 in median income.

Median income also appears to be correlated with the number of potholes found on the block (r-value = -0.43, p-value = 8.33 x 10^{-5}). The slope of regression line shows that 0.06 fewer potholes are found per block for every extra $1000 of median income in the community area. However, there was no correlation between income and response time.







	Median income
Average response time	0.087976
Annual requests	0.330591
Average number of potholes found	-0.433145

2.3.2 Mode of Transportation for Daily Commuting

We could theorize that community areas with a higher percentage of drivers (defined as those who drive alone or carpool to work) would either report potholes at a higher rate or possibly cause more potholes due to heavier traffic.

CMAP's data set includes method of transportation to work, and a "driverPct" series was created by combining those who reported driving alone and those who carpooled and dividing by the total number of commuters. However, no correlation appeared with any of the pothole metrics.

2.3.3 Percentage of Land Used as Transportation

CMAP's Land Use Inventory includes a category defined as "transportation" that includes roads, railways, and airfields. Apart from O'Hare, Clearing and Garfield Ridge (these last two each contain a half of Midway Airport), we will assume that most of the land classified as transportation in the community area consists of roadways.

Possibly, with more roadways, we'd find a larger number of potholes and more requests for repair. However, almost all community areas are fairly consistent in the amount of land devoted to transportation (25-40%), and no correlation with the number of requests, the number of potholes found, or the response time presented itself.

This dataset did have a few outliers but not necessarily those outlined above. Of the four CA's with more than 50% of land devoted to transportation, one was O'Hare and the others contain mainly railyards (South Deering and Riverside) or consist of a narrow strip of line surrounding I-90 (Fuller Park).

2.3.4 Percentage of Land Defined as Residential, Commercial and Industrial

Are homeowners or business owners responsible for more pothole reporting? Does the presence of industry result in more pothole formation due to heavy truck traffic? CMAP's Land Use Inventory also includes information on the percentage of land used residentially, commercially, and industrially, and we'll look at how these land classifications affect our pothole metrics next.

Residential

Only one pothole metric yielded a correlation. Annual requests and percentage of land zoned as residential appear to have a linear relationship (r-value = 0.41), and the slope of regression line indicated that for every extra percentage point of land zoned as residential, 17.6 more requests are submitted for pothole repair per year. This is most likely due to the strong correlation with total population, so as more residents are added to the community area, the DoT receives more service requests.

Commercial

Three community areas qualified as outliers in this analysis: The Loop, Near North Side, and Near South Side all exceeded 15% commercial land when most CA's cluster around 2-6% of land zoned as commercial. When these outliers are removed, we see possible correlations with all three pothole metrics. It appears that more commerce increases pothole requests and response time and is associated with fewer potholes per block—which connects with our earlier conclusion that prompt and frequent service by the DoT yields fewer potholes.

	Pct commercial land
Annual requests	0.355257
Response time	0.247548
Average number of potholes per block	-0.355052

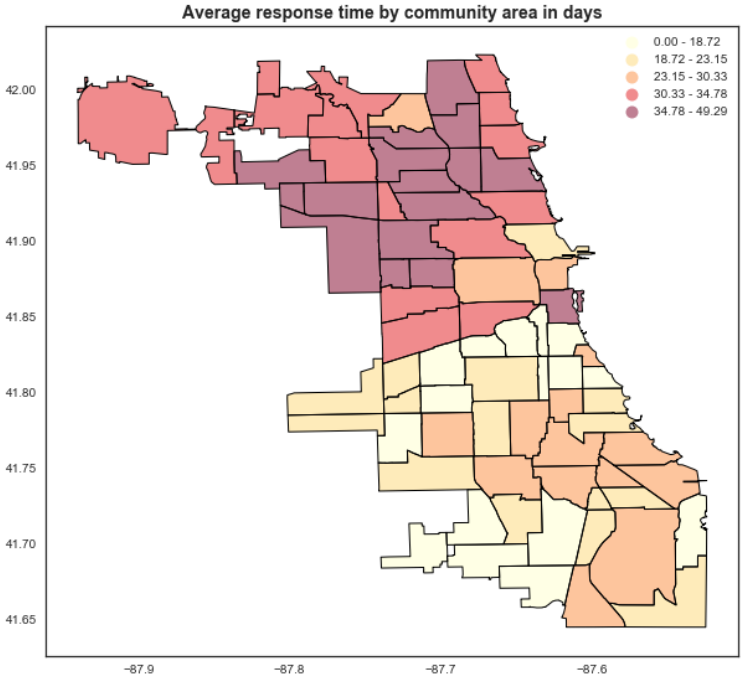
Industrial

No correlation appeared between the percentage of land zoned as industrial and any of the pothole metrics.

2.4 Northern Community Areas vs. Southern Community Areas

Response time

Chicagoans have long identified as either North-Siders or South-Siders. How does this traditional boundary translate to pothole response time? When we segregate community areas by region, we find that northern community areas (defined as North Side, Far North Side, Northwest Side) have a much longer average response time of 36.4 days compared to 21.1 days for southern community areas (defined as South Side, Southwest Side, Far Southwest Side, Far Southeast Side). Community areas defined as "Central" (3 CA's) or "West Side" (9 CA's) were excluded from this analysis. A Welch's t-test confirms that this difference in means is statistically significant (p = 0.0). The chloropleth below visualizes this difference in average response time geographically.



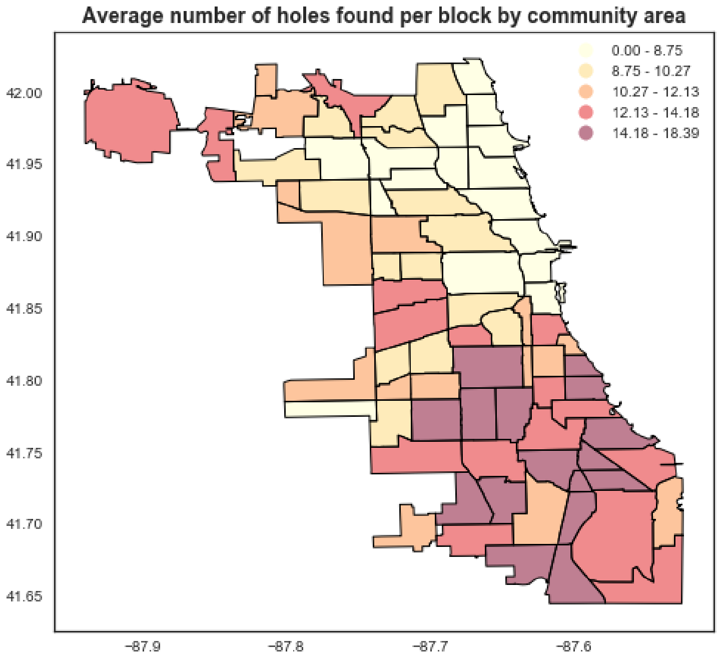
We saw in a previous analysis that response time increases in the summer. Is there a difference in response-time increase between northern and southern CA's? The average wait-time increase across all northern community areas is 7.4 days while southern areas see an average wait-time increase of 8.9 days. Another Welch's t-test confirms that the difference is statistically significant (p = 0.0025). Consequently, the difference in response time between northern and southern areas is greatest in the summer.

Requests over time

Plotting pothole repair requests over the timespan of the dataset (2011-2017) for both northern and southern areas shows a few key differences. Northern areas filed about 40% more repair requests in early 2013 than southern areas. However, southern areas requested pothole repairs 33% more than northern areas in early 2014. These differences could be due to increased snowfall in one part of the city compared to another. A bar plot of annual requests by region produced no insights.

Potholes per block

Southern CA's have significantly more potholes per block (13.1) than northern areas (9.0) on average, yielding a p-value of 0.0. Again, we can clearly see this difference in the chloropleth below.



3. PREDICTIVE MODELS

We'll next explore how this data can be modeled to predict future pothole metrics using census data. We will also investigate clustering the dataset to look for geographic patterns.

3.1 Linear Regression Analysis

To generate a linear regression model, we first need to create a new dataframe with a row for each community area. The two series containing the average number of potholes found per block by CA and the average response time by CA were merged into the original census data to ensure row alignment and then split off again to act as the dependent variable vectors for the regressions. We will attempt to build a model using only the census data to predict the pothole metrics.

3.1.1 Feature selection

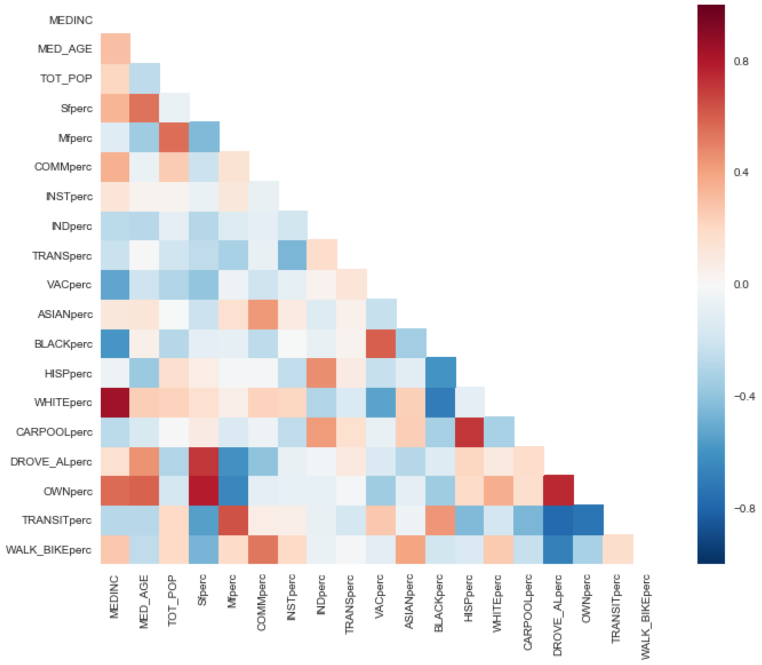
Nineteen columns were originally chosen as possible features of interest that might impact pothole response time or frequency; their names and descriptions are listed below.

Name	Description
MEDINC	Median income
MED_AGE	Median age
TOT_POP	Total population
Sfperc	Percentage of land designated as single-family homes
Mfperc	Percentage of land designated as multi-family homes
COMMperc	Percentage of land designated as commercial
INSTperc	Percentage of land designated as institutional (ex. schools)
INDperc	Percentage of land designated as industrial

TRANSperc	Percentage of land designated as transportation (ex. roadways and train lines)
VACperc	Percentage of land designated as vacant
ASIANperc	Percentage of residents who identify as Asian
BLACKperc	Percentage of residents who identify as black
HISPperc	Percentage of residents who identify as Hispanic
WHITEperc	Percentage of residents who identify as white
CARPOOLperc	Percentage of commuters who carpool
DROVE_ALperc	Percentage of commuters who drive alone to work
TRANSITperc	Percentage of commuters who take public transit
WALK_BIKEperc	Percentage of commuters who walk or bike
OWNperc	Percentage of housing units that are owner-occupied

The features regarding race (white, black, Asian and Hispanic) are not directly dependent as the census data included a fifth race category of "OTHER". Likewise, an "OTHER" category was provided for the commuting method. Several land classifications (agricultural, open space, mixed use) were excluded so the land designations in the table above do not sum to 100%.

A correlation plot of all the above features is shown below to gain a better understanding into the dependencies between these variables.



We can see that there are several highly correlated pairs of features. For example, the percentage of white residents is very positively correlated with median income while the percentage of black residents is negatively correlated. The percentage of single-family homes is associated with the percentage of owned-occupied housing units—an unsurprising result given most single-family homes are occupied by the homeowner. Additionally, we can see that Hispanics are highly correlated with carpooling and that those who own their home are associated with driving alone to work.

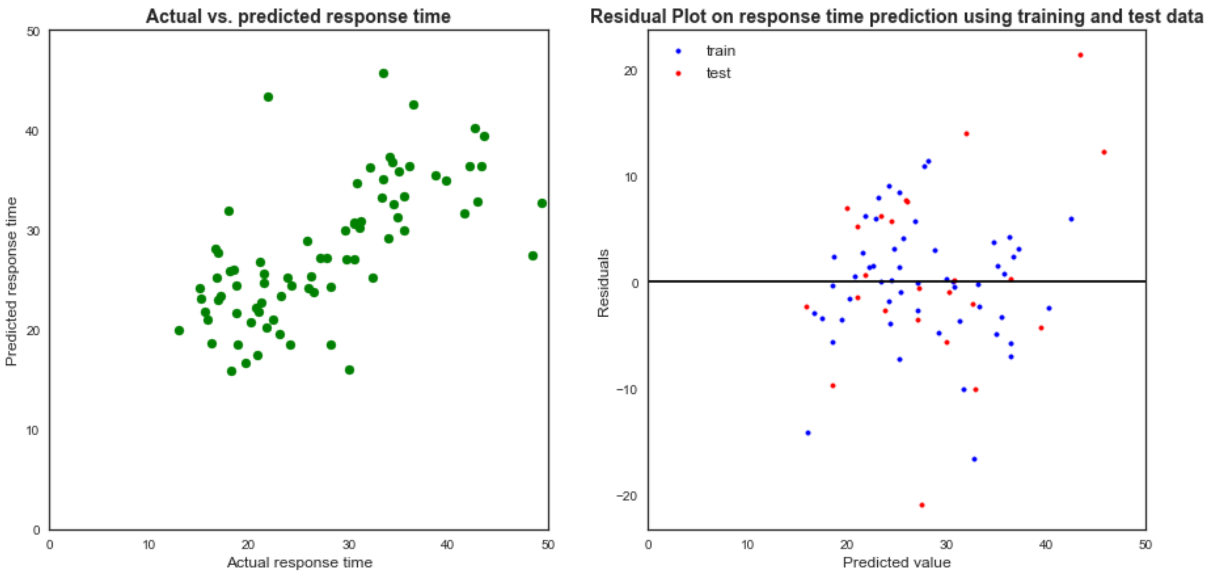
The correlation between median income and percentage of white residents is the strongest on this matrix (0.85). This relationship is not so dependent as to skew the regression results, and I chose to keep all 19 features for analysis.

3.1.2 Linear Regression Results on Response Data

The LinearRegression class from sklearn was used to fit the model with ordinary least squares (OLS) linear regression. The data showing census results from the 77 community areas and the response time vector were split into a training (70%) and a test (30%) set, and the model was fit on the training data after normalizing the features. However, the adjusted R-squared value for this model is only 0.37—a moderate number that may indicate that a linear relationship is not the right fit for this dataset.

We next look at the mean squared error (MSE), which was relatively low on the training data (MSE = 30.2) but much higher for the test data (MSE = 74.4)—a large increase that signals an overfit model. To limit the effects of overfitting, we can use k-fold cross-validation (k = 5), but this procedure calculates a high MSE of 79.5, indicating that the response time is not an excellent candidate for linear regression.

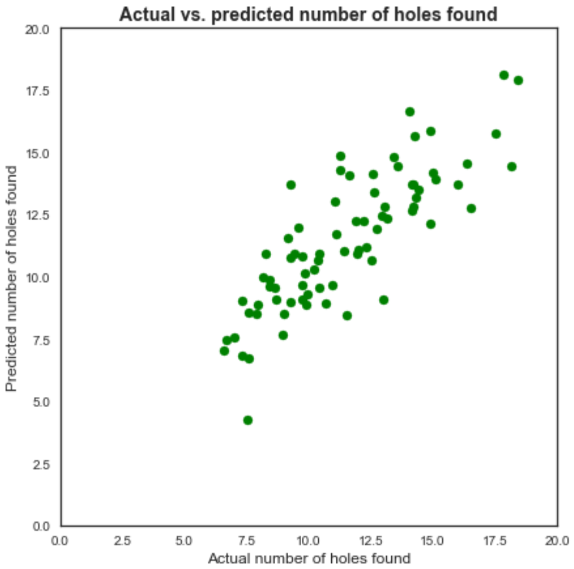
Below, we've plotted the actual vs. predicted response times using the fitted model. The scatterplot does seem to line up well in a linear fashion. On the right, we can see the residuals for both the training and test set are distributed uniformly around zero**.**

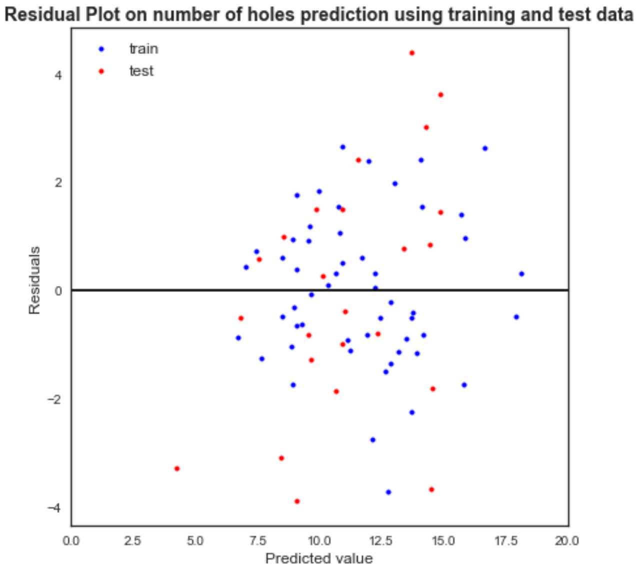


3.1.3 Linear Regression Results on Number of Holes Data

The same steps were taken with the number of holes data as with the response time data above. But the adjusted R-squared value is 0.65 for this model—much better than the response time model. The MSE on the training data set is 1.9, while the MSE on the test set is 4.9. A k-fold cross validation (k = 5) gives an MSE of 6.8. With such low MSE's, overfitting is not a concern for this model.

The graphs below confirm that the number of holes found per block is modelled well using linear regression. Actual vs. predicted values using the model are clustered tightly around the diagonal, and the residuals are uniformly distributed around zero.





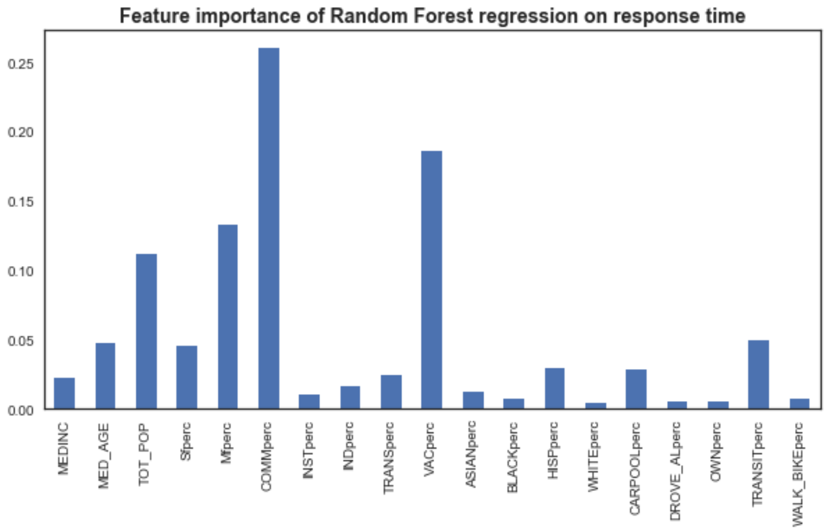
3.2 Random Forest Regression

After linear regression produced a subpar model for the response time, we can turn to random forest regression using the same training and test sets from linear regression.

3.2.1 Random Forest Regression Results on Response Data

The R-squared value from the fitted random forest regression on the response data is much higher than the linear model—0.89. However, while the MSE on the training set was 8.0, the MSE for the test set jumped significantly to 84.1. The regression for the response data is again overfit. Performing a k-fold cross validation with 5 folds still showed a higher MSE of 76.2. We can conclude that neither linear or random forest regression is a good fit to model this data.

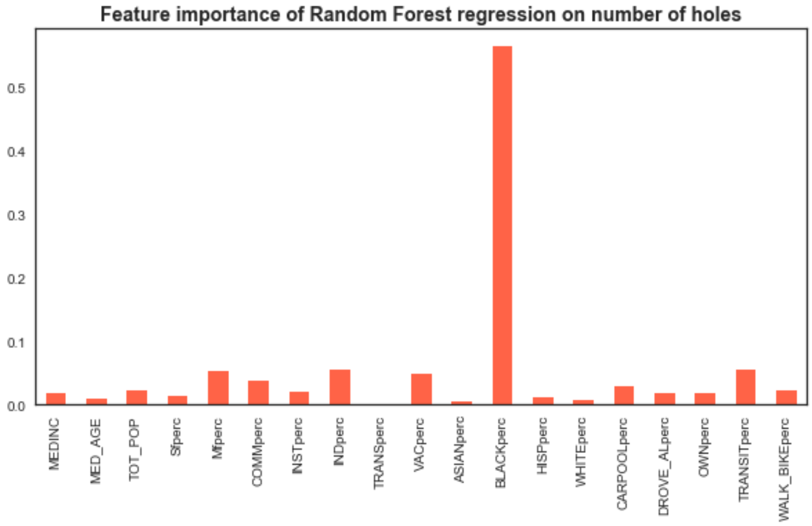
However, the feature importances are shown below to give a sense of which variables the model found to have the largest impact on mean decrease node impurity during the training. We can see that percentage of land zoned as commercial and percent of vacant land were more important than the other census variables in decreasing node impurity.



3.2.2 Random Forest Regression Results on Number of Holes Data

The random forest regression fits the number of holes per block data well with an R-squared value of 0.90, and the MSE on the training set was only 0.9. This MSE grew just slightly on the test set to 4.2 so we can say this model is not overfit. The most important feature as seen below is the percentage of residents who are black, followed distantly by the percentage of residents who are Asian.

From the chart below, we see that the percentage of black residents is by far the most important feature in this model, followed distantly by the percentage of Asian residents.



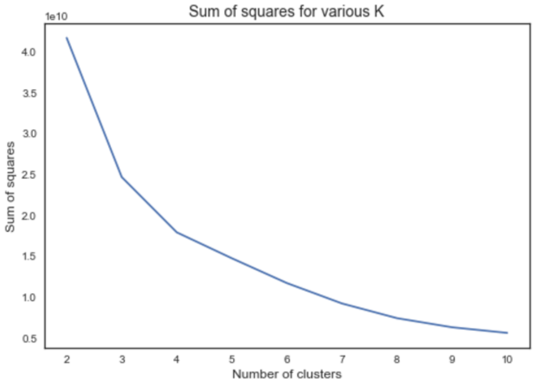
3.3 K-Means Clustering

We can apply k-means clustering to the total dataset (the features selected from the regression and the pothole metrics) to determine if optimal clustering lies along geographic lines as theorized above. We saw from the chloropleths that northern community areas have a longer response time while southern community areas experience more potholes per block. Do we see the same north-south clustering when applying the all the census features explored in this analysis?

3.3.1 Determining the Optimal K

There are several ways to determine the best number of clusters to use in k-means clustering. After first creating a new dataframe to include the features analyzed in the regressions as well as the average response time and number of holes data, I applied PCA to reduce the dimensionality of the dataset from 21 to 2 and labeled these dimensions x and y. I then applied k-means clustering for K = 2 to K = 10 and compared the color-coded clusters visually on an x-y plot. The plots showing K = 4 and K =5 seemed to best fit the dataset.

To more quantitatively determine the best K, we can use the elbow method to find the point at which the decrease in the sum of the squared distances between the points and their respective centroids starts to level out. From the graph below, I determined the optimal K to be 4.



I also plotted the silhouette score for the same K range to find the maximum, which occurred at K = 3. However, the maximum silhouette score was only around 0.45—indicating a weak structure.

The optimal K-values found from the above analyses (K = 3 and K = 4) were applied to the dataset with regional information appended. No geographic pattern emerged. I also tried K = 9 (the number of CA regions) and again saw a scattering of regions through all clusters.

However, K = 2 appeared to separate the CA’s accurately between north and south on a training set. The first cluster contained 12 northern CA's and 2 southern ones, while the second cluster contained 5 northern CA's and 27 southern ones. Western CA’s were split between the two clusters. When this model was applied to the test set, the first cluster added 6 northern CA's and 2 southern ones, while second cluster added 0 northern CA's and 11 southern ones—a very accurate generalization.

4. CONCLUSIONS AND FUTURE WORK

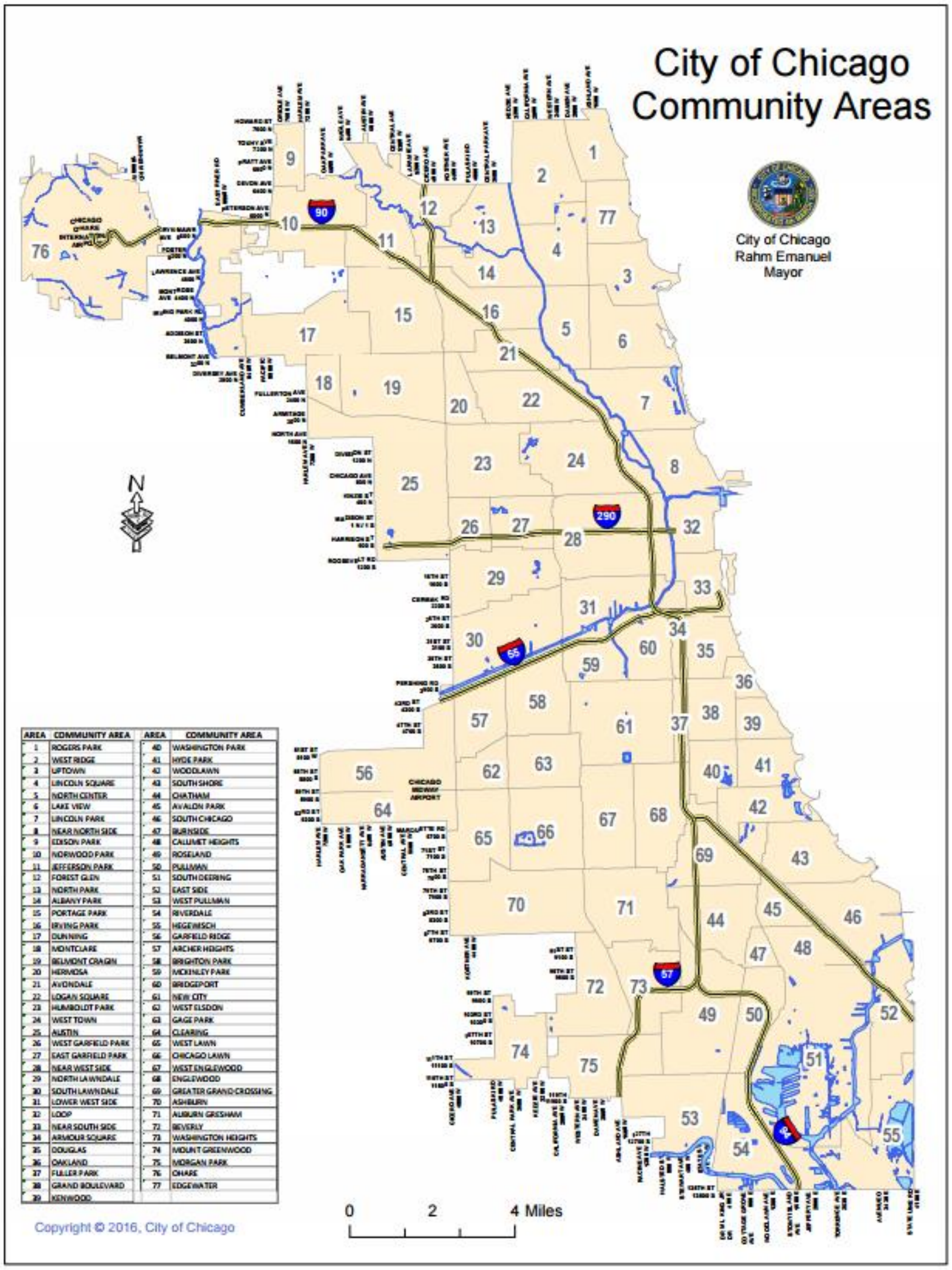
While the correlations shown in this paper are certainly intriguing, the real value of this analysis lies in the predictive models’ application to new data points. Chicago residents can use censusreporter.org to access similar census data in narrower geographies of interest (census tracts, zip codes, congressional districts, etc.) and predict pothole repair response time and the number of holes per block in those areas using the regressions.

Recommendations for further action based on this analysis:

- For the City:
    - Conduct a study to determine the root cause of the north/south divide in 311 service request response time. Are more repair crews based in the southern community areas, allowing for a faster response?
- For Chicago residents:
    - Lobby your alderman if your community area has a longer than average response time for pothole repair requests or if a large number of potholes are expected per block in your neighborhood.
    - Using smaller geographic areas as suggested above, plot your daily commute to determine if those streets are pothole-ridden. Use this data to route a new course on smoother roads.

Future opportunities to continue this work include further training of the clustering model on smaller geographic areas to determine if the north/south division holds on a larger data set. Incorporating weather data may also add accuracy to the regressions, allowing city residents to pinpoint the response time for their 311 service request based on the weather on the request's creation date. We can also determine if these correlations are standardized across metropolises by applying the regression models to similar census and pothole data from other cities.

APPENDIX



AREA	COMMUNITY AREA	AREA	COMMUNITY AREA
1	ROGERS PARK	40	WASHINGTON PARK
2	WEST RIDGE	41	HYDE PARK
3	UPTOWN	42	WOODLAWN
4	LINCOLN SQUARE	43	SOUTH SHORE
5	NORTH CENTER	44	CHATHAM
6	LAKE VIEW	45	AVALON PARK
7	LINCOLN PARK	46	SOUTH CHICAGO
8	NEAR NORTH SIDE	47	BURNSIDE
9	EDISON PARK	48	CALUMET HEIGHTS
10	NORWOOD PARK	49	ROSELAND
11	JEFFERSON PARK	50	PULLMAN
12	FOREST GLEN	51	SOUTH DEERING
13	NORTH PARK	52	EAST SIDE
14	ALBANY PARK	53	WEST PULLMAN
15	PORTAGE PARK	54	RIVERDALE
16	IRVING PARK	55	HEGEWISCH
17	DUNNING	56	GARFIELD RIDGE
18	MONTCLARE	57	ARCHER HEIGHTS
19	BELMONT CRAGIN	58	BRIGHTON PARK
20	HERMOSA	59	MCKINLEY PARK
21	AVONDALE	60	BRIDGEPORT
22	LOGAN SQUARE	61	NEW CITY
23	HUMBOLDT PARK	62	WEST ELSDON
24	WEST TOWN	63	GAGE PARK
25	AUSTIN	64	CLEARING
26	WEST GARFIELD PARK	65	WEST LAWN
27	EAST GARFIELD PARK	66	CHICAGO LAWN
28	NEAR WEST SIDE	67	WEST ENGLEWOOD
29	NORTH LAWNDALE	68	ENGLEWOOD
30	SOUTH LAWNDALE	69	GREATER GRAND CROSSING
31	LOWER WEST SIDE	70	ASHBURN
32	LOOP	71	AUBURN GRESHAM
33	NEAR SOUTH SIDE	72	BEVERLY
34	ARMOUR SQUARE	73	WASHINGTON HEIGHTS
35	DOUGLAS	74	MOUNT GREENWOOD
36	OAKLAND	75	MORGAN PARK
37	FULLER PARK	76	OHARE
38	GRAND BOULEVARD	77	EDGEWATER
39	KENWOOD		

Image from the City of Chicago https://www.cityofchicago.org/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Community_Areas/Community_Areas_w_Number.pdf