Preventing Hospitalization through 311 Services

311 service requests and NTA community health data

Citadel Data Open Princeton
Team 10

Karthik Tadepalli, Qinya Li, Nico Winata, Kevin Sun

September 22, 2018

Executive summary

How can we reduce the rate of preventable hospitalizations through 311 services? The New York Department of Sanitation receives thousands of requests every year to deal with public health issues. Their service covers problems ranging from water puddles to rat infestations, all of which have health detriments if left unaddressed. We hypothesize that an increased rate of efficacy in 311 services will decrease the rate of preventable hospitalization, all other factors kept constant. The nuance in this conclusion is that not all 311 services are equal; we hypothesize that some are more important than others, and we seek to identify those services which are most important in protecting public health.

Preventable hospitalization represents hospitalization due to causes generally caused by public safety measures failing. We chose preventable hospitalization as a metric of public health deliberately, because it best captures the middle ground of how government can affect health. It is more pliable than deep-rooted metrics like premature mortality, yet at the same time it is more general and informative than a narrow metric like rates of lead poisoning. **Thus, preventable hospitalization's middle ground allows us to draw conclusions that are both actionable and generally informative.**

We use LASSO regression to model preventable hospitalization using explanatory variables from 311 service data as well as potential confounders, and we obtain a sequence of explanatory variables that best coincide with a high rate of preventable hospitalization.

Our findings have important implications for policymakers. First, we establish an association between government inefficiency and preventable hospitalization Governments have limited resources and allocating those resources for maximum impact is the heart of government policy. Through our visualization, we also show exact neighbourhoods where important service aspects are currently lacking (see conclusion). Thus, we give policymakers a concrete guide to prioritizing 311 requests given their limited resources.

Technical exposition

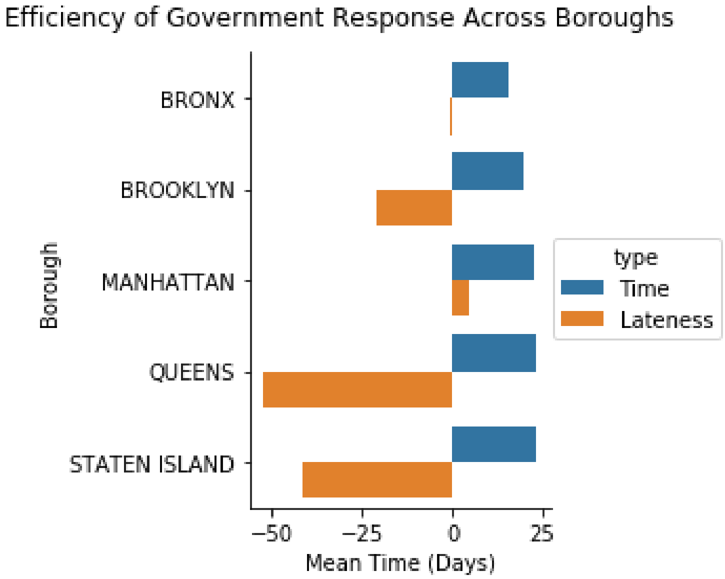
We analyzed three datasets; **311_service_requests** and **nta_community_data**, an external dataset from the NYC Department of Health which represents many demographic, socioeconomic and health features across neighborhoods. We engineered many new features from these datasets to represent the problem and serve as features for LASSO regression.

Data preparation and exploratory analysis

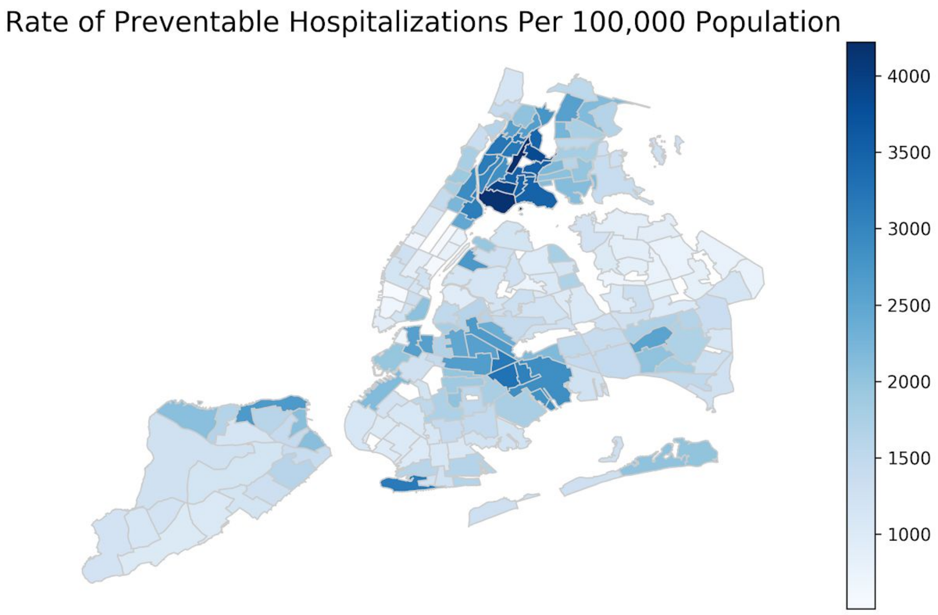
We focused on 311_service_requests and nta_community_data and ignored the other provided datasets. Our decision to focus on these two and not include the community_health dataset was based on community_health not being granular enough for our analysis, as it focused on borough-level health features rather than neighborhood-level features.

We first explored 311_service_requests, focusing primarily on the datetime data. We realized that a key metric of government performance was its efficiency at responding to 311 requests. No feature in the dataset incorporated this perspective, so we created our own features: **response_time** represented how long the government took to resolve a complaint, while **response_lateness** represented how *overdue* the resolution was (based on the government's guidelines on how long a resolution should take).

To see if this line of investigation showed promise, we grouped the incident reports by borough and plotted a bar graph of response_time and response_lateness. The great disparity in government response efficiency across boroughs showed us that we were on track to find something important; in Queens, the government response was 50 days late on average!



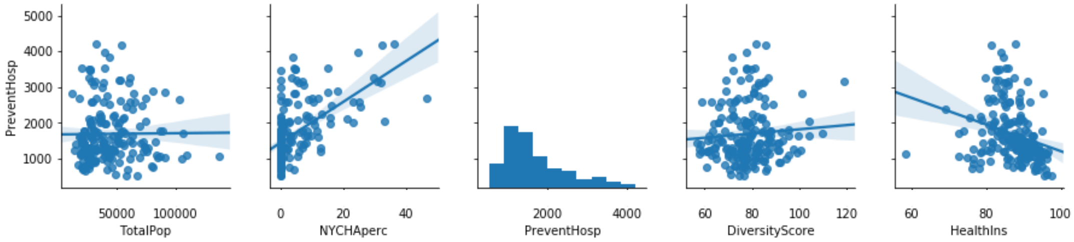
At the other end, we explored nta_community_data to see if we would find important details about preventable hospitalizations and found some stark disparities.



We filtered the data to be more relevant to our investigation; we excluded requests for non-health related departments (e.g. NYPD complaints on public nuisances), and then selected only data from 2012-2014, which is the corresponding time data available in nta_community_data. We took a representative sample of 50,000 service requests from this category (sampling because the dataset was too large to feasibly analyze and model with limited time).

Furthermore, we wanted to investigate whether certain complaint types were more important than others in predicting health outcomes, so we created 54 count variables representing the count of each complaint type. These include the count of each complaint type, average processing time, and the proportion of complaints that unresolved or are past their due date. (e.g. count of rodent complaints, average processing time of rodent complaints, and percentage of late/unresolved rodent complaints)

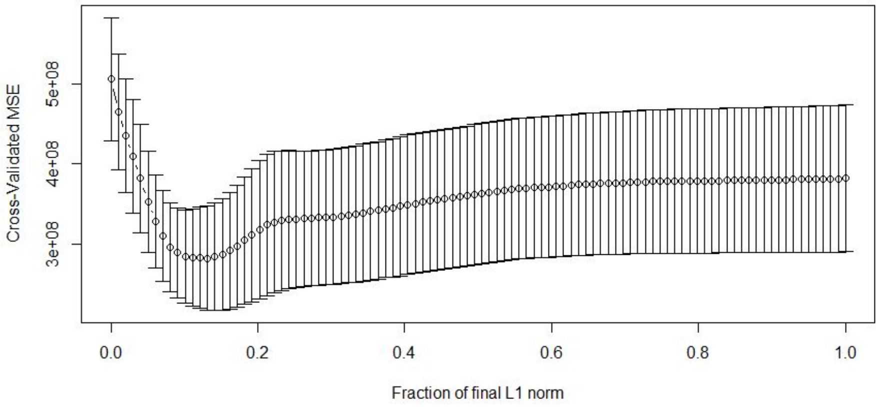
The last step in our data preparation was to take into account potential confounders which can affect our model. This is an important aspect of our modeling, as we want to discover the true effect of government efficiency on preventable hospitalization. To illustrate this, we can see in the graphs below that on average neighborhoods with high health insurance coverage rate have lower rates of preventable hospitalization, for example, because insurance coverage is a good measure for health awareness. Hence it is very important to include health insurance coverage in our model so that we can quantify how much effect is truly due to government efficiency and how much is due to confounding effects.



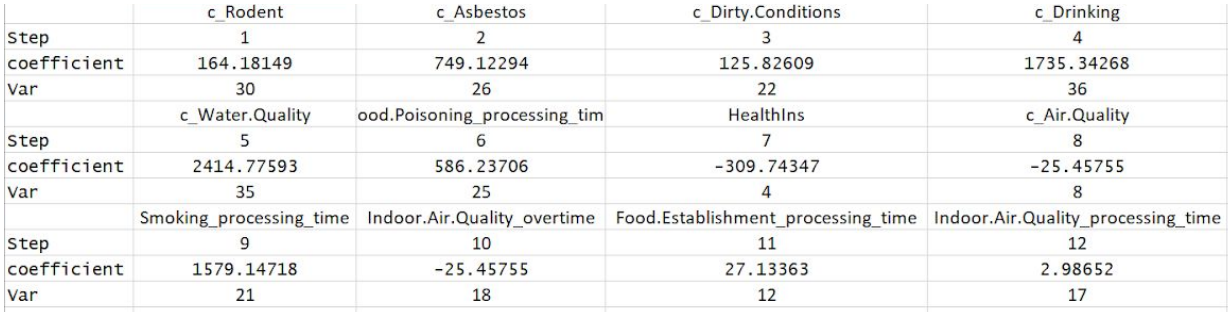
In order to establish a relationship between preventable hospitalizations and government inefficiency, however, we had to establish a corresponding variable between 311_service_requests and nta_community_data. We decided to use the Neighborhood Tabulation Area (NTA) as a unit of analysis because 311_service_requests provided coordinates for each incident which we could map to an NTA using a shapefile of NYC from NYC Open Data. Thus, we kept the NTA code as an ID that could map our observations between datasets.

Quantitative Modeling

We place preventable hospitalizations as our target variable and select a number of variables from both 311_service_requests and nta_community_data, as explained above. As discussed, we also include potential confounders in our set of explanatory variables. We perform an elastic net regression on 188 observations (188 NTAs) and 59 explanatory variables. We chose the parameter alpha based on cross-validation, minimizing the mean-squared error. This resulted in a choice of **alpha=0.1**.



We then chose variables using Least Angles Regression's sequence of forward selection, where each step is chosen to minimize Mallow's Cp, and we stop when we reach minimum Mallow's Cp. This gives us a total of 12 variables, which are shown in the table below.



	c_Rodent	c_Asbestos	c_Dirty.Conditions	c_Drinking
Step	1	2	3	4
coefficient	164.18149	749.12294	125.82609	1735.34268
Var	30	26	22	36
	c_Water.Quality	ood.Poisoning_processing_tim	HealthIns	c_Air.Quality
Step	5	6	7	8
coefficient	2414.77593	586.23706	-309.74347	-25.45755
Var	35	25	4	8
	Smoking_processing_time	Indoor.Air.Quality_overtime	Food.Establishment_processing_time	Indoor.Air.Quality_processing_time
Step	9	10	11	12
coefficient	1579.14718	-25.45755	27.13363	2.98652
Var	21	18	12	17

Mallow’s Cp is a measure of model fit that is penalized by number of parameters in the model. It is minimized with a parsimonious model that has a good predictive power. We chose to minimize Mallow’s Cp as it is in line with our objectives - we want to sequestrate the most important features among our list of engineered features. The sequence here can be interpreted as the degree of information they explain in the model. The most important variables are above, as they contribute most to minimizing Mallow’s Cp. The most important indicators here are number of complaints related to Rodent, Asbestos, and Dirty Conditions.

It is important to note that some 311 service variables outperform confounding variables such as Health Insurance coverage - showing the importance of 311 services in minimizing preventable public health conditions.

Conclusions

Our analysis identifies the 12 most important predictors of preventable hospitalization. Notably, **our features of government response time outperformed all potential confounders**, except for rates of health insurance coverage. This indicates that our features are strongly predictive and governments should allocate more resources towards addressing those cases faster and more efficiently.

Furthermore, of the 54 complaint types, only 11 appear in the model; LASSO reduces the coefficient of the other 43 variables to 0. This implies that many of those complaint types are simply irrelevant in determining preventable hospitalization, and reallocating government resources away from those complaints would benefit society as a whole.