Boston Area Research Initiative

Final Report on the Mass 211 Data Analysis Project

Using community characteristics to predict 2-1-1 demands

Jesse Yang, Research Assistant
Boston Area Research Initiative, Northeastern University
Dec 28, 2017

OVERVIEW

With the support of Mass 211, we have completed an initial analysis of the Mass 2-1-1 datasets, for both the call record data and resource database. We employed various data science techniques to uncover the relationship between 2-1-1 demands and community characteristics, built graphics and tools for easier data exploration, and produced multiple data byproducts that may be helpful for future projects on public health and human service issues in Massachusetts.

We tested different models for predicting 2-1-1 demands with community characteristics and established a solid baseline for the prediction. We examined the importance of variables and investigated how subtle patterns in community characteristics might be helpful in locating vulnerable population.

We composited the data we studied into a single interactive map[1], allowing the public to compare variables side-by-side at different geographical levels.

We also created a resource map where users could type and search instantly for any 2-1-1 resources.

[1] https://mass211.herokuapp.com/

Table of Contents

Overview 1

Background: Project Objectives and Research Methods 3

Preparation: Collect, Clean and Transform Data 4

- Data Collected 4
- Data Treatment 5
- Taxonomy Topics 5
- Data Exploration 6

2-1-1 Calls: Find a Benchmark for Predictions 8

- Linear Regression 8
- Poisson-based Models 8
- Variable Importance and Improved Models 9
- Outlier Analysis 11

Resources: Data Quality and Easier Access 11

Conclusion and Next Steps 12

Appendix: User Manual for the Mass 211 Map 14

- Demands 14
- Resources 16
- Development and Maintenance 17

BACKGROUND: PROJECT OBJECTIVES AND RESEARCH METHODS

This collaboration of Boston Area Research Initiative (BARI), National Alliance on Mental Illness of Massachusetts (NAMI) and Massachusetts 2-1-1 (Mass 211) was initiated out of concerns regarding an alarmingly high volume of housing needs for people with mental health issues. But housing and mental health issues accounted for only 13% of the 2-1-1 calls—it would be unwise to throw the other 87% of the data away. Therefore, we decided to study the general trends of all service types first, then dive into specific categories as we see fit.

The call records contained information regarding the sex, age, source of income, military status, and family characteristics of the callers, but most of the data fields were either incomplete or specific to a limited set of service types. In addition to the fact that there were no reliable ways to trace callers and their associated calls, we worked mainly on aggregated measurements, i.e., number of calls per capita—down to specific service types when appropriate—aggregated at the ZIP code, township, and county level.

The 2-1-1 calls were labeled with the AIRS/211 LA County Taxonomy, documenting types of services inquired by the callers. The Taxonomy is a hierarchical system and contains more than 9,000 terms. For easier interpretation, we handpicked the most common service types—a few selected sets of taxonomy terms—and assigned each 2-1-1 call a "topic". All breakdown analyses were based on these topics.

In consideration of the data we had access to, we hoped to achieve four major goals:

1. Create a framework to predict and evaluate human service demands based on 2-1-1 data and easily obtainable public datasets.
2. Locate underserved neighborhoods where 2-1-1 services would add the most value.
3. Identify public health problems underlying 2-1-1 calls.
4. Track geographic and demographic distribution of critical societal issues that are hard to fix (e.g., housing for people with mental health issues).

Time and resources allowed us to finish Goal 1 and worked on part of Goal 2 and 4, while remaining goals would require more corroborative data and more careful analyses to attain robust conclusions.

The framework for Goal 1 includes scripts to scrape public datasets and clean the 2-1-1 data, a website for interactive data exploration, and statistical models to predict 2-1-1 calls based on community characteristics. Goal 2 was achieved by looking at the outliers in prediction, in combination with their demographics and socioeconomic descriptors. Goal 4 was addressed by allowing anyone interested to compare different variables using the interactive map.

PREPARATION: COLLECT, CLEAN AND TRANSFORM DATA

Data Collected

We used mainly the census data (2011-2015 American Community Survey 5-year Estimate) to describe community characteristics, but have also investigated the potential of some other public health datasets. Here is a complete list of the datasets we collected and explored:

1. 2011-2015 ACS 5-year Estimate: the demographic and socioeconomic data, including race, age, education level, income, housing cost, etc. of the population in a ZIP code area or township. The main dataset we use for future analysis.

2. Massachusetts Health Status Indicators: released by Massachusetts Department of Health and Human Services, including perinatal and child health Indicators, infectious disease, chronic disease, substance abuse indicators, and hospital discharges. Underlying data were collected between 1996 and 2013.

3. 500 Cities - Local Data for Better Health: data about chronic disease risk factors, health outcomes and clinical preventive service use for the largest 500 cities in the United States, released by CDC. MA has 13 cities selected. It also offers estimation at census tract level based on demographics. Did not proceed because there were only 13 usable observations and the 2-1-1 data could not be aggregated at the census tract level.[1]

[1] Technically we could do an estimation at the census tract level for the 2-1-1 data, too. But two estimations would not give us the statistical power we want.

4. Massachusetts Opioid Statistics: statistics of overdose deaths and opioid-related EMS incidents at the city/town level.

5. US Chronic Disease Indicators: contains yearly trends of chronic disease indicators, available at the state level only, unfortunately.

6. Urban Area to ZIP Code Tabulation Area Relationship File: used to estimate the degree of urbanness of a ZIP code area.

Data Treatment

As mentioned, all measurements were aggregated at three geographic levels, of which ZIP code areas were the most granular. Our model was optimized mainly at the ZIP code level; however, Mass Health Status Indicators and Opioid Statistics were only available at the city/town and county level. Their relationships with the 2-1-1 data were examined separately. Because of the scarcity of the data (most variables contained a lot of zeros), there were no statistically sound relationships found.

Before the aggregations, we excluded certain types of calls to avoid skewness and biases:

1. Call2Talk calls were excluded as they were more prone to repetitive callers and was not incorporated into the system until Dec 2016. They also represented very different population groups than issues driven by economic issues, such as child care and housing assistance.

2. Calls before June 1st, 2016 were excluded, as data for the months before seemed incomplete. This was likely due to the migration of the 2-1-1 system.

3. Calls from outside of Massachusetts were discarded.

For the statistical models, we also excluded observations with almost no calls and a very small population. Outliers in population characteristics, such as 01063, the ZIP code for Smith College, where almost all population are females, were also excluded.

Taxonomy Topics

We converted the taxonomy into 17 topics. Ordered by the size of call volumes, they are childcare, housing, utilities, homeless, food/cloth, income, info service, legal, government, health, mental, community, disability, education, care/companion, and youth help. These topics covered 90% of the calls and might overlap with each other. For example, all "homeless" calls were also labeled with the tag "housing".

We used these topics to label the resources, too.

Data Exploration

In the end, we aggregated with *49,805* calls generated during the 14 full months between June 1, 2016, and August 31, 2017. About one-third of the calls were labeled by more than one Taxonomy term. After filtering observations with less than 200 people, we obtained *509* ZIP code areas (out of 537 in total), with the population size ranging from *237* to *60,725*, and a total number of calls per 1,000 people from *0* to *88.26*. Fig. 1 shows the distribution of the per capita call volumes across ZIP code areas.

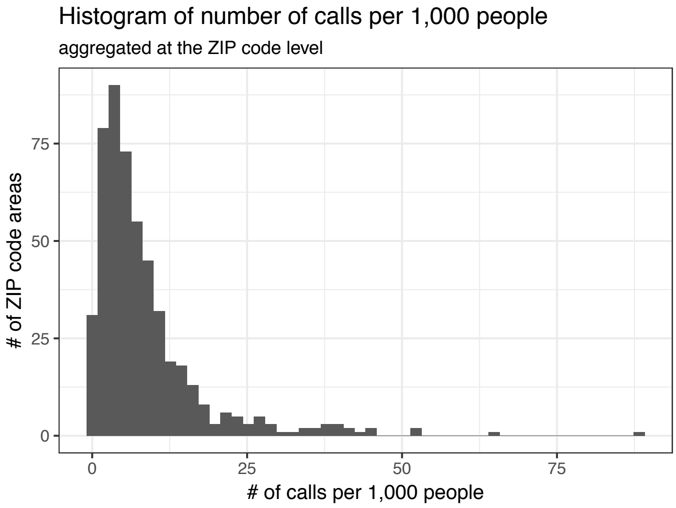


Fig. 1: histogram of calls per capita

The neighborhoods with the highest per capita number of calls are:

1. **01901** - A small fraction of Lynn around the Central Square, population *1,371*, 39.4% are foreign born (14% are naturalized citizen) and 38% are Hispanics. The neighboring ZIP code 01902 also have a fair number of Hispanics and high demand in 2-1-1 services, especially about child care.

2. **02121** - Roxbury, Boston, *64.60* calls per 1,000 people, population *28,051*. Predominantly black and Hispanic neighborhood, median house income about 26,000.

3. **01105** - Metro Center and South End of Springfield, *52.71* calls per 1,000 people, population 12,027. Large Hispanic community, about 63% are Hispanics and 27% are US citizens born out of US.

4. **01608** - West of Worcester train station. *51.60* calls per 1,000 people, population *3,275*. Another Hispanic community with more than 54% of the population being Hispanic and 47% of the residents born out of US.

These areas or neighboring areas also generated the highest amount of calls related to housing issues, too.

Aggregation at the city/town level had about a little fewer observations (348 townships with more than 200 people), but the distributions are the same.

In terms of call count distribution across topics, Fig. 2 shows the relative sizes of the call volumes. Note that since a call may be labeled with multiple topics, some of the calls were double- or even triple-counted.

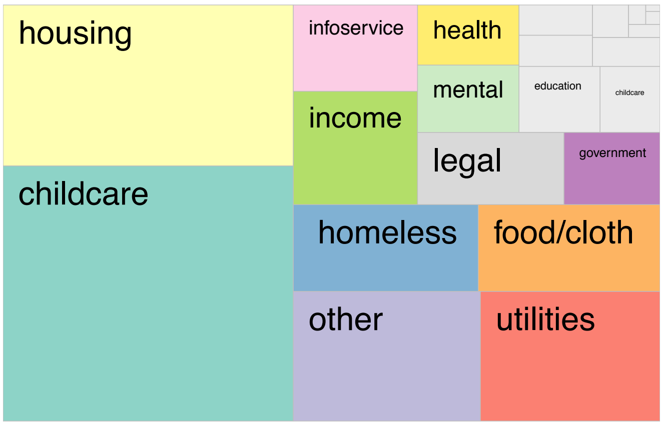
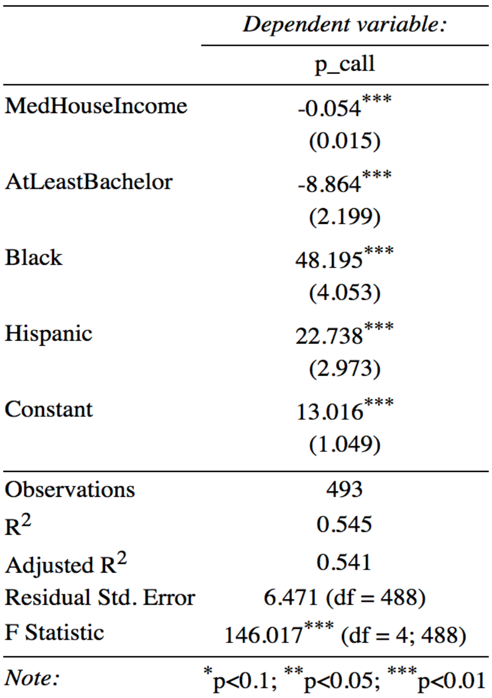


Fig. 2: Relative sizes of call volumes by topic

The sizes were decent when counting for the whole state, but when we broke them down by topic AND geographic units, the data becomes very sparse. At the ZIP code level, when breaking down by topics, about 40% of the observations had zero calls, and more than 70% of them had less than 5 calls.

2-1-1 CALLS: FIND A BENCHMARK FOR PREDICTIONS



	Dependent variable:
	p_call
MedHouseIncome	-0.054***
	(0.015)
AtLeastBachelor	-8.864***
	(2.199)
Black	48.195***
	(4.053)
Hispanic	22.738***
	(2.973)
Constant	13.016***
	(1.049)
Observations	493
R^2	0.545
Adjusted R^2	0.541
Residual Std. Error	6.471 (df = 488)
F Statistic	146.017*** (df = 4; 488)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1. Simple OLS regression

After exploring correlations between variables, we tested different regression models to evaluate how accurate a model could we possibly build to predict 2-1-1 call volumes using community characteristics.

Linear Regression

The first thing we tried was an Ordinary Least Squares (OLS) linear regression model in which we used race, education level, and house income as the predictors, and the raw number of calls per 1,000 people as the outcome variable. The model could easily explain 54% of the variance and all variables had *p*-values less than 0.001, indicating strong statistical significance.

The variables were chosen based on their relevancy to poverty measurements and simplicity for interpretation. We also used Recursive Feature Elimination (RFE), i.e., backward selection, to select the variables. RFE recommended a model with more than 10 variables, but it was more prone to overfitting and made the model harder to interpret.

Poisson-based Models

For count data, Poisson-based models are often more revealing. We fitted a Poisson regression and a negative binomial regression using the total population of the neighborhood as the offset variable, and the total number of calls as the independent variable.

Poisson regression gave odd results where almost every variable was significant, and the residual graph showed a fan shape (Fig. 3), indicating the data were over-dispersed and a Poisson model was not able to handle large values. Negative binomial regression showed more reasonable results and the range of residuals significantly narrowed.

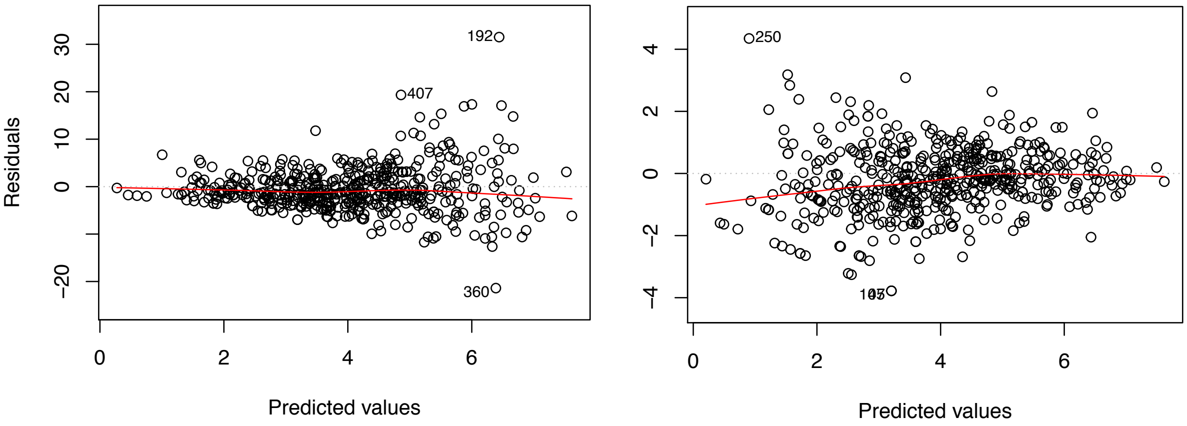


Fig. 3: Residual graphs of Poisson-based models. Left: Poisson regression; right: negative binomial regression. Variables fitted: black, ethnic heterogeneity, median

Variable Importance and Improved Models

We collected 67 census variables, including almost all poverty-related census measurements: proportion of people receiving public assistance, people with disability, household types, proportion of single-parent households, children living with grandparents, proportion of residents born out of state/US, etc.

In addition to handpicking the most commonly used predictors in social science study, we needed a smart way to identify opportunities for other variables, too. We utilized the decision-tree-based machine learning algorithm, *Random Forest,* to evaluate the importance of variables.

Random Forest evaluates the importance of a variable by calculating the percentage of increase in Mean Squared Errors of predictions when excluding the variable from the assembled decision trees. As shown in Fig. 4, percentage of people with private insurance (`HI_PrivateInsured`) and median housing value were the two most important variables.

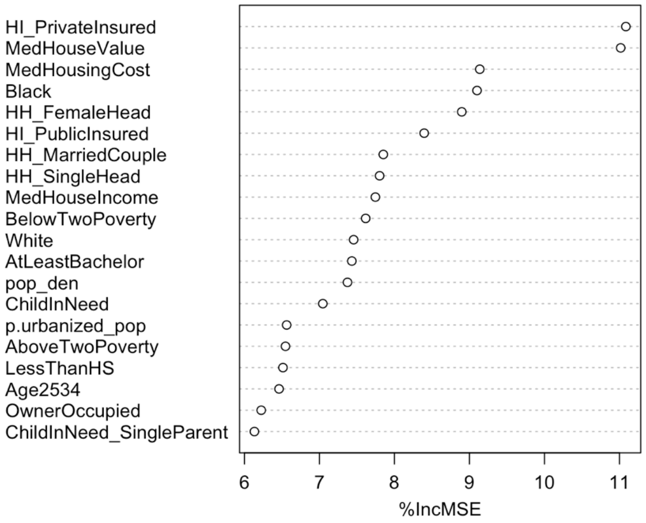


Fig. 4: variance importance as discovered by Random Forest

When we added these variables to the regression models, we did see significant improvements in the OLS regression model (Table 2), which is especially true for `HI_PrivateInsured`.

model	residual DF	RSS	DF	sum of sq	p-value
~ MedHouseIncome + Black	490	23683.93	NA	NA	NA
+ Hispanic	489	21116.11	1	2567.82	<2e-16
+ HI_PrivateInsured	488	18473.26	1	2642.84	<2e-16
+ MedHouseValue	487	17969.81	1	503.44	0.0002
+ p.urbanized_pop	486	17149.99	1	819.82	<2e-5
+ AtLeastBachelor	485	17138.82	1	11.17	0.0579

Table 2: ANOVA test showing the decrease of degree of freedom (**DF**) and residual sum of squares (**sum of sq**), when adding more variables to the model.

Adding these variables to the negative binomial model did not see such significant improvements.

Another variable we tested and improved the linear model is the proportion of population living in an urbanized area (`p.urbanized_pop`), even though it was deemed important by Random Forest.

Outlier Analysis

After getting a satisfactory model, we examined the outliers to try to understand why did they behave differently than neighborhoods with similar characteristics.

First up are the ZIP code areas with a decent number of inhabitants but zero 2-1-1 call record:

1. **02575:** 2,564 people, Town of West Tisbury on the island Martha's Vineyard, where rich people build mansions.
2. **01731:** 2,082 people, Hanscom Air Force Base.
3. **01434:** 1,658 people, Fort Devens, US Army Reserve Force training area.
4. **01262:** 1,239 people, Town of Stockbridge, located at westernmost of MA.

This alarms us that we may need to exclude military bases and some other special areas, such as university campuses, to reduce noises in our model.

Then we looked at the areas with an especially higher amount of calls than predictions. They are mostly the aforementioned ZIP code areas with the largest per capita call volumes. The fact that most of them had a large population of Hispanic population seems to suggest that residents in large Hispanic neighborhoods tend to seek help for social services more often.

RESOURCES: DATA QUALITY AND EASIER ACCESS

As the primary provider of "information and referral" for Massachusetts residents, Mass 211 maintains a comprehensive database of agencies and organizations that can help people access health, human and social services.

In an effort to analyze whether the presence of certain resources influences the behavior patterns of people seeking help, we spent a considerable amount of time to clean and transform the resources database into an easier-to-work-with format.

In the iCarol system used by Mass 211, resources are stored in 4 categories: *agency*, *program*, *sites*, and *program at sites*. An agency may have multiple sites (e.g. branches of AA), and may operate multiple programs (e.g. hospitals run an HIV/STD testing program, in addition to a behavioral health assisting program, etc.). A program may be conducted in multiple sites and covering different areas. The program at sites data indicates which programs are running on which sites.

But the exported data is just one single table, and important information such as location, coverage area, etc., can be found in either one or all of the aforementioned four types of entities. However, taxonomy code (service types) can only be found at *programs*, and geolocation is mostly present in *sites*.

We consolidated the data into one entity type–a single geographical entity with attributes explaining its service types and coverage areas:

1. *Agency* was used to obtain the website and contact info.
2. *Site* was used for the name, geolocation, address and operation hours of the sites.
3. *Program* and *Program at sites* were used to extract taxonomy code and names, as well as coverage areas of the sites.
4. A new data field *location type* (e.g. city hall, shelter, hospital, etc.) was added and inferred from taxonomy code.

After cleaning and removing duplicate records, we created a database of *8,463* human service sites that are physically located in Massachusetts. They belong to *3,488* agencies.

This data is now accessible in an easy-to-use resource map[1], with which users could search resources by name, website, city, address, ZIP code, service types, and coverage areas, all searchable in free text and come with automatic spelling correction.

CONCLUSION AND NEXT STEPS

By experimenting with different regression models, we identified the most relevant variables associated with demands for health, human and social services and achieved a robust benchmark in predicting the 2-1-1 demands with community characteristics.

[1] https://mass211.herokuapp.com/resources

We discovered that coverage of private health insurance was an important predictor of 2-1-1 demands—the more residents are covered by private health insurance, the less 2-1-1 calls a neighborhood would generate. We suspect this is because people often get private health insurance through quality employment, and having a good job means more economic and social stability and less reliance on public services. The benefit of a good job is so important it even outshines other descriptors combined.

We could also confirm that urbanized areas do generate more 2-1-1 calls, and race and community effect may indeed influence how people in need seek help.

We collected a comprehensive set of public datasets relating to health and human services in Massachusetts and incorporated them in an all-in-one interactive map. The map is useful for anyone who wants to explore the needs and resources of Massachusetts communities.

The map could be further improved by adding longitudinal data such as changes of demands over time and shifts of population characteristics.

We have consolidated the resource database into a simple sharable format, yet the data quality still has room for improvements. For example, we could build automatic tools to reverse geocoding the addresses of the sites, to improve the accuracy of the geo-coordinates; or we could adopt the *OpenReferral* initiative[1], following the steps of Maryland and Miami, use open-source tools to democratize 2-1-1 resources and boost public engagement.

Existing academic research on 2-1-1 data almost always involved some sort of experiments and required significant manpower for following up with the callers. We also see immense opportunities in following up with referral outcomes, or at least identifying and recording unique callers by phone number (could be encrypted) so we would know when they called 2-1-1 again. Being able to analyze repetitive callers, especially those requested different services, would make it easier to evaluate the evolving needs of communities.

[1] https://openreferral.org/

APPENDIX: USER MANUAL FOR THE MASS 211 MAP

The *Mass 211 Map*[1] is a web app hosted on public cloud service *Heroku*. Currently, it has only two tabs: *Demands* and *Resources*. The *Demands* tab allows users to explore and compare 2-1-1 demands, public health indicators, demographic and socioeconomic metrics; the *Resources* tab is an easy-to-use search interface for all available health and human service resources in the 2-1-1 database.

Demands

On the left top corner of the *Demands* tab are two dropdown controls for users to select the geographic level and the aggregated variable. Variables are grouped under three categories: *211 Calls*, *Demographic*, and *Public Health*. One may search for a variable by focusing variable dropdown and start typing (Fig. 5).

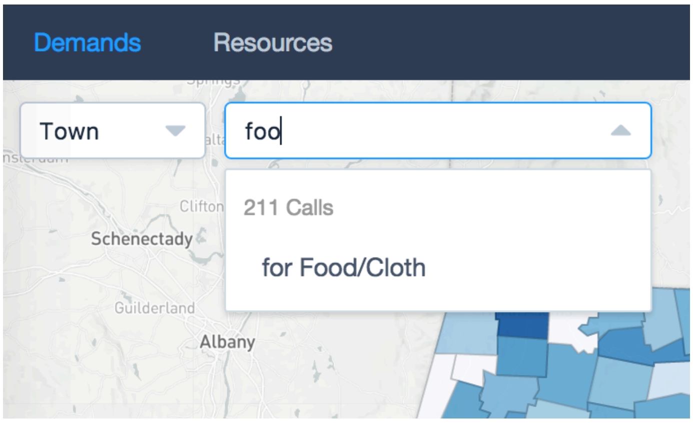
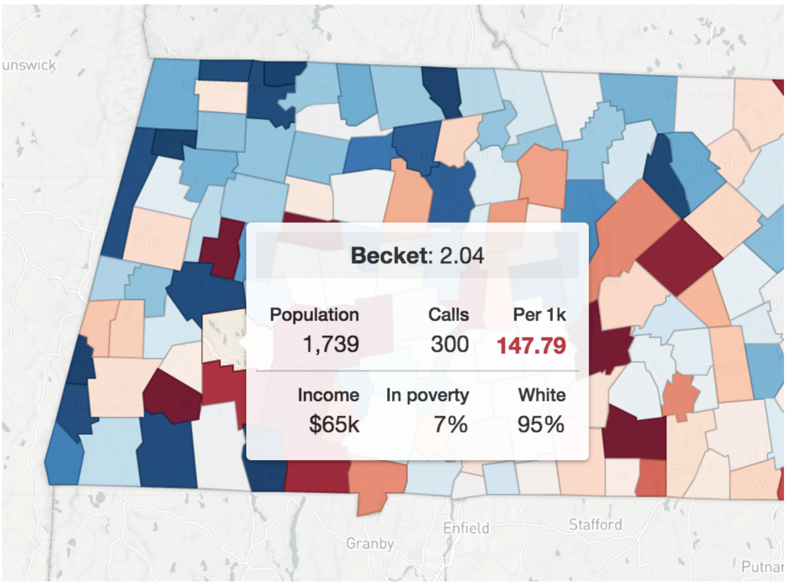


Fig. 5: Type to search a variable

The *211 Calls* are total numbers of calls per 1,000 people for a given geographic area and specific topic if selected. There is also a *Prediction Residual* which is the difference between the predicted number of calls and the actual call volume, according to a negative binomial regression model. On the residual map, dark red represents higher than prediction and dark blue represents lower than prediction.

The *Demographic* variables are important census variables describing the population characteristics of a region.

[1] https://mass211.herokuapp.com/



Becket: 2.04

Population	Calls	Per 1k
1,739	300	147.79

Income	In poverty	White
$65k	7%	95%

Fig. 6: Residual map and the overlay tooltip

The *Public Health* variables were those found in Mass Health Status Indicators and Opioid Statistics. They are available only at the township and county level.

If hovering on the map, the overlap tooltip will display basic statistics of the geographic unit under the cursor. The number behind the geographic unit name is the value of the currently selected variable. The *Per 1k* field turns red when the actual number of calls is at least 1 standard deviation higher than prediction (Fig. 6).

The bottom left of the map is the legend for the variable. For some variables, it also contains a short description of how the variable was calculated. You can fold the legend by clicking the arrow at its top if it blocks you from viewing the whole map.

At the top right corner are controls for creating new map panes. The panes can be used to compare variables and are synchronized when you drag or zoom the maps. You can disable the synchronization behavior by click on the sync switcher appeared at the left of the “add a pane” icon.

You may create up to four panes, simply hovering on the control and click on the “+” icon again. After creating a new pane, you can change the variable and geographic unit in it in order to compare different variables.

The map will update its URL when you add panes and change viewport, so you can copy and share the URL in your browser’s location bar, and anyone with the URL can see the exact same map as you do.

You may click on polygons on the map to zoom in and click again to zoom out. When the map is zoomed in, the opacity of the polygon would decrease so that you could see more details of the area.

Resources

The resource tab works very similar to Google Maps. Simply type anything about the resources you want to find and the map would automatically update.

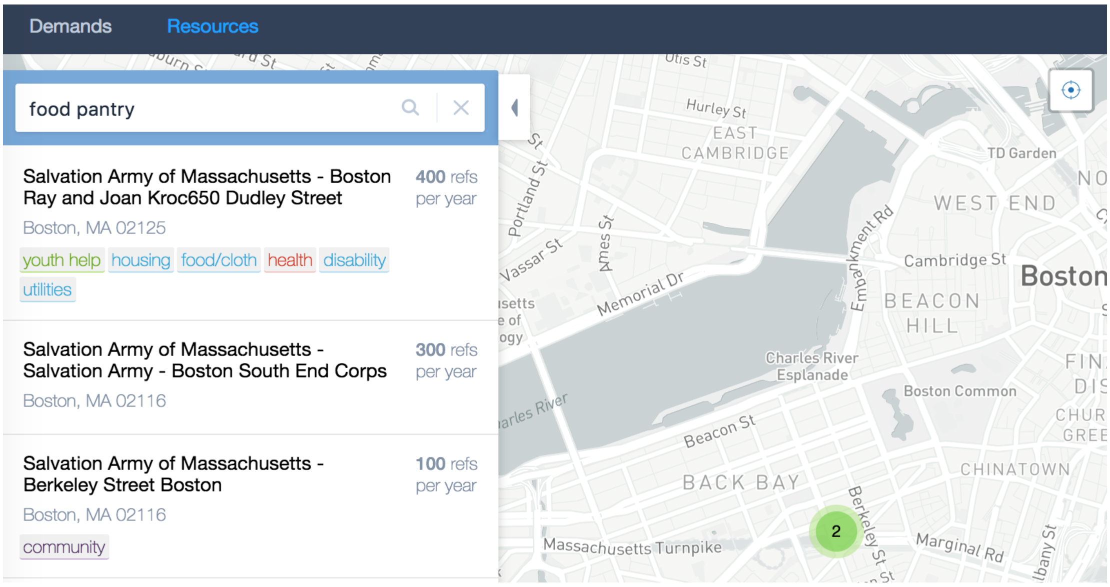
The search results list displays the name, address, tags by taxonomy topic, and referrals received per year, of the site and the agency.

The map displays at most 100 results by default. You can keep scrolling the results list to add more results to the list and more markers to the map.

Markers are clustered when they are close to each other, you can click on the clusters to zoom in, or you can also click search result items from the list.

Clicking on the markers gives you an info window of all the details about the site. You may view its alternative name, website, description, services offered, etc.

It is possible to search for resources by location (Fig. 7). Simply click the Location icon at the top right corner then the app would find resources nearest to you. You can search by keywords even when you are in *Search by Location* mode. Resources



food pantry

Salvation Army of Massachusetts - Boston Ray and Joan Kroc650 Dudley Street — 400 refs per year
Boston, MA 02125
youth help housing food/cloth health disability utilities

Salvation Army of Massachusetts - Salvation Army - Boston South End Corps — 300 refs per year
Boston, MA 02116

Salvation Army of Massachusetts - Berkeley Street Boston — 100 refs per year
Boston, MA 02116
community

Fig. 7: Search by location

near you would be ranked first. To turn off location search, click on the Location icon again.

Development and Maintenance

This web app does not require backend service, so it would be extremely unlikely for it to break. But should any improvements be considered or new features planned, one could refer to the source code on GitHub for development and deployment guides.[1]

[1] https://github.com/ktmud/mass211-map