ML4VA Proposal: determinants of neighborhood emergency service availability

A Preprint

Lanyin Zhang
University of Virginia
Charlottesville, VA 22904
lz8aj@virginia.edu

Steve Zhou
University of Virginia
Charlottesville, VA 22904
wz8ry@virginia.edu

December 11, 2023

Abstract

This study investigates the distribution and accessibility of emergency services in Virginia, with a focus on fire and health services. This project aims to assess the adequacy of current emergency service locations. Employing geospatial analysis, the research uses K-means clustering to create categories of emergency service availability based on a holistic view of the neighborhood's proximity to fire stations and hospitals as well as per-capita hospital bed availability. Then, a predictive classification model is built using the random forest model, factoring in demographic and socio-economic neighborhood characteristics to assess the importance of each factor in determining the emergency service availability. The study's outcomes are expected to contribute to the optimization of emergency service distribution and preparedness, especially in health-related crises.

1 Introduction

1.1 Motivation

Emergency services play a vital role in safeguarding the well-being and property of residents in any region. As local residents living in Charlottesville, Virginia, we value emergency services and safety in the case of emergency around us here in Virginia. We believe it is imperative to assess whether the current distribution of emergency services is optimal.

This study aims to employ geospatial analysis to evaluate the efficiency and sufficiency of emergency services in Virginia, for instance, the location of the fire station is of great significance for firemen to be able to arrive at the scene promptly. Another example would be food aid distribution or refugee for people in need of food and shelter in extreme weather such as hurricanes or flooding. Moreover, the recent COVID-19 pandemic highlighted the need for a well-distributed health infrastructure. During the peak, many regions in Virginia and other places faced hospital overflows and overwhelmed healthcare systems. This study will thus focus in the health field: when a health emergency happens, what is the availability of beds in hospitals for the neighborhood?

1.2 Literature

There are many measures and methodologies proposed to evaluate the sufficiency and efficiency services using geospatial data. Such topics are more discussed in areas like applied geography and health care management with local data, including Virginia. Kc and Corcoran [2017] discussed a spatial analysis method to model the response times for residential fire incidents. They employed quantile regression to investigate how socio-demographic, infrastructure characteristics and temporal factors influence response times. From another perspective, rather than focusing on response time to measure the performace of medical service systems, McLay and Mayorga [2010] introduced a method to assess how fixed response times impact patient survival rates. They designed an ambulance location model optimized for patient survival rates, using data form Hanover County, Virginia.

The two papers mentioned above utilized quantile regression and mathematical modeling as methods to analyze the problem. The quantile regression assumes a simple relationship between dependent and independent variables; The theoretic models are under various kinds of assumptions which may limit the its application. Thus this paper can provide a methodological innovation to employ deep learning in analyzing the problem.

1.3 Dataset

1.3.1 Data on Emergency Services

We will be measuring two types of emergency services: fire and health. The most direct measure of emergency service availability would be the actual response time of an emergent 911 call. However, due to privacy protection reasons, there is not an open dataset of response time available to us as undergraduate students. Thus, we will approximate the availability of emergent fire rescue services as a function of the neighborhood's distance to a fire station, and the availability of health services as a function of available hospital beds, weighted by the inverse of their distance to the hospital. We obtain this information from

1. ***Definitive Healthcare: USA Hospital Beds*** keeps track of 124 hospitals in the state of Virginia. It was last updated on May 22, 2023 with information on the geographic location, licensed hospital beds and ICU beds, and bed utilization rate.
2. ***Homeland Infrastructure Foundation-Level Data (HIFLD)*** is a general purpose dataset maintained by the Department of Homeland Security. We will be using the recorded spatial locations of all 52,184 fire stations in the United States, specifically those in and near Virginia.

1.3.2 Data on Neighborhood Features

We will query important socio-economic indicators of these neighborhoods from the Census Bureau's 5-year Estimate from American Community Survey (ACS) made in 2021. In general, we will define a neighborhood as a block group, which is defined by the Census Bureau as an area with a population of 600 to 3,000 that does not cross the census tract boundary. We will gather

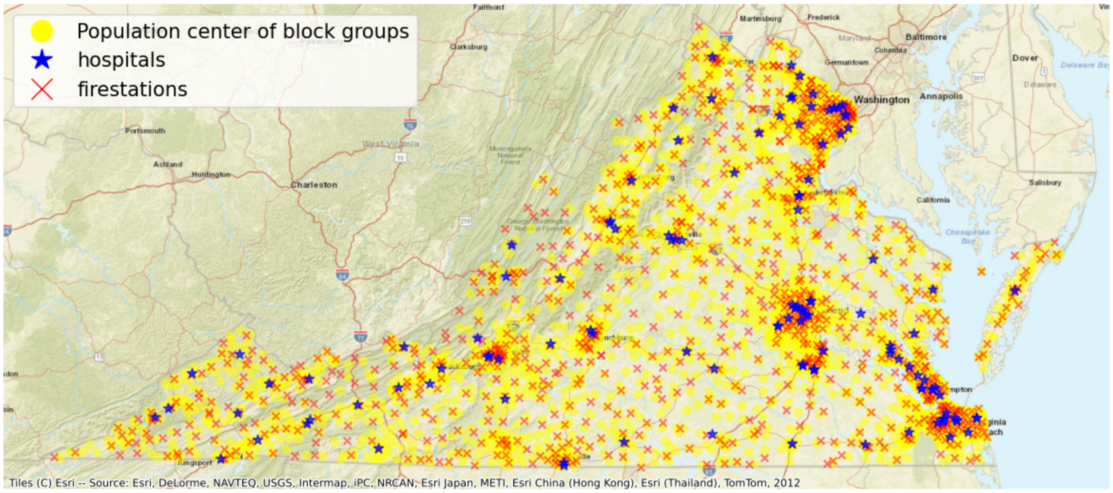
1. ***ACS Table B02001*** provides the population by race in the level of the census block group.
2. ***ACS Table B19013*** provides the median household income in the level of the census block group.
3. ***ACS Table B23025*** provides unemployment and labor-force participation statistics at block group level.
4. ***ACS Table B01002*** provides the median age at block group level.
5. ***ACS Table B15003*** provides the population with each level of educational degree at block group level.

In future research, we may decide to include more socio-economic factors in our analysis. These will be the input into our model, and we should have each census block group in Virginia as an observation, yielding a total of 5,332 observations.

2 Methodology

Firstly, we recognize that there are complicated reasons for people's decision of where to live, and setting an inappropriate granularity as our definition of neighborhood may incur unexpected flaws in our analysis. For example, although it is true that neighborhoods immediately adjacent to a hospital should have high healthcare availability, people may be reluctant to live there due to the 24/7 noise from ambulances, a factor that we cannot measure from our existing data sources. Thus, we decided to start by treating each block group as a single observation. Alternative definitions of neighborhood we will explore include census tract, which will yield a total of 1,907 observations, and census block, which will yield a total of 285,762 observations. Thanks to the versatility of ACS tables, data are easily available at all these levels from the same URLs.

With this definition of neighborhoods, we can merge all data sources into a single data and perform our analysis. With the socio-economic factors as input X and the emergency service availability (either fire rescue or health care) as the dependent variable Y. We will mainly experiment with different supervised machine learning models to predict Y. We plan to experiment with simple neural networks, random forests and boosted trees. The results of these ML models will be evaluated relative to the bench mark of an Ordinary Least Square Regression.



Tiles (C) Esri -- Source: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2012

Figure 1: Geospatial distribution of emergency services and population centers in Virginia.

2.1 Limitation

We recognize that there are a few limitations in our proposed methodology: i) Our dependent variables, the "scores" of hospital and fire service availability, is imputed by function that we define on our own. This is a possible source of biases as we do not have a ground-truth measure to tune this imputed value. ii) Due to the limit in available data, if we want to use advanced algorithms like ANN, we should not use too many features. As a result, we have selected only 5 socio-economic variables to reflect the demographics of the entire block group. In regressional analysis, this can lead to omitted variable problems. We will continue our review of the literature to find any possible ground-truth measure to solve the first problem, but due to restricted data availability, we are unlikely to find a definitive solution to the second problem. Nevertheless, with the introduction of more advanced methods, we will try to mitigate any possible bias introduced.

3 Preliminary Experiments

3.1 Manual inspection of data

After we had acquired the data from the sources listed above, we first took a manual inspection of the data after merging the tables from various different sources. Recognizing that different sources use different coordinate reference systems (CRS), we used the Python package GeoPandas to project them to the same CRS EPSG 4326 and produced the plot above. This visualization verifies an accurate projection.

3.2 A Baseline: Ordinary Least Square Regression

To continue our exploratory analysis, we fitted an intuitive OLS regression model on the block group's distance to the nearest hospital/fire station and their socio-economic features. Generally, we find the general trend reported by the OLS results not conforming with our expectations. It was to our surprise to find that blocks with higher median income actually fall further away from fire stations, and that blocks with a higher percentage of white population are not necessarily closer to emergency services.

Multiple reasons should account for this bizarre trend: for example, land might be more expensive in wealthier neighborhoods. Nevertheless, with a R^2 of only around 0.2, we do not think OLS is a good model for our research question.

4 Clustering Emergency Service Availability

We recognize the complicated nature of the problem of emergency service availability, as it is multidimensional and each dimension cannot be reduced: for example, living close to a fire station does not make up for the lack of hospital beds

Dep. Variable:	nearest_hospital_dist	R-squared:	0.219
Model:	OLS	Adj. R-squared:	0.218
Method:	Least Squares	F-statistic:	310.5
No. Observations:	5554	Prob (F-statistic):	5.87e-294
Df Residuals:	5548	Log-Likelihood:	-58828.
Df Model:	5	AIC:	1.177e+05
Covariance Type:	nonrobust	BIC:	1.177e+05

	coef	std err	t	P>|t|	[0.025	0.975]
const	3342.7280	703.094	4.754	0.000	1964.389	4721.067
median_age	182.9291	14.889	12.286	0.000	153.740	212.118
white_pct	112.5120	5.766	19.514	0.000	101.209	123.815
median_income	0.0038	0.004	0.986	0.324	-0.004	0.011
college_degree_pct	-337.5772	16.516	-20.440	0.000	-369.954	-305.200
unemploy_pct	-4.1332	24.741	-0.167	0.867	-52.635	44.369

Table 1: Linear Regression Between Block Endemic Socio-Economic Factors and Hospital Distances

Dep. Variable:	nearest_hospital_dist	R-squared:	0.185
Model:	OLS	Adj. R-squared:	0.184
Method:	Least Squares	F-statistic:	251.9
No. Observations:	5554	Prob (F-statistic):	3.16e-243
Df Residuals:	5548	Log-Likelihood:	-50637.
Df Model:	5	AIC:	1.013e+05
Covariance Type:	nonrobust	BIC:	1.013e+05

	coef	std err	t	P>|t|	[0.025	0.975]
const	1000.2677	160.875	6.218	0.000	684.890	1315.646
median_age	37.0378	3.407	10.872	0.000	30.359	43.716
white_pct	25.7126	1.319	19.491	0.000	23.126	28.299
median_income	0.0062	0.001	7.054	0.000	0.004	0.008
college_degree_pct	-75.1015	3.779	-19.874	0.000	-82.510	-67.693
unemploy_pct	-9.4021	5.661	-1.661	0.097	-20.500	1.696

Table 2: Linear Regression Between Block Endemic Socio-Economic Factors and Fire Station Distances

in the area. Therefore, we used an unsupervised clustering algorithm to create different categories of neighborhoods, each with a distinct array of emergency service availability features.

There are four features fed into the K-means clustering algorithm, namely

- the median age of the population of the neighborhood,
- the distance to the nearest hospital in kilometers,
- the distance to the nearest fire station in kilometers, and
- the average number of licensed hospital beds per one thousand population.

Note that we compute the last feature by dividing the hospital's number of licensed beds by the aggregate population in all neighborhoods that have it as the nearest hospital. All features are normalized to limit the effect of extreme outliers. For a K-means algorithm, the only hyper-parameter we need to tune is the number of clusters, K. We manually examined the distribution of different clusters with K ranging from 4 to 8 and selected $K = 6$. Figure 2 shows the distribution of each cluster across Virginia.

The radar diagram in Figure 3 shows the radar diagram for each cluster, with each spike as the mean of the normalized value of each feature of neighborhoods in the cluster. This gives us insights into the meaning of each cluster, or category, of emergency service availability. A brief summary is

- Cluster 0 consists of neighborhoods with an aged population but relatively poor level of hospital beds per capita;

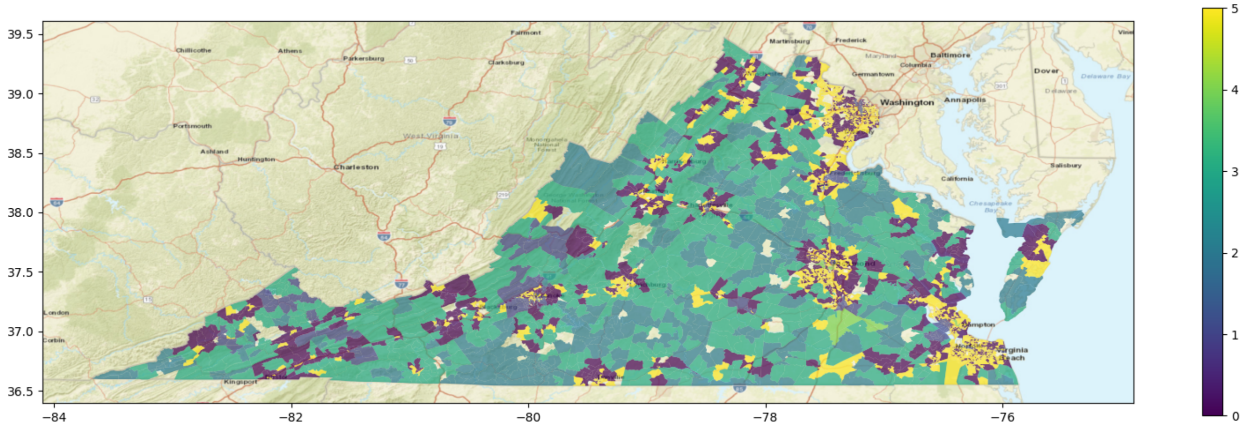


Figure 2: Color map of all clustered neighborhoods in Virginia.

- Cluster 1 consists of neighborhoods with a young population, close to a hospital, and a lot of hospital beds per capita;
- Cluster 2 consists of neighborhoods with a middle-aged population that are extremely far away from a hospital;
- Cluster 3 consists of neighborhoods with a middle-aged population that are extremely far away from a fire station;
- Cluster 4 consists of neighborhoods with extremely high number of hospital beds per capita;
- Cluster 5 consists of neighborhoods with a very young population and a moderate level of hospital beds per capita.

One can argue that the medical resources are too well-supplied in Cluster-4 and Cluster-1 neighborhoods but under-supplied in Cluster-0 neighborhoods, or that there are a lack of fire stations near Cluster-3 neighborhoods. In one way or another, this novel way we adopted constructed a more detailed and informative map compared to traditional methods of single-dimensional “score-of-emergency-service” analysis.

5 Random Forest Analysis

With the neighborhoods clustered, we have transformed the problem of predicting service availability into classifying which cluster that a neighborhood would fall in, based on an array of input features. Since we defined the neighborhoods as the Census Block Groups, whose definition updates every ten years, we would expect the definition of each cluster to be completely different when we re-fit the clustering model with the new definition, thanks to the stochastic nature of K-means algorithm. Therefore, we aim to build a predictive model so that newly defined neighborhoods can be classified using a definition consistent with our current clusters. We adopted the random forest classification model to do so.

The input into the random forest are socio-economic variables of the neighborhoods, including total population, median age, percentage of white population, median family income, unemployment rate, and the percentage of the population with a college degree or higher. All input features are numeric so we only needed to perform normalization. Since our cluster names are not ordinal, we used one-hot encoding on it. There are two hyper-parameters to tune for the random forest model, namely the number of estimators and the maximum features for each split. After a grid-search tuning with 5-fold cross-validation, we trained our final model with 150 estimators and 1 maximum feature per split to achieve an accuracy of 70.2%.

6 Conclusion and Implication

In this project, we gathered data on neighborhood fire station service availability and health care availability in Virginia. To encompass the multifaceted problem of emergency service availability as a whole, we used a K-means clustering algorithm to classify each neighborhood into one of the 6 categories based on a holistic evaluation of all of their emergency-service feature array. Then, a random forest classification model is used to ensure new neighborhoods defined in the future can be categorized using the same definition as today.

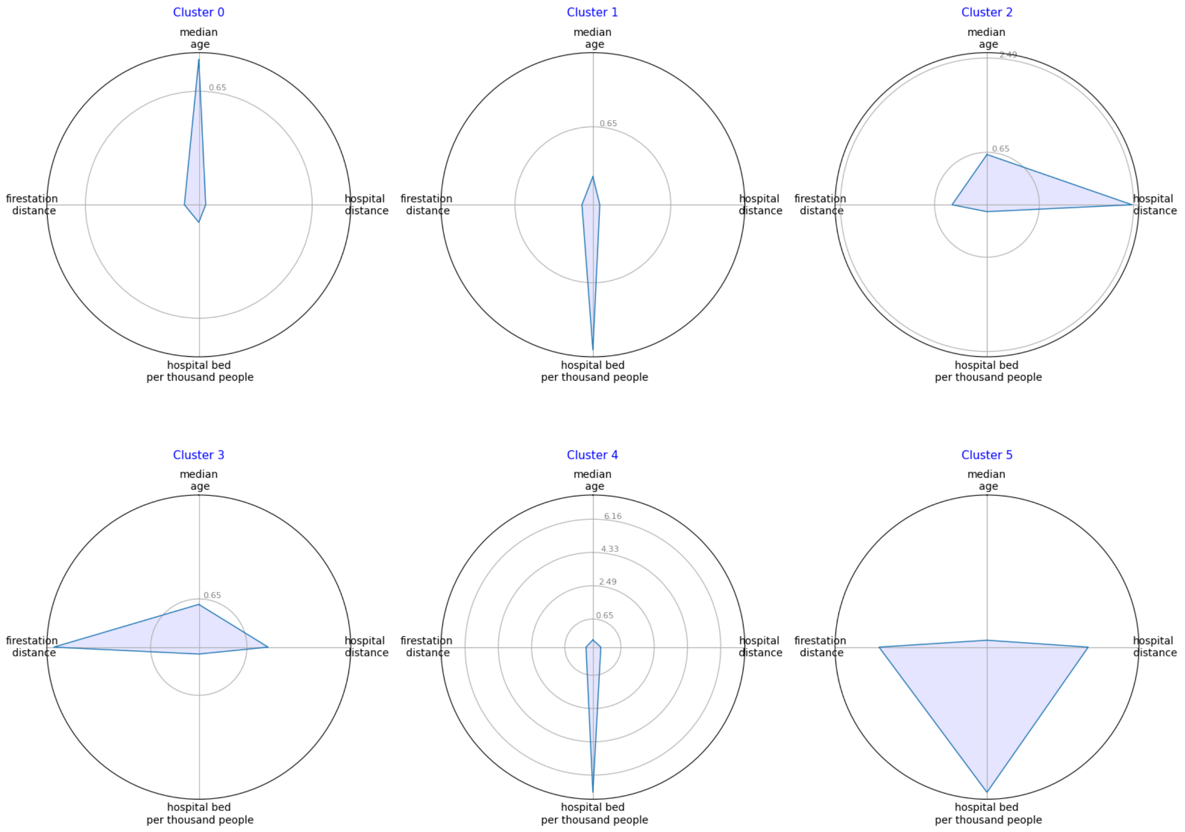


Figure 3: Radar diagram with the means of each normalized feature for clusters.

feature	median_age	total_pop	white_pct	median_income	college_degree_pct	unemploy_pct
importance	0.394264	0.115868	0.148817	0.119345	0.130212	0.091493

Table 3: Feature importance reported by the random forest classifier.

Moreover, the random forest model allows us to gauge the relative importance of each input feature, as shown in Table 3. While our input data contains variables that directly differentiate poor and wealthy neighborhoods, the percentage of the population that is white still carries considerable weight in deciding the emergency service availability - it is more important than the median income of the neighborhood. This is a worrisome signal that highlights the extent of racial discrimination embedded in our current distribution of emergency services.

7 Member Contribution

Lanyin's role encompasses examining existing literature, summarizing previous work, and developing the project's methodology. Lanyin is also tasked with organizing the data and creating the video presentation.

Steve collected datasets and completed the process of data cleaning to ensure accuracy and usability, based on the preliminary analysis. Steve also built the clustering and the random forest model, as well as visualization of the results.

References

Kiran Kc and Jonathan Corcoran. Modelling residential fire incident response times: A spatial analytic approach. *Applied Geography*, 84:64–74, July 2017. ISSN 0143-6228. doi:10.1016/j.apgeog.2017.03.004. URL https://www.sciencedirect.com/science/article/pii/S0143622816304684.

Laura A. McLay and Maria E. Mayorga. Evaluating emergency medical service performance measures. *Health Care Management Science*, 13(2):124–136, June 2010. ISSN 1572-9389. doi:10.1007/s10729-009-9115-x. URL https://doi.org/10.1007/s10729-009-9115-x.