Spring
2021

Geospatial Analytics

LEVEL STUDIO CONSULTING PROJECT

NASTASE, JAMESON RILEY

Table of Contents

Introduction

This research project will seek to explore the usefulness of data in architecture. There are several potential applications: customer surveying, sentiment analysis, social media traffic analysis, but this document will focus specifically on the use of *geospatial analytics*, data science methodologies that manipulate location data in order to drive business insights.

Geospatial Analytics: Architecture Usecases

Geospatial analytics utilizes data from a variety of sources – build location data, GPS tracking, social media usage, geographic features (mountains, rivers, etc.) and manipulates these sources to provide insight into business decisions by two primary means:

1. Data Visualization – Plotting the distribution and interactions of the data to help understand what can be grasped from it.
2. Statistical Analysis / Prediction – How does the proximity of a cell tower impact cell reception? Does a park within *x* miles of a neighborhood make families with small children more likely to move there? Are there certain locations that receive more foot traffic, and are these locations ideal for gift shops? These are all questions that geospatial analytics is well equipped to handle.

This research will attempt to use geospatial-driven statistical analysis to provide value for an architectural business by evaluating the over-representation or under-representation of particular building types across the state of Arkansas. This could help to target regions that have a market inefficiency for a building type (such as a wealthy zip code lacking a private school,) and could be used to preemptively identify demand and pursue a value-providing project.

Pre-Requisites: Data and Tools

This is the frightening question that scares many small businesses away from analytics before they get started. In the modern world of “Big Data” every analytics project seems massive and unreasonable. Data storage and the engineers needed to manage such a storage are expensive! Analytics projects, however, are often more feasible than they might seem, for a couple reasons.

First, not every data project must be a “Big Data” project. There is an abundance of public data available online, and when used wisely this can enable a remarkably helpful platform for evaluating a business domain, even if a lack of subscription/streaming makes it difficult to build the high-impact predictive products larger companies are coming to be known for.

Second, not every data project must be a data “Big Project.” The best analysis starts small, not trying to fully simulate the world or predict everything, but by helping to clarify what the landscape looks like in terms of resources, competition, and value. As with everything else, analytics must walk before it runs, and walking just might be significantly more valuable than crawling!

For this project, there are only two requirements, accessible to anyone with a laptop:

1. Arkansas GIS Data – This data is free and publicly available at *https://gis.arkansas.gov/*. It contains a massive collection of files, enabling the creation of Arkansas maps at the state, county, zip code, or even school/fire district levels. It is also subscribable, meaning a feed can easily be set up to pull data from the source when updated.
2. Any Python coding platform – The data products in this project have been created using Jupyter Notebook, a tool that creates an environment for coding in Python via Anaconda.

This can be downloaded for free as well at *https://www.anaconda.com/* and can be up and running in as little as 30 minutes.

Research Methodology

After collecting data from GIS, the next step is to formulate a process for identifying value. The primary components used in this work are *scarcity* and *population density*, and these metrics are used to determine where over or under-representation might be occurring, specifically relating to Arkansas Emergency Medical Centers (EMC.)

There are a few assumptions made that will be investigated in more detail further on, but central to this analysis is the assumption that population density influences building density. I.e. the presence of more people usually indicates more schools, more hospitals, more subway stations, etc. To quantify how this relates to medical emergency centers, this research will follow a 4-step process:

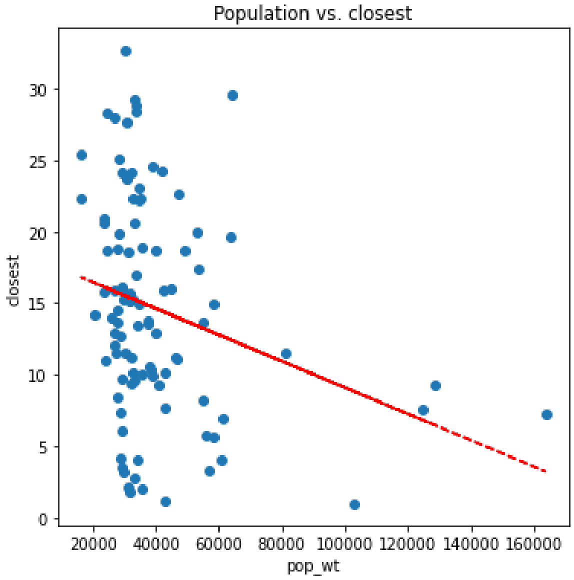
1. Randomly generate 100 points in Arkansas, and create a weighted average (by distance) of nearby counties' population density to estimate population density at that point.
2. Calculate two metrics at each of those points:
    a. The distance to the closest emergency medical center (designated by "closest.")
    b. The number of emergency medical centers within a 25 mile radius of the point (designated by "within_25.")
3. Use Linear Regression to predict expected values for these metrics by population density. This will determine a baseline for how many EMC should be in the area and how far one would have to drive to get to one.

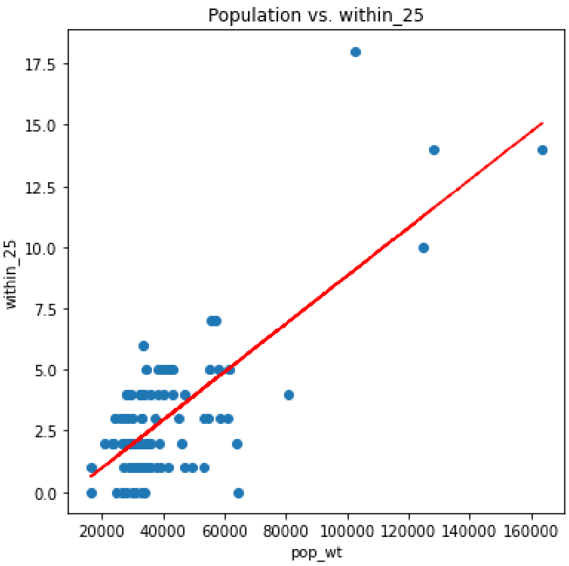
4. Compare actual metrics with the expected metrics to determine if this point has an overage or a shortage of EMC. If a shortage, the public might be more willing to pursue building a structure here.

After this is completed, the model should be able to represent each of the 100 points, with latitude/longitude coordinates and a score that demonstrates how well the point is served by the proximity of EMC.

Validating Assumptions

The first step in this research is to check the assumptions of the relationship between population density and EMC structure density. After performing the random point generation and weighted averages on the data to estimate population, the following plots were created to visualize the relationships. The results are presented in the plots below.



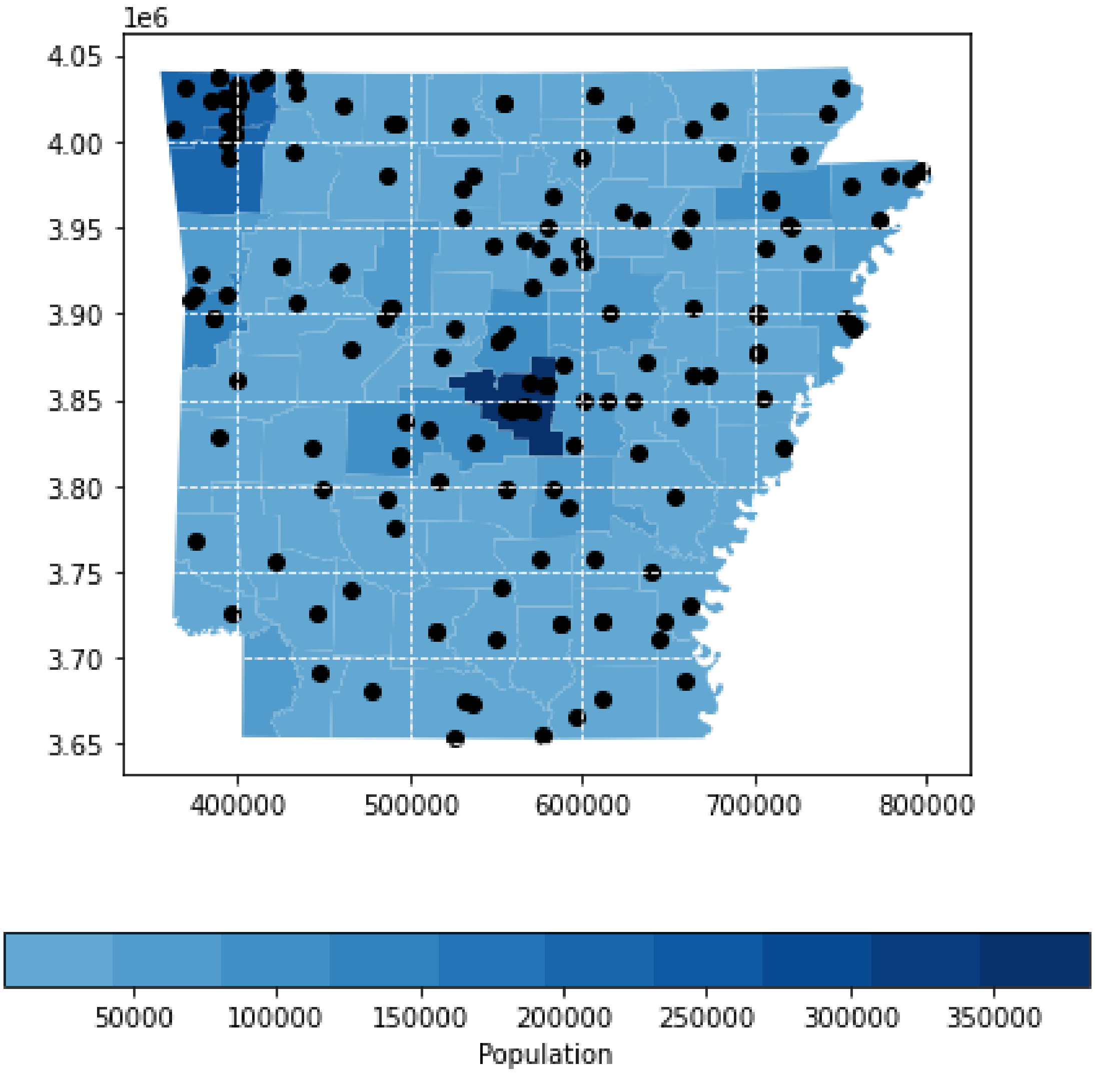


From these plots, it is clear that there is a significant correlation between population density and EMC proximity/scarcity, so we will use population as a predictive variable to identify building scarcity going forward.

Data Products

Data Product #1: Basic Visualization

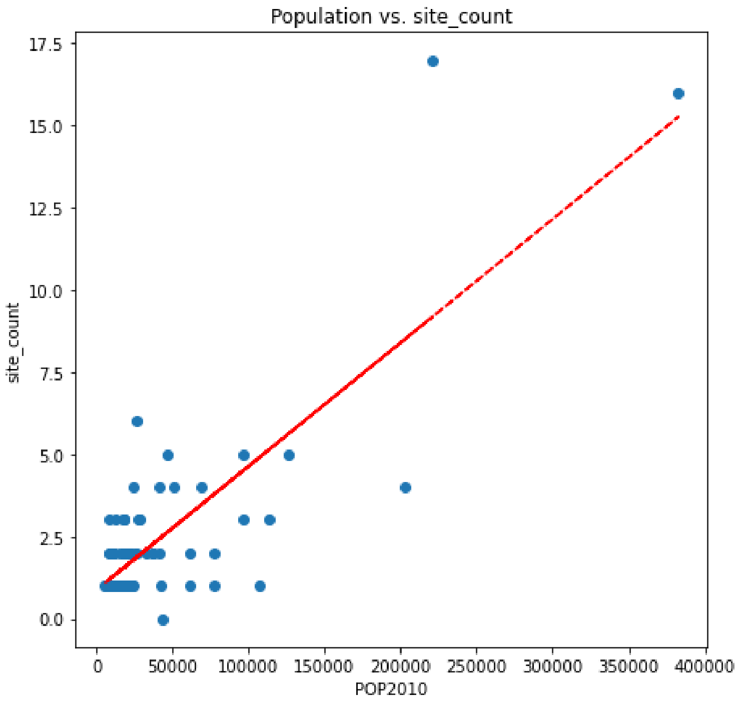
The first step here is to start slow and gain a high-level understanding of the data. Below is an overlay of EMC (represented by black dots) on a plot of Arkansas. The counties are shaded in blue, where higher populations are represented by darker colors, as reflected on the legend.



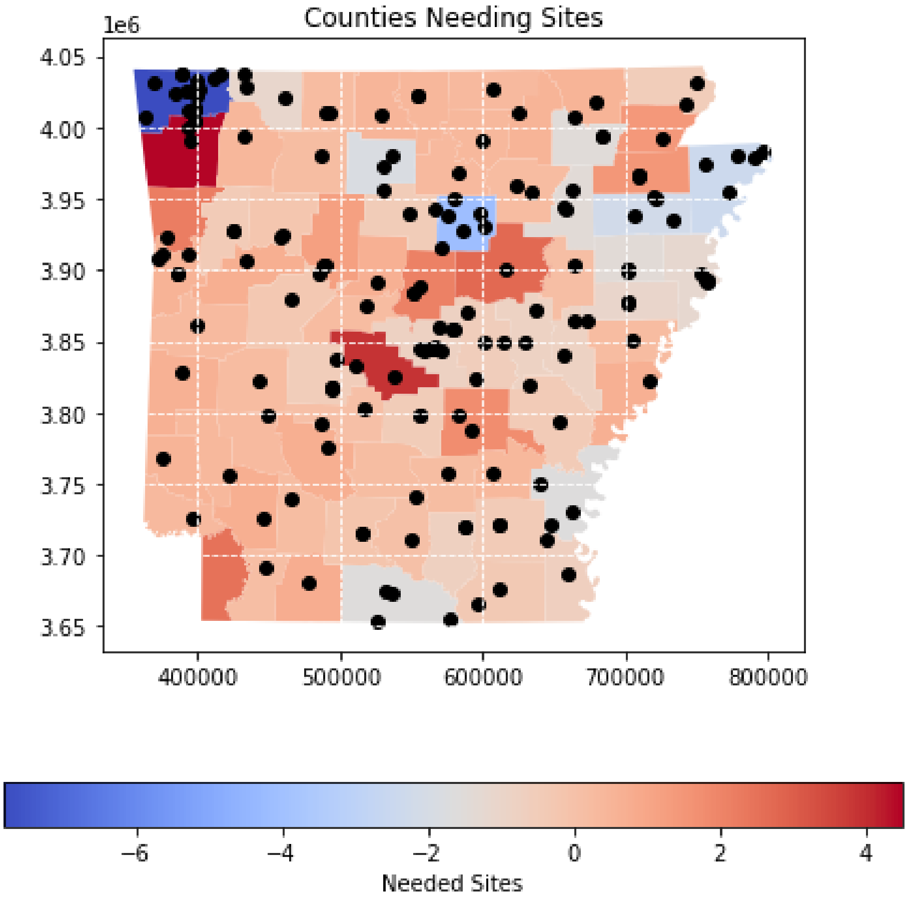
Notice Miller County in the lower left hand corner. Though it sports an average population, there is a total lack of EMC in the county. We'll make a note of this and see if the model identifies the county as a scarce or under-represented region.

Data Product #2: Heat Map Visualization

The second data product generated from the GIS data is a heat map. This model uses the population of each county to predict how many EMC should be within the county, as portrayed by the scatter plot below.



Using this regression to form an expected number of EMC per county, "sites_needed" will be calculated as (expected_EMC – actual_EMC) and plotted on the heat map.



This visual highlights in dark red the counties that are under-represented by EMC, and thus may be in need of an additional structure or two. Miller County, as expected, is one of a few counties that features prominently.

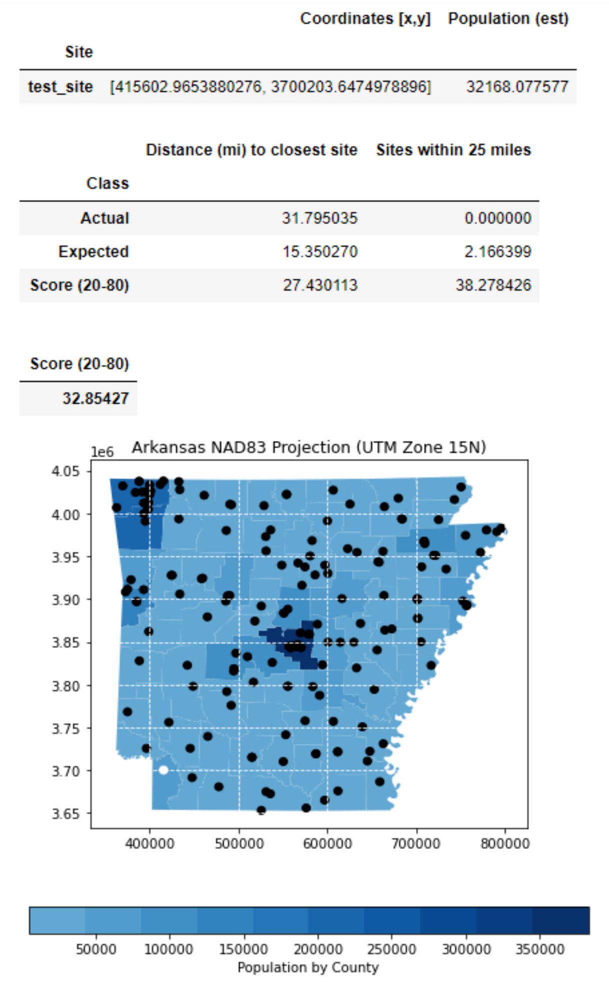
In general, this plot, as heat maps often do, may serve as a helpful launch point in future analyses, having gained some insight into where to best start digging deeper!

Data Product #3: Hypothetical Build Score

This may be my favorite tool in the entire set. What if a particular building type comes up, and builders are interested in choosing the best of several options? Given, there are many factors that influence this, but let’s give geospatial analysis a piece of that pie, using the scarcity metrics we’ve already defined.

In the Hypothetical Build Score, the program asks for an input from the user: either (*x*,*y*) NAD83 coordinates, or the easier option, a zip code. The model then performs aggregation to estimate population at that exact point, compares the location data with locations of all EMC in the state, and generates a score that represents how well the region is covered by EMC already. For example, let’s look at this comparison of results from Miller and Mississippi Counties.

Texarkana, AR, Zip: 71854

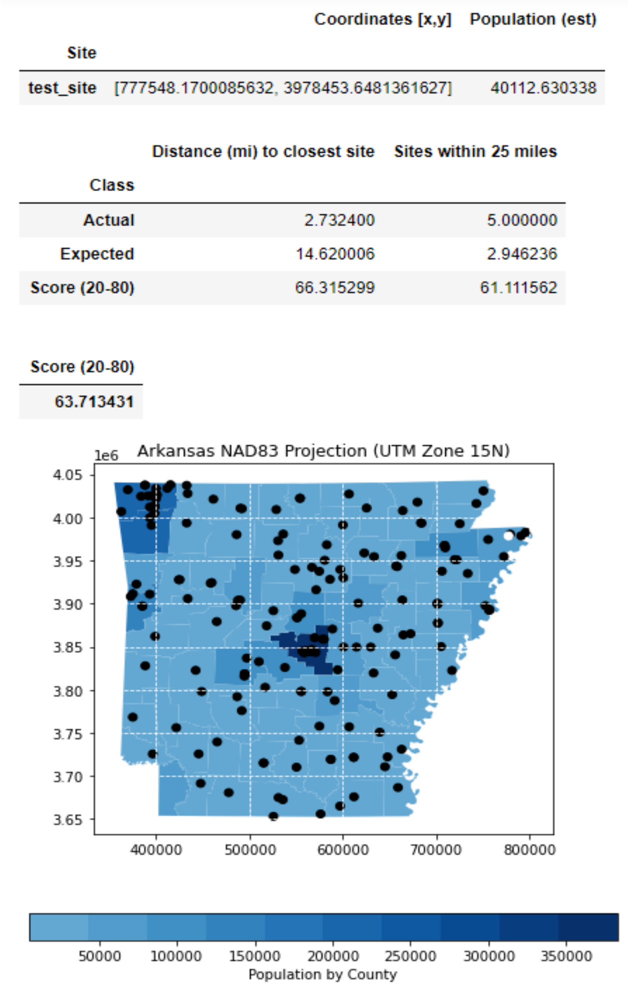


Site	Coordinates [x,y]	Population (est)
test_site	[415602.9653880276, 3700203.6474978896]	32168.077577

Class	Distance (mi) to closest site	Sites within 25 miles
Actual	31.795035	0.000000
Expected	15.350270	2.166399
Score (20-80)	27.430113	38.278426

Score (20-80)
32.85427

Blytheville, AR, Zip: 7231



Site	Coordinates [x,y]	Population (est)
test_site	[777548.1700085632, 3978453.6481361627]	40112.630338

Class	Distance (mi) to closest site	Sites within 25 miles
Actual	2.732400	5.000000
Expected	14.620006	2.946236
Score (20-80)	66.315299	61.111562

Score (20-80)
63.713431

The tables in this result lay out the metrics we have been considering: estimated population, actual/expected distance to closest EMC, and actual/expected number of EMC within 25 miles. The new addition is the Score (20-80) value appearing in the second and third tables.

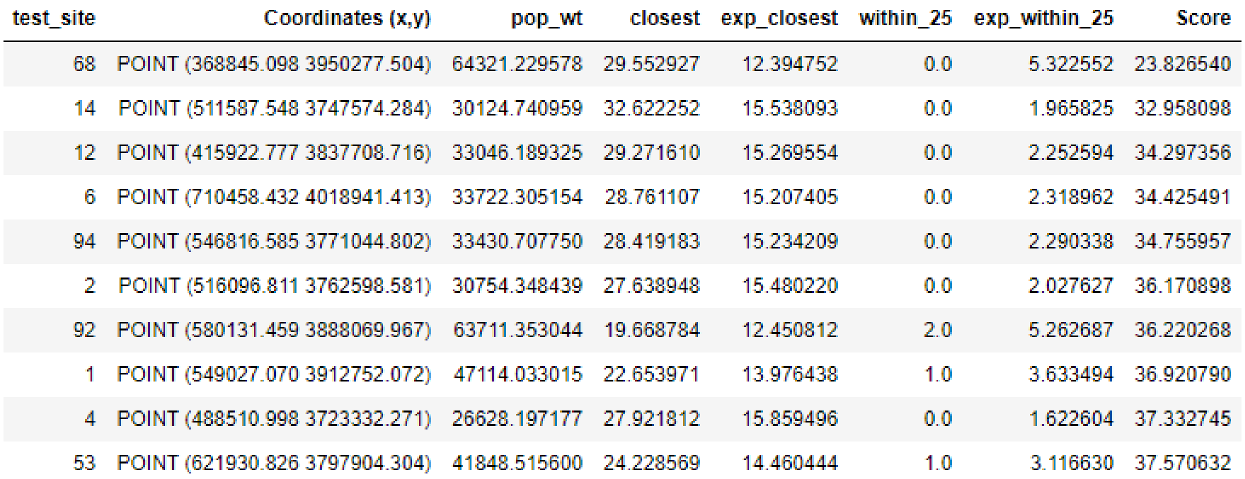
The value in the second set of tables (Class, Distance..., Sites…) calculates a coverage score for each of the calculated metrics: closest and within_25, where 20 represents terrible coverage, EMC that vastly under-represents the region. A score of 80 represents well over-represented regions, and 50 is a median expectation value. The Score (20-80) value in the third table is a unified aggregate overall score that takes into account both.

As we would expect from the heat maps, Texarkana upholds the expectation from Miller County and is significantly under-represented, coming in with a score of 32.85. Blytheville on the other hand receives an overall score of 63.71, and is likely not a source of demand for EMC at this point.

The other helpful feature in this tool is the emergence of a new data point on the plot, the white one! This is the point generated by the (*x*,*y*) or zip input, and can be used to gain a better visual understanding of where the hypothetical build site falls into the overall map and distribution of EMC.

Data Product #4: Location Prioritization

The final data product presented here is the ranking of locations with the lowest coverage scores in the randomly generated set. An important: this is not exhaustive, but iterates through the 100 points to find an upper 10th percentile of optimal build spots. This can, however, be expanded to significantly more, as the code has been standardized to handle any number of points. The results for this particular set, with coordinates, are as follows:



test_site	Coordinates (x,y)	pop_wt	closest	exp_closest	within_25	exp_within_25	Score
68	POINT (368845.098 3950277.504)	64321.229578	29.552927	12.394752	0.0	5.322552	23.826540
14	POINT (511587.548 3747574.284)	30124.740959	32.622252	15.538093	0.0	1.965825	32.958098
12	POINT (415922.777 3837708.716)	33046.189325	29.271610	15.269554	0.0	2.252594	34.297356
6	POINT (710458.432 4018941.413)	33722.305154	28.761107	15.207405	0.0	2.318962	34.425491
94	POINT (546816.585 3771044.802)	33430.707750	28.419183	15.234209	0.0	2.290338	34.755957
2	POINT (516096.811 3762598.581)	30754.348439	27.638948	15.480220	0.0	2.027627	36.170898
92	POINT (580131.459 3888069.967)	63711.353044	19.668784	12.450812	2.0	5.262687	36.220268
1	POINT (549027.070 3912752.072)	47114.033015	22.653971	13.976438	1.0	3.633494	36.920790
4	POINT (488510.998 3723332.271)	26628.197177	27.921812	15.859496	0.0	1.622604	37.332745
53	POINT (621930.826 3797904.304)	41848.515600	24.228569	14.460444	1.0	3.116630	37.570632

This result indicates a significant gap of EMC in southwestern Washington County at hypothetical site #68. The population would indicate about five EMC within 25 miles, there are none, and the closest EMC is 17 miles farther from the site then would be expected.

Granted, there is some nuance to this as well. Working with population density at the county level has its drawbacks, as this may be a lightly-populated region being pulled up by the rest of the county. Still, the tool serves as a helpful starting point in quantifying a demand/supply relationship of structures from across the state.

Short-Term and Long-Term Strategy

One benefit of the tools and methodologies offered here is that they are fairly standard, and the use of them need not change much. There is however, potential long-term value in the acquisition of datasets that better score individual regions. For instance, the only input required to do away with the potential downside of the Washington County result is more granular data, perhaps at the zip code level, which is easily available from other sources. Other factors/variables could be included as well. Crime rates might have a relationship with the need of EMC, and demographic information could also be useful.

Beyond additional features, the GIS datasets extend far beyond EMS. Public/private schools, community colleges, hospitals, fire stations, libraries, museums, trailer parks, town halls, and courthouses could be analyzed quickly and efficiently using much of the same code that developed this analysis, which also unlocks the use of more additional variables. For instance, perhaps the proximity to a cell tower is something that could help to be used in tandem with scarcity to determine optimal building locations.

Summary

In summary, I do hope to have gotten a couple points across that I feel could be helpful to any business. Analytics does not need to look like something flashing across the screen in a sci-fi movie. It can start with free, publicly available sources that are simple to use, and it can produce insight quickly and efficiently without requiring the wide resources of a massive company.

I do hope you find some ideas here helpful, and for your own use I will attach much of what was used to build out these analyses, data and code included in a zip file. If there is something here that you would like to implement, I would happily take some time to help you get Anaconda and Jupyter Notebook set up as well!

Thank you for the opportunity to work with you this semester, and I wish you the best in continuing the rise of Level Studio, hopefully with some new and shiny analytical tools in tow!

Appendices / Links

All data, code, and documentation available on Google Drive

Arkansas GIS Office – *https://gis.arkansas.gov/*

Anaconda / Jupyter Notebook Download – *https://www.anaconda.com/*