Identifying Suitable Locations for New Supermarket Investments in the Regional Unit of Thessaloniki, Greece.

Joseph Doundoulakis

January 9th, 2021

1 Introduction

1.1 Background

The regional unit of Thessaloniki is an administrative unit of Northern Greece and subdivision of the region of Central Macedonia. It has a total extent of 3,683 km2 and a population of approximately 1.1 million people. The unit is of general strategic interest due to its proximity to central Europe, its connectivity population size.

Figure 1: Regional unit of Thessaloniki, Greece

The local supermarket brand **Masoutis** dominates the market within the regional unit and the northern area of Greece in general. In fact, it is one of the few (if any other) Greek brands that refused to sell to foreign competition during the financial crisis.

1.2 Business problem

A competitor brand, **AB Vasilopoulos**, which is currently owned by Ahold Delhaize and already present in the area, wants to increase its local investment and get a better hold of the market. For that reason, the company needs to identify potential strategically important investment locations, based on demand and competition, which would help improve its position within the area.

Specifically, they asked for the following criteria to be met:

- The location has to be within town boundaries. They explain that the tolerated distance they are willing to have from a town's center according to its population, can be given by the formula:

$$d = 6 * \sqrt{population} + 700 \qquad [1]$$

in meters.

- The location has to be within 300 meters of a residential street.

2 Data description

- First, we will need the area boundaries and ancillary data such as the location of towns, their corresponding population and the road network of the area. Town locations and the road network will be used to limit the study area and the possible new locations. The area boundaries and town locations are provided by the Hellenic Statistical Authority. The road network will be extracted from OpenStreetMap using **OverPy** to call their API.

- Additionally, we will need the national census data to populate each town point with the corresponding population for the last census year (2011). Population information for each town will help us create an estimation for demand. The census data are also provided by the Hellenic Statistical Authority.

- For the solution of the given scenario, it is of utmost importance to get the locations and brand names of supermarkets and grocery stores. The **Foursquare API** will be used for the creation of this dataset.

3 Methodology

3.1 Preprocessing

3.1.1 Ancillary data acquisition and preprocessing

The area boundaries and town locations are downloaded directly from the Hellenic Statistical Authority and clipped to boundaries. The town points have also been populated with the census data. The town points were then transformed to polygons using the buffer algorithm and the population data according to formula (1).

For the downloading of the road network however, I used Shapely and OverPy to request and load the area we need directly from the OpenStreetMap Overpass API.

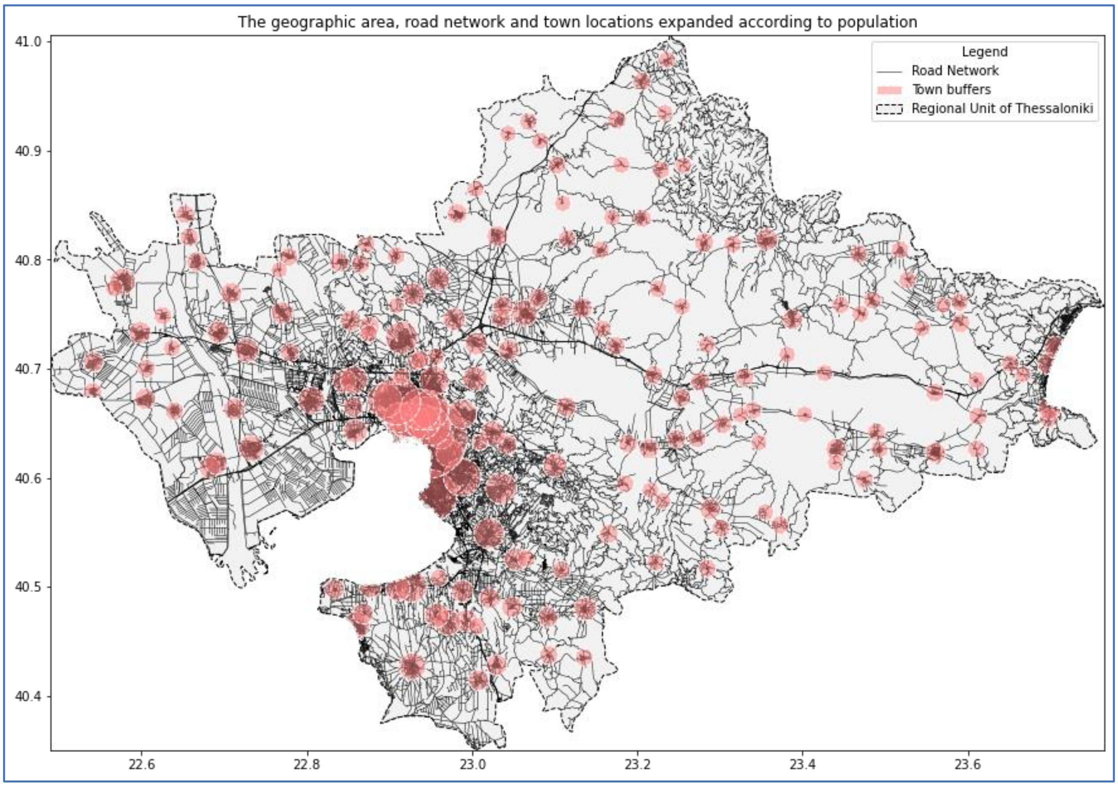

Figure 2: Visualization of the ancillary data

3.1.2 Foursquare venues collection

In order to get the venues for the entire regional unit of Thessaloniki, we will have to think of a way that is both efficient and as thorough as possible. For those unfamiliar with the API, the most convenient way to request venues is by posting location coordinates. However, a single call, even at a small neighborhood, is not guaranteed to return all the venues due to traffic regulations, at least for the free tier. For that reason, we will have to make multiple calls in different locations of close proximity, in a scan-like procedure.

A good approach would be to generate a search-grid of points, say 500 meters apart from each other, within the area boundaries. Looking at the map above, though, we can see that this approach would make a large number of additional calls in empty uninhabited space. Something like that would mean a waste of time and resources, since the API only permits 5000 calls per hour for the free tier.

Instead, keeping in mind that any sort of venue of interest would require road access, we can use the road network as a guide for the generation of call locations. The locations below have been generated in 500-meter intervals along the entire road network:

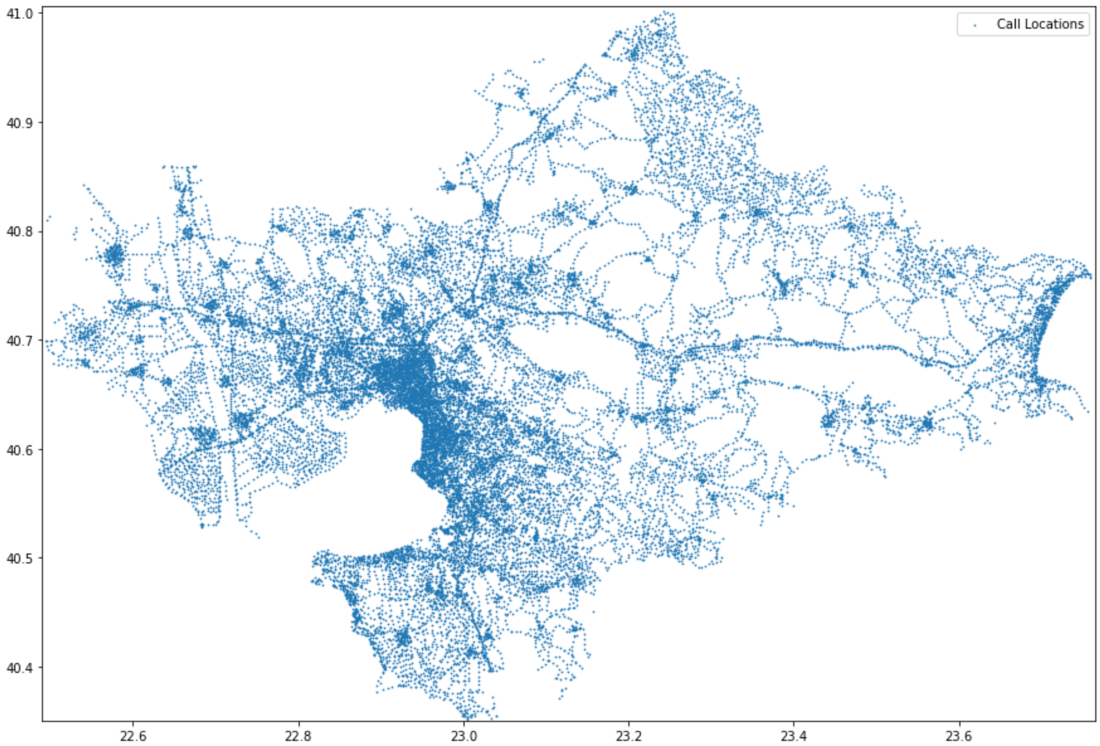

Figure 3: Call locations generated using the road network

The entire process was finished in approximately 6 hours. Below, the venues that were extracted for the entire regional unit of Thessaloniki.

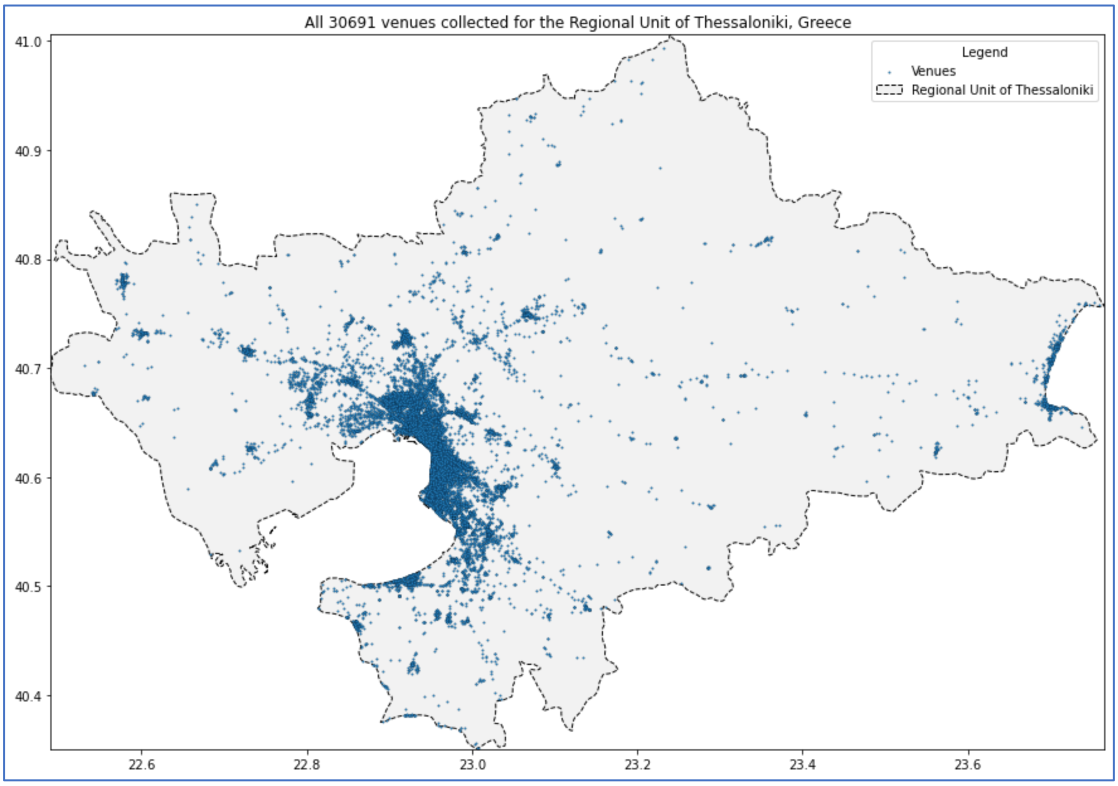

Figure 4: Map of extracted Foursquare venues for the regional unit of Thessaloniki

From these venues, the grocery stores and supermarkets were isolated, while the main brands in the area were identified and cleaned. Individual stores and lesser-known local brands were all grouped in a category named "Other". It can be seen at the store count, on the left, Masoutis has 3 times the number of stores of AB Vasilopoulos in the area.

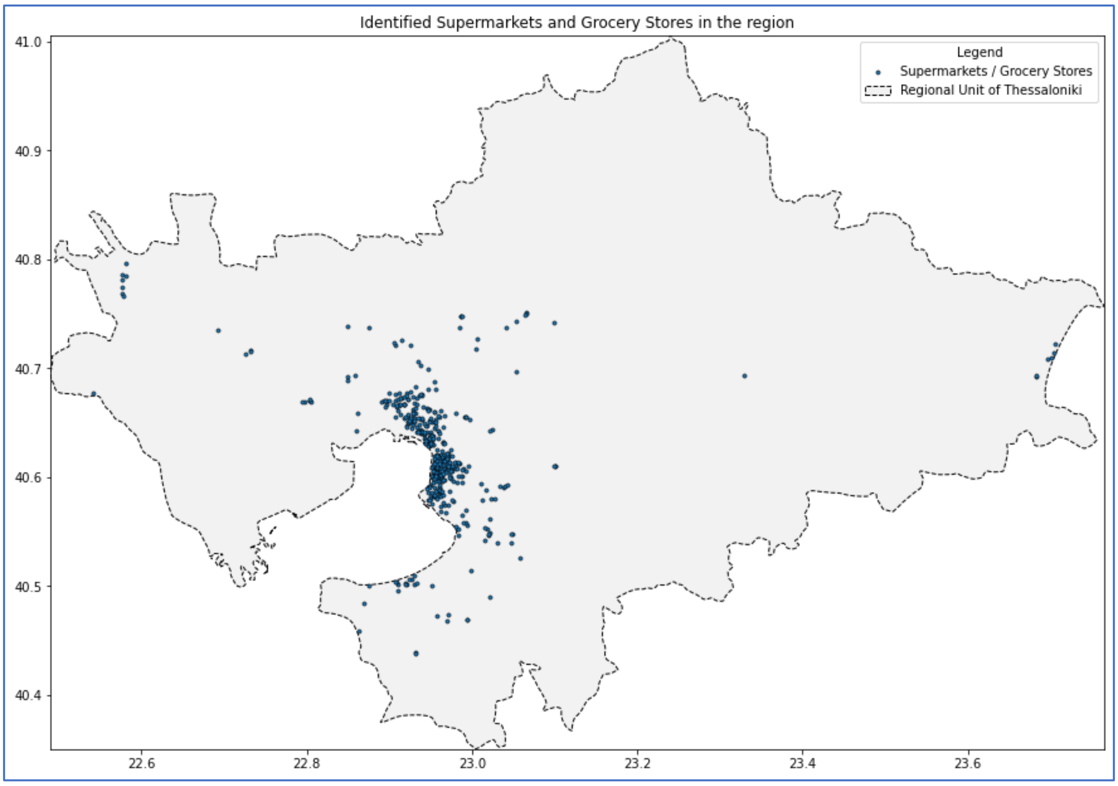

Figure 5: Supermarkets and grocery stores in the area

```
Other              124
Μασούτης           123
AB Βασιλόπουλος     43
LIDL                30
Αρβανιτίδης         27
Σκλαβενίτης         22
Γαλαξίας            16
Discount Markt      15
Αφροδίτη            13
Bazaar              12
Ελληνικά Market     12
Άριστα              11
MyMarket             6
Name: name, dtype: int64
```

Figure 6: Store count per brand

3.1.3 Residential area and supermarket dataset compilation

Furthermore, all town polygons were unified, and the populations merged for intersecting geometries. All streets classified as "residential" in the road network were isolated and transformed into polygons using the buffer algorithm with a distance of 300 meters. The resulting polygons indicate all the eligible areas for the location of a new AB Vasilopoulos store.

Every existing supermarket was attributed to its corresponding town cluster and residential polygon, initially by the rule of intersection. For the supermarkets that were outside the town/polygon limits, they were assigned to their nearest neighbors. The reason for establishing which supermarket belongs to each area is that the number of stores will be used to quantify competition.

Result:

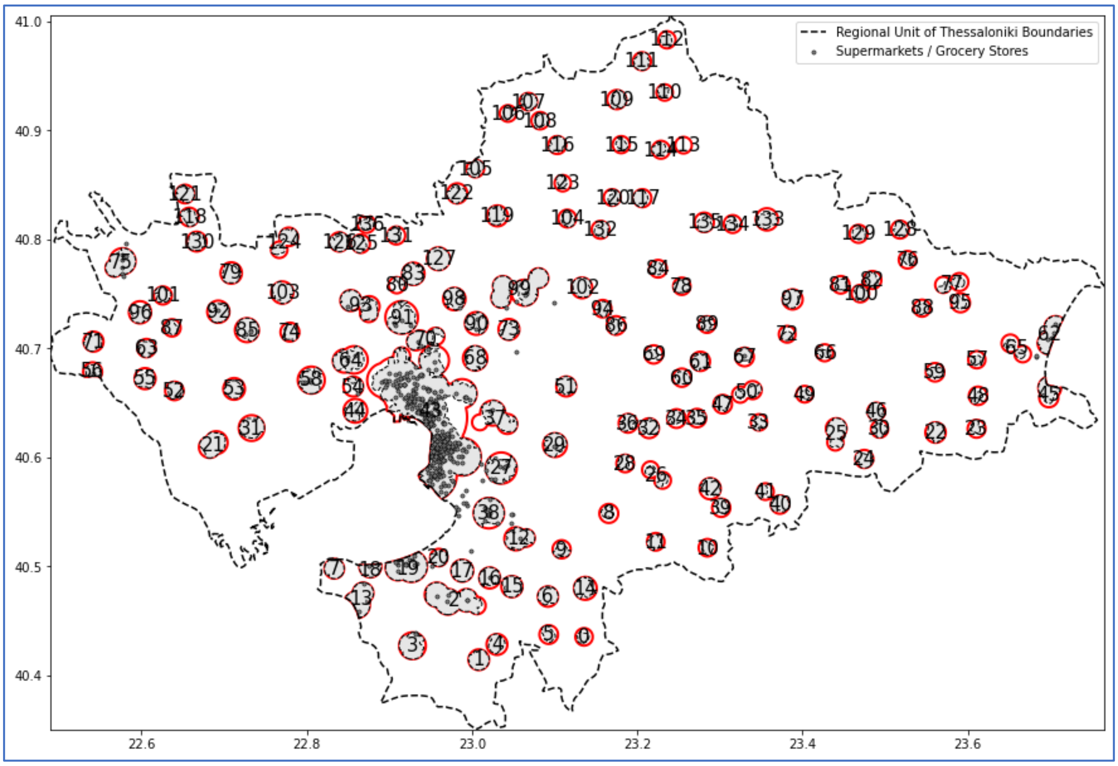

Figure 7: The final sub-areas for analysis

3.2 Current state assessment

3.2.1 Defining the assessment criteria

For the assessment of the suitability of each supermarket location we will need some sort of numerical measurements, that is easily visualized on our map. That measurement should be a result of demand, as much as competition.

For example, we know that a favored location would serve as many people as possible. Hence, a higher population should give increased indicator values. On the other hand, we would like to have as little competition as possible. Therefore, an increased number of stores within that polygon should return lower indicator values. Expressed mathematically:

$$a = \frac{population}{number\ of\ stores} \qquad [2]$$

However, the above indicator would not give any further geospatial context from within the town boundaries. In order to create a sub-polygon efficient measurement, we could leverage the distance each store has to competition, as well as the distance it has from the center of the city or town.

For example, a location relatively further from competing stores could exploit previously unattended areas and attract people wanting to travel shorter distances for groceries, while the opposite could indicate increasing competition.

Additionally, since we do not have information regarding population distribution within the towns, we could use the center of mass of existing supermarket stores for an estimation. Meaning, the existing supermarket stores are bound to have been risen near the highest demand and for that reason, being closer to the center mass should be considered preferable. An index to satisfy the above requirements could be the following:

$$b = \frac{average\ distance\ to\ competition}{distance\ to\ center\ mass} \qquad [3]$$

which would give higher values for better locations and lower values for worse. In particular, it maximizes for locations that are close to the center and far from competition. Therefore, from (2), (3), and since both give better locations at high values, multiplying a and b should give a more representative result. Additionally, b is smoothed and mapped to [0.5, 1] in order to produce a more normalized result, by being passed to a sigmoid function. Since we are using the town population to demonstrate demand, values of b greater than 1 could interfere with our scale.

$$suitability\ index = a * \frac{1}{1 +\ e^{-b}} \qquad [4]$$

3.2.2 Assessment results

After computing all the required parameters for the equation (4), the suitability indicator is calculated for each supermarket store. The average value for all brands is calculated as well. AB Vasilopoulos appears to achieve an average value of 1854. Eventually, we will be looking for locations that improve this number.

```
Mean value:  1790.23
```

name	Mean Suitability Indicator Value per Brand
Bazaar	1952.709575
Μασούτης	1902.155807
ΑΒ Βασιλόπουλος	1854.198195
Σκλαβενίτης	1830.626420
Άριστα	1828.536546
Discount Markt	1783.628285
Αρβανιτίδης	1752.108168
Ελληνικά Market	1747.352974
Αφροδίτη	1745.061086
Other	1742.455522
Γαλαξίας	1604.103541
LIDL	1565.625746
MyMarket	1472.660356

Figure 8: Average suitability value per brand

Below you can see the distribution of values on the map.

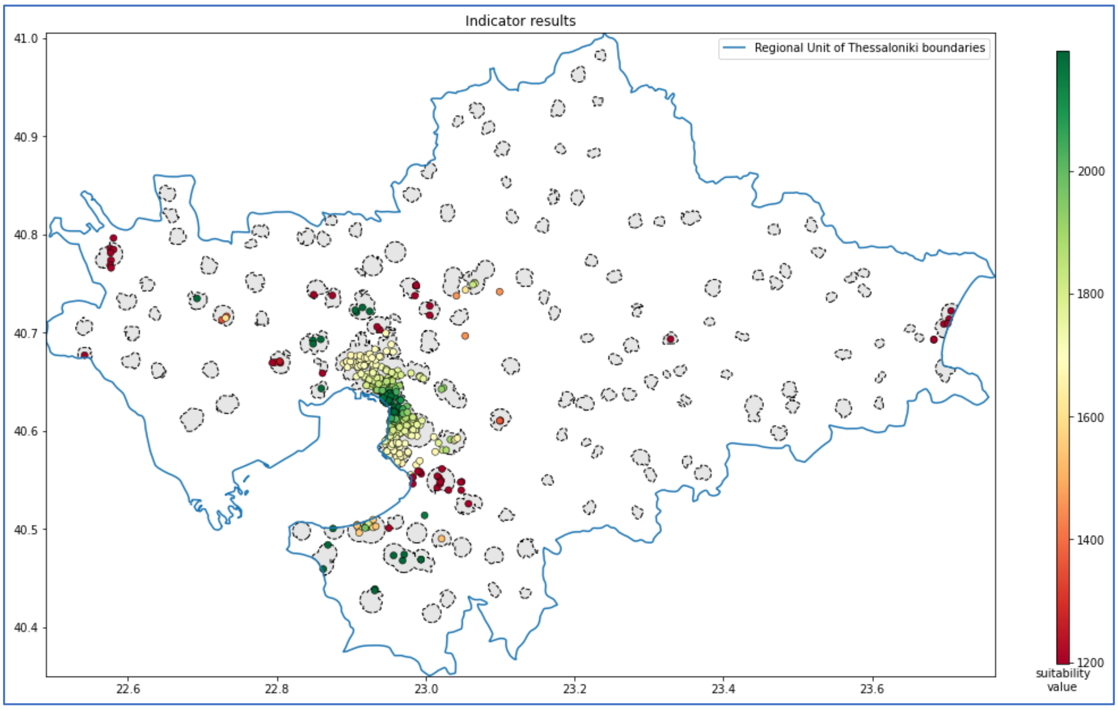

Figure 9: Suitability indicator results

Examining the relationship of the chosen parameters for the calculation of our indicator, we can observe that the result behaves roughly as expected. However, there is no significant variation for locations in similar distances from the center:

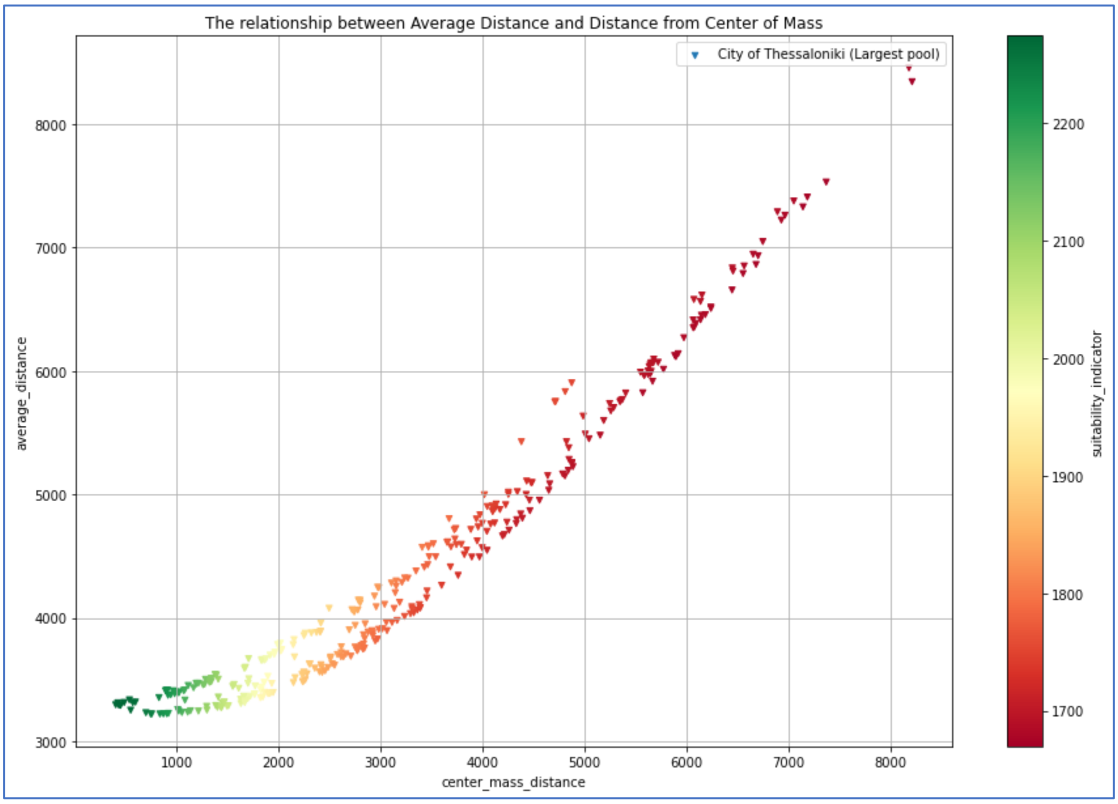

Figure 10: Average distance and center mass distance relationship to result

Perhaps, the average distance used, could be restricted to the average distance of the 5 or 10 nearest stores. Providing that way an increased sensitivity to our result.

3.3 Grid search

Now, for the effective search of the area for the most suitable locations, we will need to create a dense point grid, or alternatively, rasterize our polygons with all of their information, including locations, in separate layers, and perform array calculations. However, since our indicator is not highly sensitive to location, a simple 300-meter point-grid should suffice, saving as from technical trouble.

3.3.1 Creating the grid

We can create the grid using **NumPy's** linspace and meshgrid methods, to the vertical and horizontal extent of our map. Afterwards, the rectangular point grid can be easily clipped to our desired areas, the defined residential polygons.

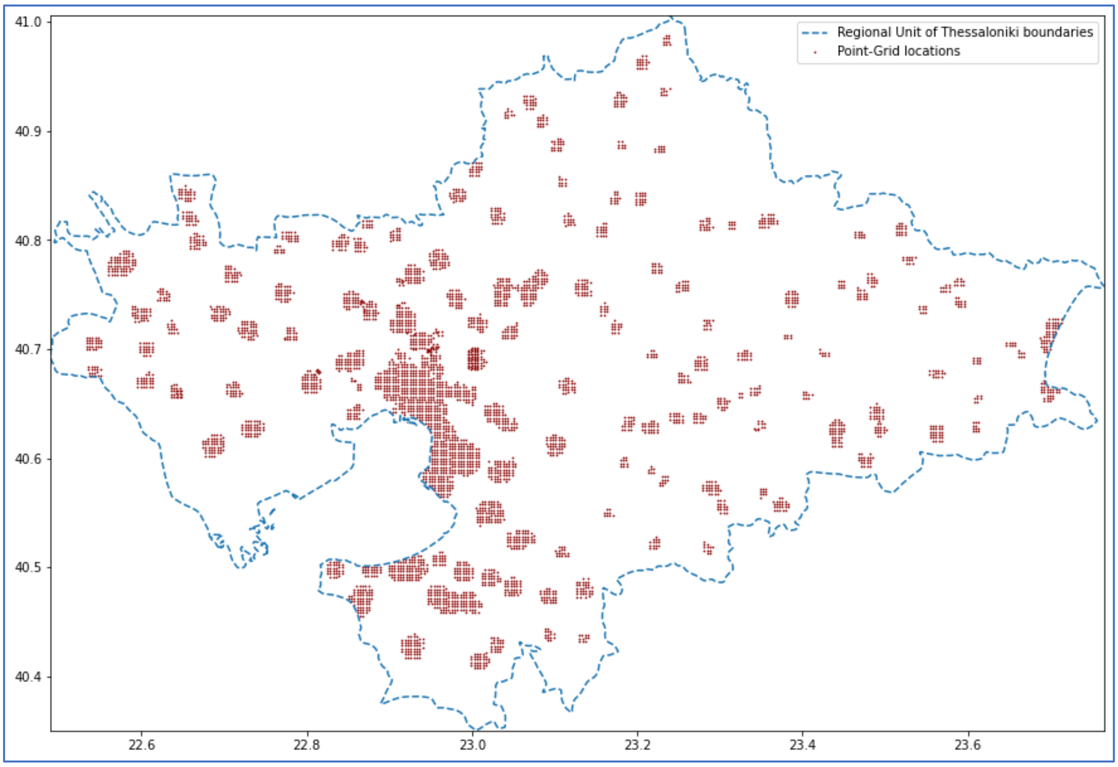

Figure 11: The produced point-grid

3.3.2 Evaluating the grid locations

After the grid's creation, all necessary attributes for the computation of the suitability indicator will need to be passed to each point. That includes, the population it will serve, the number of stores it will be operating against, as well as the average distance from them and the distance from their center mass. For polygons without competition, we will use the centroid of the town cluster polygon, in an attempt to replicate it.

Then we can calculate the suitability indicator for each point, assign the average value for AB Vasilopoulos as vmin in the matplotlib colormap and produce the following map:

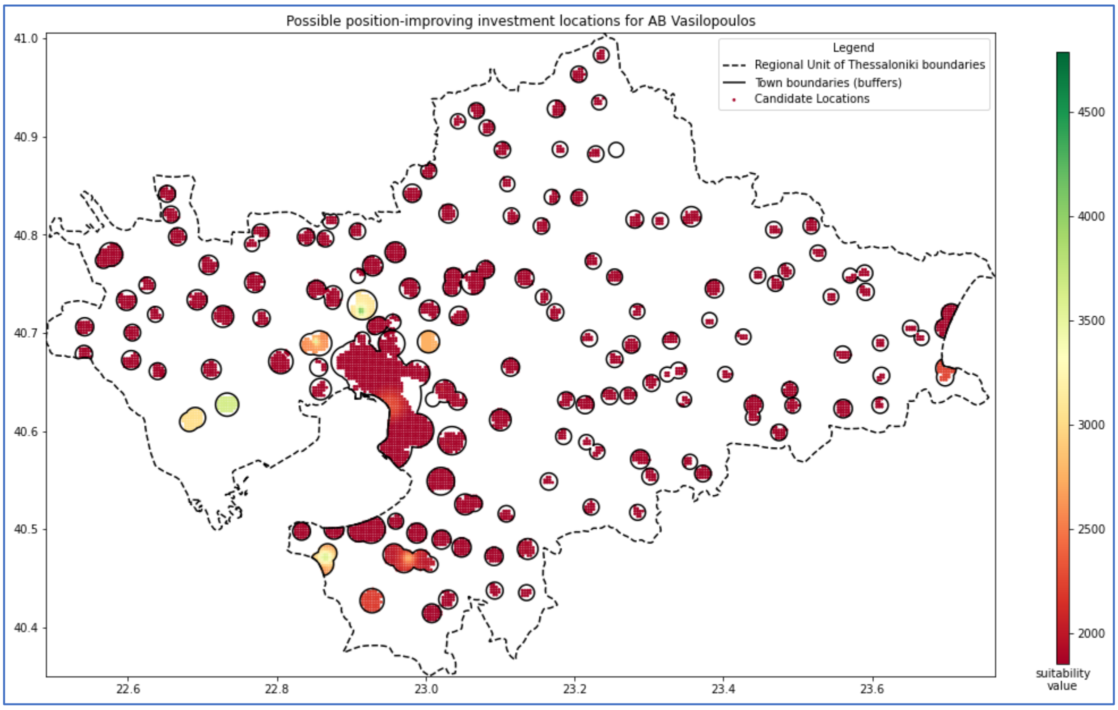

Figure 12: Evaluated point-grid

Furthermore, we can filter out all the values less than the desired ones. Lower than the value of **1854**:

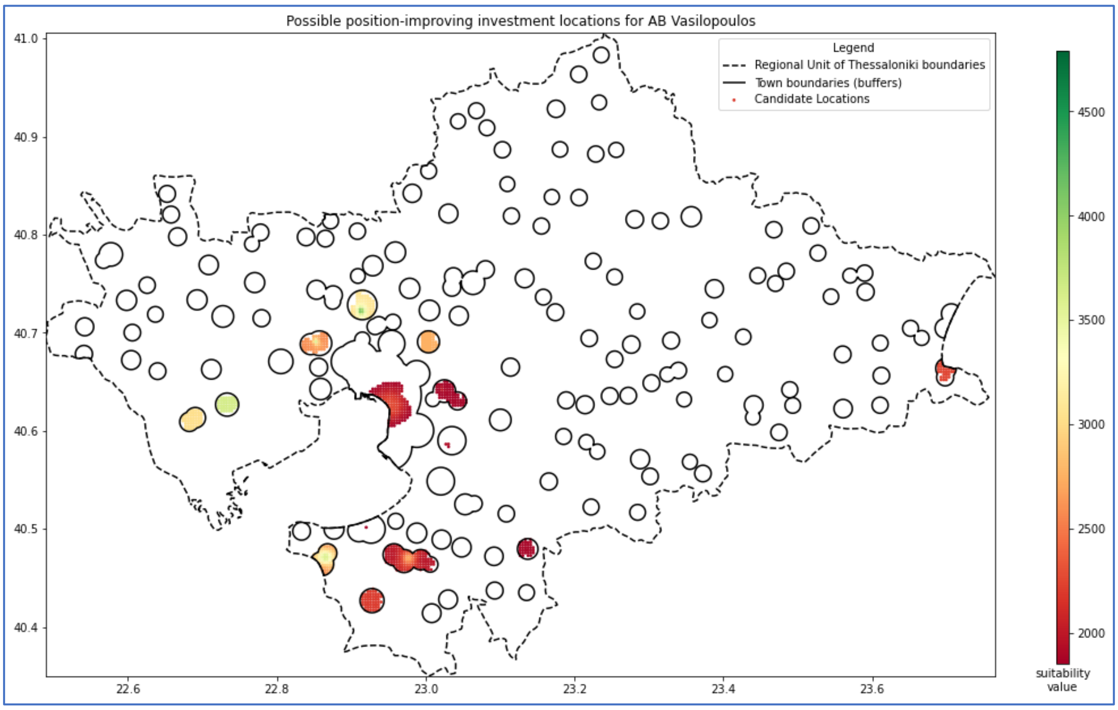

Figure 13: Candidate locations

These locations can now be used for more elaborate and more financially oriented studies, sustainability, land search, etc.

4 Discussion

Foursquare relies heavily on crowdsourcing and while we did a great effort to extract all the venues possible, we cannot be certain that all supermarket and grocery stores would be actually listed.

During data cleaning, several (but not many) supermarkets and grocery stores of known brands were looked for and identified in categories other than the expected. That is rather impossible to do with unknown local brands. Therefore, some local markets could most likely be omitted, in case they were listed under a wrong category.

As previously discussed in 3.2.2. , while the indicator used appears to be functioning as expected, it does not demonstrate significant variance for locations with similar distances from the center of mass. A different approach to calculating the average distances to competition could provide us with better results.

The resulted locations do not indicate what is ideal, but they could be instead used as a narrower area to focus for further and more elaborate research, including cost assessments, sustainability, etc.

5 Conclusions

The economic center and largest cluster on our map, the city of Thessaloniki, seems to have an increased competition that is not comparing very well to demand. An investment in Thessaloniki would be considered of higher risk.

Instead, an investment towards the western side of the area appears to be safer and more promising. Additionally, while not visible on the map, those areas highlighted are located in close proximity to the country's national road artery. That makes them more favorable with regards to supply costs and connectivity in general.