Site Selection for Amazon Go in Philadelphia

Xiaoyuan Sun (Layla)

Abstract

The primary goal of this project is to provide recommendations on site selection for Amazon.com when it brings Amazon Go stores into the grocery retail market in Philadelphia. According to Amazon.com[1], in 2016, Amazon has opened a testing Amazon Go store to Amazon employees in the Beta program and this store is located at 2131 7th Ave, Seattle, WA. Assuming that Amazon Go would be able to pass the testing phase and start the national expansion phase, Philadelphia would eventually become a target city when the store comes to East Coast. For any store entering a new market, site selection is one of the most important issues. Instead of solely relying on qualitative analysis, this study tries to use k-means cluster analysis to narrow the possible target sites from the whole City to a few specific census blocks; further, a qualitative analysis, accompanied with a Poisson regression model, was conducted to evaluate the proposed blocks.

Introduction

By implementing technologies, such as computer vision, sensor fusion and deep learning, Amazon Go is created to provide the "fastest" – no lines, no checkouts – shopping experience in an upscale convenience store. However, there is only one testing store opened inside the building named Amazon Tower II in downtown Seattle. Given the nature of the store (i.e. prototype store only open to employees in the beta program), there are three question regarding the site selection and the fundamental assumptions of this study:

1) Is it the optimal location for a prototype store?
2) Would it still be an optimal location when the testing store is open to public?
3) Are the current site selection criteria generalizable? In other words, can we use the similar criteria to select some sites in other cities?

In order to conduct clustering analysis, we need positive answers to all three questions. However, in the reality, little information on the decision making process of the site selection for this very first Amazon Go store could be found; hence, a quick site evaluation for the current testing store is performed.

First, let us look at some unique features for the Amazon Go testing store:

- It is inside Amazon Tower II (also known as Day 1 and Rufus 2.0 Block 19), one of the high-rise and newly opened Amazon Headquarters. The new building could imply that the infrastructure and amenities are superior enough to ensure the daily operation of the high-tech convenience store implemented with the "most advanced shopping technology"[2] in the world.
- There are or soon will be (since the direct neighbor, "The Spheres", is expected to be in use in early 2018[3]) roughly 19 other amazon office buildings within a 0.6-mi, or 15-min walk, radius up to the North or to the East of the testing store [as shown in Figure 1.Above (on next page)]. In addition, according to geekwire.com[4], Amazon could expand its total working space from 8.5 million SQFT (as of the middle 2016) to 12 million SQFT across 40 buildings in Seattle by 2020. Hence, we can assume that the internal demand for Amazon Go would be sufficient after the prototype is open to all other employees in the future.

[1] "Amazon.com: Amazon Go." Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & More. Accessed April 5, 2017. https://www.amazon.com/b?node=16008589011.

[2] "Amazon.com: Amazon Go." Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & More. Accessed April 5, 2017. https://www.amazon.com/b?node=16008589011

[3] "Amazon's Spheres: Lush Nature Paradise to Adorn $4 Billion Urban Campus." The Seattle Times. Accessed April 26, 2017. http://www.seattletimes.com/business/amazon/amazons-spheres-are-centerpiece-of-4-billion-effort-to-transform-seattles-urban-core/.

[4] "Yes, Another Amazon Building: 17-story Office Project Will Expand New Seattle Campus to 5th Block." GeekWire. Accessed April 26, 2017. http://www.geekwire.com/2017/yes-another-amazon-building-17-story-office-project-will-expand-new-seattle-campus-5th-block/.

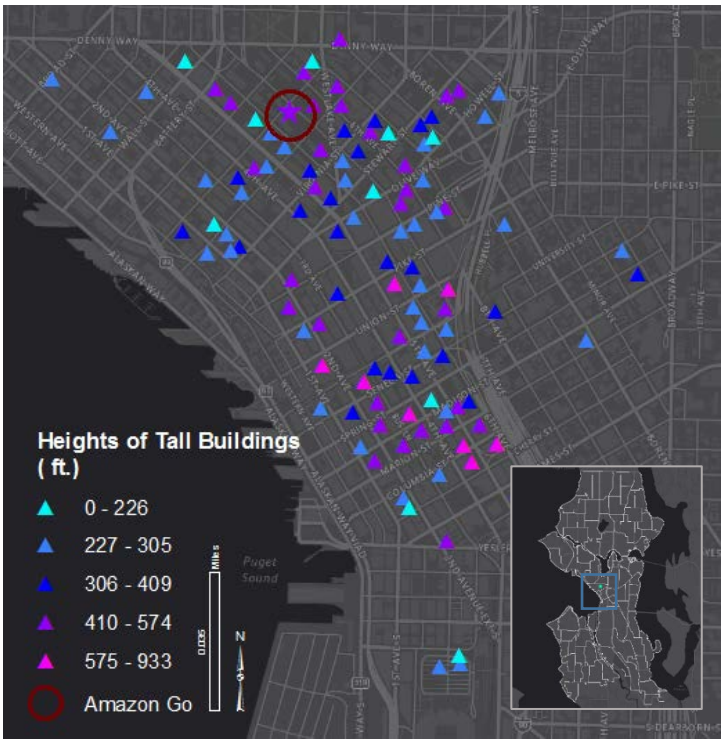


Figure 1. Quick Surrounding Checking
Left: Tall Buildings in Vicinity
Above: 19 Amazon Offices

Based on the observed unique site features of the testing Amazon Go store, *it is reasonable to assume that the current location of the prototype is the optimal choice.* In order to answer the second question, we need to look at the external demand.

Figure 1.Left (shown above) shows a map of tall buildings (including residential, office and hotel) in downtown. Amazon Go is not only surrounded by mid-rise buildings but also located within 0.7-mi from the aggregation of skyscrapers on the South in the central business district (CBD). The concentration of tall buildings implies that large number of working population in downtown could become potential customers who need fast shopping experience during breaks on workdays. Further confirmed by the 2016 daytime population per census tracts[5], the census tract harboring the cluster of tall buildings in CBD has a daytime population of 73,452, 71,348, or 97.14% of which are workers and the rest 2.86% are residents.

Therefore, it is also reasonable to expect that the demand for fast-shopping experience at Amazon Go from external customers would be high given the base of customers is large. Although traditional grocery or convenience stores would also be densely distribution in such areas, but the fast and high-tech shopping experience could give Amazon Go a significant competitive advantage over the traditional grocery stores. *In this case, we can also assume that this current location of the testing Amazon Go would still be optimal when the store is officially open to public.*

So far, with a quick site evaluation, we can assume that the location of the prototype is optimal for now and in the future. Then, there is a third question to answer – are the selection criteria generalizable to other locations in other states? Although the criteria are unknown to us, we still could assume that *if the new market in the region that is similar to the region where Amazon Go is located, the same criteria (with needed modification) will be generalizable enough to use for new site selection.*

In this study, Philadelphia is the scope of the new market. First, we need to check whether the two cities – Seattle and Philadelphia – have similar regions, and then, whether the region that Amazon Go is located has the matching similar regions in Philadelphia. If the two cities or certain groups of regions are similar, it will imply that the matching regions in Philadelphia would have a higher propensity of hosting Amazon Go stores.

[5] Esri. Esri Living Atlas Layer. Accessed April 27, 2017. https://laylasun.maps.arcgis.com/home/webmap/viewer.html?webmap=6d22431ac6e14bcf84ee2d38094c508b.

For this step, we will combine all the census block groups into one data set and apply k_means clustering analysis to divide the all the census block groups of both cities into different groups given a set of variables.

If the block group level clustering renders matching results, we will use cluster analysis but with a different set of variables as the pseudo criteria at a granular level – blocks, in order to further narrow down the candidate pool of possible sites. If the first-step clustering provides no matching block groups (i.e. block groups from different city in the same cluster), we will then pick center city area of Philadelphia as the focused study area to analyze and evaluate.

Finally, a qualitative raster overlay will be conducted to accompany the evaluation of the proposed sites from the previous steps. A Poisson regression is used to serve as a reference of the association of the predictors (the features) and the dependent variable – the count of the traditional grocery stores per unit in the selected neighborhoods where the proposed sites are located. Despite the possibility that all the predictors – the additional site features in Philadelphia – could be not statistically significant, we might still learn something about the relationship based on the model when manually assigning and adjusting the weights for each features to find the matching sites in the selected study area.

The report will then briefly talk about the definition of clustering and Poisson regression and the reason for choosing the methods. After that, the specific dataset used and the outcomes of each analysis will be presented and examined in detail. Finally, a summary of findings, dataset selection, methods implementation, possible improvement and interactive data visualization on results will also be discussed.

Methods

Cluster Analysis[6]

The goal of the census-block-group-level cluster analysis is to find meaningful segments for a dataset that contains both 481 census block groups (two block groups of water areas are excluded) of Seattle and 1338 (those with population less than 50 are removed) of Philadelphia. The cluster analysis takes into consideration of all the 26 variables, describing different characteristics of both cities, to generate a set of partitioning criteria and to assign each of the block group to a cluster/group if the values of all 26 variables meet the criteria of that particular group.

After identifying the clusters, we need to examine their characteristics, such as the mean values of all variables for each cluster and the spatial distribution of the clusters. A successful classification displays clearly separated clusters both numerically and spatially so that the difference between each two different clusters can be well interpreted.

There are three commonly used algorithms for cluster analysis: k-means, hierarchical and density-based with noise. Given the large dataset with continuous variables in this study, we will use k-means as the primary method.

K-Means Clustering

In k-means clustering, each observation finds its nearest cluster center and becomes part of that cluster; hence, the resulting clusters are non-overlapping. In this algorithm, a difficult part is to pick the "proper" number of clusters by ourselves. Since the optimal number of clusters is unknown, we need to run the algorithm multiple times and each time with a different number of clusters. Based on the output – within-cluster sum of square – or other criteria, we will then decide on the optimal number.

In this study, the maximum possible number of clusters is set to be 40 and the minimum is two. K-means algorithms will run 39 times. Each time, a different number within the interval is used and a within-cluster sum of squared error (SSE) is calculated by adding up the squared distance between each observation and the center of its assigned cluster. By plotting all the 39 within-cluster SSE, we obtain a Scree plot to help us decide

[6] Brusilovskiy, Eugene. "K-Means Clustering." Presentation, MUSA 501 Spatial Statistics & Data Analysis, University of Pennsylvania, 210 South 34th Street; Philadelphia PA 19104-6311, November 21, 2016.

on the optimal number for partitioning. In the Scree plot, we look for an elbow – a significant drop of the within-cluster SSE value.

In addition to the Scree plot, we will also use the 26 statistical methods available in the NbClust package in R. Each method will suggest an optimal number of clusters; therefore, the most frequently suggested number or numbers of clusters would consider being optimal.

The biggest limitation of k-meaning clustering is the way of finding the optimal number of clusters due to the possibility that the real optimal number is even not included in the defined interval [2, 40]. This method does not exclude noises and outliers and is only appropriate for numeric variables.

Alternative – Hierarchical Clustering

This method forms a hierarchy in either a "top-down" or a "bottom-up" fashion. The "top-down" or divisive approach starts with overarching cluster including all observations and keeps dividing it into pairs until reaching the individual data point. While the "bottom-up" approach starts with each individual observation and keeps pairing up clusters until group every data point into one single cluster. Then, based on the nature of study, the preferred level of hierarchy will determine the optimal number of clusters. *However, this method is more appropriate for smaller data sets, especially regarding the visualization of the clustering process*.

Alternative – Density-based Clustering

Density-based clustering is one of the most common clustering algorithms: it can identify non-globular shaped clusters, groups data points based on densities and excludes the identified outliers that lie too far from their nearest neighbors[7]. Although density-based clustering seems to overcome the limitations of k-means, we do not want to use it as the primary method given the probability that this method might treat all of our observations as outliers.

Poisson Regression

Poisson regression is useful to predict the response variable representing counts from a set of categorical and continuous independent variables[8].

In the final step of this study, the study area is set to be the 60 block groups in the center area of Philadelphia. A "fishnet" or a lattice containing 41,408 equally sized quadrat (50 ft. x 50 ft.) is created in ArcMap to cover the study area. In this case, the *dependent variable (DV)* is the number of grocery stores (i.e. retails with NAISC code 4451) falling into the quadrat. Given the confined study area and the limited number of grocery stores in the center city area, it is expected to see that the majority of the quadrats would have zero store.

The generalized linear model for positive integer values (counts) as the response Y is[9]:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k = \mathrm{x}_i^T \beta$$

The response has a Poisson distribution

$$p(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \quad Y = 0,1,2,\ldots \quad \mu > 0$$

$$Properties: E(Y) = V(Y) = \mu$$

[7] Sun, Xiaoyuan. *MUSA500 Assignment 5 - K-Means Clustering*. Philadelphia, PA: unpublished homework, n.d.

[8] Kabacoff, Robert I. "Generalized linear models." In *R in Action: Data Analysis and Graphics with R*, 2nd ed., 312. Shelter Island: Manning Publications Co., 2015.

[9] "9.1 - Poisson Regression Model | STAT 504." Redirect. Accessed April 27, 2017. https://onlinecourses.science.psu.edu/stat504/node/168.

The properties of Poisson distribution indicate: "any factor that affects one will also affect the other;"[10] hence, the usual assumption of homoscedasticity for standard OLS regression would not be appropriate for Poisson data[11].

To interpret the parameter estimate, we will take a single predictor *X* as an example.

$$\log(\mu) = \alpha + \beta_x$$

And the above function can also be written as:

$$\mu = exp(\alpha + \beta_x) = exp(\alpha)\, exp(\beta_x)$$

$exp(\alpha)$ represents the effect on μ, the mean of response *Y*, when $X = 0$.

$exp(\beta_x)$ indicates the multiplicative effect on μ, when *X* increases one unit:

- If $\beta = 0$, then $exp(\beta) = 1$, the expected count $\mu = exp(\alpha)$, and *X* and *Y* are not related
- If $\beta > 0$, then $exp(\beta) > 1$, the expected count $\mu = exp(\alpha)$ is $exp(\beta)$ times larger than when $X = 0$
- If $\beta < 0$, then $exp(\beta) < 1$, the expected count $\mu = exp(\alpha)$ is $exp(\beta)$ times smaller than when $X = 0$

Maximum Likelihood Estimation (MLE) is used for parameter estimation. That is, with MLE, a set of parameters for which "the probability of the observed data is greatest."[12] MLE entails finding the coefficients (i.e. β_0 , β_1 … β_k) that make the log of the likelihood function as large as possible.

The assumptions of Poisson regression model include the followings:

- Distribution of response, the counts, follows a Poisson distribution[13]
- The response must be non-negative integer
- No severe multicollinearity, that is two or more predictors are not strongly correlated with each other
- At least 50 observations per explanatory variable are needed because MLE is used for parameter estimation
- In addition, the common OLS assumptions of linearity (i.e., linear relationship between DV and each predictor), homoscedasticity and normality of residuals – do not hold in logistic regression

Goodness of model fit can be assessed by chi-square statistic, deviance and Likelihood ratio test, which will not be discussed in detail. Since we are dealing with count data, overdispersion is often encounter[14]. Overdispersion occurs in Poisson regression when the observed variance of *Y* is larger than the assumed variance, i.e., $Var(Y) = \varphi\mu$ where φ is a scale parameter. Overdispersion is suggested when "the ratio of the residual deviance to the residual degrees of freedom is much larger than 1."[15] When overdispersion is present, we will "adjust for overdispersion where we estimate $= \frac{x^2}{N-p}$ and adjust the standard errors and test statistics."[16]

We will use R to run and test the Poisson regression model and use quasi-Poisson approach provided by package qcc to adjust the standard errors and test statistics.

[10] GR's Website. Accessed April 27, 2017. http://data.princeton.edu/wws509/notes/c4.pdf.
[11] GR's Website. Accessed April 27, 2017. http://data.princeton.edu/wws509/notes/c4.pdf.
[12] Czepiel, Scott A. "Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation." czep.net. Accessed November 30, 2016. http://czep.net/stat/mlelr.pdf.
[13] "How to Perform a Poisson Regression Analysis in SPSS Statistics | Laerd Statistics." SPSS Statistics Tutorials and Statistical Guides | Laerd Statistics. Accessed April 27, 2017. https://statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php.
[14] Kabacoff, Robert I. "Generalized linear models." In *R in Action: Data Analysis and Graphics with R*, 2nd ed., 315. Shelter Island: Manning Publications Co., 2015.
[15] Kabacoff, Robert I. "Generalized linear models." In *R in Action: Data Analysis and Graphics with R*, 2nd ed., 315. Shelter Island: Manning Publications Co., 2015.
[16] "9.1 - Poisson Regression Model | STAT 504." Redirect. Accessed April 27, 2017. https://onlinecourses.science.psu.edu/stat504/node/168.

Raster Overlay Analysis

The final step of the analysis, a raster overly is used to create a single layer, which combines all the local characteristics of the narrowed study area – the 60 block groups around the center city neighborhood in Philadelphia. Each cell or grid of the final layer has a numeric value that represents an overall score considering all the input raster layers. The rules of interpreting the score vary based on different weight assignments; in our case, the higher the score, the more suitable or the more likely the cell will host a general type grocery store.

In raster overlay, each cell of each layer references the same geographic location. That makes it well suited to combining characteristics for numerous layers into a single layer. Usually, numeric values are assigned to each characteristic, allowing you to mathematically combine the layers and assign a new value to each cell in the output layer.

We rasterize each local feature from a vector layer, then use zonal statistics to burn the data of each feature onto the abovementioned lattice layer that contains the grocery store count data for Poisson regression. Hence, each feature will have a raster layer that is necessary for raster calculator, as well as a field in the lattice feature (vector) layer that for both regression model building and data visualization in the further steps.

Since "each cell of each [raster] layer references the same geographic location"[17] and contains a numeric or categorical values inherited from the original vector layer, we can use raster calculator to mathematically combine the layers together. An example[18] is shown below to illustrate the function of raster calculator in terms of combining the layers with weight equal 1.

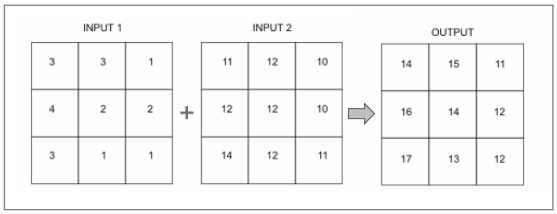


Figure 2. Overlaying Layers with Weight = 1

We use weighted raster overlay for two purposes:

1) If the Poisson regression model in the previous step shows that the local features are statistically significant to predict the count of a future grocery store in a quadrat, we will use the estimated coefficients of the model as the weights to construct the score map in raster calculator. The resulting raster map will show the likelihood of each grid to become a potential site for a new grocery store. This ***likelihood score map*** is a useful visualization tool to compare the results produced by the block-level clustering; that is, we can check whether the suggested blocks are located near or at the area with a likelihood in the top 5% percentile of the overall scores.

2) If the Poisson regression model shows that only one or two predictors are statistically significant, it will be less meaningful to use the parameters of the model to build a likelihood map. However, we might take the sign of the association between each of the predictors and the response as reference when manually assigning weights, based on our own theories, in the raster calculator to build a ***suitability map***. The higher the score of a grid, the more suitable it implies for a new grocery store to open. This suitability map will still be a comparison tool to check on whether the blocks chosen by clustering analysis using general criteria could be suitable when using local specific criteria.

However, when empirically assigning weights to different local features, the resulting map would be highly arbitrary and might not be representing the real suitability defined by the decision makers of the location of the prototype Amazon Go. Nevertheless, such a map inspires some interesting data visualization projects to allow people pick the predictors themselves and assign the weights they would prefer to build an interactive suitability map.

[17] "Overlay Analysis." ArcGIS Resource Center. Accessed April 27, 2017. http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/geoprocessing/overlay_analysis.htm.
[18] "Overlay Analysis." ArcGIS Resource Center. Accessed April 27, 2017. http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/geoprocessing/overlay_analysis.htm.

Data & Results

K-means Analysis at Census Block Group Level

As mentioned before, we use k-means to explore meaningful clusters for all census block groups of Seattle and Philadelphia for the purpose to identify the block groups located in different cities but share similar or comparable features.

Data

In order to obtain better clustering results, we need to find comparable data that are measured according to the same set of rules, recorded in the same format and are relatively updated, reliable and accurate for both cities. We have found that *2011-2015 American Census Survey* provides a wide variety of data related to the characteristics of population and housing. Furthermore, thanks to the easy-to-use query design and a clear organization of the topics in census data at ***nhgis.org***, we have quickly collected the needed 2015 census data at block groups level for both Washington and Pennsylvania. A detailed breakdown of variables is listed below:

Population

POP15	Total population in 2015
POP15DEN	Population density (# of people per square-mile)
PCTPOPCHA	Percentage change in total population from 2010 to 2015

Age

MDAGE	Median age of total population in 2015

Per Capita Income

PCINC15	Per capita income in 2015 (“in 2015” will be omitted in the following description)
PCTINCCHA	Percentage change in per capita income from 2010 to 2015

Education

PCTBACH15	Percentage of population at least with a bachelor degree

Rent-related

MDCRENT15	Median contract rent*
PCTMDCR	Percentage change in median contract rent from 2010 to 2015
PCTHSAFF15	Percentage of households with affordable housing**
PCTHSBRD15	Percentage of households with housing burden***

Commute Time to Work

PCTCTLT15	Percentage of total working population with commuting time less than 15 mins
PCTCT1540	Percentage of total working population with commuting time between 15 mins and 40 mins
PCTCTMT40	Percentage of total working population with commuting time more than 40 mins

Means for Commuting

PCTPT	Percentage of workers using public transportation to work for workers older than 16 year-old
PCTMBW	Percentage of workers using motorcycle, bicycle or walked
PCTCAR	Percentage of workers using cars

Geographical Mobility in the Past Year

PCTSAMEHS	Percentage of population staying in the same house 1 year ago in 2015
PCTMETROPC	Percentage of population moving from a principal city in metropolitan statistical area

Housing

MDHV15	Median value of owner-occupied housing units
MDHVCHA	Percentage change in median value of owner-occupied housing units from 2010 to 2015
PCTVA15	Percentage of vacant housing units

PCTFAM15	Percentage of family household
HSUNIT15	Total housing units
HS15DEN	Housing units per square mile
MDBLDAGE15	Median year of structure built

* *Contract rent* is the monthly cash rent agreed to, regardless of any furnishings, utilities, fees, meals, or services that may be included. In comparison, gross rent is the contract rent plus the estimated average monthly cost of utilities and fuels if these are paid by the renter.
** Percentage of households that spend less than 35% of household income on rent [19]
*** Percentage of households that spend more than 35% of household income on rent
All data are measured at census block group level.

We assume that these census data are able to summarize the characteristics of population and housing in both cities to help k-means produce meaning clusters. Before revealing the results, let us look at the mean values of each variable for both cities to grasp a basic understanding how similar the two cities are, solely based on census data. Without detailing each variable, we will focus on the variables whose values are considerably different.

	Philadelphia		Seattle	
Variable	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
General population				
POP15	1,171.38	537.54	1,366.59	418.48
POP15DEN	23,176.60	14,485.44	12,335.15	12,433.23
PCTPOPCHA	10.79	41.28	13.34	31.07
Age				
MDAGE	35.50	8.84	38.02	7.56
Per Capita income				
PCINC15	22,734.33	14,893.50	46,476.07	20,103.87
PCTINCCHA	11.75	44.60	15.31	31.12
Education				
PCTBACH15	23.31	21.78	58.46	18.99
Rent-related				
MDCRENT15	666.43	363.79	1,060.20	451.44
PCTMDCR	19.83	52.55	23.14	40.40
PCTHSAFF15	43.89	24.25	60.58	20.58
PCTHSBRD15	39.78	23.30	27.67	18.79
Commute Time to Work				
PCTCTLT15	13.80	11.61	18.88	9.82
PCTCT1540	52.86	15.30	60.74	10.54
PCTCTMT40	33.34	16.13	20.37	9.64
Means for Commuting				
PCTPT	29.49	18.00	19.51	9.80
PCTMBW	9.38	13.47	12.47	12.30
PCTCAR	57.31	21.15	60.25	17.44
Geographical Mobility in the Past Year				
PCTSAMEHS	86.27	11.37	78.19	13.52
PCTMETROPC	10.49	9.10	15.81	10.27
Housing				
MDHV15	148,973.51	122,886.22	451,086.28	204,549.75
MDHVCHA	12.20	51.71	(0.11)	18.72
PCTVA15	13.68	10.87	5.64	5.51
PCTFAM15	54.59	17.16	48.96	19.97
HSUNIT15	504.78	232.31	660.84	280.05
HS15DEN	10,254.38	7,466.50	6,584.44	8,930.08
MDBLDAGE15	67.16	11.62	54.20	18.71

Figure 3. Variable Mean Values for Both Cities

For general population, Seattle (abbreviated as "SEA") has a higher average block-group population than Philadelphia (shortened as "PHI"), but SEA's average block group population density is only half of PHI's. In order to find out the reason, we need information at city level.

Seattle has an estimated total population of 686,800 in 2015 and its total land area is 83.78 sq-mi divided into 481 unequal-sized block groups. The citywide population density is 8,197.66 per sq-mi and average size of block group is 0.17 sq-mi.

In comparison, PHI has an estimated total population of 1,526,006 in 2015, total land area of 141.7 sq-mi. and 1338 block groups (as mentioned before, 10 block groups with total population less than 50 are excluded). The citywide population density in PHI is 10,769.27 per sq-mi, or around 1.3 times denser than that of SEA and the average size of block group is 0.11 sq-mi. Hence, a denser population but smaller block groups on average in PHI result in the larger average block-group population density.

Another big difference appears in terms of the average block-group-level (BGL) per capita income. In Seattle, not only the average per capita income (at block group level) is more than twice the value of that in PHI, but also the average BGL percentage increase over the past five years is larger than that in PHI.

[19] "Housing's 30-Percent-of-Income Rule Is Near Useless - Bloomberg." Bloomberg.com. Accessed April 28, 2017. https://www.bloomberg.com/news/articles/2014-07-17/housings-30-percent-of-income-rule-is-near-useless.

The large difference in per capita income could also be reflected and justified in the rent-related variables. The average BGL median contract rent of SEA is $1,060.20, almost 60% higher than that ($666.43) of PHI. However, the average BGL percentage of households that spend more than 35% of the household income in SEA is 27.67%, which is 30% lower than that in PHI. In other words, on average, people living in SEA can afford higher contract rent with their higher income; while people in PHI do not earn as much, hence the housing burden is higher even with lower median contract rent in general.

The higher average BGL per capital income might be a result of education. In SEA, the BGL average percentage of population received at least Bachelor's degree is 2.5 times higher than that in PHI.

In terms of commuting to work, employed people spend less time on traveling to work; and this phenomenon can be partially explained by the difference in the sizes of the cities. When it comes to the commuting means to work, compare to working people in PHI, people in SEA seem to take less public transits to work and prefer cars and other means, such as motorcycles, bicycles and feet (to walk…). In addition, according to the mean values for both BGL percentage of population staying in the same house a year ago (PCTSAMEHS) and percentage of people who move from a principal city in 2015 (PCTMETROPC), people living in SEA seem to be somewhat more mobile than people living in PHI. Also, more people, who previously have lived in a principal city in either same or different metropolitan areas, chose to come to SEA.

Based on the difference in per capita income, education level and contract rents, we would expect the average housing values in Seattle to be higher than that in Philadelphia. It is still surprising to see that SEA has triple the BGL average median housing value of PHI's, although this value has declined by 11% on average from that in 2010. The BGL average housing density difference can be justified by the population difference explained before. However, the BGL average vacancy rate (PCTVA15) of PHI is still 2.5 times the value of that in SEA, despite of the fact that PHI's houses on average are 1.23 times the median age of houses in SEA.

Overall, Seattle has less dense population who are more educated, have higher per capita income and can afford to live in houses with higher rent than the population in Philadelphia. In terms of housing, generally speaking, Seattle's houses are slightly younger, much less vacant and considered more valuable.

Results

The number of the clusters, denoted as K, falls within an interval [2, 40]; hence, the algorithms run each time with a different K and calculate the corresponding within-cluster SSE. Figure 4 (shown below) shows a Scree plot for the 39 SSE values (on the left), as well as a bar plot indicating some commonly suggested optimal partitioning number from the 30 methods in the NbClust package.

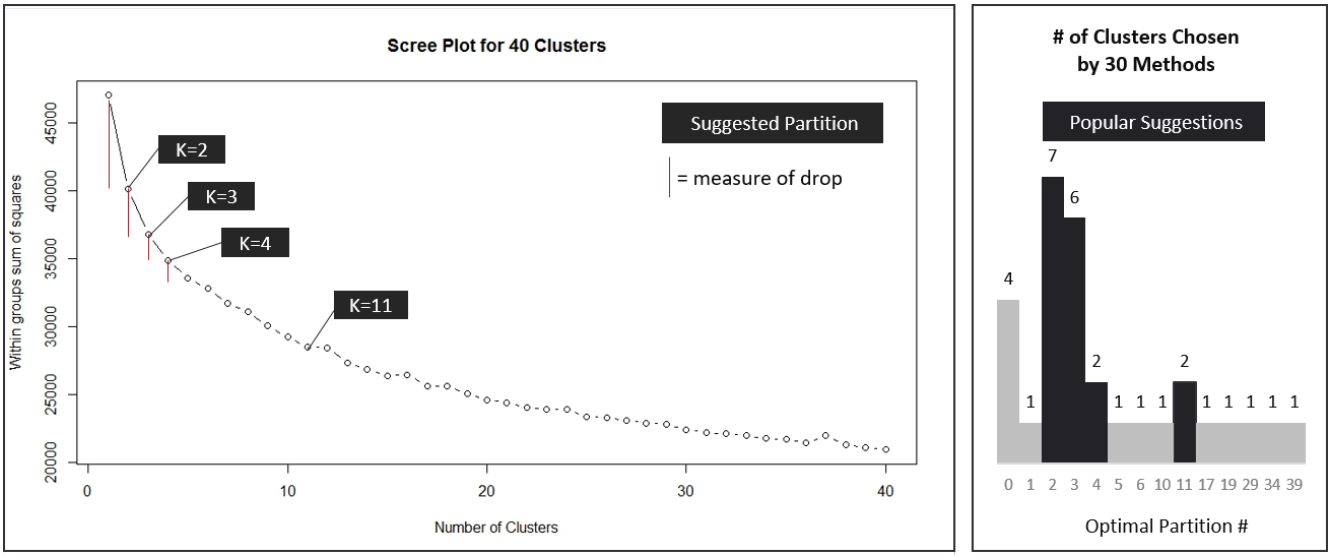


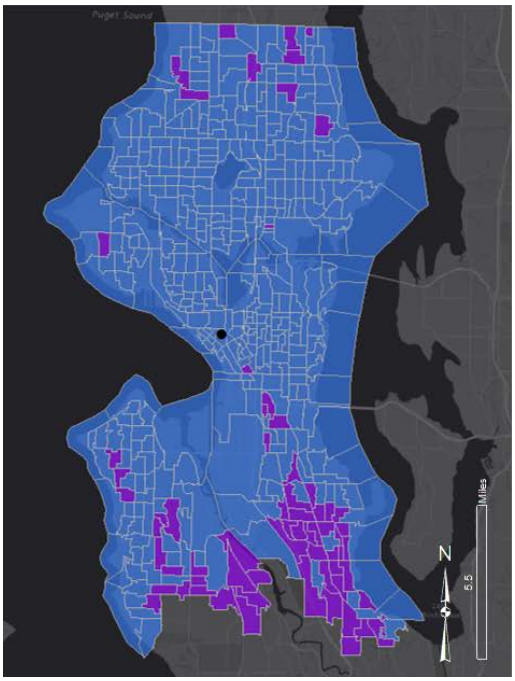
Figure 4. Scree Plot and Bar Chart for Popular Optimal Clustering Numbers

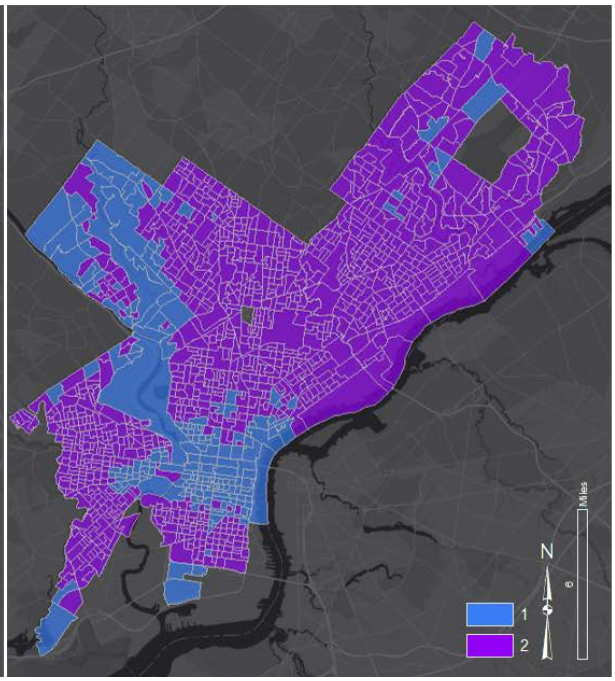
Based on the visual examination of Scree plot, we would suggest that K=3 is the optimal partitioning number. Based on the suggestions from the 30 methods, K=2 is the most frequent results and K=3 just falls behind by one. The majority rule suggests us choose the winner, K=2; but combined with the suggestion from Scree plot, we would be more inclined to K=3 as the optimal number of clusters.

However, given the nature of spatial data, we should not only rely on the mean value of each cluster, but also need to look at the spatial distribution of the clusters to decide whether k-means has generated clearly separated and well defined clusters. In this case, we also keep K=4 and K=11 as comparing groups for map visualization; if the suggested optimal clustering does not appear meaningful, the comparing groups would serve as the supporting groups if they could better display the results both in value and on map .

Furthermore, we use the term "target block group" to indicate the block group housing the prototype Amazon Go. Likewise, we say "target cluster" to suggest the cluster, which the target block group is assigned to.

Figure 4a (shown below) displays the spatial distribution of clusters when K=2 with a table of mean values for all the variable in each cluster. Visually, the clustering seems to be reasonable in PHI but not in SEA. In PHI, the block groups in the areas that have relatively higher-income people are assigned to the target cluster, cluster #1. However, in SEA, according to the same clustering criteria, the majority (414 out of 481, or, 86%) of the block groups meets the requirements that only "richer" block groups, only 16% of total block groups in study, could in PHI. Nevertheless, such a result is not surprising after we have looked at the average characteristics at city level in the previous step. Hence, we need to use other more commonly suggested partitioning numbers as comparison to find the best clustering results. Overall, with K=2, the spatial distribution of the two clusters tells a story: only the rich part of Philadelphia can compete with the city Seattle as a whole.

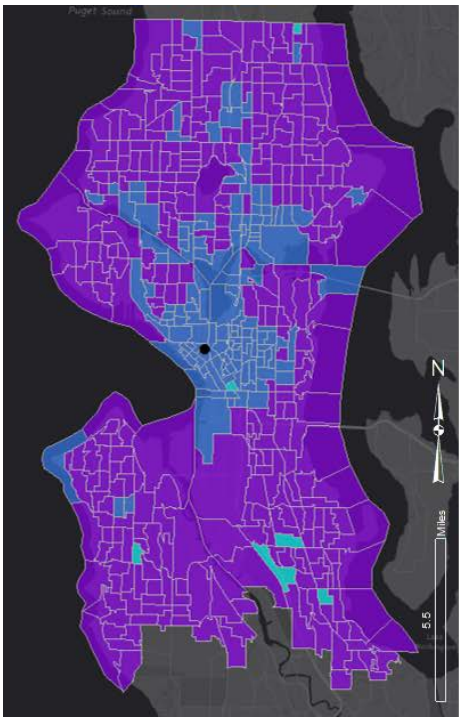


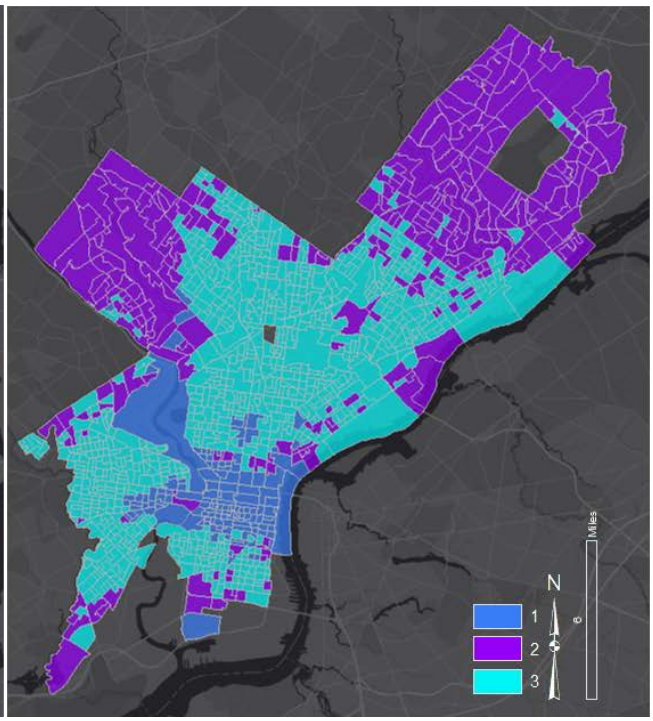


Cluster	1 (Target)	2
POP15	1,149.6	1,361.7
POP15DEN	21,984.0	17,133.4
PCTPOPCHA	9.6	14.9
MDAGE	35.6	37.2
PCINC15	19,062.2	47,895.3
PCTINCCHA	9.8	18.1
MDHV15	129,048.3	418,330.4
MDHVCHA	11.6	3.8
MDCRENT15	589.1	1,114.4
PCTMDCR	19.2	23.6
PCTHSAFF15	40.8	62.5
PCTHSBRD15	42.2	25.9
PCTBACH15	16.9	62.4
PCTVA15	13.8	7.3
PCTFAM15	58.1	43.5
HSUNIT15	475.3	680.3
HS15DEN	9,256.0	9,336.3
MDBLDAGE15	66.7	58.2
PCTCTLT15	12.5	20.2
PCTCT1540	52.3	59.9
PCTCTMT40	35.2	19.9
PCTPT	30.5	19.8
PCTMBW	5.9	18.3
PCTCAR	60.1	54.4
PCTSAMEHS	88.0	76.8
PCTMETROPC	9.6	16.3

Figure 4a. Map of Clusters Distribution in SEA and PHI and Means Values of Variables for Each Cluster | K=2

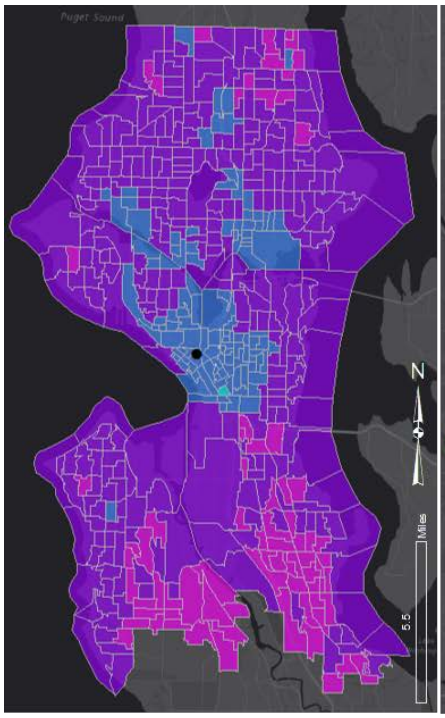
Figure 4b (shown on next page) illustrates the spatial distribution of the three clusters in both cities, as well as a table of all variables with mean values in each cluster. *Based on visually judgement of the distribution, the clustering seems to be well executed given the clearly separated groups. Amazingly, simply based on 26 variables created from BGL census data, the k-means algorithms managed to identify the downtown areas as well as the regions with universities of both cities with such an accuracy*. Figure 4c (shown on the next page) provides the same type of map and table but with K=4. Likewise, clustering with K=4 is able to separate the city areas with denser population (in either downtown or university districts) from the rest.

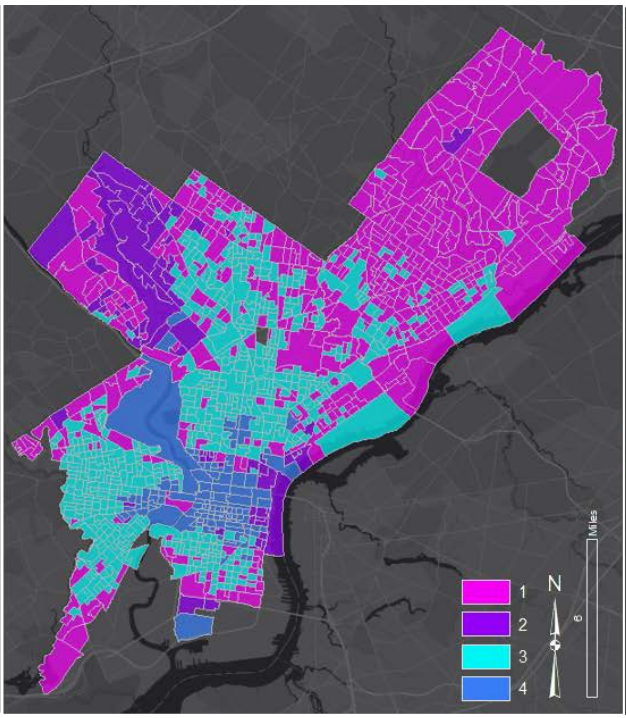




Cluster	1 (Target)	2	3
POP15	1,452	1,327	1,078
POP15DEN	27,732.3	10,665.1	24,830.1
PCTPOPCHA	22.3	8.1	10.5
MDAGE	33.3	40.3	34.1
PCINC15	46,445.2	38,890.1	16,625.3
PCTINCCHA	19.0	13.8	9.9
MDHV15	356,709.3	352,705.5	101,730.7
MDHVCHA	8.2	0.1	15.4
MDCRENT15	1,142.4	880.6	577.3
PCTMDCR	26.0	17.9	21.1
PCTHSAFF15	60.2	57.5	38.1
PCTHSBRD15	27.9	27.9	45.5
PCTBACH15	65.1	44.8	13.9
PCTVA15	9.4	5.8	16.3
PCTFAM15	29.5	57.8	57.1
HSUNIT15	790.6	572.4	451.5
HS15DEN	15,967.7	4,612.9	10,514.2
MDBLDAGE15	55.2	58.9	69.8
PCTCTLT15	23.6	16.7	11.4
PCTCT1540	58.7	57.8	51.7
PCTCTMT40	17.6	25.5	36.9
PCTPT	23.3	15.7	35.8
PCTMBW	31.9	5.6	6.7
PCTCAR	36.9	72.9	54.2
PCTSAMEHS	68.0	86.8	87.2
PCTMETROPC	21.7	9.5	10.6

Figure 4b. Map of Clusters Distribution in SEA and PHI and Means Values of Variables for Each Cluster | K=3

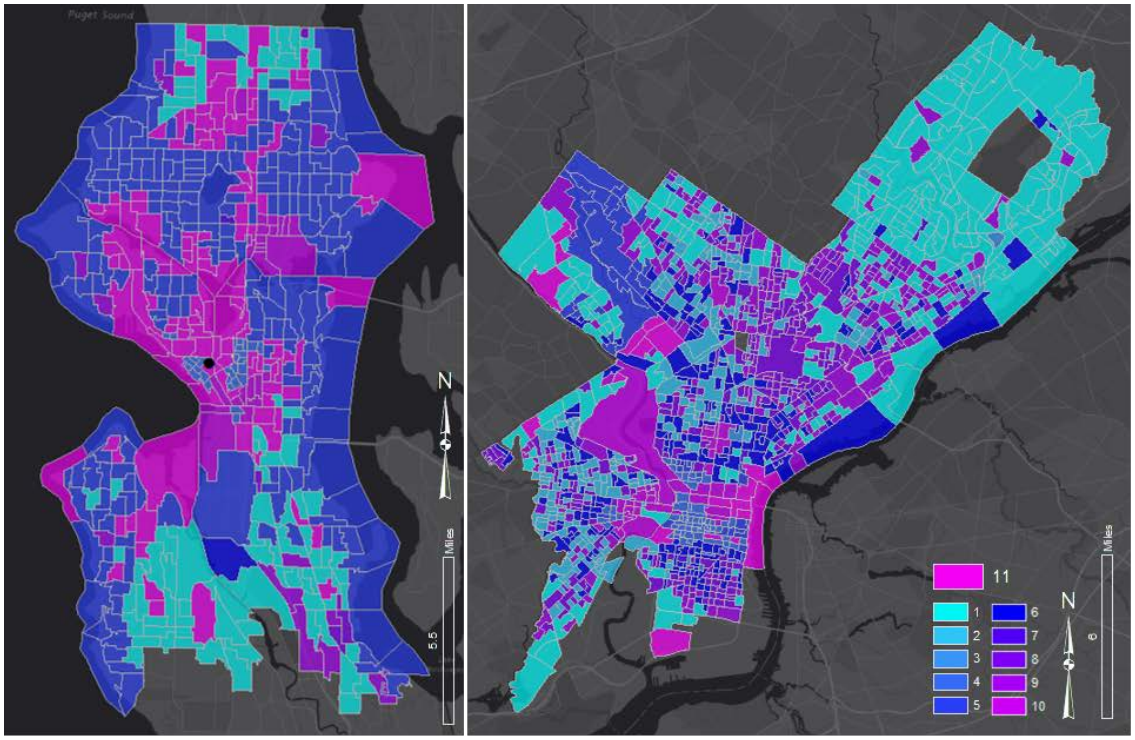




Cluster	1	2	3	4 (Target)
POP15	1341.6	1217.8	1011.8	1517.9
POP15DEN	17072.5	9520.4	25981	30205.8
PCTPOPCHA	12.5	7.8	8.4	23.1
MDAGE	37.6	39.7	34.1	32.5
PCINC15	22711.3	52154	16286.2	44292.4
PCTINCCHA	5.1	19.7	13.9	17.8
MDHV15	173873	494430	93952.4	327043.3
MDHVCHA	0.9	2.8	20.3	7.7
MDCRENT15	626.5	1128.6	578.4	1114.8
PCTMDCR	10.8	23.5	26.4	26
PCTHSAFF15	44.4	65.7	38.4	58.6
PCTHSBRD15	38.3	23.4	45.2	28.9
PCTBACH15	22.1	64	13.3	64.2
PCTVA15	8.3	6.1	18.5	9.4
PCTFAM15	60.4	53.9	55.1	26.8
HSUNIT15	535.3	566.1	444.1	831.7
HS15DEN	6790.2	4502.9	11373.7	17661.8
MDBLDAGE15	60.5	62.2	71.3	53.3
PCTCTLT15	14.2	17.7	11.1	25.1
PCTCT1540	55.5	61.2	49.9	57.5
PCTCTMT40	30.3	21.2	38.9	17.4
PCTPT	16.9	18	42.5	23.1
PCTMBW	4.5	9.4	7.3	34.7
PCTCAR	74.8	64.9	47.1	34.5
PCTSAMEHS	89.5	82.5	86.3	66.3
PCTMETROPC	7.8	12.8	11.4	22.7

Figure 4c. Map of Clusters Distribution in SEA and PHI and Means Values of Variables for Each Cluster | K=4

Without going much in details to describe each of the variables in all clusters, we could easily summarize the results. When K=3, our **target cluster** (cluster #1) represents the block groups hosting the business districts, university centers, as well as a thin layer of the adjacent block groups. *Cluster #2* indicates the block groups in the similar suburban areas in both cities. However, it is still very interesting to see that the majority of non-downtown and non-university-center areas in SEA share the same living standards as the rich neighborhoods in PHI. *Cluster #3* seems to *accentuate the block groups in PHI's neighborhoods where people from middle-class*



Variable	Cluster 11 (Target)	Ranking (From H-L)
POP15	1,450.2	2
POP15DEN	39,290.3	1
PCTPOPCHA	22.6	4
MDAGE	36.5	5
PCINC15	59,197.1	1
PCTINCCHA	22.2	2
MDHV15	385,607.1	2
MDHVCHA	8.4	5
MDCRENT15	1,191.5	1
PCTMDCR	26.3	3
PCTHSAFF15	65.6	2
PCTHSBRD15	23.5	10
PCTBACH15	71.2	1
PCTVA15	9.0	8
PCTFAM15	25.8	11
HSUNIT15	965.6	1
HS15DEN	26,537.3	1
MDBLDAGE15	54.7	10
PCTCTLT15	22.8	2
PCTCT1540	60.2	4
PCTCTMT40	17.1	11
PCTPT	21.4	8
PCTMBW	38.3	1
PCTCAR	32.7	10
PCTSAMEHS	70.8	10
PCTMETROPC	20.5	2

Figure 4d. Map of Clusters Distribution and Means Values of Variables for Each Cluster | K=11

and lower reside; however, according to the criteria that separate the block groups in PHI into, plainly speaking, rich and not-rich neighborhoods, only 6 out of 481, or 1.2% of the block groups, in SEA are considered as the "suburbs for middle-class people." Said differently, the cluster #2 and #3 (when K=3) explicitly show us the fact: the overall living standard and qualities in Seattle as a whole represent the lives of people in richer neighborhoods in Philadelphia.

Importantly, this phenomenon implies that K-means with K=3 succeeds in identifying the city parts that are more comparable, since the downtown areas as well as the educational centers of comparable metropolitan cities might have highly similar features across the country. As a result, it is easy to separate these parts from the rest based on the same criteria. However, when it comes to suburban areas where the substantial difference of each city lies, using the same criteria, even with good adjustment to corporate the distinguish features of the cities, might not be successful in describing the uniqueness of each city. As in our case, the criteria used to separate the block groups into different social class works well for PHI, but apparently, this particular set of criteria could not properly partition the block groups into SEA's own social classes.

Interestingly, based on the observation above, *we can conclude* that the target cluster, cluster #1 when K=3, represents the most comparable parts of two cities. Hence, in the further steps with narrowed study areas, we should focus on those high comparable city parts. That is because, as mentioned at the beginning of the report, it is more likely that the same criteria, used to select the site for the current testing Amazon Go store in Seattle, would succeed in targeting the sites as the potential store locations in Philadelphia.

Finally yet importantly, let us look at the clustering results when K=11 specifically in terms of the mean values of each variables. Without detailing, the mean values of the target cluster #11 appear to be more persuasive when being compared to the other ten clusters. The ranking for each variable (in terms of mean value from highest to lowest) seems to be at either extremes for most cases.

Without having to choose a particular optimal clustering number, we are curious to know whether there are block groups that are always belongs to the target cluster no matter which of the four values for K is used. We use Intersection function in ArcMap to find out the block groups that never changed sides (shown in Figure 5 below).

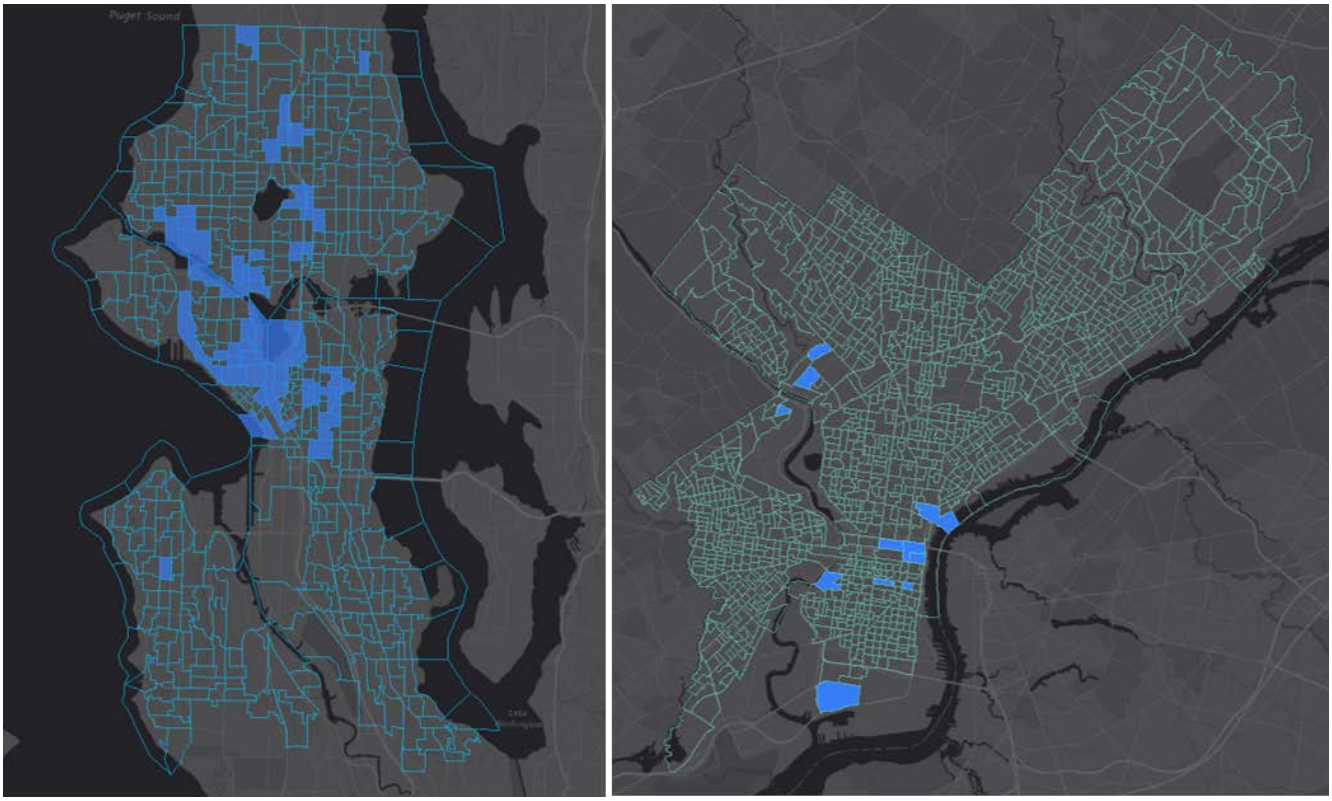


Figure 5. Map of Block Groups for Narrowed Cluster Analysis at Census Block Level

In conclusion, k-means algorithms produced meaningful clusters, and, in both cities, there are block groups that always stay in the target cluster regardless the number of the clusters used to partition. These block groups suggest that they share highly similar characteristics in terms of population and housing. In next step, we will narrow our study scope from the whole city to these census block groups. Additionally, we have chosen a different set of variables as pseudo-criteria to perform a new cluster analysis at a census block level, which is a level lower than census block group.

Cluster Analysis at Census Block Level

Data

In the block-level cluster analysis, we did not merely choose census data at block level but we selected a set of landmarks and facilities as the raw input features for two reasons:

1) Using the same set of data to run k-means at a deeper level does not ensure the meaningful clustering results, as the selected block groups seems to be the ones that prove to be highly similar/comparable regardless the number of clusters (among the four choices).
2) At a deeper level, we switch focus to the spatial features on the map, such as distance to certain features as well as a deeper level of data regarding the same variables in the previous step, such as more detailed but information on buildings/built structures.

For geospatial features, we used Near Table function in ArcMap to calculate the average distance to the nearest five observations for each feature if the feature contains more than five observations. For the features with less than five data point or the polyline features, we use Near function to find the nearest data point or the line segment in the feature layer. And the following is a list of spatial features used in this step:

Utilities

- ND_WWF: Near distance to waste water facilities
- ND_BCF: Near distance to broadcast facilities

- ND_OR: Near distance to oil refineries
- ND_PS: Near distance to power stations

Transit System

- ND_HWS: Near distance to highway segments
- ND_RWS: Near distance to railway segments
- ND_RF: Near distance to railway facilities
- ND_LRS: Near distance to light railway segments
- ND_LRF: Near distance to light rail facilities
- ND_BUS: Near distance to bus stations
- ND_PORT: Near distance to ports
- ND_FERRY: Near distance to ferry facilities

Emergency Facilities

- ND_SH: Near distance to schools
- ND_MEDC: Near distance to medical centers
- ND_FS: Near distance to fire stations
- ND_POLIS: Near distance to police stations

The reasons for choosing the above features are:

It is very hard to find the dataset that contain the same type of features for both cities; hence, we set our priority of this step to be finding the optimal dataset with high-quality data. Fortunately, the inventory database from HAZUS ("HAZUS is a nationally applicable standardized methodology that contains models for estimating potential losses from earthquakes, floods, and hurricanes."[20]) contains the same set of feature layers and other census-block-level data in terms of building stocks and vehicles for both cities. Rather than searching for comparable datasets manually, we decided to select the data that are suitable for the second cluster analysis directly from HAZUS.

Secondly, we do not know exactly the criteria used to choose the current location of Amazon Go, hence, we intend to include all the spatial relationships between each block in the study area and the available features in HAZUS dataset. Although some of it might not seem to be relevant by our own judgement, such as distance to oil refineries, we should still include such a feature given the possibility that it might be included in the original criteria.

The other part of the dataset contains of detailed information on building and vehicles as listed below. For the same reasons mentioned above, we incorporate those variables.

Daytime Vehicle Counts by Types	**Night-time Vehicle Counts by Types**
TotalDay	TotalNight
DayCars	NightCars
DayLightTrucks	NightLightTrucks
DayHeavyTrucks	NightHeavyTrucks
Daytime Vehicle Dollar Exposure by Types	**Night-time Vehicle Dollar Exposure by Types**
DayTotalNewCarsUSD	NightTotalNewCarsUSD
DayTotalNewLightTrucksUSD	NightTotalNewLightTrucksUSD
DayTotalNewHeavyTrucksUSD	NightTotalNewHeavyTrucksUSD
DayTotalUsedCarsUSD	NightTotalUsedCarsUSD
DayTotalUsedLightTrucksUSD	NightTotalUsedLightTrucksUSD
DayTotalUsedHeavyTrucksUSD	NightTotalUsedHeavyTrucksUSD
DayTotalVehiclesUSD	NightTotalVehiclesUSD

[20] "FEMA Flood Map Service Center | Hazus." FEMA Flood Map Service Center | Welcome!. Accessed April 28, 2017. http://msc.fema.gov/portal/resources/download#HazusDownloadAnchor.

Building Blocks - Dollar Exposure by Occupational Types
TotalExposureOccu
ResiExp (residential)
CommExp (commercial)
IndusExp (industrial)
AgriExp (agricultural)
ReligExp (religious)
GovExp (government)
EduExp (education institution)

Building Blocks - Inventory Counts by Occupational Types
TotalCount
ResiCnt
CommCnt
IndusCnt
AgriCnt
ReligCnt
GovCnt
EduCnt

Building Blocks - Inventory Counts by Material Types
TotalExposureMType
WoodExp
SteelExp
ConcreteExp
MasonryExp
ManuHsExp

Building Blocks - Inventory Counts by Material Types
TotalMTypeCnt
WoodCnt
SteelCnt
MasonryCnt
ConcreteCnt
ManuHsCnt

Given 66 variables used in the dataset, we do not want to spend too much time study each input, rather, we want to look at the results of the cluster analysis.

Results

From the Scree plot (in Figure 6 shown below), we would choose K=5 as the optimal partitioning number. And based on the suggested optimal numbers, we select K=2, 3 and 5 as primary candidates while keeping 34 for comparing and supporting. It is noticeable that the 11 out of the 30 methods failed to produce results.

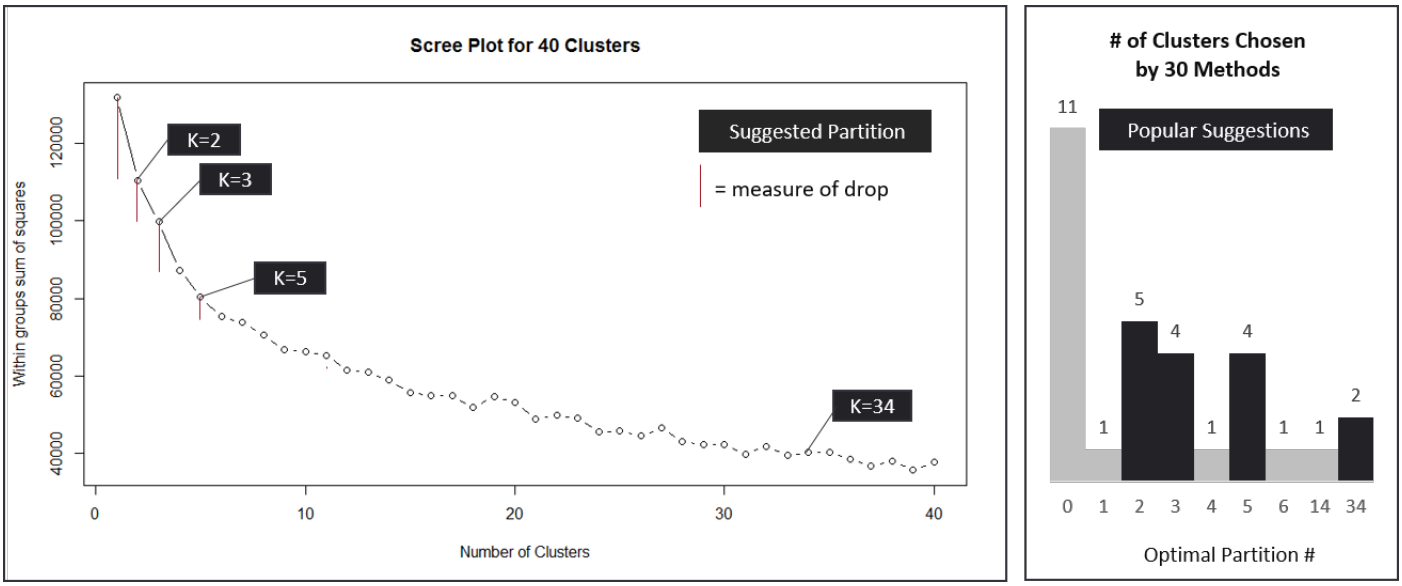


Figure 6. Scree Plot and Bar Chart for Popular Optimal Clustering Numbers

Figure 6a-6d display a close-up map for the target census block and its surrounding census blocks since we are more interested in the target cluster than the true meaning of each cluster given different optimal partitioning number used. Likewise, we are more focused on the spatial distribution of the clusters than the mean values for each clusters, because the focus of this step is to interpret the spatial relationships of input features used as pseudo-criteria.

When K=2, the target cluster – cluster #2 – contains the majority of the census blocks in both cities. That is, 1856 out of 1884 blocks in the focused study area, or 98.5% of the total census blocks in the study in SEA have been classified into the same group. Similarly, in PHI, 381 out of 387 census blocks, or 98.4% of total census blocks have fallen into the target cluster. Either the criteria of partitioning do not seem to be appropriate or the blocks in both areas are highly similar. This result justifies the reason we had to choose a completely different set of variables.

When K=3, the target cluster – cluster #3 – contains 1003 census blocks, or 53.24% of the total census blocks in the study in SEA; however, 370 out of 387 census blocks, or 95.6% of total blocks in PHI are assigned to the target cluster. When K=5, the target cluster (#5) still contains 85.5% of total census blocks in PHI while the distribution of cluster in SEA seems to divert. Similar to the results in the BGL clustering in the previous step, the criteria used to partition the clusters seem to be largely influenced by the city that dominates the input dataset: 1884 blocks belong to Seattle while only 387 belong to Philadelphia. With an overwhelming presence in the analyzing database, we can then expect the partitioning criteria are skewed to be more representing regarding the spatial features in Seattle. Hence, we see that diversification of the distribution of the clusters as well as the numbers of the census blocks in each cluster in Seattle when using a large K. Hence, we use the "extreme" number, 34, to see how responsive the census blocks in PHI would be to such a large number.

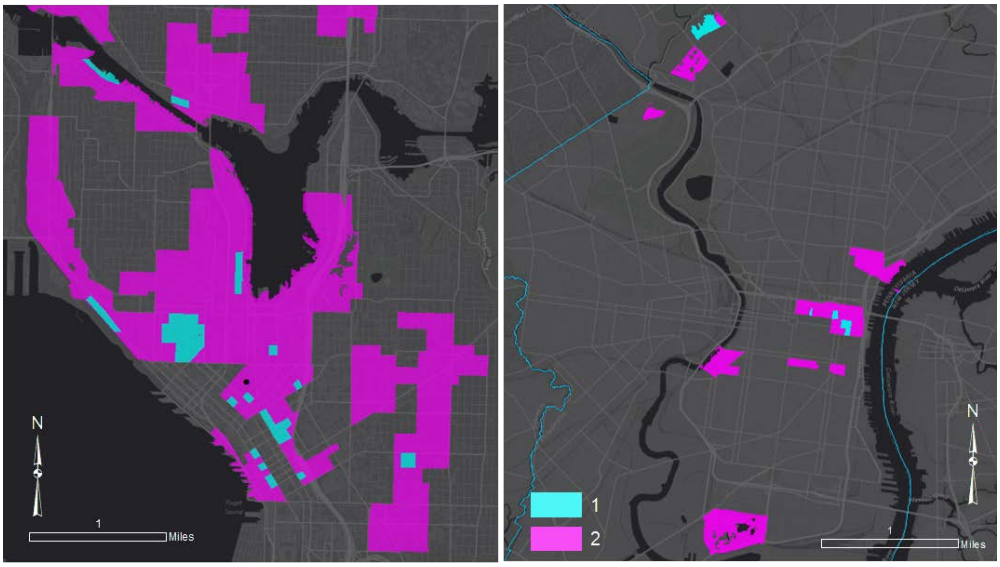


Figure 6a. Map of Clusters Distribution | K=2

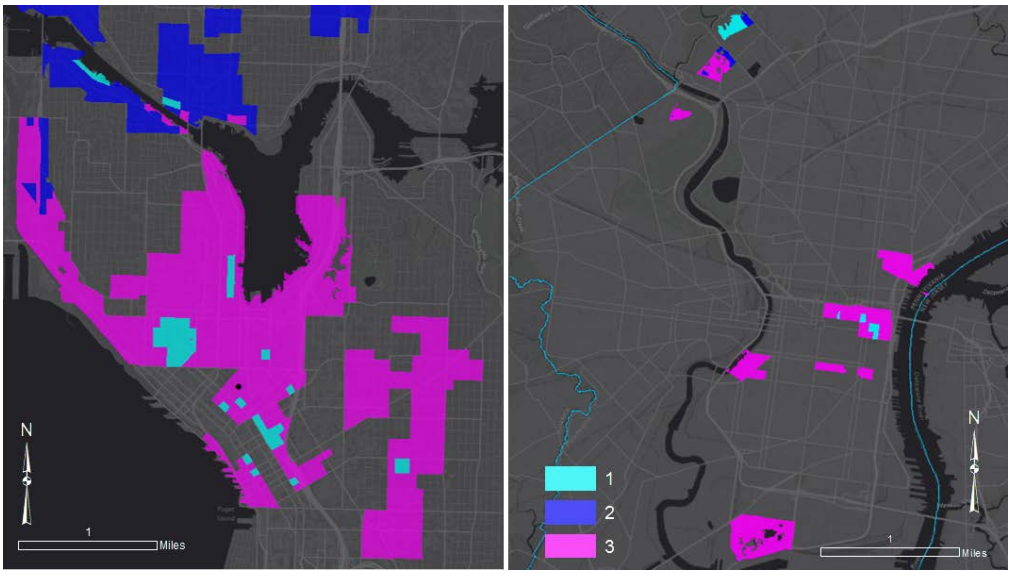


Figure 6b. Map of Clusters Distribution | K=3

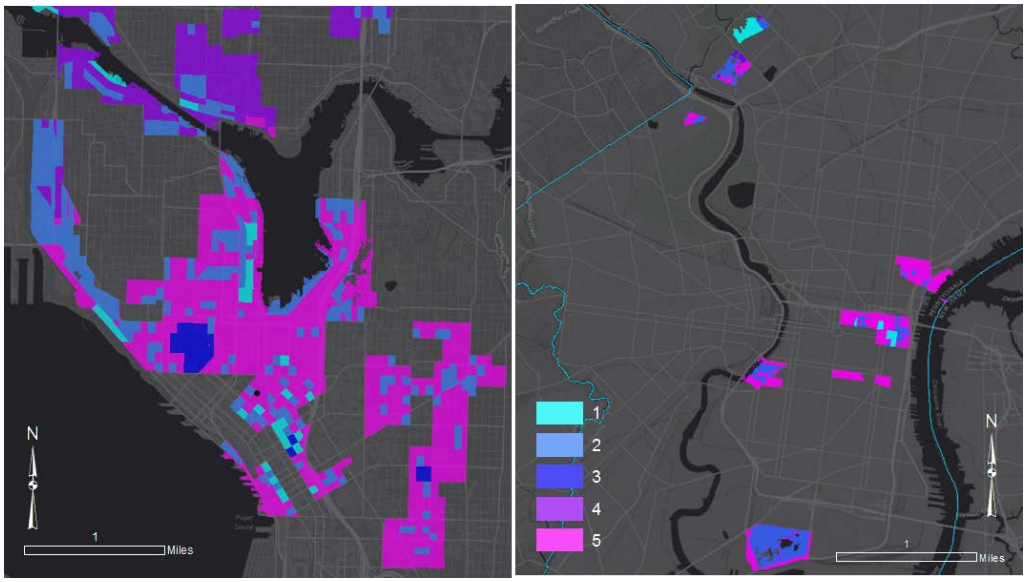


Figure 6c. Map of Clusters Distribution | K=5

Interestingly, the target block #6 (when K=34) does not even exist in Philadelphia while the map of Seattle seems to be increasingly diversified since no empty cluster present in this city. Nevertheless, in PHI, no apparent cluster with dominant amount of census blocks is shown on the map. So, we need to check whether the clusters that the adjacent blocks in SEA contain the census blocks in PHI.

From the map in the middle in Figure 6d, we identify cluster #22 is the nearest direct neighbor of Amazon Go and the target census block. And there are 32 census blocks in PHI assigned to this cluster. By a little checking on the zoning codes in Seattle, we have been assured that the surrounding census block all fall in the similar zoning areas. Although cluster #6, the original target cluster, falls in both mixed-use commercial/office zoning area and in office/business district, we still prefer use clusters that are located at the areas with the exact same zoning code as the block where the Amazon Go store is located. In this case, we consider cluster #4 and #28 as the Tier 1 clusters and #12 and #22 as Tier 2 clusters. Luckily, all of the four clusters contain some census blocks in PHI as illustrated in Figure 6f below.

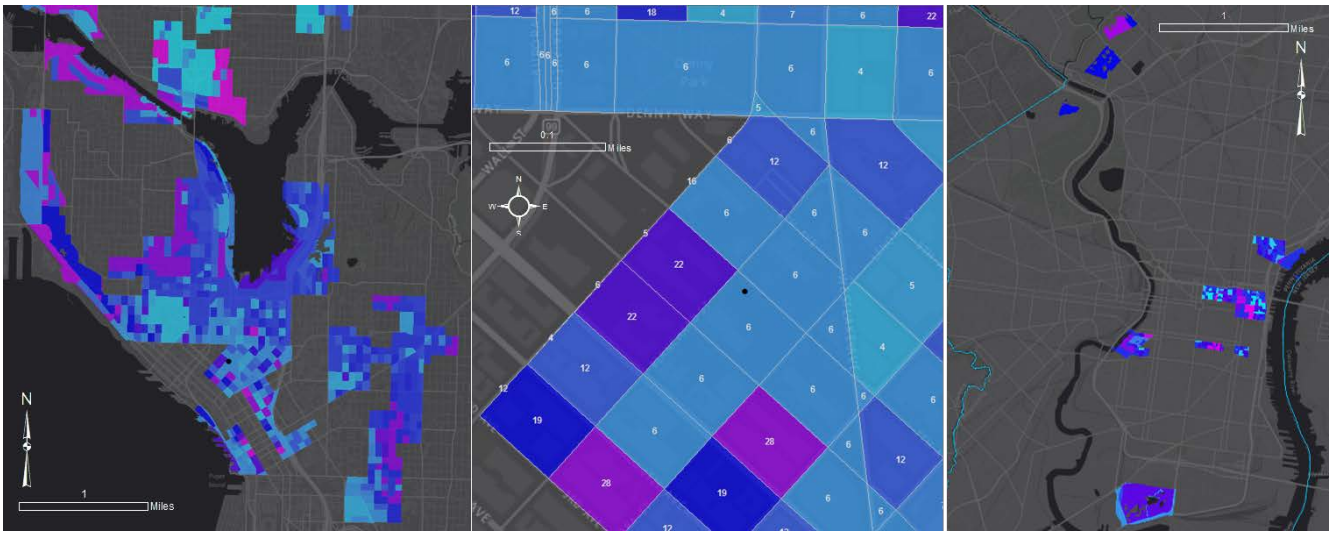


Figure 6d. Map of Clusters Distribution with Focus on Target Block| K=34

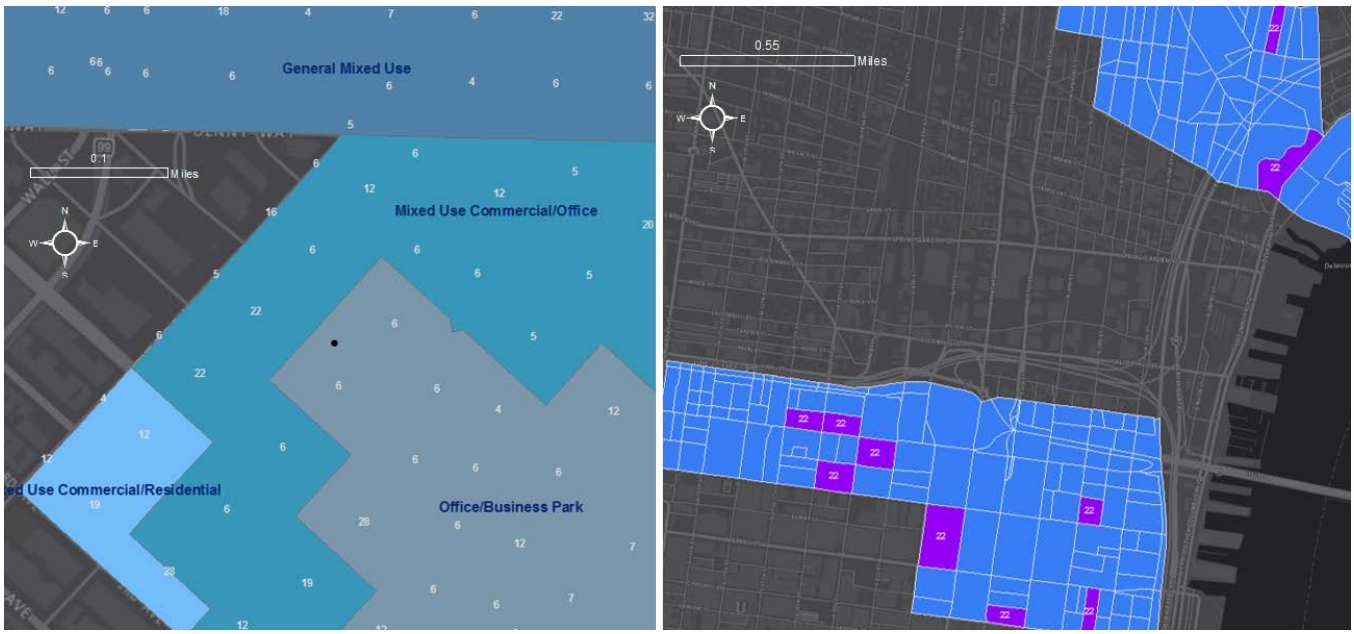
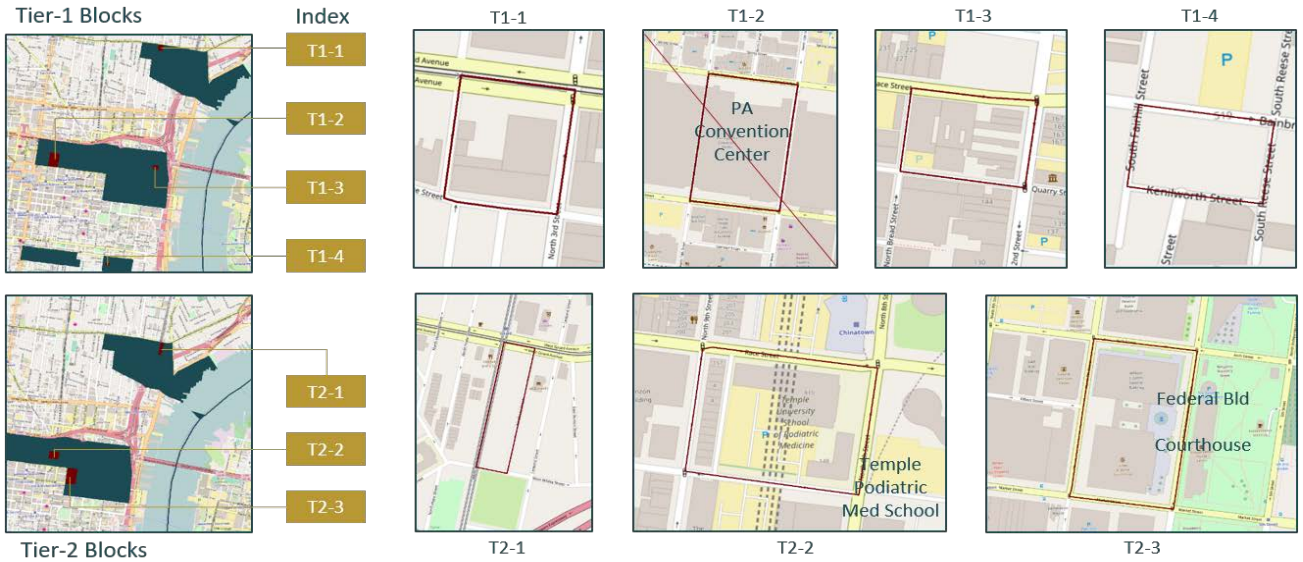


Figure 6e. Zoning in Amazon Go Area in SEA and Matching Cluster#22 in PHI| K=34

Figure 6f. Tier 1 (4 & 28) and Tier 2 (12 & 22) Clusters in PHI| K=34

Given the less meaningful results produced for K=2, 3 and 5, we decided to use the modified clustering results using K=34 as the partitioning number. However, given different tiers when using K=34, we will intersect the census blocks in the first three target clusters separately with the census blocks in different tiers. When intersecting the census blocks in cluster #2 (when K=2), #3 (when K=3), #5 (when K=5) and #12 and #22 (when K=34), respectively, three census blocks (shown in the map on the left in Figure 6g) become the final products of the intersection. Similarly, three different census blocks (shown in map on the right in Figure 6g) are indicated as the final products after intersecting with the census blocks in the Tier 2 clusters.

Figure 6g. Tier 1 and Tier 2 Census Blocks in PHI



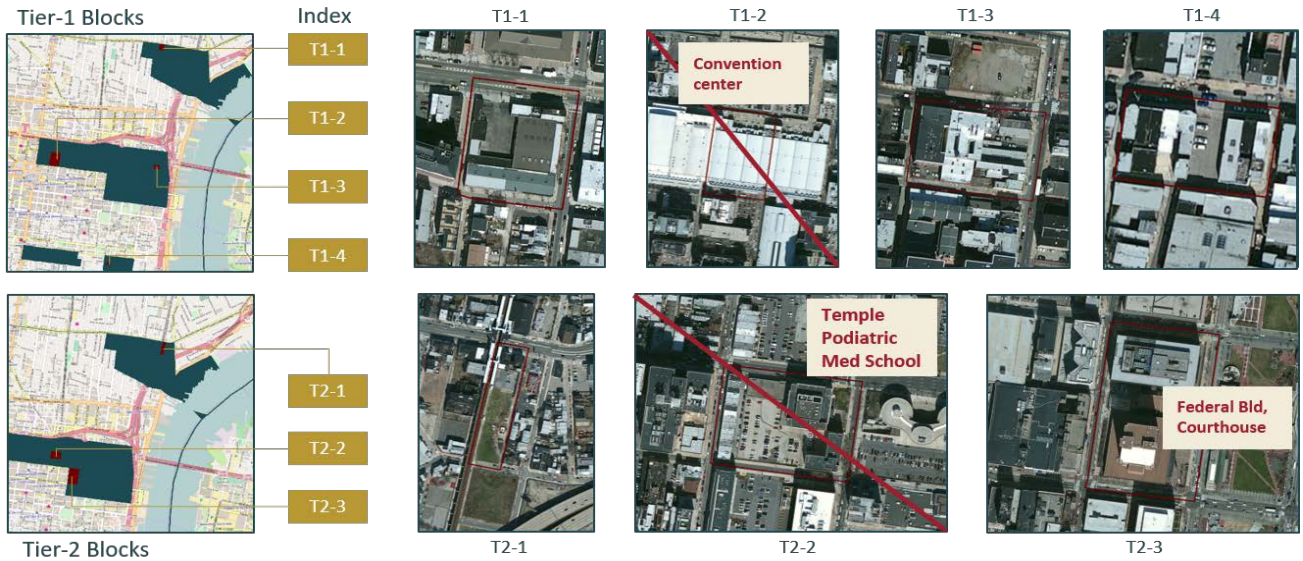


Figure 6h. Using OpenMap and Satellite Imagery to Check on Each Census Block

Figure 6h shows both a street map image and a satellite image of each resulting census block. First, we need to exclude the blocks that seem to be less likely to host a general grocery store. T1-2 block is the middle part of the Philadelphia convention center and we do see a high-tech grocery to open in such a building, so we crossed it out. T2-2 block is located close to the center business district, but it belongs to the Podiatric Med School of Temple University. To us, Amazon Go and Med school just do not match.

T1-1, T2-1, T1-3 and T1-4 are relatively far from the center city area. Although they both seem to be proper to host a grocery store, we still look for some place near downtown within a 15-min walking distance. T2-3, with a federal building and courthouse right on top, does not seem to have more space to host a high-tech convenience store.

In other words, we do not obtain a very satisfactory block as a final suggestion; but, the two cluster analysis indeed show us the areas that have the highest propensity to host a high-tech convenience store. Those areas are all located in or near the center business district in Philadelphia. Instead of further narrowing down the study scope, in the last step, we want to scale up the study focus to the 60 census blocks in the center city area and apply a mixed local dataset, containing data at different level and from different aspects, to conduct a raster overlay analysis. Then, comparing the results across the methods to see how different the results can be.

Raster Overlay with Poisson Regression

Data & Results

Data that mainly focus on the retail market at both census tract and census block levels are semi-manually collected from esri's living atlas layers online[21]. The dataset with all retail stores by NAISC in Philadelphia in 2015 was obtained in Professor Landis's Real Estate Development class in fall 2016.

We select these datasets to explore the current supply condition of retail market in the study area. In addition, we incorporate features, such as a comprehensive set of bus shelters, bike and pedestrian traffic counts, locations of all tall buildings in the study area as well as the day/night population and size of working population, to represent the potential demand for a grocery store. The following list shows a detailed descriptions of all variables used in this step.

dist_comp	Average distance to the nearest 5 competitors (NAISC 4451 & 44611; 2k~6k sqft)
dist_sales	Average distance to the nearest 5 retail stores with no less than avg sale volume ($1,294,487.41) in 2015
budget	2016 average annual budget expenditures per block group
hsspent	2016 average amount spent per household on retail goods per block group
resales	2016 total annual retail sales (Supply) per block group
redemand	2016 total annual retail potential sales (Demand) per block group
relsf	2016 market opportunity (leakage/surplus factor) per block group
lsf4411	2016 market opportunity for automobile dealers per block group
lsf4451	2016 market opportunity for grocery stores per block group
lsf448	2016 market opportunity for clothing/accessories stores per block group
lsf722	2016 market opportunity for food, service and drink stores per block group
lsf445	2016 market opportunity for food and beverage stores per block group
lsf4452	2016 market opportunity for specialty food stores per block group
lsf4453	2016 market opportunity for bear/wine/liquor per block group
pot4451	Annual Retail Sales Potential (Demand)for grocery stores per block group
trct_dpop	Day population per census tract
trct_npop	Night population per census tract
trct_pctdw	The percentage of day population that are workers per census tract
dis_tallbd	Average distance to the nearest 5 tall buildings
den_tallbd	Kernel density of tall buildings

[21] Esri. Esri Living Atlas Layer. Accessed April 27, 2017. https://laylasun.maps.arcgis.com/home/webmap/viewer.html?webmap=6d22431ac6e14bcf84ee2d38094c508b.

dist_trans	Average distance to the nearest 5 public transit stations (subway and bus)
kd_bikes	Kernel density of bike counts
kd_pedes	Kernel density of pedestrians
z_ind1	Industrial zones: 1=yes; 0=no
z_cm1	Mixed-used commercial zones: 1=yes; 0=no
z_spp1	Special purpose zoning code: 1=yes; 0=no
z_hres1	Higher-density residential or mixed use residential: 1=yes; 0=no
z_lres1	Lower-density residential: 1=yes; 0=no

All the above feature layers are rasterized for raster overlay analysis.

As mentioned before, we created a lattice that is cropped into the shape of the study area as shown in Figure 7 below. Each cell in the lattice is a 50 ft. x 50 ft. quadrat and the modified lattice contains 41,408 quadrats. Then, we burn data of each variable onto this lattice to prepare for Poisson regression analysis. The response is the current grocery store counts per cell. However, before running the regression, we should expect that the model would not be robust given the highly confined study area. Nevertheless, we want to look at the estimated coefficients as a reference for raster overlay.

As expected, when we construct a Poisson regression model with response and all the variables, only distance to nearest competitors is a highly statistically significant predictor for grocery store count. Two other variables with p-values less than 0.05 are special purpose zoning and lower-density residential zoning. Moreover, all of the statistically significant variables are negatively associated with the dependent variable.

Figure 7. Cropped Lattice

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-2.7342727	1.5480228	-1.766	0.0773	.
dist_comp	-0.0072925	0.0006812	-10.706	<0.0000000000000002	***
dist_sales	0.0002257	0.000471	0.479	0.6318	
budget	5.796E-05	0.0001035	0.56	0.5753	
hsspent	-0.0001517	0.0003558	-0.426	0.6698	
resales	5.297E-10	2.647E-09	0.2	0.8414	
redemand	4.368E-08	1.797E-07	0.243	0.8079	
relsf	-0.0037209	0.0101074	-0.368	0.7128	
lsf4411	0.0010822	0.0035681	0.303	0.7617	
lsf4451	-6.28E-05	0.0117428	-0.005	0.9957	
lsf448	0.0014716	0.0044868	0.328	0.7429	
lsf722	0.0016063	0.0048574	0.331	0.7409	
lsf445	-0.0013122	0.015806	-0.083	0.9338	
lsf4452	0.0034022	0.0036563	0.93	0.3521	
lsf4453	-0.0003796	0.0039048	-0.097	0.9226	
pot4451	-2.953E-07	1.109E-06	-0.266	0.79	
trct_dpop	3.002E-06	1.189E-05	0.252	0.8007	
trct_npop	0.0002215	0.0001484	1.492	0.1357	
trct_pctdw	-0.3320157	1.2639521	-0.263	0.7928	
dis_tallbd	-2.368E-05	0.0001994	-0.119	0.9055	
den_tallbd	-0.0027331	0.002711	-1.008	0.3134	
dist_trans	0.0001044	0.0002249	0.464	0.6424	
kd_bikes	2.611E-06	7.456E-06	0.35	0.7262	
kd_pedes	1.999E-06	1.183E-06	1.69	0.091	.
z_ind1	-13.414902	426.75699	-0.031	0.9749	
z_cm1	-0.2367014	0.2122368	-1.115	0.2647	
z_spp1	-2.1531289	1.0307244	-2.089	0.0367	*
z_hres1	-0.4239308	0.5061693	-0.838	0.4023	
z_lres1	-1.2510815	0.4008347	-3.121	0.0018	**

Figure 8. Summary of Estimated Parameters

Although we could not use Poisson regression to build a model with sufficient predictive power, we could look at use the coefficients as a reference when we assign weights in the raster overlay analysis. Also, since the model suggests that most of our variables are not statistically significant as predictors, it seems to be more appropriate to empirically "set" the coefficients if we still believe that the information will help identify meaningful sites. But, different people with different theories with come up with different score maps. This inspires us to create an interactive map in R that allows people to select variables and assign different values as coefficient. Then, a customized score map is produced.

However, before the app creation, we still need to look at whether raster overly can produce some meaningful results. So, among all the 28 variables, we pick 9 variables that we believe more important than the rest to re-run the Poisson regression. Although the results are still not improved (as shown in Figure 9, only 2 variables are statistically significant), we still use the exponentiated coefficients to construct a map (in Figure 10).

```
Call:
glm(formula = cnt4451 ~ dist_comp + relsf + lsf4451 + trct_pctdw +
    kd_bikes + den_tallbd + kd_pedes + dist_trans + z_lres, family = "poisson",
    data = retaildata1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4191  -0.0814  -0.0410  -0.0171   3.8723

Coefficients:
               Estimate   Std. Error z value            Pr(>|z|)
(Intercept) -1.808394290  0.937541671  -1.929              0.0537 .
dist_comp   -0.006441444  0.000558291 -11.538 <0.0000000000000002 ***
relsf        0.002424632  0.002483875   0.976              0.3290
lsf4451      0.000353562  0.002121587   0.167              0.8676
trct_pctdw  -0.280983372  0.975053536  -0.288              0.7732
kd_bikes     0.000005369  0.000005906   0.909              0.3633
den_tallbd  -0.002308966  0.001643699  -1.405              0.1601
kd_pedes     0.000001896  0.000001172   1.617              0.1059
dist_trans   0.000166331  0.000212862   0.781              0.4346
z_lres1     -0.847726464  0.385655538  -2.198              0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1443.9  on 41407  degrees of freedom
Residual deviance: 1190.5  on 41398  degrees of freedom
AIC: 1457.1

Number of Fisher Scoring iterations: 10
```

Figure 9. Summary of Estimated Parameters with 9 Variables

From the resulting map, we find that the clusters of quadrats with higher scores are actually very close to the blocks in the previous step. Moreover, the clusters on this map shows the areas that we might actually prefer over the suggested blocks from the deeper level clusters.

Discussion

In conclusion, combined the results from both cluster analysis and raster overlay, we suggest that keep the four blocks (T1-1, T2-1, T1-3 and T1-4) from the cluster analysis as the primary potential sites. However, a further exploration of Philadelphia’s central area by using the interactive raster overlay map app and by physically visiting the potential sites is highly necessary.

In this study, we attempt to use statistical methods to find potential sites for a high-tech convenience store, Amazon Go, with the assumptions that the current location of the testing store is optimal and the criteria used for selecting this particular site can be used to select sites in other cities or regions that have highly similar or comparable features.

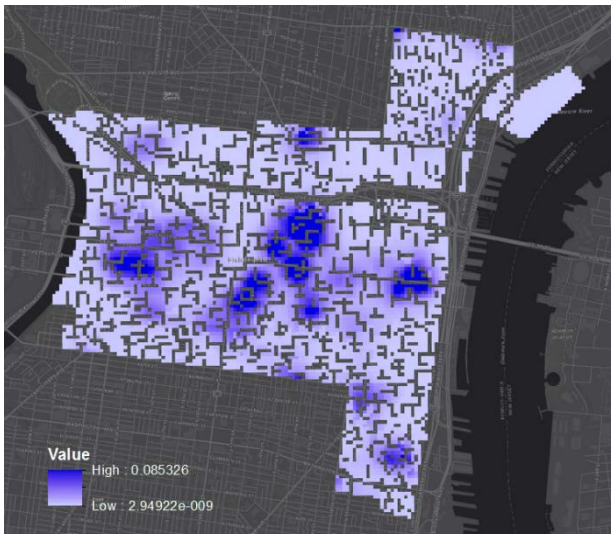


Figure 10. Score Map with 9 Variables

Both of the statistical methods have their own strengths and limitations, but they have helped us identify the highly comparable parts of both cities and implied other possibilities around the primary blocks.

However, we found ourselves overly rely on data. For each analysis, we would have to collect and build a different dataset to ensure the quality of the analysis. Furthermore, we have to question on the stability of R, since we just found that R has produced a different set of results in the block group level clustering when we finalized the report. Given limited time, we could not re-build the dataset for the block-level cluster analysis, which might be able to provide us with more satisfactory results than the current ones.

Hence, a second edition of the report is in on the way, along with the interactive map in R.

Bibliography

2017 Council on Tall Buildings and Urban Habitat. "Seattle - The Skyscraper Center." The Skyscraper Center. Accessed April 26, 2017. https://www.skyscrapercenter.com/city/seattle.

"9.1 - Poisson Regression Model | STAT 504." Redirect. Accessed April 27, 2017. https://onlinecourses.science.psu.edu/stat504/node/168.

"Amazon.com: : Amazon Go." Amazon.com: Online Shopping for Electronics, Apparel, Computers, Books, DVDs & More. Accessed April 5, 2017. https://www.amazon.com/b?node=16008589011.

"Amazon's Spheres: Lush Nature Paradise to Adorn $4 Billion Urban Campus." The Seattle Times. Accessed April 26, 2017. http://www.seattletimes.com/business/amazon/amazons-spheres-are-centerpiece-of-4-billion-effort-to-transform-seattles-urban-core/.

Brusilovskiy, Eugene. "K-Means Clustering." Presentation, MUSA 501 Spatial Statistics & Data Analysis, University of Pennsylvania, 210 South 34th Street; Philadelphia PA 19104-6311, November 21, 2016.

Esri. Esri Living Atlas Layer. Accessed April 27, 2017. https://laylasun.maps.arcgis.com/home/webmap/viewer.html?webmap=6d22431ac6e14bcf84ee2d38094c508b.

"FEMA Flood Map Service Center | Hazus." FEMA Flood Map Service Center | Welcome!. Accessed April 28, 2017. http://msc.fema.gov/portal/resources/download#HazusDownloadAnchor.

"Four forces shaping competition in grocery retailing." Strategy& - the Global Strategy Consulting Team at PwC. Accessed January 25, 2017. http://www.strategyand.pwc.com/media/file/Four-forces-shaping-competition-in-grocery-retailing.pdf.

GR's Website. Accessed April 27, 2017. http://data.princeton.edu/wws509/notes/c4.pdf.

"Housing's 30-Percent-of-Income Rule Is Near Useless - Bloomberg." Bloomberg.com. Accessed April 28, 2017. https://www.bloomberg.com/news/articles/2014-07-17/housings-30-percent-of-income-rule-is-near-useless.

"How to Perform a Poisson Regression Analysis in SPSS Statistics | Laerd Statistics." SPSS Statistics Tutorials and Statistical Guides | Laerd Statistics. Accessed April 27, 2017. https://statistics.laerd.com/spss-tutorials/poisson-regression-using-spss-statistics.php.

Kabacoff, Robert I. "Generalized linear models." In *R in Action: Data Analysis and Graphics with R*, 2nd ed., 312. Shelter Island: Manning Publications Co., 2015.

"Overlay Analysis." ArcGIS Resource Center. Accessed April 27, 2017. http://resources.esri.com/help/9.3/arcgisdesktop/com/gp_toolref/geoprocessing/overlay_analysis.htm.

Sun, Xiaoyuan. *MUSA500 Assignment 5 - K-Means Clustering*. Philadelphia, PA: unpublished homework, n.d.

"Yes, Another Amazon Building: 17-story Office Project Will Expand New Seattle Campus to 5th Block." GeekWire. Accessed April 26, 2017. http://www.geekwire.com/2017/yes-another-amazon-building-17-story-office-project-will-expand-new-seattle-campus-5th-block/.