**Lab Report**

Title: <Draft 1 Final Project >

Notice: Dr. Bryan Runck

Author: Rongxuan Zhu

Date: 11.10.2021

**Project Repository:**< https://github.com/zhux0474/GIS5571/tree/main/FinalProject>

**Google Drive Link:** <N/A>

**Time Spent:** <48 hours>

**Abstract**

The machine learning method k means clustering algorithm and density-based spatial clustering of applications with noise (DBSCAN) method will be used to group sets of latitude and longitude coordinate points from Travel Behavior Inventory (TBI) Household Survey Interview Data for 2010 in Minnesota to determine the clustering pattern of household/individual interaction (and trip origins and destination) location. The interaction between survey participants is defined by the time geography concept and space-time coupling constraints. The latitude and longitude coordinate points will be grouped into k groups using the unsupervised k means algorithm and it will be tested in DBSCAN to compare the results of these two different methods. This project will combine spatial analysis with the power of machine learning to determine and obtain the clustering pattern of TBI data to understand how the household and trip locations are segregated spatially and to test the performance of two different clustering methods.

**Problem Statement**

This project will apply and compare the machine learning method k mean clustering algorithm and DBSCANE clustering method in Jupyter notebook python environment to find the household/individual interaction clustering pattern using the Travel behavior Inventory Household Survey Interview Data for 2010 and to figure out how the household and trip locations are segregated spatially in Minnesota. Table 1 shows the broken-down elements of the problem statement.

Table 1. <Table of Elements Break Down from Problem Statement >

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **#** | **Requirement** | **Defined As** | **(Spatial) Data** | **Attribute Data** | **Dataset** | **Preparation** |
| 1 | Travel Behavior Inventory Household Interview Survey | Raw input dataset for k mean clustering algorithm from MNDOT | Household survey data | Travel Behavior Inventory Household Interview Survey | [Mn GeoSpatial Commons](https://gisdata.mn.gov/dataset/us-mn-state-metc-society-tbi-home-interview2010) | Downloaded from website and import into Jupyter Notebook |
| 2 | Traffic Analysis Zoned Data | Raw input dataset as spatial boundary for result verification in ArcGIS Pro | TAZ Boundary Shapefile | Traffic Analysis Zone Data | [Mn GeoSpatial Commons](file:////Users/rongxuan/Desktop/GIS%205571/Final%20Project/Draft/The%20traffic%20analysis%20zone%20is%20a%20commonly%20used%20geography%20unit.%20The%20shapefile%20includes%20the%20most%20current%20Metropolitan%20Council%20Transportation%20Analysis%20Zone%20(TAZ)%202020-2040%20forecasts%20for%20cities%20and%20townships%20within%20the%207-County%20Metropolitan%20Area.) | Downloaded from website and import into ArcGIS Pro |
| 3 | K Mean Clustering Algorithm | Unsupervised Machine Learning Method | N/A | N/A | N/A | N/A |
| 4 | DBSCANE | Clustering method | N/A | N/A | N/A | N/A |

**Input Data**

Travel behavior inventory of household interview survey data is a travel survey of households in the Twin Cities region that has been conducted every other year. The data collection is based on smartphone GPS applications. The data was collected from fall 2010, through spring 2012 Smartphone participants completed a 7-day travel diary. The data file includes Trip, Household, and Person tables.

The traffic analysis zone is a commonly used geography unit. The shapefile includes the most current Metropolitan Council Transportation Analysis Zone (TAZ) 2020-2040 forecasts for cities and townships within the 7-County Metropolitan Area.

Table 2. < Table of Data Downloaded from Websites with Links >

|  |  |  |  |
| --- | --- | --- | --- |
| **#** | **Title** | **Purpose in Analysis** | **Link to Source** |
| 1 | Travel Behavior Inventory Household Interview Survey | Raw input dataset for k mean clustering from MNDOT | [Mn GeoSpatial Commons](https://gisdata.mn.gov/dataset/us-mn-state-metc-society-tbi-home-interview2010) |
| 2 | Traffic Analysis Zoned Data | Raw input dataset as spatial boundary for result verification in ArcGIS Pro | [Mn GeoSpatial Commons](The%20traffic%20analysis%20zone%20is%20a%20commonly%20used%20geography%20unit.%20The%20shapefile%20includes%20the%20most%20current%20Metropolitan%20Council%20Transportation%20Analysis%20Zone%20(TAZ)%202020-2040%20forecasts%20for%20cities%20and%20townships%20within%20the%207-County%20Metropolitan%20Area.) |

**Methods**

The potential interaction is defined and calculated between two persons using space-time coupling constraints, that is, if and only if two persons are present at the same location during the same time can they potentially interact. (Wong & Shaw, 2010) An individual is exposed to interaction with others in each traffic analysis zone he or she visits. If person A and person B spend time at two different locations, there is no interaction between them. The interaction starts when they enter the same location at the same time. Person A is in contact with person B at the same location during the same time. The total time of the interaction between these two persons is defined by the overlapping time period that they spend at that location. The interaction stops once they exit the location. (Farber et al., 2015)

Python is used to clean and process the merged table from TBI data. Python code is used to handle the null value and missing data and to identify participants that were presented at the same traffic analysis zone at the same time and then extract their activity information into new tables.

For now, activity information includes the trip ID, person ID, household ID, start time and end time of the trip, latitude, and longitude of the destination of an activity.

After the latitude and longitude data is ready in CSV format, the k means clustering algorithm will be applied to the data and generate a clustering pattern graph. DBSCAN will then be applied to the data again to compare the clustering pattern results from the k mean clustering algorithm.

The general methods are summarized into Data Flow Diagram(Figures 1)

**Diagram

Description automatically generated**

Figure 1. Data flow diagram

**K mean pseudo code:**

import libraries

load csv

create elbow curve

plot elbow curve

define k mean parameter (randomly initialize k centroids)

assign each point to its closet centroid

compute the new centroid of each cluster (repeat until the centroid position do not change)

visualize the results

**DBSCAN pseudo code:**

import libraries

load csv

define epsilon parameter (max distance that points can be from each other to be considered a cluster) and min points (minimum cluster size)

start with arbitrary point that has not been visited

retried this points neighborhood and

if it contains sufficient points(equal or greater than min points)

a cluster is started

otherwise

the point is labeled as noise(might later be found as part of other cluster)

visualize the results

**(Preliminary)Results**

This section includes (very rough and almost incorrect) preliminary results (Figures 2, 3, and 4) from k mean and DBSCAN which are applied to the origin location of the trip data directly from TBI.

The final results will come from these two methods applied to the destination location of an interactive activity which comes from the data that is processed and extracted using python code. (This part has not been finished)Chart, line chart

Description automatically generated

Figure 2. Elbow Curve (to select the optimal number of cluster) of Trip Origin

The line levels off after 3 clusters which means that addition of more clusters will not be helpful, so the optimal number of clusters is 3 based on Figure 2.

Chart, scatter chart

Description automatically generated

Figure 3. Preliminary results after running k mean algorithm on the GPS data of trip origin location with k = 3 based on Figure 2

Chart, scatter chart

Description automatically generated

Figure 4. Preliminary (reduced) results after running DBSCAN on the GPS data of trip origin location

Figure 4 is the reduced results of DBSCAN which only centroids of each clusters are plotted.

**Results Verification**

The data will be imported and mapped in ArcGIS Pro and a clustering analysis tool such as Hot Spot analysis will be run to generate a clustering pattern which will be compared with the results generated in Python. The results will also be compared visually for result verification.

**Discussion and Conclusion**

K means clustering algorithm and DBSCAN works differently to find clustering patterns.

Kmeans algorithm tries to partition the dataset into k pre-defined clusters where each point will be assigned to only one cluster. (Pietro, 2021) It tries to make the intra cluster data as similar as possible by assigning data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. (Kurniawan, 2021)

Density-Based Clustering identifies distinctive groups or clusters in the data based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. DBSCAN finds core samples of high density and expands clusters from them. It groups the densely clustered data points into a single cluster, and it can identify clusters in large spatial datasets by examining the local density of the data point. It requires a function to calculate the distance between values and guidance for what amount of distance is considered a cluster. (Boeing, 2018)

K means clustering algorithm is likely not ideal for spatial data because it minimizes the variance so that it does not work well with too many outliers, and it is sensitive to the number of clusters k that are defined in advance. (Baruah, 2021) Even with the help from the elbow curve, the number of clusters still seems odd for the trip origin, and it could be hard to determine a reasonable k value in advance. Unlike k mean, DBSCAN does not require the user to define the number of clusters in advance and it determines them automatically within the algorithm. DBSCAN is also better at detecting outliers. (Magiya, 2019) Based on the preliminary results, DBSCAN already detects more clusters than the k-mean clustering algorithm. (The discussion/conclusion of the final results is not included because it has not been finished yet)

After the k means algorithm and DBSCAN method, the latitude, and longitude coordinates of interaction location, trip origins and destination will be grouped into clusters that help to describe the data and find the spatial clustering pattern. The results from these two methods will be helpful to compare the performance of these two different methods and the results will help to analyze the respondent's travel behavior as to which locations are more likely to have interactions between households based on their origins and destination location.

**References**

Baruah, I. D. (2021, July 21). *K-means, DBSCAN, GMM, agglomerative clustering - mastering the popular models in a segmentation...* Medium. Retrieved November 9, 2021, from https://towardsdatascience.com/k-means-dbscan-gmm-agglomerative-clustering-mastering-the-popular-models-in-a-segmentation-c891a3818e29.

Boeing, G. (2018, March 22). *Clustering to reduce spatial data set size*. Geoff Boeing. Retrieved November 9, 2021, from https://geoffboeing.com/2014/08/clustering-to-reduce-spatial-data-set-size/.

Farber, S., O'Kelly, M., Miller, H. J., & Neutens, T. (2015). Measuring segregation using patterns of daily travel behavior: A social interaction based model of exposure. *Journal of Transport Geography*, *49*, 26–38. https://doi.org/10.1016/j.jtrangeo.2015.10.009

Kurniawan, T. B. (2021, April 2). *The clustering algorithm with geolocation data*. Medium. Retrieved November 9, 2021, from https://medium.com/thelorry-product-tech-data/the-clustering-algorithm-with-geolocation-data-d6dd07ed36a.

Magiya, J. (2019, November 23). *Clustering GPS coordinates and forming regions with python*. Medium. Retrieved November 9, 2021, from https://levelup.gitconnected.com/clustering-gps-co-ordinates-forming-regions-4f50caa7e4a1.

Pietro, M. D. (2021, August 28). *Clustering Geospatial Data*. Medium. Retrieved November 9, 2021, from https://towardsdatascience.com/clustering-geospatial-data-f0584f0b04ec.

Scikit Learn (2007-2021). Retrieved September 29, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.

Sharma, P. (2021). *K means Clustering simplified in Python: K Means Algorithm*. Analytics Vidhya. Retrieved September 29, 2021, from https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/.

Wong, D. W., & Shaw, S.-L. (2010). Measuring segregation: an activity space approach. *Journal of Geographical Systems*, *13*(2), 127–145. https://doi.org/10.1007/s10109-010-0112-x

**Self-score**

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Description** | **Points Possible** | **Score** |
| **Structural Elements** | All elements of a lab report are included **(2 points each)**:  Title, Notice: Dr. Bryan Runck, Author, Project Repository, Date, Abstract, Problem Statement, Input Data w/ tables, Methods w/ Data, Flow Diagrams, Results, Results Verification, Discussion and Conclusion, References in common format, Self-score | 28 | 28 |
| **Clarity of Content** | Each element above is executed at a professional level so that someone can understand the goal, data, methods, results, and their validity and implications in a 5 minute reading at a cursory-level, and in a 30 minute meeting at a deep level **(12 points)**. There is a clear connection from data to results to discussion and conclusion **(12 points)**. | 24 | 24 |
| **Reproducibility** | Results are completely reproducible by someone with basic GIS training. There is no ambiguity in data flow or rationale for data operations. Every step is documented and justified. | 28 | 28 |
| **Verification** | Results are correct in that they have been verified in comparison to some standard. The standard is clearly stated **(10 points)**, the method of comparison is clearly stated **(5 points)**, and the result of verification is clearly stated **(5 points)**. | 20 | 20 |
|  |  | 100 | 100 |