Journal of Urban and Regional Analysis 15–2 (2023): 221–237, https://doi.org/10.37043/JURA.2023.15.2.3

CLUSTERING AND MULTINOMIAL LOGIT ANALYSIS OF FACTORS INFLUENCING HOUSEHOLD RESIDENTIAL LOCATION CHOICE IN THE WASHINGTON METROPOLITAN AREA

Hamid Mirzahossein[a], **Ali Bakhtiari**[a], **Mahdi Nosrati**[a], **Xia Jin**[b]

[a] Imam Khomeini International University, Qazvin, Iran

[b] Florida International University, Miami, USA

Abstract: The vast range of location alternatives and the preference heterogeneity have made it challenging to model, analyse, and predict the location choice process. In this study, we propose a two-step analytical model to focus on lowering the magnitude of these choices. The Transportation Planning Board's 2007-2008 household survey was used in the Washington metropolitan area, consisting of 3722 Traffic Analysis Zones (TAZ). First, location choice alternatives were clustered based on TAZs into homogeneous groups. These TAZs were categorised based on accessibility to public transport, population density, and employment density. Then, the Multinomial Logit (MNL) model was employed to allow the interpretation of the relationship between the clustered areas and the socio-economic characteristics. Four clustering algorithms were compared in terms of efficiency, and the mini-batch k-means performed the best based on the silhouette coefficient. Overall, households tend to prefer suburban areas as household size and the number of owned vehicles increase. Urban areas were selected with an increase in income, number of household workers, number of unemployed looking for a job, number of part-time employees, number of retirees, and the presence of university students. This paper contributes to the current trend of using unsupervised algorithms in the urban planning literature.

Keywords:

residential location choice;
clustering;
mini-batch k-means;
discrete choice

Email: mirzahossein@eng.ikiu.ac.ir

Initial submission: 03.02.2023; Revised submission: 13.06.2023; Final acceptance: 13.07.2023

Introduction

Transportation and the built environment are directly and bilaterally associated (Schwanen and Mokhtarian 2007). Land use patterns and availability of transportation facilities create mobility, and transportation as the primary mobility source can lead to land-use transformation. The residence of households directly impacts their travel behaviour and daily activities (Næss et al. 2018). Various criteria can influence location choices, such as density, accessibility, and distance to public transportation (Cox and Hurtubia 2022).

One of the challenges among most of the studies conducted in this field is the vast number of residence alternatives (Heldt et al. 2014, Zolfaghari 2013). Based on the size of the choice span, household options can range from hundreds to hundreds of thousands. Neighbourhoods, traffic analysis zones, parcels, and buildings are a few examples of different choice scales for Residential Location Choice (RLC) models. Such humongous choice sets make it challenging for urban planners to model and analyse residential location choices. Some studies were based on the property of an alternative or on selecting random subsets of all available options (Lee and Waddell 2010). Another substitute solution is dividing a city into contrasting subregions, then modelling and analysing the output (Bagley and Mokhtarian 1999). Clustering can group the wide range of alternatives into small categories and thus address the ample choice availability (Saxena et al. 2017). With clustering, urban and transportation planners' modelling and interpretation can be facilitated.

Discrete choice models, especially the Multinomial Logit (MNL), have been utilised to model and analyse the factors influencing RLC. MNL model does not provide accurate results considering the unobserved preference heterogeneity solely. The Mixed Logit Model (MXL) with random parameters or Latent Class Analysis (LCA) has been used to resolve preference heterogeneity problems (Habib and Miller 2009, Liao et al. 2015, Lee et al. 2019). This paper identifies and analyses household features and socio-economic factors affecting RLC and it aims to cluster alternatives into homogeneous differential groups.

In previous studies, the dominant factors affecting RLC have been studied in four general categories: (1) residence characteristics (Lee et al. 2019); (2) residence vicinity (Cao 2008); (3) lifestyle and attitude criteria (Smith and Olaru 2013); and (4) socioeconomic features (Gurrutxaga 2023). The extensive list of available choices and preferences heterogeneity are the main challenges in estimating residential household demand. In this section, some related works in this field are reviewed and sequenced from 1959 to 2019.

After World War II, as income and household size increased, households' preferences changed in favour of living in low-density neighbourhoods. The impact of household

characteristics, such as economic, social, and demographic factors, outweighed transportation and public service issues (Weisbrod et al. 1980). RLC modelling was first proposed by McFadden (1978), who introduced an MNL for this purpose. In the proposed model, the consumer was assumed to select the most desirable alternative while examining the features of the available location options. The influential parts of RLC were divided into two classes: (1) features related to the house itself; and (2) features regarding the house vicinity, such as accessibility to public services. The vast availability of available locations was a challenge in modelling location choice.

Hansen (1959) introduced public services availability and the potential productivity of land use based on residential land-use models. Results suggested that as accessibility and mobility expand, the region's growth potential increases consequently. With further distance from the urban area due to fewer job opportunities and higher transportation costs, the land value decreases, and peripheral areas are less attractive for households (Alonso 1964).

Household characteristics, lifestyle, and attitudes toward residence choice can be analysed using the binary logit model, considering household characteristics in central and suburban areas. Attitudinal criteria such as transportation pricing policies and lifestyle criteria, such as participation in cultural events, differ significantly in central and suburban residents, and residence choice using the features observed at the individual level causes errors. It should be done more significantly (Bagley and Mokhtarian 1999). On the one hand, socio-economic factors play a pivotal role in relocation; for example, a shift in the number of people or a change in a household occupation creates a powerful incentive to relocate (Clark and Davies Withers 1999). On the other hand, travel time and commuting expenses impel households to relocate (Van Ommeren et al. 1999).

The nested logit model is another method to identify and analyse the factors influencing RLC. Using the nested logit model, Kim et al. (2005) assessed the impact of transportation and amenities in the UK Oxford area on the location decision-making process. The results showed that a combination of factors such as shorter travel time, lower transportation costs, lower density, and higher school quality determines the preference of individuals in RLC. Travel time, housing rent, household income, and neighbourhood conditions such as air pollution, crime, and distance to school are the most influential variables in modelling RLC (Molugaram and Rao 2005).

Many aspects of RLC cannot be explained by modelling with observed variables. Combining socio-economic factors as observed variables and lifestyle as an unobserved variable makes it possible to identify RLC preferences more insightfully. In this regard, based on lifestyle and employment of the latent class model, Levine and Frank (2007) categorised the city into three homogeneous groups, including city dwellers, suburb dwellers, and public transportation users, and they identified the

impact of socio-economic variables in each category.

Population and employment density are the main variables for separating neighbourhoods, and household income is the pivotal factor in RLC (Bhat and Guo 2007). Heterogeneity in data is one of the many formidable challenges in estimating housing demand. An MXL framework can be used to minimise the unobserved heterogeneity in RLC preferences; as the house unit price rises, the possibility of choosing that alternative decreases, regardless of the household income (Habib and Miller 2009). Marcucci et al. (2011) used the MXL and the MNL to perform an analysis and comparison to highlight the role of every household member in RLC, finding that the wife, adolescents, and the husband are respectively the most influential in the household preference on residence selection. Adolescents are more sensitive to accessibility, noise, and losing their current position.

The set of household choices based on the average commuting distance could create an accurate selection set and lead to a more efficient model without bias (Rashidi et al. 2012). More accurate results are typically attained by classifying choices into homogeneous subsets and by applying the LCA (Liao et al. 2015). Likewise, Ardeshiri and Vij (2019) employed the LCA to label households as six classes based on different preferences, residence characteristics, and household characteristics. They found that high-income immigrants and white families live more in the suburbs, and suburban households depend more on private cars.

To address the problem of choice vastness observed in previous studies, and consequently the modelling complications, the present study aims to develop a cluster-based multinomial logit model to limit the number of choices available and to provide interpretability at the TAZ scale. The results of the model are expected to allow the analysis of similar TAZ, even though they might be scattered in the studied area, and to scrutinise how the socioeconomic features of the household impact their choice to live in a TAZ. In order to achieve this aim, first, the location preferences were clustered based on the zonal percent-walk-to-transit (PWT), population density, and employment density into homogeneous groups; next, the MNL was employed to interpret the results. The results of this model can be used as a tool in urban planning, specifically for the classic problem of location choice.

Methodology

Data

The household data used for modelling is the transportation planning board transportation survey (TPB) 2007-2008. The Transportation Planning Board conducts periodic surveys in the Washington metropolitan area to collect data on demographic information and travel behaviour. This area encompasses Washington DC and several

US states, including Virginia, Maryland, and West Virginia encoded at the TAZ level. The survey data includes more than 11,000 household records, 25,000 individual records, 16,000 vehicle records, and 130,000 travel records. The data was pre-processed to extract the features for the modelling. Then each feature was scaled in the range of 0 to 1. Table 1 shows a description of the features used for this study.

Table 1. Household feature description

Variable	Description
hhsiz	Household size
hhwrk	Number of workers
hhveh	Number of vehicles
	Low<= 50000 $
incom	Mid 50000 $ 100000$
	High >= 100000 $
bikes	Number of bikes
has_wk_at_home	The household has people working at home (if any = 1 else =0)
n_retired	Number of retirees (numeric value)
unempl_look_job	Number of unemployed people looking for a job (numeric value)
has_uni_stu	The family has at least one university student (if any = 1 else =0)
n_part_time_worker	Number of part-time employees (numeric value)

TAZ Feature Data

There are a total of 3722 TAZs in the Washington metropolitan area. Table 2 shows the TAZ features considered for clustering in this study, including access to public transportation, total population, total employment, population density, employment density, and employment numbers concerning job categories.

Table 2. TAZ features considered for clustering

Variable	Description
MTLRTSHR	The zonal percent walk to transit (PWT) within a short (0.5 mile) walk of Metrorail or LRT service
MTLRTLNG	The zonal percent walk to transit (PWT) within a long (1.0 mile) walk of Metrorail or LRT service
ALLPKSHR	The zonal percent walk to transit (PWT) within a short (0.5 mile) walk of any transit service (including Metrorail and LRT) in the AM peak period
ALLPKLNG	The zonal percent walk to transit (PWT) within a long (1 mile) walk of any transit service (including Metrorail and LRT) in the AM peak period
ALLOPSHR	The zonal percent walk to transit (PWT) within a short (0.5 mile) walk of any transit service (including Metrorail and LRT) in the off-peak period
ALLOPLNG	The zonal percent walk to transit (PWT) within a long (1 mile) walk of any transit service (including Metrorail and LRT) in the off-peak period
Popden	Population density
Empden	Employment density

Analysis framework

The methodology of this analysis is a combination of unsupervised machine learning methods and discrete choice models (Figure 1). The entire Washington metropolitan area consists of 3722 TAZs. Once TAZs were clustered based on accessibility to public transportation, population densities, and employment densities, each sample's corresponding cluster coupled with demographic features were fed into MNL to allow the interpretation of the reasons for the households' choices.

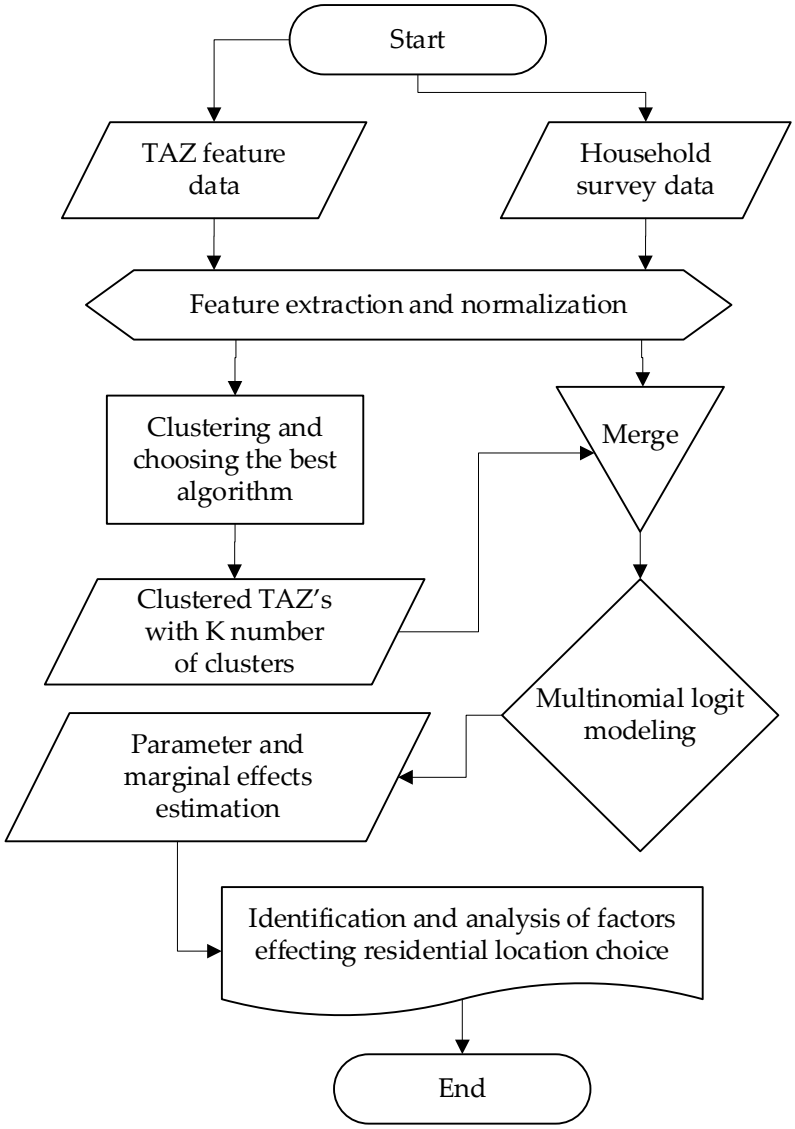

Figure 1. Analysis framework

Clustering

Clustering is an unsupervised learning method that identifies natural clusters in multidimensional data into a chosen number of categories, where samples within each group (cluster) are akin compared to other groups (Jain et al. 1999). With the aid of clustering algorithms, a handful of TAZ clusters with the most similar features can be achieved instead of dealing with thousands of alternatives (in this study, traffic analysis zones). TAZs can be distinguished based on accessibility, land use, employment density, or other features. TAZs were clustered based on the zonal PWT, population density, and employment density. Similarities and distinguishing differences in these features were discovered by comparing the results of clusters. Data visualisation and numerical observation of statistical descriptions made the cluster comparison possible.

The TAZ data was clustered using four clustering algorithms. The algorithms employed and compared in this paper comprise Birch (Zhang et al. 1996), Agglomerative (Ackermann et al. 2014), Spectral (Ng et al. 2001), and mini-batch K-means (Sculley 2010). The K-means algorithm tends to be the most popular clustering algorithm. The objective of k-means optimization is to find the set of C of cluster centres $\in R^m$, given $|C| = k$ to minimise a set of X of examples by the objective function:

$$\sum_{x \in X} \|f(C, x) - x\|^2 \quad (1)$$

where $f(C, x)$ is the function to return the nearest cluster centroid $c \in C$ to x using Euclidean distance (Sculley 2010).

After clustering with different algorithms, the optimised number of clusters was found, along with the best-performing algorithm. The silhouette coefficient was used as the evaluation metric to make this possible (Rousseeuw 1987). The silhouette coefficient is calculated from:

$$Silhouette\ Score = \frac{(b(i) - a(i))}{max(a(i), b(i))} \quad (2)$$

where here $b(i)$ is the minimum distance between recognized data patterns i from the selected features and patterns in all the other dissimilar clusters not containing pattern i, and $a(i)$ is the mean distance between the data pattern and every other pattern in the same cluster (Zhou and Gao 2014). The best value for the silhouette coefficient is 1, and the worst is -1. If one is obtained for this coefficient, clusters are distinctly separated, while when this coefficient is 0, clusters are indistinguishable, or the distance between clusters is insignificant. If this coefficient is -1, it can be concluded that the clusters are wrongly determined.

Multinomial Logit Model

Based on discrete choice modelling, households choose the alternative with the highest utility. In the MNL, alternative utility is defined as a function of the factors influencing RLC. This model suits the situation where the purpose is to predict the residence location of households concerning the factors affecting it. The utility of the household RLC is defined as:

$$U_{ij} = \beta X_{ij} + \varepsilon_{ij} \quad (3)$$

where the utility function U_{ij} is created for the respondent i who selects alternative j. β is the coefficient of the descriptive variable and ε_{ij} is the unobserved error term of the utility function. The probability that the respondent i selects alternative j is defined with the following function:

$$P_{ij} = \frac{exp\ exp(\beta X_{ij})}{\sum_{i=1}^{j} exp\ exp(\beta X_{ij})} \quad (4)$$

The use of multinomial logit in our study has one limitation. The alternatives of the model are selected by a clustering algorithm, and because the clustering algorithms are expected to work based on the similarity of features, this may create some uncertainty about how the identified groups relate to socioeconomic features in the results of the multinomial logit. To minimise this issue, we used a range of clustering algorithms and evaluation criteria to choose the best possible clustering model in the process.

Results

Clustering Comparison

Figure 2 compares the silhouette coefficient from the four clustering algorithms and four choices for the number of clusters (two to five). The mini-batch K-means algorithm had the highest silhouette coefficient value in all cluster numbers except for three, where it performed only better than the agglomerative algorithm. In dividing TAZs into three clusters, Spectral had the highest silhouette coefficient. The agglomerative algorithm showed the worst performance, and its silhouette coefficients were minimal for all numbers of clusters. Finally, the Birch algorithm had a moderate performance compared to other algorithms.

The two clusters and mini-batch K-means algorithm had the most differentiation and the most significant silhouette coefficient. Therefore, they were used for the final modelling.

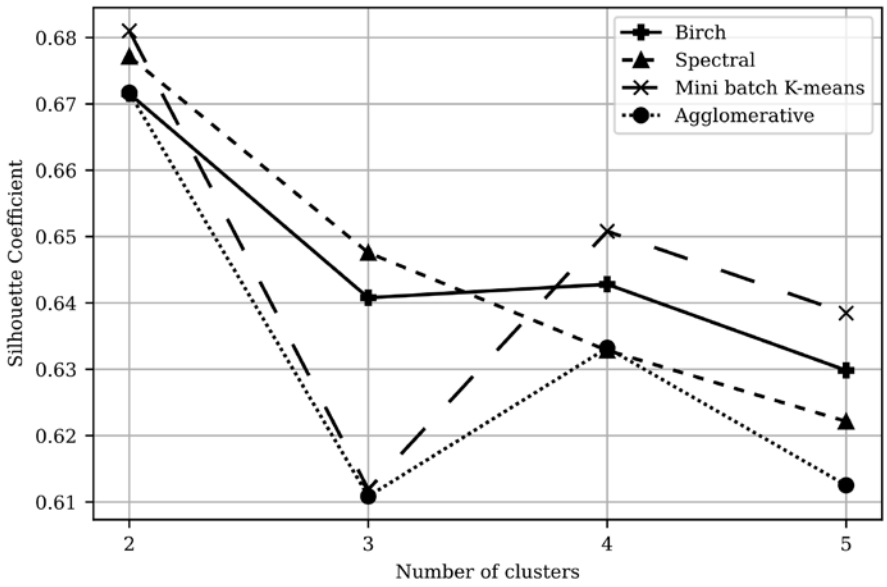

Figure 2. Performance comparison of different clustering algorithms

Analysis of the Clusters

The characteristics of each cluster were discovered by data visualisation and statistical description. Employment-related parameters and population density were compared and discussed.

Accessibility to public transport in cluster 2 is much easier than it is in cluster 1. To illustrate this, Figure 3 compares the two clusters for a public transport accessibility parameter. Cluster 1 (58.03% of the TAZs) includes peripheral and less developed areas, whereas cluster 2 (41.97% of the TAZs) covers more central and developed areas. Access to public transport in cluster 1 is less than in cluster 2. In addition to the example parameter, a similar difference was observed in other parameters related to public transport accessibility.

Comparing the values in Table 3, it is noticeable that the mean of total employment in cluster 1 is close to half of the mean for the same feature in cluster 2. Population and employment in TAZs of cluster 2 are by a substantial amount more densely distributed.

The difference between clusters is also comparable, pondering the redundancy of each employment category. It is assumed that industrial jobs are located in the suburbs, and office employment lies closer to or in the urban area. The average industrial employment in cluster 1 is higher than in cluster 2, and office employment in cluster 1 is much lower than in cluster 2. Given that these variables were not considered in the clustering process, this confirms cluster 1 as the suburbs and cluster 2 as the urban areas.

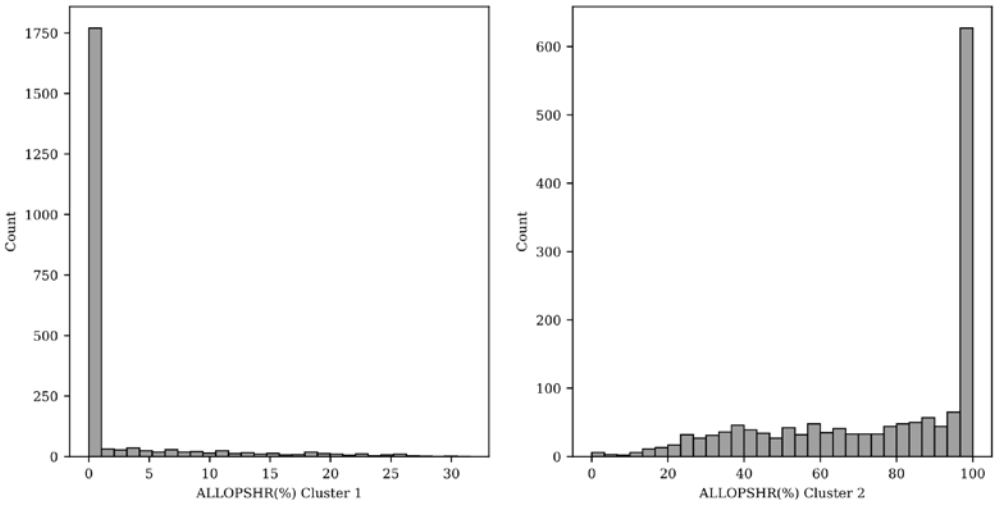

Figure 3. Comparing value counts of ALLOPSHR (the zonal percent walk to transit – PWT within a short, 0.5 mile, walk of any transit service in the off-peak period) for the 2 clusters

Table 3. The statistical description of employment and population characteristics in clusters

	Cluster 1			Cluster 2		
	mean	Std.	max	mean	Std.	max
Total Employment (Person)	788.72	1902.36	42638	1343.339	2508.41	22290
Industrial Employment (Person)	161.49	541.19	14111	127.2644	334.5192	4103
Retail Employment (Person)	148.41	360.02	4695	220.6159	411.3774	4873
Office Employment (Person)	306.71	936.8	20298	725.6876	1903.869	17639
Other Employment (Person)	172.1	517.02	14110	269.7714	782.4411	18097
Population Density	1.5169	2.7899	55.869	6.65616	8.48945	77.5862
Employment Density	1.189308	4.913772	135.2821	11.2768	36.2743	413.754

Multinomial Logit Analysis

As mentioned in the previous sections, cluster 1 is for households with low access to public transportation and living in suburban areas. In this cluster, the positive sign of coefficient β for household size, and the number of vehicles, indicates a positive effect, while the negative sign of coefficient β for other variables indicates their negative impact on selecting this cluster. The results for cluster 1 are reported in Table 4.

The same signs of upper and lower coefficient intervals and the t-test value for all variables, except for the number of household bicycles and the presence of a family member working at home, show the significance of the variables. The parameter of descriptive variables can only express its positive or negative impact on the likelihood

of choosing an alternative. Therefore, the marginal effects analysis is required to analyse the variables affecting RLC.

Table 4. Estimation of MNL for cluster 1 (base cluster = cluster 2)

Variable Name	Coefficient	Std. Err.	t	95% Conf.	
hhsiz ***Interval***	0.4462516	0.0939333	4.75	0.2621457	**0.6303574**
hhwrk	-0.5651202	0.0965597	-5.85	-0.7543737	**-0.3758666**
hhveh	0.941377	0.0307236	30.64	0.8811598	**1.001594**
bikes	-0.0236511	0.0169711	-1.39	-0.0569138	**0.0096116**
n_part_time_worker	-0.2546975	0.0908127	-2.80	-0.4326871	**-0.0767079**
n_retired	-0.0974938	0.0457508	-2.13	-0.1871638	**-0.0078238**
income_1	0.5707335	0.0764166	7.47	0.4209598	**0.7205072**
income_2	0.4671024	0.0656184	7.12	0.3384927	**0.595712**
income_3	0.3262629	0.0657375	4.96	0.1974198	**0.455106**
unempl_look_job	-0.4567923	0.1218809	-3.75	-0.6956744	**-0.2179101**
has_uni_stu	-0.8604996	0.1292721	-6.66	-1.113868	**-0.607131**
has_wk_at_home	-0.0436464	0.0698691	-0.62	-0.1805873	**0.0932945**
ASC	-2.140847	0.0897519	-23.85	-2.316758	**-1.964937**

Marginal Effects Analysis

The descriptive variables of the utility function were estimated. The marginal effects were calculated in percentages to determine how much the probability of selecting each group differs with one more unit of another variable. Table 5 shows the average marginal effects for cluster 1 (households in the suburbs with low access to public transportation). Adding one person or vehicle to the household, choosing cluster 1 increases by 11% and 23%, respectively. A decrease in income increases the likelihood of selecting suburban areas. Adding a vehicle to the household is the most influential factor in choosing a suburban residence.

Cluster 2 embodies households with better access to public transportation. The marginal effects of cluster 2 variables, such as household size, number of vehicles, and household income, are identical to cluster 1 with only the opposite sign.

With the addition of an employed member or a member who has a part-time job to the household, the probability of selecting the urban area increases by 14% and 6%, respectively. Moreover, adding an unemployed member looking for a job to the household increases the likelihood of choosing the urban area.

The presence of students in the household has the most significant impact (21%) on choosing the urban area. The presence of retirees in the household effectively increases the probability of choosing cluster 2 (2%). The number of bicycles and the number of members working at home have no effect on RLC in the studied case.

Table 5. The average marginal effects for cluster 1

Variable Name	dy/dx	Std. Err.	t	95% Conf. Interval	
hhsiz	0.1108379	0.0233306	4.75	0.0651107	**0.1565651**
hhwrk	-0.1403619	0.0239849	-5.85	-0.1873714	**-0.0933524**
hhveh	0.2338148	0.0076564	30.54	0.2188086	**0.2488211**
bikes	-0.0058743	0.0042152	-1.39	-0.0141359	**0.0023872**
n_part_time_worker	-0.0632606	0.0225558	-2.80	-0.1074691	**-0.0190521**
n_retired	-0.024215	0.0113643	-2.13	-0.0464886	**-0.0019415**
income_1	0.1417561	0.0189868	7.47	0.1045426	**0.1789696**
income_2	0.1160167	0.0163015	7.12	0.0840664	**0.1479671**
income_3	0.0810357	0.0163288	4.96	0.0490319	**0.1130394**
unempl_look_job	-0.1134559	0.0302713	-3.75	-0.1727866	**-0.0541253**
has_uni_stu	-0.2137269	0.0321071	-6.66	-0.2766557	**-0.1507981**
has_wk_at_home	-0.0108407	0.0173538	-0.62	-0.0448536	**0.0231722**

Discussion

This study aimed to tackle the issue of having too many location choices in previous research, which made modelling complicated. To do this, we developed a cluster-based multinomial logit model that limits the number of choices available and it provides interpretability at the TAZ scale in the Washington metropolitan area. The model's results are expected to allow for the analysis of similar TAZ, even if they are scattered in the studied area, and to examine how household socioeconomic features affect their choice to live in a TAZ. The knowledge of location choice preferences can be a decisive factor in the development of an area, as policymakers and urban planners need to know why socioeconomic groups choose their residence location in order to be able to allocate local and government funding to the building of areas as they find appropriate. Thus, the results of this study can play an important role in the process of urban planning.

The availability of vehicles for households and the lack of dependence on public transportation can influence RLC, meaning that residents buy cars to be able to live in the suburbs. With a rise in the number of owned vehicles, households are more likely to choose a location far from the urban area; and previous studies reached the same conclusion (Schwanen and Mokhtarian 2007, Cao 2008, Ardeshiri and Vij 2019). From the coefficients related to household size and vehicles, it can be deduced that cluster 1 belongs to households far from the urban area with little access to public transportation. The coefficient related to the household size and the number of household vehicles in this cluster is positive. That is, increasing the size of the family and the number of vehicles positively affects the likelihood of choosing cluster 1. The increase in the number of household members also seems to be a factor in the increased

number of vehicles in the household. It can be inferred that as the household size and the number of owned vehicles increase, selecting cluster 2 is less likely.

Income is one of the most critical factors affecting RLC in this study and it has always been a key factor (Weisbrod et al. 1980, Molugaram and Rao 2005). Income coefficients indicate that households choose to move to the urban area as income increases.

The RLC of workers depends mainly on their job location (Clark and Davies Withers 1999, Van Ommeren et al. 1999, Liao et al. 2015, Lee et al. 2019). The number of household workers and the number of people with part-time jobs are among the most critical descriptive variables in RLC. This may be because household workers prefer more affordable travel costs and lower commuting time and distance.

University students choose residences that offer less travel time and expenses because universities are typically located in urban areas. This study, like previous studies, considers travel time and cost as influential factors in choosing a residence (Guo and Bhat 2002, Kim et al. 2005). This result might be due to the educated people's tendency to move from the suburbs to urban areas over time. The same results were found in Costa and Kahn (2000). However, some studies have shown that people with low education tend to live in urban areas (Cao 2008).

The unemployed household members looking for a job probably are more considerate about their transportation expenses due to the absence of income. They are likely to prefer to live in the urban area since there are ample affordable public transportation facilities in developed areas. These people are probably more in need and deprived. Thus, they can benefit from affordable public transportation, accessible market, and amenities.

The marginal effects for the number of retirees in the household show that with the presence of a retiree in the household, the probability of choosing the urban area with better accessibility increases. This could be because retirees are likely to have mobility and physical challenges due to senescence, making them prefer to be closer to amenities such as leisure and shopping centres.

Despite our efforts to conduct solid research, we should acknowledge several limitations. Although the use of clustering algorithms solves the problem of a large number of alternatives, there is no approach to evaluate the accuracy of clustered results. The clustering algorithm groups unlabelled data based on similarity criteria without any prior baseline, so no comparison can be made to determine precision. Future work could experiment with some labelled data and compare the results of this model against some baseline truth. Additionally, the data used in this study belong to an old survey as we did not have access to newer data. The same model should be implemented on more recent data to discover the preference changes in the area.

Conclusions

This study examined the role and effects of factors influencing RLC in the Washington metropolitan area. Households choose their residence location based on environmental criteria such as density, accessibility, and proximity to public transportation. One of the issues that researchers have faced since the beginning of RLC studies is the wide range of possible alternatives for RLC and the heterogeneity in the preferences of households. Discrete choice modelling and reliable estimations are infeasible in the presence of such an extensive number of alternatives. Clustering was performed based on location attribute data, including 3722 traffic analysis zones in the Transportation Board Planning Survey 2007-2008, splitting the data into two distinct categories: central areas of the city with high accessibility to public transport, and suburban areas with low accessibility to public transport. Then, the MNL model was employed to analyse the features that influenced RLC.

The results showed a greater tendency to choose suburban areas with increasing household size and the number of owned vehicles. A growth in income, the number of household workers, the number of unemployed looking for a job, the number of part-time employees, the number of retirees, and the presence of university students leads to a greater desire to choose urban areas.

Since the influential factors in choosing RLC depend on the research case, and since there was a challenge with accessing newly collected data, it is necessary to be careful in generalising the results obtained from this study. Future research could do a more concrete analysis of the factors influencing RLC using a more comprehensive and up-to-date dataset. It is also possible to conduct the clustering of traffic analysis zones based on criteria not considered in this study, such as the area of various land use types, quality and number of educational centres, land value, and level of service.

References

ACKERMANN M. R., BLÖMER J., KUNTZE D., SOHLER C. (2014) Analysis of agglomerative clustering, Algorithmica 69, 184-215, https://doi.org/10.1007/s00453-012-9717-4.

ALONSO W. (1964) Location and land use: toward a general theory of land rent, Harvard University Press, Cambridge, MA.

ARDESHIRI A., VIJ A. (2019) Lifestyles, residential location, and transport mode use: A hierarchical latent class choice model, Transportation Research Part A: Policy and Practice 126, 342-359, https://doi.org/10.1016/j.tra.2019.06.016.

BAGLEY M. N., MOKHTARIAN P. L. (1999) The role of lifestyle and attitudinal characteristics in residential neighborhood choice, UC Berkeley: University of California Transportation Center, Retrieved from: escholarship.org.

BHAT C. R., GUO J. Y. (2007) A comprehensive analysis of built environment characteristics on household residential choice and auto ownership levels, Transportation Research Part B: Methodological 41 (5), 506-526, https://doi.org/10.1016/j.trb.2005.12.005.

CAO X. (2008) Is alternative development undersupplied?: Examination of residential preferences and choices of Northern California movers, Transportation Research Record 2077 (1), 97-105, https://doi.org/10.3141/2077-13.

CLARK W. A. V., DAVIES WITHERS S. (1999) Changing jobs and changing houses: mobility outcomes of employment transitions, Journal of Regional Science 39 (4), 653-673, https://doi.org/10.1111/0022-4146.00154.

OSTA D. L., KAHN M. E. (2000) Power couples: changes in the locational choice of the college educated, 1940-1990, The Quarterly Journal of Economics 115 (4), 1287-1315, https://doi.org/10.1162/003355300555079.

COX T., HURTUBIA R. (2022) Compact development and preferences for social mixing in location choices: Results from revealed preferences in Santiago, Chile, Journal of Regional Science 62 (1), 246-269, https://doi.org/10.1111/jors.12563.

GUO J., BHAT C. (2002) Residential location choice modeling: Accommodating sociodemographic, school quality and accessibility effects, University of Texas, Retrieved from: www.caee.utexas.edu.

GURRUTXAGA M. (2023) A life-stage approach for decomposing spatiotemporal population changes along an urban-rural gradient: implications for regional planning, Geographical Review 113 (1), 134-155, https://doi.org/10.1080/00167428.2021.1906669.

HABIB M. A., MILLER E. J. (2009) Reference-dependent residential location choice model within a relocation context, Transportation Research Record 2133 (1), 92-99, https://doi.org/10.3141/2133-10.

HANSEN W. G. (1959) How accessibility shapes land use, Journal of the American Institute of Planners 25 (2), 73-76, https://doi.org/10.1080/01944365908978307.

HELDT B., GADE K., HEINRICHS D. (2014) Challenges of data requirements for modeling residential location choice: The case of Berlin, Germany, Proceedings of European Transport Conference, 1-17.

JAIN A. K., MURTY M. N., FLYNN P. J. (1999) Data clustering: a review, ACM Computing Surveys 31 (3), 264-323, https://doi.org/10.1145/331499.331504.

KIM J. H., PAGLIARA F., PRESTON J. (2005) The intention to move and residential location choice behaviour, Urban Studies 42 (9), 1621-1636, https://doi.org/10.1080/00420980500185611.

LEE B. H. Y., WADDELL P. (2010) Residential mobility and location choice: a nested logit model with sampling of alternatives, Transportation 37, 587-601, https://doi.org/10.1007/s11116-010-9270-4.

LEE Y., CIRCELLA G., MOKHTARIAN P. L., GUHATHAKURTA S. (2019) Heterogeneous residential preferences among millennials and members of

generation X in California: A latent-class approach, Transportation Research Part D: Transport and Environment 76, 289-304, https://doi.org/10.1016/j.trd.2019.08.001.

LEVINE J., FRANK L. D. (2007) Transportation and land-use preferences and residents' neighborhood choices: the sufficiency of compact development in the Atlanta region, Transportation 34, 255-274, https://doi.org/10.1007/s11116-006-9104-6.

LIAO F. H., FARBER S., EWING R. (2015) Compact development and preference heterogeneity in residential location choice behaviour: A latent class analysis, Urban Studies 52 (2), 314-337, https://doi.org/10.1177/0042098014527138.

MARCUCCI E., STATHOPOULOS A., ROTARIS L., DANIELIS R. (2011) Comparing single and joint preferences: a choice experiment on residential location in three-member households, Environment and Planning A: Economy and Space 43 (5), 1209-1225, https://doi.org/10.1068/a43344.

MCFADDEN D. (1978) Modeling the choice of residential location, Transportation Research Record 673, 72-77.

MOLUGARAM K., RAO K. V. (2005) A stated preference residential location choice model in [the] Indian context, Proceedings of the Australian Transport Research Forum 28, 1-17.

NÆSS P., PETERS S., STEFANSDOTTIR H., STRAND A. (2018) Causality, not just correlation: Residential location, transport rationales and travel behavior across metropolitan contexts, Journal of Transport Geography 69, 181-195, https://doi.org/10.1016/j.jtrangeo.2018.04.003.

NG A. Y., JORDAN M. I., WEISS Y. (2001) On spectral clustering: Analysis and an algorithm, NIPS'01: Proceedings of the 14th International Conference on Advances in Neural Information Processing Systems, 849-856.

RASHIDI T. H., AULD J., MOHAMMADIAN A. K. (2012) A behavioral housing search model: Two-stage hazard-based and multinomial logit approach to choice-set formation and location selection, Transportation Research Part A: Policy and Practice 46 (7), 1097-1107, https://doi.org/10.1016/j.tra.2012.01.007.

ROUSSEEUW P. J. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20, 53-65, https://doi.org/10.1016/0377-0427(87)90125-7.

SAXENA A., PRASAD M., GUPTA A., BHARILL N., PATEL O. P., TIWARI A., ER M. J., WEIPING D., CHIN-TENG L. (2017) A review of clustering techniques and developments, Neurocomputing 267, 664-681, https://doi.org/10.1016/j.neucom.2017.06.053.

SCHWANEN T., MOKHTARIAN P. L. (2007) Attitudes toward travel and land use and choice of residential neighborhood type: Evidence from the San Francisco bay area, Housing Policy Debate 18 (1), 171-207, https://doi.org/10.1080/10511482.2007.9521598.

SCULLEY D. (2010) Web-scale k-means clustering, WWW '10: Proceedings of the 19th international conference on World wide web, 1177-1178, https://doi.org/10.1145/1772690.1772862.

SMITH B., OLARU D. (2013) Lifecycle stages and residential location choice in the presence of latent preference heterogeneity, Environment and Planning A: Economy and Space 45 (10), 2495-2514, https://doi.org/10.1068/a45490.

VAN OMMEREN J., RIETVELD P., NIJKAMP P. (1999) Job moving, residential moving, and commuting: a search perspective, Journal of Urban Economics 46 (2), 230-253, https://doi.org/10.1006/juec.1998.2120.

WEISBROD G., BEN-AKIVA M., LERMAN S. (1980) Tradeoffs in residential location decisions: Transportation versus other factors, Transport Policy and Decision Making 1 (1), 13-26.

ZHANG T., RAMAKRISHNAN R., LIVNY M. (1996) BIRCH: an efficient data clustering method for very large databases, ACM Sigmod Record 25 (2), 103-114, https://doi.org/10.1145/235968.233324.

ZHOU H. B., GAO J. T. (2014) Automatic method for determining cluster number based on silhouette coefficient, Advanced Materials Research 951, 227-230, https://doi.org/10.4028/www.scientific.net/AMR.951.227.

ZOLFAGHARI A. (2013) Methodological and empirical challenges in modeling residential location choices, Imperial College London, Retrieved from: spiral.imperial.ac.uk, https://doi.org/10.25560/12565.