

Developing a Parsimonious Classification of Traffic Analysis Zones Using a Large Number of Accessibility Indicators and Transportation Level of Service

Konstadinos G. Goulias(✉)

University of California, Santa Barbara, CA 93106, USA
kgoulias@ucsb.edu

Abstract. Accessibility, the ease or difficulty with which activity opportunities can be reached from a given location, can be measured using the cumulative number of opportunities from an origin within a given amount of travel time and/or distance. Estimating accessibility indicators is used to show the level of service quality offered by a transportation system and as explanatory factors of travel behavior equations in transportation modeling and simulation. With the availability of large quantities of data, detailed representation of transportation networks (a set of nodes and links forming a graph), and level of service often measured in movement speed on network links, analysts can compute hundreds of accessibility indicators. Reporting of all these indicators is tedious, cumbersome, and too complicated. In this paper, the combination of principal component analysis to extract summaries of service quality from 190 indicators is demonstrated. Each of these indicators is computed for 1822 traffic analysis zones (TAZ - a convenient geographic subdivision used to model transportation systems) covering the entire country of Qatar in the Arabian Peninsula. Then, a small number of principal components (e.g., 27 components capturing 91.6% of the variance of 190 variables) are used to classify each TAZ in a group with clearly identifiable characteristics using hierarchical clustering. In this paper experiments with different subsets of the 190 indicators are also used to show how one can distinguish among different levels of service offered at each of the TAZs analyzed by time of day and means of travel.

Keywords: Accessibility · Opportunities · Principal Components · Clustering

1 Introduction

Accessibility is the ease (or difficulty) with which activity opportunities may be reached from a given location using one or more means of travel (transportation modes). Defined in this way, accessibility indicators (or measures) incorporate the spatial distribution of land-use activities and the performance of a transportation system within a given area. Accessibility indicators are needed for a variety of regional planning and modeling purposes, such as characterizing quality of life [1], describing transportation quality of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14111, pp. 418–436, 2023.
https://doi.org/10.1007/978-3-031-37126-4_28

service [2], measuring jobs and housing lack of balance [3], as a tool to assess inequity in spatial organization of activities and transportation [4], the dynamics of inequity evolution [5], and the study of reachable opportunities using non-motorized modes such as walking and bicycling [6]. A review of many different accessibility estimation methods employed in the past to serve a variety of purposes can be found in [7].

In this paper emphasis is given to accessibility indicators using the idea of "intervening opportunities." In the context of developing a large-scale activity-based microsimulation model for travel demand forecasting [8], accessibility indicators are computed by enumerating and cumulating the opportunities that are found within the area covered by a certain travel time from an origin point (in essence building buffers around points in space). This formulation is particularly intuitive, because it allows one to identify different travel time thresholds (e.g., 10 min, 20 min, and 50 min) and create geographic "buffers" within which activity opportunities that can be reached are counted. Travel time is measured using the speed of travel on the network by any mode of travel. This accessibility formulation is both a function of land use patterns and the performance of the transportation system [9]. The main application envisioned for the techniques presented here is functional classification of hierarchically classified zones to distinguish the center of a region or city from suburbs, exurbs, and rural geographies to then be used in simulation [10]. In this paper a new method is presented that first develops indicators similar to [9] that were developed for Southern California and expanded to include additional level of service data to achieve a more comprehensive representation of opportunities and level of service.

The geographic area of interest in this paper is the State of Qatar. However, the methods developed and illustrated use data that are often available from agencies that routinely provide information about the spatial distribution of employed persons by industry type and the transportation network that serves them. The unique aspect of our current method is that "static" sources of data are combined with secondary data on the temporal (within a day) availability of opportunities and the condition of the transportation network to develop accessibility indicators that vary by time of day. One outcome of the research here is a set of time-sequenced maps of activity opportunity availability and a classification of the TAZs by their access to these opportunities that are then used to create the final zonal classification. To achieve this, network and socioeconomic data at a fine level are prepared, and a methodology that integrates spatial/temporal factors reflecting the variation in available opportunities and travel time during the day is created. To the best of my knowledge this is the first time an analysis using this type of data and this type of zonal classification has ever been tested.

Section 2 describes the data used in the current research effort and outlines data processing and preparation details. Section 3 discusses the methodology for extracting the variance using principal components from accessibility indicators and then through hierarchical clustering classify geographical areas that are mapped for comparisons. Section 4 presents the analysis and results, and the last section provides a brief summary and next steps.

2 Data Used Here

The data used to develop traffic analysis zone classification are the accessibility indicators that are computed for each zone using a variety of databases available by different agencies in Qatar [11] aiming at the provision of accessibility indicators to be used in the models described in [12]. A first set of attributes are assigned to each of the 1822 zones and they are:

1. Employment by type from values in a land use database
2. Dwelling units by type from an inventory in parcel data
3. Lane kilometers by roadway type from a network database
4. Transit stops from an inventory in a shape file
5. Parking locations and size from an inventory with estimates from buildings and roadways
6. School types and capacity from an inventory of schools by type
7. TAZ area - computed from a GIS shape file

In this paper focus is on employment, travel time by roadway and public transportation, travel distance by roadway and public transportation, and the area of each TAZ. Travel times by car and public transportation are the outputs from a transportation simulation. The attributes are also divided into three groups depending on time of day corresponding to approximately morning peak travel period (AM peak from 6:00 to 8:59), midday (Midday 12:00 to 14:59), and after noon peak period (PM peak 17:00 to 19:59). Opportunities change by time of day due to opening and closing hours of stores and/or other restrictions of operations. In this application information was used from place-based diaries of households (attempting to interview every person in a household) and immigrant workers (called laborers in Qatar) of arrival at work and departure from work of respondents to build time of day profiles of workers in different industry sectors. Then, these profiles are converted into percentages of service availability offered by each TAZ and used to modify the availability of opportunities in each zone by each type of industry. This approach in essence attempts to capture the opening and closing hours of business establishments and the portion of the jobs capacity that is available to people for services throughout a day. A similar approach was used successfully in Southern California [9].

The basic process of accessibility calculation is as follows:

1. Start with a TAZ database that contains the attributes we want to convert to accessibilities as columns and the TAZs as rows.
2. Create a TAZ to TAZ database that contains the criteria for buffers (e.g., travel times from zone to zone or travel distances from zone to zone using the output of the transportation simulator.
3. Define a set of buffers to use (e.g., 0–10, 10–20, and 20–50 min travel time). These are “rings” around the centroid of each zone.
4. Select for each origin zone the zones that are within each of the set of buffers by creating a neighborhood matrix.
5. Sum the attributes of the zones that are within the neighborhoods defined in points 2 and 3.

6. Allocate these counts of attributes to each origin zone.

The process above allows any type of buffer and neighborhood decay functions of attractiveness or proximity and any type of zonal attributes that can be counted. This is also the portion of the accessibility computation that is designed to be eventually integrated in a transportation simulation system in Qatar called the Qatar Activity Based Model [12]. In this example of application, the accessibility computation is based on an aggregated type of jobs that are four categories: a) Construction; b) Leisure that includes accommodation, foodservice, arts, entertainment, and recreation; c) Service that includes finance and insurance, real estate, administrative and support services, public administration, health and social work, and other services; and d) All Other not included in the three groups above. These categories can be modified at will depending on the available data. In summary, we have 4 types of employment in 3 time of day periods within 3 buffers of travel time and travel distance from the centroid of each TAZ. Travel time buffers are 10 min, 10 to 20 min, and 20 to 50 min based on a distribution of trips observed in surveys. Similarly, the travel distance buffers are 2 km, 2 to 5 km, and 5 to 20 km. Table 1 shows a small sample of the 190 indicators that are used here reporting only indicators for the morning peak using travel time of 10 min buffer zones and buffer zones at the distance of 2 km.

3 Method

The method used here belongs in the cluster analysis family of methods that aim to divide observations into groups according to the values of variables these groups have in common. The ultimate objective is to partition the data in such a way that zones with high density and diversity of opportunities and a good roadway system and public transportation service end up in the same group because they are similar. We also know there is some degree of continuity in similarity in space for land use and for the transportation service provided. In this way zones that are adjacent to each other also have similar characteristics, but we also know we may have abrupt differences in space (see photos in Appendix A). The classification method used needs to be sensitive to both continuity and abrupt differences.

It is usual to differentiate between methods that are algorithmic, and they discern patterns in multidimensional data clouds (these are also labeled unsupervised methods) and methods that are model-based and require assumptions about the data generating mixing of distributions. The algorithmic approaches are the earliest clustering techniques and include the well-known K-means, K-medoids, and hierarchical clustering among many others [13]. Early concerns about important questions on deciding about the right number of clusters, treatment of outliers, and uncertainty about the right partition motivated the second family of methods that specify a probability model for the data (e.g., the likelihood function). To fill this void model-based cluster methods were developed that make assumptions about the probability distributions of the data analyzed (in reality a mixture of these distributions as in [14] and evolved into a major field of data that span the entire taxonomy of data types from simple continuous variables to text [15].

One of the thorniest problems in model-based clustering, however, is the specification of joint distributions of data that are fundamentally different such as binary variables

Table 1. A selection of the variables used to create principal components and clusters

Characteristic	N = 1,822 Mean [Median] (SD)
Number of persons employed in the construction sector that can be reached within 10 min by roadway in the AM morning peak	26,411.0 [18,638.3] (27,439.9)
Number of persons employed in the leisure sector that can be reached within 10 min by roadway in the AM morning peak	4,725.7 [816.6] (9,627.1)
Number of persons employed in the service sector that can be reached within 10 min by roadway in the AM morning peak	5,482.7 [1,350.2] (7,190.5)
Number of persons employed in all other sectors that can be reached within 10 min by roadway in the AM morning peak	6,875.7 [1,847.7] (21,500.1)
Total lane kilometers within 10 min by roadway in the AM morning peak	325.7 [311.9] (167.3)
Total number of public transport (transit) stops that can be reached in 10 min by roadway in the AM morning peak	60.4 [66.0] (36.7)
Number of persons employed in the construction sector that can be reached within 2 km by roadway in the AM morning peak	9,741.2 [4,280.8] (13,948.0)
Number of persons employed in the leisure sector that can be reached within 2 km by roadway in the AM morning peak	1,572.2 [122.6] (3,892.7)
Number of persons employed in the service sector that can be reached within 2 km by roadway in the AM morning peak	2,040.4 [133.8] (4,061.7)
Number of persons employed in all other sectors that can be reached within 2 km by roadway in the AM morning peak	1,457.1 [397.9] (7,485.4)
Total lane kilometers within 2 km by roadway in the AM morning peak	100.4 [102.6] (52.5)
Total number of public transport (transit) stops that can be reached within 2 km by roadway in the AM morning peak	23.5 [20.0] (20.0)

Note: Note: Travel time and distance buffers around a TAZ centroid are computed using time and distance measured along the network of roadways, area is the sum of the square kilometers of each TAZ reached, AM morning peak period is from 6:00 to 8:59

with multicategory variables and continuous variables and the restrictive assumptions needed to write the likelihood function that will be used to estimate model parameters [16, 17]. To avoid imposing restrictive assumptions on our data and to avoid issues with skewed distributions (see Table 1), a model-free algorithmic method in creating clusters is used here. This method takes advantage of more recent developments in data analysis of mixed types of variables (discrete and continuous) and a variety of diagnostics about cluster quality. The method used here is a distance-based clustering of mixed data [18]. The word "mixed" refers to quantitative (continuous numeric data of any distribution) and qualitative (categorical factor data). In this accessibility application we consider all indicators to be quantitative and continuous. Cluster analysis in this approach is the second step after a data combination and reduction step based on principal component analysis (PCA). The data are viewed as a table of rows representing the TAZs and columns representing the attributes standardized (subtract their mean and divide by their standard deviation). This decreases the possibility that one variable "dominates" all the other variables due to its size. The resulting matrix of rows of TAZs and columns of the transformed variables is then used to derive principal components projecting the observed data on an orthogonal coordinate system (axes that yield uncorrelated components) that captures variation in the data in a hierarchical way with the first component having the highest variation, the next component having the second highest variation and so forth until 100% of the observed variation is represented in the new coordinate system (in essence an eigenvalue/eigenvector computation). This allows to identify components that map the entire variation in the data from all the variables jointly. Then, retain for the subsequent step, which is cluster analysis, the desired number of components based for example on the percent of variance that is considered to be the signal (the variation we want to explore further) versus noise in the data.

The PCA in the way that is presented in [19] and implemented in FactoMineR [20] uses the matrix of the transformed variables described above. This is equivalent with the use of a pairwise distance (dissimilarity) between observations as the Euclidean distances [18, 19]. This technique produces principal components representing dimensions of the desired zonal characteristics (e.g., number of employees in construction that can be reached in the AM within 10 min, number of public transportation stops available within 10 min in the AM, and so forth). These are the examples of the variables presented in Table 1. After all the variables are transformed in the new coordinate system (the principal components) we retain the number of principal components that capture the desired variance. Then, these principal components are used in a hierarchical clustering routine to extract the clusters with systematic differences and similarities in their principal components and by reflection of the projection on the principal components of the original accessibility indicators. The principal components and their correlation with each of the accessibility indicators can be studied to assign role played by each component and to discern the most important variables in the indicators used. In this paper, I skip the presentation of these components favoring the clustering portion of the analysis.

Before presenting the findings it is worth mentioning that deciding on the number of clusters to retain in analyses of this type cannot be done exclusively based on indicators of "good" clusters. For example, one popular indicator is the total within cluster sum of

squares. This tells us that we minimize the amount of variation of the spatial variables within each cluster by increasing the number of clusters. There are many types of indicators (statistical criteria) one can use to identify the "optimal" number of clusters and they are available in an R library NbClust [21]. As in most clustering methods to partition the observations here we aim to find groups (classes) with small within-class variability and large between-class variability. When we work with inertia we can achieve both aims contemporaneously because total inertia (which is a constant for the data we have) is the sum of within-class inertia and between-class inertia as shown in Eq. 1.

$$\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{N}(x_{iqk}-\bar{x}_k)^2=\sum_{k=1}^{K}\sum_{q=1}^{Q}\sum_{i=1}^{N}(x_{iqk}-\bar{x}_{qk})^2+\sum_{k=1}^{K}\sum_{q=1}^{Q}(\bar{x}_{qk}-\bar{x}_k)^2 \quad (1)$$

i = 1,…,N observations
q = 1,…,Q classes
k = 1,…,K variables
x_{iqk} = value of variable k in class q for individual i
$\bar{x}_k$ = value of overall average for variable k
$\bar{x}_{qk}$ = value of average for variable k within class q

Hierarchical clustering when it is agglomerative starts will all the observations in one cluster (one class). This means the between inertia is equal to the total inertia. Then at a sequence of steps observations first and clusters afterward are combined to minimize the decrease in between-class inertia. In this way, observations are classed in the same cluster by minimizing the decrease in the between-class inertia. The difference between two cluster steps (one step with q clusters and the next with q + 1 clusters) is computed and if appearing to be a large difference the q + 1 number of clusters solution is accepted. This is shown in Fig. 1 left-hand side with an abrupt decrease from 1 to 2 clusters (the first bar in the graph) and then smaller changes in inertia as the number of clusters increases. This appears to be leveling off at 6 clusters (fifth dark bar on the figure). The right-hand side figure shows this 6-cluster solution in which the observations are classed in one of the six groups. Note that going from 6 clusters to 7 would not yield any major gain in maximizing between-cluster inertia. However, this is not a sufficient criterion to decide on the number of clusters as our final solution. Two additional criteria for deciding on the number of clusters are the relative balance of cluster membership and the interpretability of the solution [16, 17]. This is also demonstrated in [18] using the same method used here to illustrate its functionality.

At the end of this process the composition of these derived classes of zones can be presented in terms of the principal component scores and the original variables from Table 1 that are the zonal characteristics.

4 Analysis and Results

There are four examples in this section. The first analysis includes all 190 accessibility indicators to use as the baseline and retains 27 principal components capturing 91.6% of the variance in the 190 accessibility indicators. The second includes the public transportation (PT) accessibility indicators only to explore if higher access to public transportation

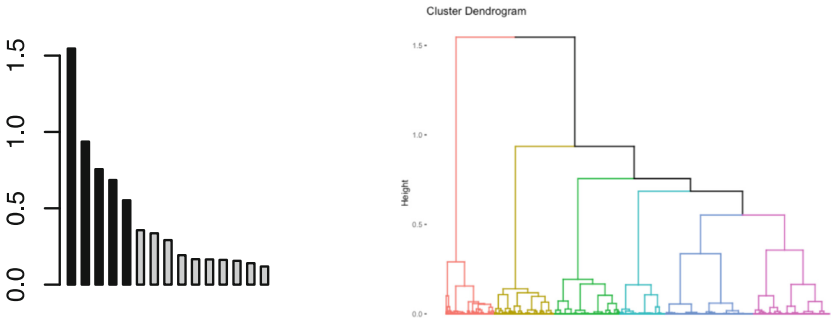

Fig. 1. Inertia gain by increasing number of clusters and dendrogram with six clusters highlighted by different colors and the height representing gain in difference of inertia (Color figure online)

is identified along corridors or areas in the center of Doha (the capital of Qatar and the most populous region). The third and fourth analyses are about the AM peak accessibility of which the first is clustering based on 23 principal components capturing 91.9% of the variance in the AM accessibility indicators and the second cluster analysis uses the rule of retaining principal components with eigenvalue > 1 yielding 13 components and capturing 80.4% of the variance in the AM accessibility indicators. Figures 2, 3, 4, and 5 show the Qatar maps and zoomed in the center of Doha. Tables 2, 3, and 4 contain a selection of accessibility indicators used to obtain the principal components and then clustering based on the components. There are many additional accessibility indicators used in each analysis but not shown in the paper and available upon request.

Figures 2 and 4 show the TAZs are classified in the typical core-periphery of a city (center of the city, suburbs, exurbs, and rural). Rural in Qatar is the dessert (mostly captured by the red color on the maps). Table 2 (all accessibility indicators) and Table 3 (AM retaining more than 90% of variance) have close similarity in the classification of zones and the number of employees that can be reached within the buffers used here. They have similarities in terms of the travel time by car in the AM peak period and travel distance by car in the AM peak period. This points out to the possibility of not needing to use all 190 different indicators to identify the area of a city that is well served by private transportation because the relative level of service by car does not change considerably between AM peak and PM peak. Major differences emerge in Figs. 2 and 4 and Tables 2 and 4, however, as we move away from the center of Doha to the zones of cluster 2 and cluster 4 that are very different in these two classifications.

Figure 3 and Table 3 show the maps of public transportation clusters and their characteristics respectively. This analysis shows clearly that there are public transportation corridors and an area that has the highest level of service by public transportation in the city center of Doha. This area includes the Souq al Wakif (an open-air market and the neighborhoods around this market to the South and East of the market – see also Figure A2). West of the market are zones with very low level of service by public transportation (yellow color) and those are parks, parking garages, a mosque, and other government buildings. The analysis here is also notable for its capability of identifying major points of interest that have a high level of accessibility within them but not around

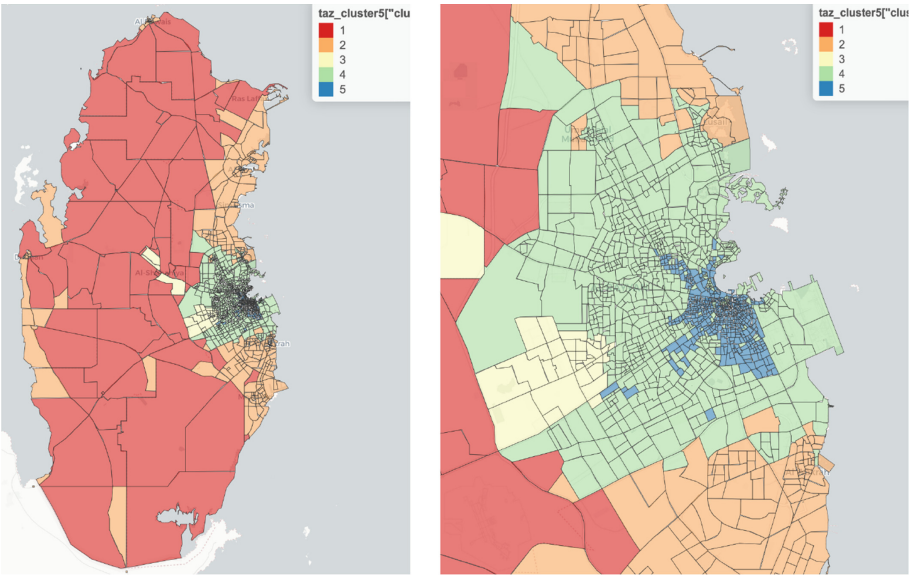

Fig. 2. Maps of the 5 clusters (numbers 1,…,5) using all accessibility indicators and 27 principal components capturing 91.6% of the variance in accessibility indicators (Color figure online)

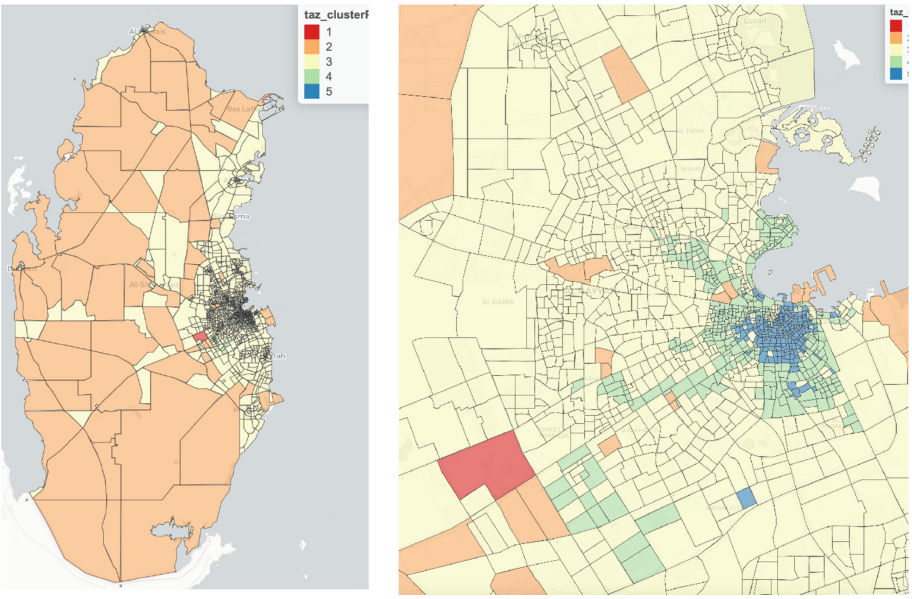

Fig. 3. Maps of 5 clusters (numbers 1,…,5) using only the public transportation accessibility indicators and 9 principal components capturing 90.2% of the variance in accessibility indicators (Color figure online)

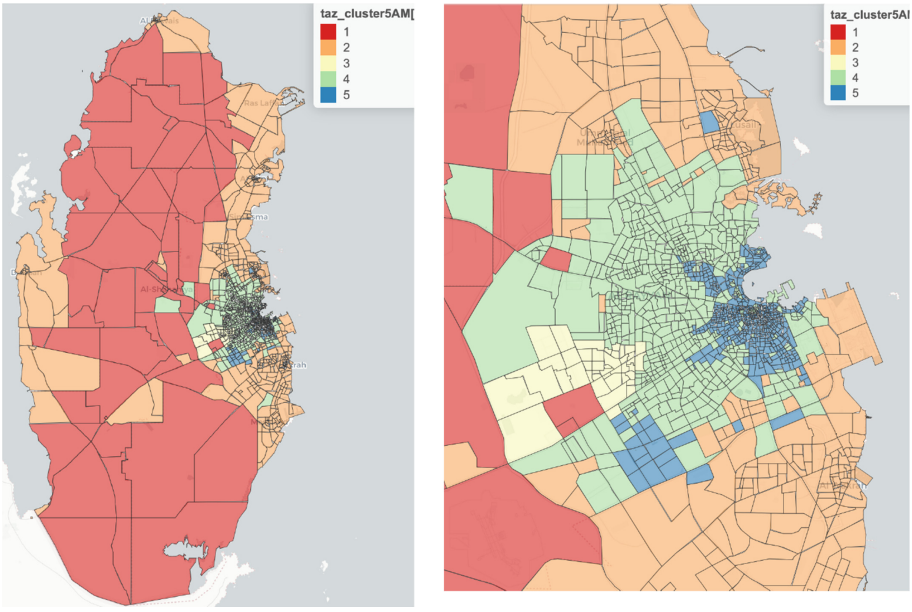

Fig. 4. Maps of the 5 clusters (numbers 1,…,5) using only the AM accessibility indicators and 23 principal components capturing 91.9% of the variance in accessibility indicators (Color figure online).

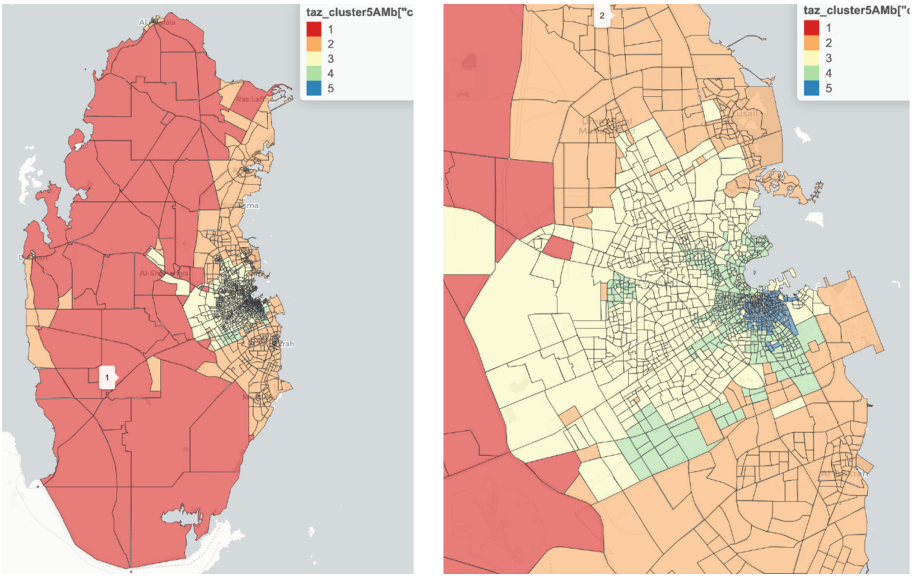

Fig. 5. Maps of the 5 clusters (numbers 1,…,5) using only the AM accessibility indicators and 13 principal components capturing 80.4% of the variance in accessibility indicators (Color figure online)

Table 2. Cluster characteristics selection of the AM peak accessibility classification using 23 principal components

Accessibility Indicator	Cluster 1 N = 79	Cluster 2 N = 285	Cluster 3 N = 41	Cluster 4 N = 894	Cluster 5 N = 523
Number of persons employed in the service sector that can be reached within 10 min by roadway in the AM morning peak	3,503.6 [195.6] (7,374.7)	560.2 [78.5] (1,118.4)	5,518.8 [334.7] (7,261.6)	6,524.1 [2,129.0] (8,236.2)	6,681.1 [5,190.3] (5,861.2)
Number of persons employed in the service sector that can be reached within a ring of 10 and 20 min by roadway in the AM morning peak	3,503.6 [275.1] (6,752.9)	496.6 [87.3] (1,470.3)	40,106.4 [44,119.9] (10,106.1)	19,456.1 [20,189.2] (12,814.4)	22,678.6 [22,484.4] (11,804.1)
Number of persons employed in the service sector that can be reached within a ring of 20 and 50 min by roadway in the AM morning peak	26,358.3 [26,111.9] (29,217.8)	22,484.6 [11,176.1] (23,875.5)	86,106.5 [88,899.8] (19,500.6)	105,493.2 [107,989.3] (18,289.7)	108,721.2 [109,779.4] (11,675.4)
Number of persons employed in the service sector that can be reached within 2 km minutes by roadway in the AM morning peak	735.6 [0.0] (3,310.6)	230.9 [0.1] (851.3)	622.8 [0.9] (3,207.2)	2,366.6 [83.9] (4,775.1)	2,777.1 [716.5] (3,576.9)
Number of persons employed in the service sector that can be reached within a ring of 2 and 5 km by roadway in the AM morning peak	756.7 [0.0] (3,262.3)	233.4 [13.8] (662.7)	10,137.9 [4,773.6] (11,487.5)	10,183.7 [9,331.0] (9,784.1)	13,639.2 [14,764.5] (7,408.9)
Number of persons employed in the service sector that can be reached within a ring of 5 and 20 km by roadway in the AM morning peak	9,853.9 [1,005.0] (15,486.3)	7,793.4 [1,549.9] (12,409.9)	102,385.1 [109,570.1] (25,328.1)	95,348.1 [101,470.2] (23,165.9)	108,816.4 [107,479.3] (8,679.9)
Total lane kilometers within 10 min by roadway in the AM morning peak	223.8 [210.3] (120.3)	168.4 [160.9] (96.2)	364.5 [407.5] (98.4)	376.3 [367.6] (189.2)	337.3 [329.5] (93.6)

(*continued*)

Table 2. (*continued*)

Accessibility Indicator	Cluster 1 N = 79	Cluster 2 N = 285	Cluster 3 N = 41	Cluster 4 N = 894	Cluster 5 N = 523
Total number of public transportation stops that can be reached within 10 min by roadway in the AM morning peak	9.5 [6.0] (9.4)	15.7 [14.0] (13.7)	54.3 [57.0] (25.8)	60.5 [63.0] (29.8)	92.9 [91.0] (21.9)
Total area in square km that can be reached within 10 min in the AM morning peak	442.4 [396.9] (302.8)	68.2 [44.5] (82.5)	75.4 [70.0] (39.2)	32.2 [28.2] (22.2)	18.6 [17.8] (6.6)

Note: Mean [Median] (SD); N is the number of TAZs in each cluster

them. In Table 3 this is the cluster 2 that has 80 zones and includes the harbor (right most side of Figure A1 and the Hamad International airport (right most side of Figure A2).

Table 3. Cluster characteristics selection of the public transportation (transit) accessibility classification

Accessibility Indicator	Cluster 1, N = 1[1]	Cluster 2, N = 80[1]	Cluster 3, N = 1,191[1]	Cluster 4, N = 336[1]	Cluster 5, N = 214[1]
Number of persons employed in the service sector that can be reached within 10 min by transit in the AM morning peak	0.0 [0.0] (NA)	1,900.9 [0.1] (3,897.5)	30.9 [0.0] (174.5)	26.5 [1.0] (85.2)	51.0 [25.6] (99.9)
Number of persons employed in the service sector that can be reached within a ring of 10 and 20 min by transit in the AM morning peak	0.0 [0.0] (NA)	2.1 [0.0] (13.3)	10.1 [0.0] (70.6)	656.5 [61.3] (1,953.8)	406.0 [310.8] (340.7)
Number of persons employed in the service sector that can be reached within a ring of 20 and 50 min by transit in the AM morning peak	0.0 [0.0] (NA)	246.8 [0.0] (1,265.0)	634.7 [0.0] (2,366.9)	9,543.7 [11,732.1] (5,969.7)	19,386.9 [23,385.6] (6,897.7)
Number of persons employed in the service sector that can be reached within 10 min by transit in the PM after noon peak	0.0 [0.0] (NA)	698.7 [0.1] (1,399.0)	16.8 [0.0] (101.7)	11.1 [0.3] (31.2)	18.6 [10.7] (31.9)

(*continued*)

Table 3. (*continued*)

Accessibility Indicator	Cluster 1, N = 1[1]	Cluster 2, N = 80[1]	Cluster 3, N = 1,191[1]	Cluster 4, N = 336[1]	Cluster 5, N = 214[1]
Number of persons employed in the service sector that can be reached within a ring of 10 and 20 min by transit in the PM after noon peak	0.0 [0.0] (NA)	0.7 [0.0] (4.3)	3.7 [0.0] (22.3)	254.4 [23.9] (816.7)	151.0 [116.3] (126.8)
Number of persons employed in the service sector that can be reached within a ring of 20 and 50 min by transit in the PM after noon peak	0.0 [0.0] (NA)	83.3 [0.0] (421.7)	260.8 [0.0] (998.7)	3,550.4 [4,036.4] (2,410.3)	8,027.3 [9,888.3] (3,232.5)
Total number of public transportation stops that can be reached within 10 min by transit in the AM morning peak	6.0 [6.0] (NA)	1.7 [1.0] (2.6)	0.9 [0.0] (1.5)	2.9 [2.0] (3.4)	4.8 [4.0] (3.2)
Total number of public transportation stops that can be reached within a ring of 10 to 20 min by transit in the AM morning peak	0.0 [0.0] (NA)	0.4 [0.0] (1.7)	1.7 [0.0] (4.1)	13.3 [12.0] (7.3)	30.9 [29.0] (12.7)
Total number of public transportation stops that can be reached within a ring to 20 to 50 min by transit in the AM morning peak	0.0 [0.0] (NA)	8.3 [0.0] (36.3)	19.0 [0.0] (37.5)	159.8 [161.0] (37.0)	248.2 [251.0] (45.9)

[1] Mean [Median] (SD); N is the number of TAZs in each cluster

In summary, the method used here classifying TAZs is able to pick up differences in zones based on low level of service (e.g., small number of bus stops in the zones but also surrounding zones through the distance and travel time buffers). It is important to note, however, that all this analysis is based on public transportation in the way it was a few years ago and before the opening of the Doha metro and Lusail tram (https://www.qr.com.qa/home). One should expect major differences in the accessibility of public transportation after their opening. Nevertheless, the analysis here provides important evidence about the method's capability of pinpointing places with low and high level of service and access to opportunities.

Figure 4 and Fig. 5 represent the same type of accessibility (AM peak by any mode) with the key difference that Fig. 4 shows the five clusters based on 23 principal components (capturing 91.9% of the accessibility indicator variance) and Fig. 5 shows the five clusters based on 13 principal components capturing only 80.4% of the variance. If one has evidence or suspicion the data are noisy Fig. 5 could be more appropriate. However, the number of clusters to use should have been based on Eq. 1 and not defined a priori

Table 4. Cluster characteristics selection of the AM peak accessibility classification using 13 principal components

Accessibility Indicator	Cluster 1, N = 71[1]	Cluster 2, N = 448[1]	Cluster 3, N = 46[1]	Cluster 4, N = 684[1]	Cluster 5, N = 573[1]
Number of persons employed in the service sector that can be reached within 10 min by roadway in the AM morning peak	4,446.3 [199.8] (8,276.0)	520.1 [78.5] (1,235.2)	8,025.2 [4,734.0] (8,856.7)	8,021.4 [6,574.1] (8,523.6)	6,256.6 [4,792.9] (5,818.7)
Number of persons employed in the service sector that can be reached within a ring of 10 and 20 min by roadway in the AM morning peak	4,999.9 [258.5] (9,583.5)	1,640.6 [212.0] (3,765.2)	41,497.1 [44,125.7] (7,336.4)	23,211.0 [23,505.9] (11,207.2)	21,714.0 [21,433.7] (11,839.5)
Number of persons employed in the service sector that can be reached within a ring of 20 and 50 min by roadway in the AM morning peak	24,659.6 [21,715.3] (28,902.4)	46,226.1 [39,270.6] (39,388.8)	92,115.9 [92,737.3] (7,687.1)	108,983.3 [110,940.4] (16,769.0)	108,116.6 [109,401.5] (12,414.3)
Number of persons employed in the service sector that can be reached within 2 km minutes by roadway in the AM morning peak	967.7 [0.0] (4,118.5)	171.8 [1.2] (715.8)	914.9 [6.0] (4,411.7)	3,074.6 [135.4] (5,238.6)	2,490.1 [666.8] (3,359.0)

(*continued*)

Table 4. (*continued*)

Accessibility Indicator	Cluster 1, N = 71[1]	Cluster 2, N = 448[1]	Cluster 3, N = 46[1]	Cluster 4, N = 684[1]	Cluster 5, N = 573[1]
Number of persons employed in the service sector that can be reached within a ring of 2 and 5 km by roadway in the AM morning peak	1,123.9 [0.0] (3,766.0)	257.3 [10.5] (954.4)	14,085.2 [13,539.5] (12,511.1)	12,229.8 [12,229.9] (9,609.8)	13,213.3 [14,051.7] (7,426.2)
Number of persons employed in the service sector that can be reached within a ring of 5 and 20 km by roadway in the AM morning peak	12,060.7 [843.9] (22,475.9)	25,134.4 [11,641.1] (29,854.8)	108,149.1 [109,005.7] (9,473.8)	103,192.3 [105,270.3] (13,983.8)	107,634.6 [107,305.3] (9,993.9)
Total lane kilometers within 10 min by roadway in the AM morning peak	231.6 [222.9] (123.2)	205.3 [201.9] (124.1)	393.5 [409.8] (83.8)	406.5 [396.9] (190.1)	329.7 [325.4] (104.3)
Total number of public transportation stops that can be reached within 10 min by roadway in the AM morning peak	10.8 [8.0] (10.4)	18.3 [16.0] (15.6)	64.1 [67.5] (25.2)	67.9 [69.0] (26.1)	90.3 [89.0] (23.7)

(*continued*)

Table 4. (*continued*)

Accessibility Indicator	Cluster 1, N = 71[1]	Cluster 2, N = 448[1]	Cluster 3, N = 46[1]	Cluster 4, N = 684[1]	Cluster 5, N = 573[1]
Total area in square km that can be reached within 10 min in the AM morning peak	477.5 [431.2] (302.0)	58.7 [38.8] (68.8)	60.7 [55.3] (20.1)	33.0 [28.4] (23.0)	18.3 [17.5] (6.5)

[1] Mean [Median] (SD); N is the number of TAZs in each cluster

as we did here for comparison with Fig. 4. The difference between the two AM analyses shows that this method is very sensitive to the number of components retained for the cluster analysis and one needs to be careful about the number of components to retain. In fact, using the rule of retaining eigenvalues that are larger than 1 would not work well here and this should be replaced by data verification with field surveys. In this analysis verification is facilitated by online mapping information and points of interest. Similar issues are found in other fields of inquiry too [22].

5 Summary and Conclusion

In this paper accessibility, the ease (or difficulty) with which activity opportunities can be reached from a given location, is used to classify the usual traffic analysis zones in transportation planning, modeling, and simulation. Many accessibility indicators used in a recent project in the State of Qatar are used here to illustrate a method that is able to extract zonal diversity in level of service and access to opportunities by two different transportation modes (private car and transit) using principal components analysis. Then, the scores of the principal components are used in a hierarchical clustering step to identify relatively homogeneous groups of zones. The analysis here uses 1822 zones and 190 indicators that reflect the differences in accessibility across different times of a day to distinguish travel times in the AM peak, midday, and PM peak periods. In addition, the illustration of the method here uses different subsets of these indicators to identify different facets and possibilities in this analysis and to identify recommendations for practice. For example, the number of principal components can be a small number (e.g., 27 components capture 91.6% of the 190 indicators variation) but practice should not use rules of thumb found often in PCA textbooks. Instead, makes more sense to perform sensitivity analysis based on the amount of variance captured by different sets of principal components retained for the subsequent cluster analysis. In the paper one example is the comparison for the AM peak period showing that selecting 23 components (capturing 91.9% of the accessibility variance) versus selecting 13 principal components (capturing 80.4% of the variance) yields significantly different classification. Moreover, using the public transportation accessibility indicators and their derived clusters a corridor and a central area of higher accessibility by public transportation is clearly identified. Overall, this method is precise enough to pinpoint specific areas where level of service can be

improved, and accessibility increased substantially for the whole day or portions of it (AM as it is done in this paper).

There are however a few limitations in this method that point to possible next steps in research. The public transportation analysis is based on data before a major infrastructure component was added (the Doha Metro and Lusail tram). Moreover, walking and biking are absent as stand-alone modes (walking is implicit in the public transportation travel times and distances). Also, the use of large TAZs (outside the core of Doha) may decrease the accuracy of travel times and distances here. However, these are more prevalent in the desert where activity opportunities are scant.

Appendix A Additional Maps and Photos from Doha, Qatar

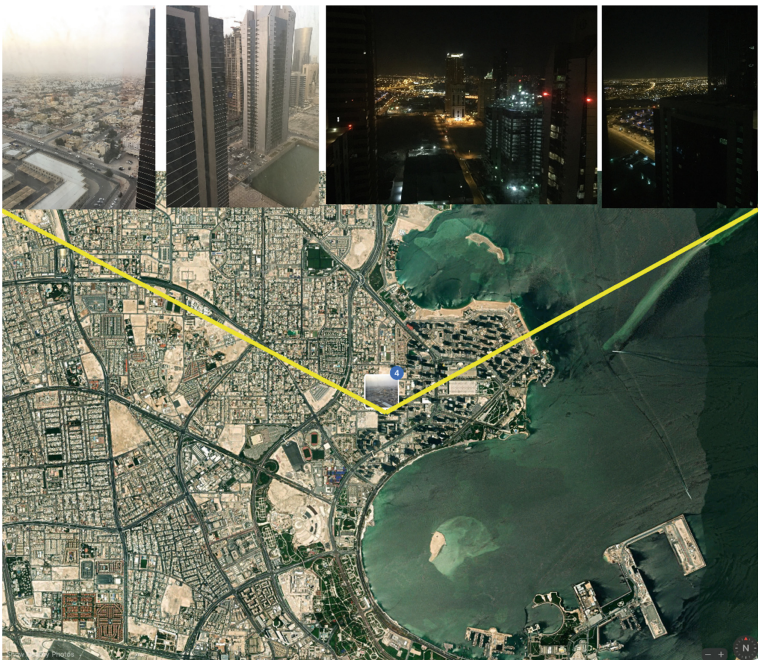

Fig. A1. Georeferenced photos showing the high differences in density among TAZs in the center of Doha, Qatar (photos by K.G. Goulias)

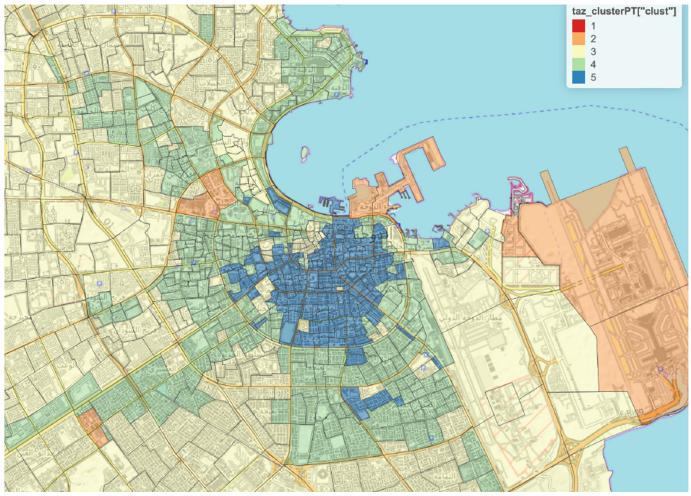

Fig. A2. Center of Doha and public transportation based zonal clusters – blue is highest level of service and green second highest (see also Table 3) (Color figure online).

References

1. Wachs, M., Kumagai, T.G.: Physical accessibility as a social indicator. Social –Econ. Planning Sci. **7**, 437–456 (1973)
2. Handy, S.L.: Regional Versus Local Accessibility: Variations in Suburban Form and the Effects on Non-Work Travel. Doctoral Dissertation, University of California, Berkeley (1993)
3. Blumenberg, E., Manville, M.: Beyond the spatial mismatch: welfare recipients and transportation policy. J. Plan. Lit. **19**(2), 182–205 (2004)
4. Pinna, F., Garau, C., Annunziata, A.: A literature review on urban usability and accessibility to investigate the related criteria for equality in the city. In: Gervasi, O., et al. (eds.) ICCSA 2021. LNCS, vol. 12958, pp. 525–541. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87016-4_38
5. Russo, A., Campisi, T., Tesoriere, G., Annunziata, A., Garau, C.: Accessibility and mobility in the small mountain municipality of Zafferana Etnea (sicily): coupling of walkability assessment and space syntax. In: Computational Science and Its Applications–ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part I, pp. 338-352. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-10536-4_23
6. Pellicelli, G., Caselli, B., Garau, C., Torrisi, V., Rossetti, S.: Sustainable mobility and accessibility to essential services. an assessment of the san benedetto neighbourhood in Cagliari (Italy). In: Computational Science and Its Applications–ICCSA 2022 Workshops: Malaga, Spain, July 4–7, 2022, Proceedings, Part VI, pp. 423-438. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-10592-0_31
7. Geurs, K.T., van Wee, B.: Accessibility evaluation of land-use and transport strategies:review and research directions. J. Transp. Geogr. **12**, 127–140 (2004)
8. Goulias, K.G., et al.: Simulator of activities, greenhouse emissions, networks, and travel (SimAGENT) in Southern California: design, implementation, preliminary findings, and integration plans. In: 2011 IEEE Forum on Integrated and Sustainable Transportation Systems, pp. 164–169. IEEE (2011)

9. Chen, Y., et al.: Development of opportunity-based accessibility indicators. Transportation Research Record: J. Trans. Res. Board **2255**, 58–68 (2011)
10. McBride, E.C., Davis, A.W., Lee, J.H., Goulias, K.G.: Incorporating land use into methods of synthetic population generation and of transfer of behavioral data. Transp. Res. Rec. **2668**(1), 11–20 (2017)
11. Goulias, K.G., Davis, A.W., McBride, E.C., Bhat, C.R., Pendyala, R.M.: Traffic analysis zone attributes and opportunity-based accessibility indicators in qatar activity-based model – QABM accessibility. Final Report 3 Submitted to ItalConsult and Qatar Ministry of Transportation and Communication, March 31, p. 29 (2019)
12. Goulias, K.G,, Bhat, C.R., Pendyala, R.M.: Qatar Activity-Based Model - Overview Final Report 1 Submitted to ItalConsult and Qatar Ministry of Transportation and Communication, March 29, p. 19 (2019)
13. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons (2009)
14. Vermunt, J.K., Magidson, J.: Latent class cluster analysis. Applied Latent Class Anal. **11**(89–106), 60 (2002)
15. Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E.: Model-based clustering and classification for data science: with applications in R, **50**. Cambridge University Press (2019)
16. Nylund-Gibson, K., Choi, A.Y.: Ten frequently asked questions about latent class analysis. Translational Issues in Psychological Sci. **4**(4), 440 (2018)
17. Weller, B.E., Bowen, N.K., Faubert, S.J.: Latent class analysis: a guide to best practice. J. Black Psychol. **46**(4), 287–311 (2020)
18. Van de Velden, M., Iodice D'Enza, A., Markos, A.: Distance-based clustering of mixed data. Wiley Interdisciplinary Rev.: Comput. Statistics **11**(3), e1456 (2019)
19. Pagès, J.: Analyse factorielle de donnees mixtes: principe et exemple d'application. Revue de Statistique Appliquée **52**(4), 93–111 (2004)
20. Lê, S., Josse, J., Husson, F.: FactoMineR: an R package for multivariate analysis. J. Stat. Softw. **25**, 1–18 (2008)
21. Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A.: NbClust: an R package for determining the relevant number of clusters in a data set. J. Stat. Softw. **61**, 1–36 (2014)
22. Björklund, M.: Be careful with your principal components. Evolution **73**(10), 2151–2158 (2019)