Application of Cluster Analysisin Trip Attraction Models of Campinas Metropolitan Region

Alexandre Frazão D'Andrea and Maria Teresa Francoso

***Abstract*—This paper discusses the application of thetravel demand model of similar cities groups from the CampinasMetropolitan Region by means of their economic characteristics. The cities groups were obtained from the statistical technique of multivariate cluster analysis which resulted in 4 clusters. This technique required the choice of a similarity measure and the clustering algorithm. The result obtained were represented by a dendogram that represents clusters of seminaries cities. Demand models correlating trips and job variableswere developed for each cluster. Finally, this article has pointed out that the partition of cities using variables such as jobs has revealed to be the most coherent with the observed data, rather than a model that considers cities in the CMR indistinctively.**

***Index Terms*—Cluster, demand, modelling, transport.**

I. Introduction

In his book, [1], states that travel demand modeling have both been important elements of transportation planning for many decades and until this day remains the foundational approach for demand forecasting.

The same author also states that in order to understand the demand for transport, we must understand the way in which these activities are distributed in space.

Data relating to economic activities in the Campinas Metropolitan Region (CMR) indicate that cities have different economic vocations, which means that the various types of jobs are heterogeneously distributed in the urban space of CMR [2] .

These characteristics suggest that a trip attraction model should contemplate this spatial heterogeneity of economic activities, with the risk of presenting simplified conclusions on the relationships between economic dynamics and attracted trips in each municipality.

The hypothesis of this paper is that individual trip attraction models for each group of economically similar cities may help to better understand the characteristics of trips throughout the CMR.

Define cluster analysis as a methodology for multivariate analysis, with the aim of dividing the set of observations in a number of homogeneous groups, according to some criterion of homogeneity [3].

According Cluster analysis is applied in many fields such as the natural sciences, the medical sciences, economics, marketing [4].

Manuscript received May 16, 2014; revised August 4, 2014.

Alexandre Frazão D'Andrea is with transport planning coordinator at Sistran Engenharia (GPO Group), UNICAMP, São Paulo, Brazil (e-mail: frazao.ale@gmail.com)

The purpose of this paper is to obtain clusters of cities through cluster analysis and verify whether the trip attraction models of these clusters have greater or lesser adherence to observed survey data (OD2011) than a model developed for all areas of traffic CMR irrespective of groups of cities.

II. Methodology

The methodology of this work involved the construction of trip attraction models using linear regression models based on data available from CMR's2011 origin - destination survey.

This survey presents data that represents all trips between 185 zones of CMR and the socioeconomics characteristics of them.

The dependent variable used in the models was the total number of trips in each zone. This variable was represented in two income classes: high-middle class and low class.

The independent variable used in the models was the total employment in each zone.

Two kinds of models were prepared:

The first one was named non-cluster based model (NCBM) and represents a general trip attraction model based on all 185 zones of CMR. This model was applied without distinction to all zones; therefore this model was built without taking into account the economics differences between the zones.

The second kind of models was named based cluster models (BCM) and these models were built from groups obtained by a multivariate cluster analysis technique.

This statistical techniquewasused to enable the grouping similar cities from totals of jobs in different economics sectors. The same income classes used in cluster based model were used in non-based cluster models.

The cluster analysis were developed in 3 steps described as similarity measures, clustering algorithm and den do gram representation.

Finally, the results of both types of models (NCBM and CBM) were compared with observed data from the survey origin destination 2011.

A. Non-Cluster Based Models (NCBM)

According [5], the traditional approach to trip demand modeling consists of four major steps: trip generation, trip distribution, mode choice and trip assignment.

According [1], These four-steps answer the following questions: "How many people travel?", "What are the travel patterns for the study area?","What travel modes are used?" and "What trip paths will be followed through the transportation network?"

For [5], most traditional trip attraction models use linear

DOI: 10.7763/IJET.2015.V7.822

regression to estimate the number of trips attracted to a zone, establishing a link between attracted trips and variables associated with employment or commercial activities (for example, the number of employees).

According [6], information from land use, population and economic forecasts are used to estimate how many trips will be made to and from each zone.

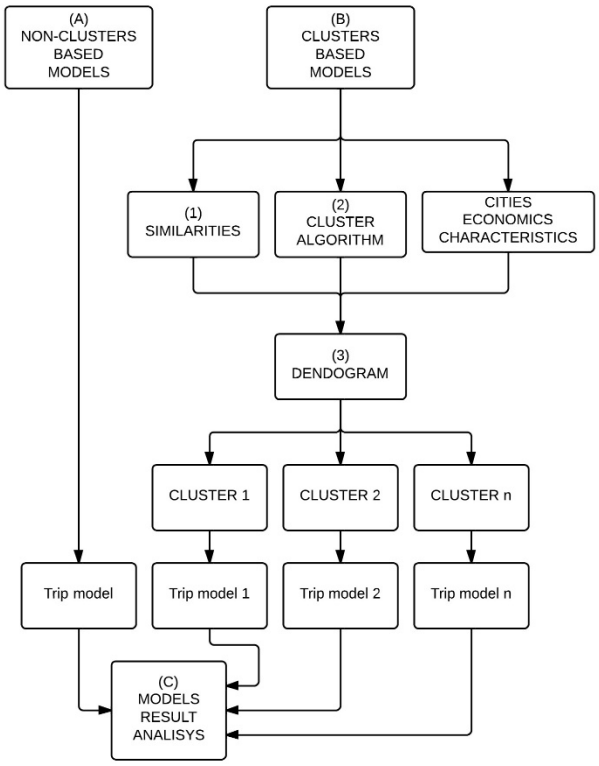


Fig. 1. Methodology.

In this paper, the trip attraction data used for the assembly of the CMR's general model were extracted from the 2011 origin destination survey and it has been separated between two kinds of data, high-middle income class and low class.

Employment data were obtained from the same origin and destination survey.

The non-cluster based models for the CMR are presented below in equations (1) and (2):

$$Trips_{HM} = -454 + 0{,}154 \times E_{HM} \quad (1)$$

$$Trips_{L} = -456 + 0{,}138 \times E_{L} \quad (2)$$

where:

HM = high and middle income class;

L = low income class;

E= employment.

B. Cluster Based Models (CBM)

According to [7], Cluster analysis seeks to divide a set of objects into a small number of relatively homogeneous groups on the basis of their similarity over N variables.

In a similar tone, [8]states that the purpose of cluster analysis is to gather objects (individuals, elements) in groups where there is homogeneity within the group and heterogeneity between groups, aiming to propose classifications.

Another definition is given by [9], who states that cluster analysis seeks dividing objects into groups that maximize the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown.

In this work we adopted the hierarchical clustering method procedures. The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. In this case, the similarities has represented by economics sectors of all CMR's cities.

According to [10], this type of procedure starts with each object representing an individual cluster. These clusters are then sequentially merged according to their similarity. First, the two most similar clusters are merged to form a new cluster at the bottom of the hierarchy. In the next step, another pair of clusters is merged and linked to a higher level of the hierarchy, and so on.

The development of clusters was performed in three steps that are described below.

1) Similarity measures

For [4] similarity measure is defined to measure the "closeness" of the objects. The "closer" they are, the more homogeneous they are. The similarity measures are defined as:

Euclidean distance. This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space;

Squared Euclidean distance. You may want to square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart;

City-block distance. This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared);

Chebychev distance. This distance measure may be appropriate in cases when we want to define two objects as "different" if they are different on any one of the dimensions;

Power distance. Sometimes we may want to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different;

Percent disagreement. This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature.

The Table I shows the distance formulation of each distance measure described previously.

TABLE I: SIMILARITY MEASURE

Kind of Distance	Measure (x, y)
Euclidean	$\left\{\sum_i (x_i - y_i)^2\right\}^{1/2}$
Squared Euclidean	$\sum_i (x_i - y_i)^2$
City-block	$\sum_i \lvert x_i - y_i \rvert$
Chebychev	Maximum $\lvert x_i - y_i \rvert$
Power	$\sum_i (\lvert x_i - y_i \rvert^p)^{1/r}$
Percent	$(numberof x_i \neq y_i)/i$

Euclidean distance was used in order to evaluate similarities why the distance between any two objects is not

affected by the addition of new objects to the analysis, which may be outliers.

2) Clustering algorithm

According to [10], after having chosen the distance or similarity measure, we need to decide which clustering algorithm to apply.

When each object represents its own cluster, the distances between those objects are defined by the chosen distance measure.

However, once several objects have been linked together, it is important determine the distances between those new clusters. The rules that determine when two clusters are sufficiently similar to be linked together are presented in clustering algorithms.

Following there is a description of the main clustering algorithms found in literature.

Complete linkage (furthest neighbor). In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters. This method usually performs quite well in cases when the objects actually form naturally distinct "clumps." If the clusters tend to be somehow elongated or of a "chain" type nature, then this method is inappropriate.

Unweighted pair-group average. In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This method is also very efficient when the objects form natural distinct "clumps," however, it performs equally well with elongated, "chain" type clusters.

Weighted pair-group average. This method is identical to the unweighted pair-group average method, except that in the computations, the size of the respective clusters (i.e., the number of objects contained in them) is used as a weight. Thus, this method (rather than the previous method) should be used when the cluster sizes are suspected to be greatly uneven.

Unweighted pair-group centroid. The centroid of a cluster is the average point in the multidimensional space defined by the dimensions. In a sense, it is the center of gravity for the respective cluster. In this method, the distance between two clusters is determined as the difference between centroids.

Weighted pair-group centroid (median). This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them). Thus, when there are (or we suspect there to be) considerable differences in cluster sizes, this method is preferable to the previous one.

Ward's method. This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. In general, this method is regarded as very efficient, however, it tends to create clusters of small size.

In this work we used the average linkage method algorithm since. According [11], this method is less affected by outliers than other methods.

According [12] in some cases, the choice of clustering variables is apparent from the nature of the task at hand and always depends on contextual influences such as data availability or resources to acquire additional data.

In this study the clustering variables used were obtained from the state data. These variables are the economic sector (see Table II) and characterize the amount of formal jobs by type of activity in the 19 cities that make up the CMR.

TABLE II: FORMAL JOBSAND ECONOMIC ACTIVITY - CMR

Cities	Sector 1	Sector 2	Sector 3	Sector 4
Americana	67	28.186	3.376	49.630
ArturNogueira	815	3.297	193	5.039
Campinas	1.693	61.630	20.758	322.454
Cosmópolis	1.187	1.964	413	6.427
Engenheiro Coelho	496	1.450	31	1.853
Holambra	3.216	1.023	287	2.687
Hortolândia	27	17.318	1.762	25.348
Indaiatuba	613	25.956	3.751	37.061
Itatiba	646	15.051	1.391	19.095
Jaguariúna	503	11.314	570	18.832
Monte Mor	539	5.161	988	4.801
Nova Odessa	130	11.307	869	6.278
Paulínia	226	11.605	5.043	27.558
Pedreira	94	7.080	30	5.924
Santa Bárbara	205	21.389	888	22.870
Santo A.de Posse	1.156	1.903	64	3.904
Sumaré	489	18.645	3.020	29.192
Valinhos	330	14.676	1.536	26.771
Vinhedo	121	6.588	172	13.967

Ref. [10] recommendsa sample size of at least 2^m, where m means the number of clustering variables.

These variables were used to calculate the distance matrix and perform the remaining steps of the algorithm described above.

3) Dendogram

The results of the cluster analysis are presented in the dendrogram box Fig. 1 and represent the distances between the similarity measures, allowing the grouping of similar cities according to predetermined economic variables.

[13] defines that the dendrogram, or tree diagram, is a mathematical and pictorial representation of the complete clustering procedure. The nodes of the dendrogram represent clusters, and the lengths of the stems (heights) represent the distances at which clusters are joined.

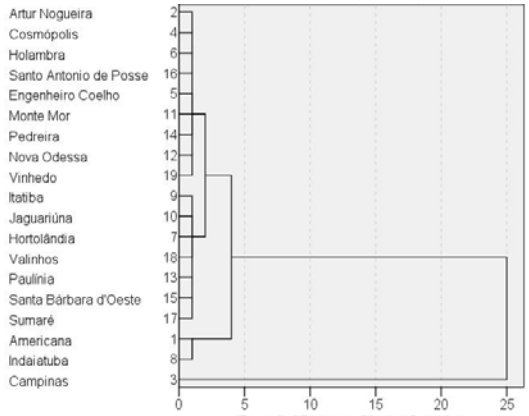


Fig. 2. Dendogram using average linkage.

The clusters obtained are summarized below:

Cluster 1: Campinas.

Cluster 2: Americana and Indaiatuba;

Cluster 3: Hortolândia, Itatiba, JaguariúnaPaulínia, Santa Barbara, Sumaré, Valinhos;

Cluster 4: Arthur Nogueira, Cosmópolis, Engenheiro Coelho, Holambra, Monte Mor, Nova Odessa, Pedreira, Santo Antonio da Posse and Vinhedo.

Tripattraction models were developed for each cluster resulting in the previous analysis and for income classes.

The following cluster based models are presented.

$$Trips\ C1_{HM} = -386 + 0{,}17 \times E \quad (3)$$

$$Trips\ C1_{L} = -475 + 0{,}16 \times E \quad (4)$$

$$Trips\ C2_{HM} = -570 + 0{,}14 \times E \quad (5)$$

$$Trips\ C2_{L} = -738 + 0{,}12 \times E \quad (6)$$

$$Trips\ C3_{HM} = -276 + 0{,}11 \times E \quad (7)$$

$$Trips\ C3_{L} = -171 + 0{,}09 \times E \quad (8)$$

$$Trips\ C4_{HM} = -172 + 0{,}10 \times E \quad (9)$$

$$Trips\ C4_{L} = -203 + 0{,}09 \times E \quad (10)$$

where:

C_1, C_2, C_3 and C_4, represents the clusters 1 to 4;

HM = high and middle income class;

L = low income class;

E= employment.

According to [14], statistic tests allows to check the accuracy of the models from two important issues were conducted.

The first is the diagnosis models, which verifies that the models adhere well to the data and are influenced by a small number of cases.

The second is the generalization of the models, which verifies that the models can be generalized to other samples.

The results of these tests meet these issues and are presented in Table II.

TABLE III: STATISTICS

Statistics	General model		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	ABC	DEF	ABC	DEF	ABC	DEF	ABC	DEF	ABC	DEF
Previsor	1	1	1	1	1	1	1	1	1	1
Casos	115	117	52	54	13	12	35	35	15	16
R2 (%)	89,2	88,1	96,1	96,9	94,1	91,2	86,5	87,8	76,0	67,1
F	935	849	1.236	1.625	175	103	211	236	41	29
p - value	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Leverage	0,009	0,009	0,0190	0,0190	0,0770	0,8300	0,0290	0,0290	0,0670	0,0620
Mahalanobis	0,991	0,99	0,98	0,98	0,92	0,92	0,97	0,97	0,93	0,94
Durbin-Watson	1,494	1,762	1,54	1,97	2,17	1,49	1,69	2,13	1,55	1,73

C. Models Results Analysis

The outcomes of trip attraction estimation from cluster based models produce less square error than the outcomes of the non-cluster based models.

The following graphs (Fig. 3-Fig. 6) shows the comparisons between observed and modeled data and it is observed that the results of non-cluster based model exhibit higher scattering than the cluster based models.

It is concluded that there is more grip when the results are obtained from cluster based models.

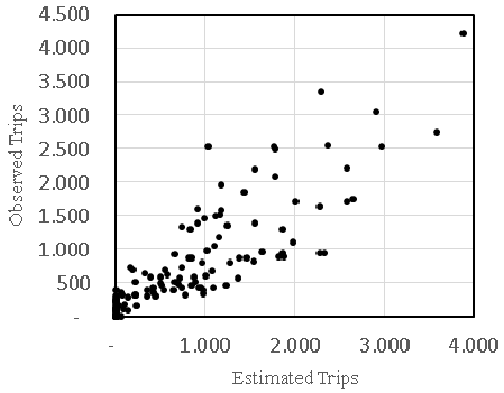
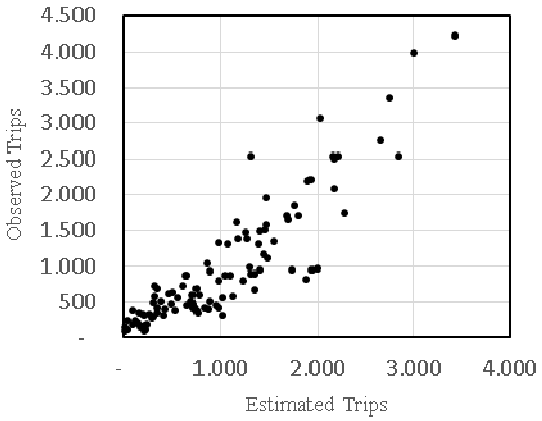


Fig. 3. Observed × Estimated clusters non-based model high and middle class.



4. Observated × Estimated clusters based model high and middle class.

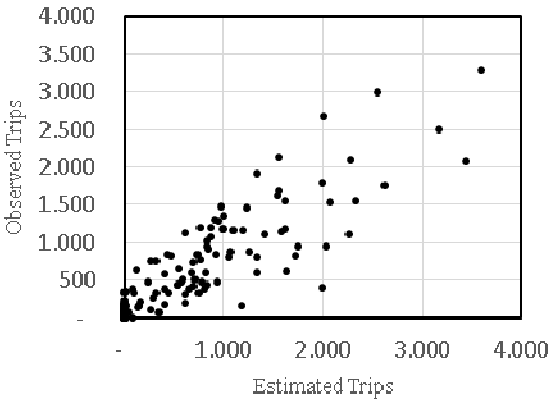


Fig. 5. Observated × Estimated non-clusters based model low class.

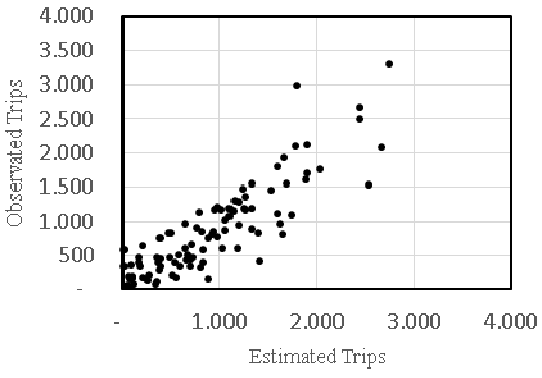


Fig. 6. Observated × Estimated clusters based model low class.

REFERENCES

[1] J. D. D. Ortuzar and L. G. Willumsen, *Modelling Transport*, 4th ed. John Wiley and Sons, Ltd., 2011, p. 608.

[2] Portal de Estatísticas do Estado de São Paulo. [Online]. Available: http://www.seade.gov.br/index.php?option=com_jce&Itemid=39&ta=1.

[3] A. J. Regazzi, *Análise Multivariada*, Viçosa: Universidade Federal de Viçosa, Centro de Ciências Exatas e Tecnológicas, Departamento de Informática, 2001, p. 166.

[4] K. Wolfgang and L. Simar, "Applied multivariate statistical analysis," *Vasa*, p. 515, 2008.

[5] *Transportation Planning Handbook*, Washington, D.C.: Institute of Transportation Engineers, 2009, p. 1067.

[6] H. T. Dimitriou and R. Gakenheimer, *Urban Transport in the Developing World: A Handbook of Policy and Practice*, 2011.

[7] D. Bailey, "Cluster analysis," *Professional Practice of Behavior Analysis*, vol. 6, no. 1975, pp. 59–128, 2014.

[8] G. C. Mead, A. P. Norris, and N. Bratchell, "Cluster analysis," *Professional Practice of Behavior Analysis*, vol. 6, pp. 105–125, 1989.

[9] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks.," *IEEE Trans. Neural Netw.*, vol. 4, no. 4, pp. 570–90, Jan. 1993.

[10] E. Mooi and M. Sarstedt, *A Concise Guide to Market Research The Process*, Data, and Methods Using IBM SPSS Statistics, Springer Berlin Heidelberg, 2011, p. 324.

[11] F. H. Joseph, *Multivariate Data Analysis*, 6th ed. 2006, p. 899.

[12] B. Everitt, "Cluster analysis," *Qual. Quant.*, vol. 14, no. 1, pp. 75–100, Jan. 1980.

[13] S. E. Brian, *Cluster Analysis*, 5th ed. London: John Wiley and Sons, Ltd., 2011.

[14] F. Andy, *Descobrindo a Estatistica Usando o SPSS*, 2th ed. Artmed Editora Ltda, 2009, p. 688.

Alexandre Frazão D'Andrea was born in 1971 at São Paulo, SP, and Brazil. He received his B.S. degree in civil engineering from Faculdade de Engenharia Industrial (FEI) at São Bernardo do Campo, SP, and Brazil in 1998.

He is a master's degree student in transport engineering at School of Civil Engineering Architecture and Urban Design, State University of Campinas.

He has been working as a transport planning coordinator at Sistran Engenharia (GPO Group) since 2005 until these days (Rua Santa Isabel 160, Santa Cecília, Centro, and São Paulo, Brazil).

He has 16 years of professional experience. He has worked for both the private and the public sectors. He has been involved in an extensive list of transportation planning projects in several cities in Brazil. Over the years he has focused on demand studies involving modeling and trips forecasting for impact analysis of new systems and the influences of traffic growth, travel behavior, payment systems, traffic and transit demands, socioeconomic changes and land use characteristics. Professor of Civil Engineering at UNIBAN University from 2004 to 2005.

He has presented the article "Decision-making support regarding the link between santos and guarujá cities focused on the environmental impacts" in 17th IRF World Meeting & Exhibition Riyadh, Saudi Arabia, November 10-14, 2013.