Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 20 (2011) 455–464

Procedia
Social and Behavioral Sciences

14th EWGT & 26th MEC & 1st RH

Data Mining Methods for Traffic Monitoring Data Analysis: A case study

Gregorio Gecchele[a], Riccardo Rossi[a,*], Massimiliano Gastaldi[a], Armando Caprini[a]

[a]*Department of Structural and Transportation Engineering, University of Padova, Via Marzolo, 9 35131 Padova, Italy*

Abstract

Presented in this paper is a comparative analysis of various Data Mining clustering methods for the grouping of roads, aimed at the estimation of Annual Average Daily Traffic (AADT). The analysis was carried out using data available from fifty-four Automatic Traffic Recorder (ATR) sites in the Province of Venice (Italy) and separated adjustment factors for passenger and truck vehicles in the grouping process. Errors in AADT estimation from 24-h sample counts indicate that model-based clustering methods give slightly better results compared to other tested methods, identifying a significant ATRs classification.

Selection and/or peer-review under responsibility of the Organizing Committee.
Keywords: Clustering analysis; AADT Estimation; Factor Approach; Traffic Monitoring

1. Introduction

In recent years the collection of information on traffic volumes has become a significant portion of the work of road planning programs in terms of both cost and personnel. In the U.S.A. the Federal Highway Administration (FHWA) provides guidance for improving these fields by way of its Traffic Monitoring Guide (TMG) (FHWA, 2001), which has a reference role for other similar agencies in the world.

The TMG provides specific recommendations on the number, extent, and duration of monitoring efforts, in order to achieve objectives of quality and integration of the information. In particular the guide recommends the design of monitoring programs consisting of a combination of portable short period traffic counts (SPTC) and permanent traffic counts (PTC), aimed at the definition of the Annual Average Daily Traffic (AADT), whose estimates are key inputs to pavement design analyses, trend analyses, revenue studies, accident analyses, and other studies of high importance. The estimates of AADT are based on both types of counts: short duration counts ensure geographic diversity and coverage, while continuous counts help to understand the time-of-day, day-of-week, and seasonal travel patterns. Different groups of similar roads, from geometric and functional points of view, are defined, and for each group adjustment factors are developed to convert short duration counts into accurate estimates of annual conditions, removing their temporal bias.

The TMG suggests three main methods for the definition of road groups: cluster analysis, geographical/functional classification or "same road" application of the factors. The choice of the "best" method is mainly related to the

* Corresponding Author. *E-mail address:* riccardo.rossi@unipd.it.

Selection and/or peer-review under responsibility of the Organizing Committee
doi:10.1016/j.sbspro.2011.08.052

availability of data and to the knowledge of the roadway system, since drawbacks exist for each of them, and it has relevant consequences in terms of errors propagation in the factoring procedure. In the TMG the term "cluster analysis" is specifically related to the implementation of a least-squares minimum distance algorithm which determines the most similar sets of factors, finding the variation patterns in the data. Nevertheless in recent years new methods of Data Mining for clustering have been developed, and many of them are now available in statistical and data mining software, both commercial and open-source.

This paper represents the first step of a more comprehensive analysis of TMG procedure, aimed at improving its critical aspects obtaining more affordable and precise results. The focus here is on a comparative analysis of clustering methods, which can be adopted to address the definition of road groups with similar traffic patterns.

Differently from previous studies the adjustment factors were calculated separately for passenger and truck vehicles in order to better understand temporal traffic patterns. The analysis is carried out using traffic data available from fifty-four Automatic Traffic Recorder (ATR) sites of the SITRA (TRAnsportation Information System) monitoring program, maintained by the Transportation Office of the Province of Venice, in north-eastern Italy.

The paper is organized as follows. Section 2 briefly summarizes literature on the development of road groups using clustering approach. Section 3 describes clustering methods adopted in this paper for road grouping. In Section 4 a description of the experimental data used in this study is provided, together with the development of the factor approach. Section 5 describes the computation process adopted for AADT estimation errors and main results obtained are included in Section 6. Concluding remarks are presented in Section 7.

2. Review of Literature

The procedure for traffic monitoring and AADT estimation proposed in the TMG involves four main steps:

1. Grouping of ATR sites characterized by similar patterns of temporal traffic volume variations;
2. For each road group, determining the average adjustment factors for traffic volume;
3. Assigning road sections monitored with SPTCs to one of the these groups;
4. Applying the appropriate adjustment factors to the sample counts to produce the AADT estimate for the road section considered.

The definition of road groups represents the first step of the procedure and is one of the main sources of errors in the estimation of AADT, together with the assignment of sample counts to road groups. Different approaches have been tested to address this issue, including Genetic Algorithms (Lingras, 2001), Artificial Neural Networks (Faghri and Hua, 1995), (Lingras, 1995), (Sharma et al., 1999), (Sharma et al., 2000), Regression Analysis (Faghri and Hua, 1995); the adoption of clustering methods represents a solution well established in this specific context.

Sharma and Werner (Sharma and Werner, 1981) first proposed the use of agglomerative hierarchical grouping (Ward's Minimum-Variance method) together with Scheffe's S-method of multiple group comparisons, in order to classify 45 PTC sites in Alberta on the basis of 12 monthly traffic factors. The same method was intensely adopted in other studies on different datasets, giving reliable results (Sharma et al., 1986) (Sharma and Allipuram, 1993). Other types of clustering were used, both agglomerative (average linkage method and centroid linkage method (Faghri and Hua, 1995)), and partitioning (k-means (Flaherty, 1993)), but in some cases the results were not considered affordable by the authors. In particular, it has been pointed out that:

- The clusters computed could not be consistent across years, meaning that ATR sites changed groups year by year (Ritchie, 1986) (Faghri et al., 1996);
- The clusters formed could not have a clear definition in terms of geographical/functional characteristics, given the purely mathematical nature of the process (FHWA, 2001);
- For hierarchical-based methods is often difficult to establish an "optimal" number of groups from a mathematical point of view (FHWA, 2001).

These criticisms must to be considered with attention, having clear that the road groups definition is a part of the procedure highly related to the characteristics of the site considered, in a positive or negative sense. At the same time the availability of new methods of cluster analysis in commercial and open source software has opened new possibilities to the cluster analysis to address effectively this particular step of the procedure.

The objective of this paper is comparing clustering methods (both traditional and more recent ones), analyzing the effects of road groups identification on AADT estimates. According to the TMG on the necessity to give particular attention to the differences in traffic patterns for passenger and truck vehicles, in this study the adjustment factors were calculated separately for passenger and truck vehicles and used together to define road groups.

3. Description of Cluster Analysis Methods

Cluster Analysis is a wide group of Data Mining methods, which have as main purpose grouping objects belonging to a certain dataset on the basis of data similarities. Clustering methods differ on how calculating the similarities between objects and how grouping these objects given the similarities. A common classification makes a distinction between non-parametric methods, including agglomerative hierarchical clustering and partitioning clustering, and parametric model-based clustering, which is becoming more popular in many application fields.

In the comparative analysis developed in this paper both kinds of methods were adopted for road groups identification, where objects were ATRs and adjustment factors represented their attributes; for this purpose some packages freely available in R (R Development Core Team, 2010), an open-source environment for statistical computing, were used together with Rapid Miner software (Mierswa et al., 2006).

3.1. Hierarchical Methods

Agglomerative hierarchical methods have a common algorithmic structure which produces a hierarchical grouping of objects by gradually agglomerating them. As a starting point each object is assigned to a group and at each step the two most similar groups are merged in a new group that substitutes the previous ones; merging is repeated until only one cluster exists.

Various measure of similarity were defined and can be used, but the most common one is the p-dimensional Euclidean distance, where p is the number of attributes used for the clustering. Once that the similarity measure is chosen, agglomerative methods differ on how Euclidean distance between objects is calculated. The R package hclust (Maechler et al., 2005) adopted for the analysis allows the use of these methods:

- Average Link = average distance between pair of objects, each in a different cluster;
- Single Link = minimum distance between two objects, belonging to different clusters;
- Centroid Link = distance between centroids of two clusters;
- McQuitty = distance between clusters is weighted using arithmetic averages;
- Ward = minimizes the increase in total within-cluster sum of squared error (Ward, 1963)

An extension of this package called energy (Rizzo and Szekely, 2010), allows to use an agglomerative method based on a measure of energy between clusters (Szekely and Rizzo, 2005). Given c_i and c_j two clusters containing n_i and n_j objects respectively, the energy between these clusters can be computed as:

$$e\left(c_i, c_j\right) = \frac{n_i n_j}{n_i + n_j}\left(2M_{ij} - M_{ii} - M_{jj}\right) \tag{1}$$

where x_{ip} is the p–th object of cluster i and x_{jq} the q–th object of cluster j and M_{ij} is computed as:

$$M_{ij} = \frac{1}{n_i n_j}\sum_{n_i}\sum_{n_j}\left\|x_{ip} - x_{jq}\right\|^2 \tag{2}$$

These methods are widely adopted in practice, even if they present some well-known shortcomings. In particular there is not a clear guidance on how to choose the optimal number of groups, even if some performance indices were defined and can help to determine the optimal number of clusters. In this study Euclidean distance was chosen as a measure of similarity and all the methods available in R were used.

3.2. Partitioning Methods

The main characteristic of partitioning methods is that they place each object of the dataset in only one group. Usually the starting point is a random assignment of each object into one of the k groups; at each step the algorithm reassigns the objects to the nearest cluster on the basis of a certain measure of proximity. The value k can be defined a priori at the beginning of the clustering, or determined by the algorithm itself.

K-means method is probably the most famous partitioning method; it aims to partition the objects into k groups such that the sum of squared distances from objects to the assigned cluster centers is minimized. K-means method is widely adopted as a clustering method, but suffers of some shortcomings, in particular a poor computational scalability, the necessity of giving a priori the number of clusters and a search prone to local minima.

One of the solutions proposed is the X-means algorithm (Pelleg and Moore, 2000). This algorithms adopts K-means algorithm as a core-algorithm and using a process of selection driven by the optimization of the Bayesian Information Criteria (BIC) gives at the same time the number of clusters and their parameters.

Another more robust version of K-means is Partitioning Around Medoids (PAM) method, which is based on the search of k representative objects, called medoids, among the objects of the dataset. These objects are chosen in order to represent and synthesize the structure of the data: once the algorithm find the k medoids, clusters are constructed by assigning each object to the nearest one. The iterative searching of the k medoids is guided by the objective of minimizing the sum of the dissimilarities of the objects to their closest medoid. Once the algorithm has found a good initial set of medoids (the so-called “build phase”), it finds a local minimum for the objective function, that is, a solution such that there is no single switch of an object with a medoid that will decrease the objective function (the so-called “swap phase”). Compared to K-means the main improvements provided by PAM is a main robust structure which minimizes a sum of dissimilarities instead of a sum of squared Euclidean distances.

In this study the three methods were adopted, using R package cluster and Rapid Miner in the case of X-means.

3.3. Model-based Methods

Model-based clustering has a completely different approach compared to non-parametric methods and have particular attractiveness for the capability of determining the optimal number of groups, employing the Bayesian Information Criteria (BIC). It assumes that each cluster could be represented by a density function belonging to a certain parametric family (e.g. the multivariate normal) and that the associated parameters could be estimated from observations (Fraley and Raftery, 2009). This means that the main aim of model-based clustering is determining the probabilistic density function for the k-th group estimating the p dimensional mean vector and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}_k$, where p is the number of attributes used for the clustering.

The covariance matrix can be decomposed as $\boldsymbol{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T$, where λ_k, $\mathbf{D}_k$, and $\mathbf{A}_k$ control the volume, the orientation and the shape of the k-th group (Banfield and Raftery, 1993); considering these geometric features as separate and independent parameters, different models can be built and analyzed. Identifiers codes can be used to describe the geometric characteristics of the model (volume, orientation and shape): E for equal, V for variable and I for Identity. For example a model classified as VEI means that the clusters are assumed to have a variable volume, equal shape and orientation on the coordinate axes.

The merging process of clusters is controlled by a maximum-likelihood criterion, following two alternative approaches. The first one, called the classification approach, aims at maximizing the likelihood over the mixture parameters and identifying the group to which each object in the dataset belongs, while the second one is the so-called mixture approach, which only aims at maximizing the likelihood over the mixture parameters.

In this study classification approach was adopted, in particular the model-based hierarchical clustering method as a good approximation of the exact but impractical maximization of likelihood (Fraley and Raftery, 2009), testing all the available models with R package mclust (Fraley and Raftery, 1997) (Fraley and Raftery, 2009).

4. Study Data

4.1. Data Source

The main data source for this study was the ATR sites located on the rural road network of the Province of Venice in 2005. The network links are mainly two-lane roads, and each ATR monitors traffic volumes on a single lane, meaning that for each site two ATRs were installed, one for each direction of traffic flow. The Transportation Office of the Province of Venice (the agency responsible for traffic monitoring) maintained 70 ATRs, which registered traffic volumes following a 7-class classification scheme of vehicles.

For the purpose of this study a total of 54 ATRs were selected, excluding the ATRs affected by considerable amounts of missing data because of equipment failures. Moreover a simple 2-class scheme has been adopted, dividing between passengers vehicles (PV) and truck vehicles (TV) with reference to a 5 m-length threshold.

4.2. Development of ATR Groups. Factors Definition

Data collected from the ATR sites were used to determine the classification of roads considering 3 day-types (Weekdays, Saturdays, Sundays), 6 two-month periods (January-February, March-April, May-June, July-August, September-October, November-December) and two classes of vehicles (PV and TV). This choice reflects the structure of SITRA monitoring program currently used by the Province of Venice, in order to compare the results obtained with the cluster analysis. Moreover the analysis was carried out considering 24-h short counts.

For the *i*-th day of week (Weekday, Saturday or Sunday) of the *j*-th two-month period, for vehicle class *c* (PV or TV), the seasonal factor for a road group can be calculated as:

$$f_{ij,c} = \frac{1}{n}\sum_{k=1}^{n}\frac{AADT_{k,c}}{ADT_{ijk,c}} \tag{3}$$

where $AADT_{k,c}$ is the AADT for the *k*-th section included in the group for vehicle class *c* calculated using the AASHTO method (AASHTO, 1992), $ADT_{ijk,c}$ is the average daily traffic recorded in the *i*-th day of week of the *j*-th two-month period in the *k*-th section for vehicle class *c* and *n* is the number of sections in the group. ATR sites were divided in different groups by the clustering methods on the basis of the reciprocal of the 36 composite adjustment factors, $rf_{ij,c}$, defined as:

$$rf_{ij,c} = \frac{1}{f_{ij,c}} \tag{4}$$

The AADT estimation for vehicle class *c* on a road section, based on 24-hours short count, is straightforward. Once that each section has been assigned to a group, AADT can be estimated multiplying the daily traffic count $DT_{ij,c}$ for vehicle class *c* obtained in a generic *i*-th day of week of the *j*-th two-month period for the corresponding seasonal factor (or dividing it for the corresponding reciprocal factor):

$$AADT_{Estimate,c} = DT_{ij,c} \cdot f_{ij,c} = \frac{DT_{ij,c}}{rf_{ij,c}} \tag{5}$$

4.3. Development of ATR Groups With Cluster Analysis Methods

The clustering methods presented in Section 3 were used in order to determine the road groups on the basis of the reciprocals of the 36 composite adjustment factors. Excluding X-means and model-based hierarchical methods, the main issue was determining the optimal number of groups. A combination of criteria was adopted for this purpose:

1. Pseudo F Statistic (Calinski and Harabasz, 1974) (Milligan and Cooper, 1985). This statistic analyses the hierarchy at each level; the peak value of PSF reveals the optimal number of clusters since it is defined as:

$$PSF = \frac{T - P_G}{G-1}\frac{n-G}{P_G} \tag{6}$$

where *G* is the number of groups and *n* is the number of ATRs. *PG* is calculated as the sum of W_j for the *G* groups at the *k*–th level of the hierarchy, being W_j calculated as in equation 7, where $\overline{x}_k$ is the mean vector for cluster C_k and *T* is calculated as in equation 7 where x_i is the *i*–th ATR and $\overline{x}$ is the mean vector for all ATRs:

$$T = \sum_{i=1}^{n}\left|x_i - \overline{x}\right|^2 \qquad W_j = \sum_{i \in C_k}\left|x_i - \overline{x}_k\right|^2 \tag{7}$$

2. Analysis of the variance of clusters. As done by other authors (Faghri and Hua, 1995) the optimal number of clusters could be found looking at the decrease in the variance of clusters while groups increase, choosing the number of groups for which the improvement in the accuracy due to the increase of groups becomes marginal.
3. Davies-Bouldin Index (Davies and Bouldin, 1979). This index assigns the best score (minimum value) to the structure that produces groups with high similarity within a group and low similarity between groups:

$$DB = \frac{1}{n}\sum_{i=1, i\neq j}^{n} max\left(\frac{\sigma_i + \sigma_j}{d\left(c_i, c_j\right)}\right) \tag{8}$$

where c_x is the centroid of group x, σ_x is the average distance of all elements in group x to centroid c_x, $d(c_i, c_j)$ is the distance between centroids c_i and c_j, and n is the number of groups.
4. Practical considerations. The number of groups must be limited in order to have a reasonable number of ATRs for each of them, avoiding to have groups with only one ATR.

The jointly application of these criteria allowed to identify the optimal number of groups as seven, excluded X-means method for which this number was four. The differences in ATR grouping between methods affect only a relatively small number of ATRs and three main groups could be defined combining some common patterns: a group of ATRs located on the coastal line, characterized by recreational traffic with peaks of PV in summer weekends, a big group of ATRs with relatively stable commuting traffic volumes during the year, and a group of ATRs with intermediate characteristics.

Since some methods gave the same results, a selection was made before the computation of AADT estimation errors. Sigle link, Average link and McQuitty link produced the same clusters of Centroid link, therefore only the latest one was tested; EEI model gave the same grouping of EVI and for this reason it was excluded.

The methods considered for the computation of AADT estimation errors were centroid linkage (Centroid), energy calculation (Energy) and Ward's (Ward) clustering as hierarchical clustering, K-means, X-means and PAM as partitioning methods, EII, EVI, VII, VEI and VVI models for model-based clustering. For comparison purpose a single group classification were included in the analysis, that means that all ATRs were put in a single group.

5. Study Method

Following a procedure already adopted in past studies (Sharma et al., 1999) (Sharma et al., 2000), 24-h sample counts were generated from the dataset and, for each classification made by clustering methods, were used as follows:

1. An ATR was removed from the group to which he belonged to create sample counts;
2. The ATR removed was called sample ATR and the group from which it was taken home group;
3. Average factors were calculated for the home group, excluding the sample ATR;
4. The average factors thus created were used to estimate AADT from samples generated by the sample ATR;
5. The process is repeated using each section at a time as a sample ATR, generating a large number of samples and AADT estimates.

Being Δ the absolute AADT estimation error, the percent error was calculated as:

$$\Delta = \left|\frac{AADT_{Estimate} - AADT_{Actual}}{AADT_{Actual}}\right| \times 100 \tag{9}$$

6. Results

6.1. Comparison of AADT Estimation Errors

The analysis of the errors obtained with the various methods was developed considering frequency distribution, divided on the basis of the differences between passenger and truck vehicles, the various day-types and periods of the year.

Table 1 shows the average percent errors resulting from the application of factors under the various classifications schemes, divided into total and day-type cases. Some common patterns emerge from the errors distribution.

Comparable performances were obtained by clustering methods, considering both the total average errors and the distribution of errors on the days of the week. Total errors in AADT estimates for truck vehicles were higher than AADT estimates for passenger vehicles (range 16,26% - 18,09% for TV and 9,99% – 11,07% for PV) and in both cases the distribution of errors changes by day of the week, increasing from weekdays to Saturdays and Sundays.

This trend is particularly strong in the case of truck vehicles, for which the values of errors in Sundays are almost the double of the same ones for passenger vehicles. This fact could be related to small truck flows observed during Saturdays and Sundays in many road sections: a small change in the number vehicles observed in various weeks could lead to percent errors higher than roads with high traffic volumes with bigger changes in traffic counts.

It is clear from Table 1 that the estimation errors for single group case are the highest, since adjustment factors for each sites could be very much different from the factors calculated on the whole set of sites.

Table 12 shows the average estimation errors divided for two-month periods of the year.

As can be observed, the distribution of errors follows a common pattern in the case of different methods, which has low values in winter time (particularly January-February) and high values in summer times (July and August). This fact could be related to the touristic activities in the coastal part of the Province of Venice, which induce high variations in traffic volumes on many roads of the network, that factors can not cover and forecast completely. The analysis of errors associated to each road group gives the chance to support this hypothesis and understanding with a higher detail the performances of the factor approach with the various road classifications.

Table 3 presents a more detailed view about errors distribution in the groups for a subset of the methods tested, chosen for their better performances. Average errors for PV an TV are subdivided for groups, sorted in decreasing order of errors with reference to the passenger vehicle case.

Table 1 Average AADT estimation errors. Total and day-type errors for passenger and truck vehicles

Method	Total		Weekdays		Saturdays		Sundays	
	PV	TV	PV	TV	PV	TV	PV	TV
Centroid	10,96	17,11	10,36	14,40	11,50	18,70	13,47	29,19
Energy	10,65	16,26	9,95	13,92	11,59	18,09	13,22	26,22
EII	10,08	16,35	9,44	14,16	10,81	17,51	12,58	26,27
EVI	9,99	16,40	9,37	14,22	10,55	17,56	12,57	26,26
VII	10,09	16,44	9,44	14,30	10,76	17,68	12,56	26,99
VEI	10,07	16,58	9,46	14,19	10,78	17,52	12,56	26,70
VVI	10,00	16,56	9,39	14,37	10,53	17,74	12,50	26,37
K-means	10,95	16,49	10,38	14,14	11,38	18,52	13,38	26,32
K-medoids	10,66	16,92	10,10	14,50	10,73	18,45	13,44	27,52
PAM	10,66	16,92	10,10	14,50	10,73	18,45	13,44	27,52
Ward	10,94	16,65	10,45	14,19	11,31	18,55	13,03	27,11
X-Means	11,07	18,09	10,33	15,58	12,03	19,89	13,82	28,88

Single Group	16,64	24,50	15,93	20,63	17,20	25,71	19,63	42,81

Table 2 Average AADT estimation errors. Period of year errors for passenger and truck vehicles

Method	January-February		March-April		May-June		July-August		September-October		November-December	
	PV	TV	PV	TV	PV	TV	PV	TV	PV	TV	PV	TV
Centroid	10,67	16,37	10,47	16,58	10,26	14,36	11,96	20,50	11,11	16,87	11,30	17,95
Energy	9,66	15,15	9,79	15,49	10,25	14,69	12,00	19,42	10,93	15,53	11,12	17,12
EII	9,06	14,98	9,22	15,56	10,21	14,46	11,27	19,22	10,58	16,31	9,99	17,42
EVI	9,03	15,14	9,10	15,51	10,15	14,54	11,27	19,28	10,41	16,32	9,86	17,47
VII	9,07	15,39	9,20	15,83	10,19	14,56	11,26	19,44	10,57	16,47	10,03	17,63
VEI	9,08	15,10	9,24	15,63	10,21	14,48	11,26	19,29	10,58	16,48	10,02	17,47
VVI	9,06	15,54	9,11	15,76	10,15	14,50	11,26	19,35	10,39	16,40	9,89	17,64
K-means	10,36	15,33	10,13	15,52	10,68	14,84	12,00	19,44	11,24	16,16	11,20	17,50
K-medoids	10,02	15,94	10,03	16,21	10,40	14,87	11,70	19,76	10,84	16,60	10,89	17,98
PAM	10,02	15,94	10,03	16,21	10,40	14,87	11,70	19,76	10,84	16,60	10,89	17,98
Ward	10,41	15,65	10,27	15,67	10,34	14,79	11,93	19,77	11,15	16,23	11,49	17,61
X-means	10,64	17,97	9,73	17,16	10,46	15,23	12,95	20,41	10,86	17,52	11,66	20,20
Single Group	18,65	23,65	15,97	13,34	13,34	21,25	19,44	27,89	14,17	23,00	18,41	27,18

Table 3 gives useful information about the grouping process performed by clustering methods. Definition of same group (i.e. groups containing the same sites) is clearly highlighted by the same error values, for example Group1 for the first six methods of the table or Group 5 for EEI and VEI. Particular groups could be highlighted, such as Group7 as identified by Energy, EEI, VEI, PAM and Ward. This group is composed by a single site (2 ATRs), and could be interpreted as an outlier compared to the other groups: the difficulty in defining common patterns for PV and TV is well highlighted by the high difference between errors for the two classes.

Moreover analyzing the spatial distribution of sections, the errors observed for AADT estimation in different groups could be related to their functional characteristics. In fact, Group 1 includes sites on coastal roads, where high variations in traffic volumes were observed in summer period and AADT estimates are more prone to high errors; conversely roads characterized by commuter trips are included in Groups 5 and 6, showing lower errors.

Table 3 Average AADT estimation errors. Group errors for passenger and truck vehicles

Method	Group 1		Group 2		Group 3		Group 4		Group 5		Group 6		Group 7	
	PV	TV	PV	TV	PV	TV	PV	TV	PV	TV	PV	TV	PV	TV
Energy	17,44	20,13	14,57	20,02	12,06	23,06	10,56	14,35	8,34	13,70	7,28	12,59	6,24	24,89
EEI	17,44	20,13	12,14	14,66	11,69	25,51	8,77	14,20	7,93	10,37	7,47	12,72	6,24	24,89
VEI	17,44	20,13	12,67	14,37	11,87	27,09	8,55	14,45	7,93	10,37	7,48	12,96	6,24	24,89
PAM	17,44	20,13	13,02	22,21	12,06	23,06	12,00	14,83	9,35	14,50	7,52	12,97	6,24	24,89
Ward	17,44	20,13	14,22	21,15	12,06	23,06	10,45	14,62	10,13	15,18	7,34	13,18	6,24	24,89
X_Means	17,27	23,61	10,45	23,94	8,68	15,58	7,96	13,07						
Single Group	16,64	24,50												

Looking at the overall results of the analysis, the choice of a "best" clustering method seems not to be clear, since slight differences in performances are shown between the methods. Nevertheless, a preference should be accorded to model-based clustering methods, which seem to outperform the others under the various points of view considered and present a strong and reliable mathematical structure. In particular VEI model (clusters with variable volumes, equal shape and orientation on the coordinate axes) distinguishes itself for better performances.

As revealed by the analysis of tables 2 and 3 some groups (commuter roads in weekdays) show low estimation errors, comparable to the findings of other similar studies (Sharma et al., 1999) (Sharma et al., 2000). Nevertheless other groups show high values of errors, in particular for TV class, highlighting the presence of some AADT estimates with extremely high errors, that further research efforts should address and reduce.

7. Conclusions

This paper has focused on a comparative analysis of cluster analysis methods, which can be adopted for the definition of groups of roads with similar traffic patterns in FHWA factor approach procedure for AADT estimation. Differently from previous studies, factors were calculated separately for passenger and truck vehicles, allowing for a deeper understanding of factor approach use and differences in temporal traffic patterns of vehicles.

The analysis was carried out using traffic data available from fifty-four ATR sites located in the Province of Venice, considering frequency distribution of AADT estimate errors. The main results obtained are that:

- Clustering methods identify a common basic structure of ATR groups which has a functional and geographical significance. The differences concentrate on the attribution of a relatively small number of ATRs.
- Clustering methods show common errors patterns and give comparable results in term of average percent errors distribution. These results are better compared to a single group attribution.
- A preference should be accorded to model-based clustering methods, in particular VEI model, which show slightly better performances compared to the other methods and have a robust mathematical structure;
- Errors patterns show worse results in AADT estimation in some specific cases:
  - Analyzing day-type, Saturdays and Sundays show higher errors than Weekdays;
  - Considering the period of the year, summer period shows higher errors than winter period;
  - Looking at the vehicle types, the errors are higher for truck vehicles than passenger vehicles.
- The errors distribution in case of passenger vehicles could be considered acceptable, while high errors in truck vehicles case should be reduced to obtain reliable estimates. Nevertheless the highest errors concentrate on road groups with very low truck volumes, for which the effects in practice become negligible.

Starting from these results, there are some directions in which this work could be extended:

- Since a strong difference was observed in AADT estimation for passenger and truck vehicles, creating separated road groups for passenger and truck vehicles trying to better reproducing the temporal traffic patterns;
- Testing the use of sample 48-h counts, which are suggested by TMG as preferable short traffic counts.
- Testing the influence of socio-economic and land-use characteristics in grouping process, in order to understand if this additional information could help to find more reliable and significant groups.

References

AASHTO. (1992). Joint Task Force on Traffic Monitoring of the AASHTO Highway Subcommittee on Traffic Engineering Guidelines for Traffic Data Programs. Washington, D.C.

Banfield, J., & Raftery, A. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics , 49* (3), 803-821.

Calinski, R., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics , 3*, 1-27.

Davies, D., & Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence , 1* (4), 224-227.

Faghri, A., & Hua, J. (1995). Roadway Seasonal Classification Using Neural Networks. *Journal of Computing in Civil Engineering , 9* (3), 209-215.

Faghri, A., Glaubitz, M., & Parameswaran, J. (1996). Development of Integrated Traffic Monitoring System for Delaware. *Transportation Research Record: Journal of the Transportation Research Board , 1536*, 40-44.

FHWA. (2001). *Traffic Monitoring Guide.* U.S. Department of Transportation. Washington, D.C.

Flaherty, J. (1993). Cluster Analysis of Arizona Automatic Traffic Record Data. *Transportation Research Record: Journal of the Transportation Research Board , 1410*, 93-99.

Fraley, C., & Raftery, A. E. (2009). *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering.* Technical Report No.154, University of Washington, Department of Statistics. Seattle, Washington.

Fraley, C., & Raftery, A. E. (1997). Model-based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association* , 611-631.

Lingras, P. (1995). Hierarchical Grouping Versus Kohonen Neural Networks. *Journal of Transportation Engineering , 121* (4), 364-368.

Lingras, P. (2001). Statistical and Genetic Algorithms Classification of Highways. *Journal of Transportation Engineering , 127* (3), 237-243.

Maechler, M., Rousseeuw, P., Struyf, A., & Hubert, M. (2005). *Cluster Analysis Basics and Extensions.*

Mierswa, I. a., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Milligan, G., & Cooper, M. (1985). An examination of procedures of determining the number of cluster in a data set. *Psychometrika , 50* (2), 159-179.

Pelleg, D., & Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *Seventeenth International Conference on Machine Learning*, (p. 727-734).

R Development Core Team. (2010). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ritchie, S. (1986). A Statistical Approach to Statewide Traffic Counting. *Transportation Research Record: Journal of the Transportation Research Board , 1090*, 14-21.

Rizzo, M. L., & Szekely, G. J. (2010). *energy: E-statistics (energy statistics). R package version 1.2-0.*

Sharma, S. C., & Werner, A. (1981). Improved Method of Grouping Provincewide Permanent Traffic Counters. *Transportation Research Record: Journal of the Transportation Research Board , 815*, 12-18.

Sharma, S. C., Gulati, B., & Rizak, S. (1996). Statewide Traffic Volume Studies and Precision of AADT Estimates. *Journal of Transportation Engineering , 122* (6), 430-439.

Sharma, S., & Allipuram, R. (1993). Duration and Frequency of Seasonal Traffic Counts. *Journal of Transportation Engineering , 119* (3), 344-359.

Sharma, S., Lingras, P. J., Hassan, M. U., & Murthy, N. A. (1986). Road Classification According to Driver Population. *Transportation Research Record: Journal of the Transportation Research Board , 1090*, 61-69.

Sharma, S., Lingras, P., Liu, G., & Xu, F. (2000). Estimation of Annual Average Daily Traffic on Low-Volume Roads. Factor Approach Versus Neural Networks. *Transportation Research Record: Journal of the Transportation Research Board , 1719*, 103-111.

Sharma, S., Lingras, P., Xu, F., & Liu, G. (1999). Neural Networks as Alternative to Traditional Factor Approach to Annual Average Daily Traffic Estimation form Traffic Counts. *Transportation Research Record: Journal of the Transportation Research Board , 1660*, 24-31.

Szekely, G., & Rizzo, M. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification , 22* (2), 151-183.

Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association , 48*, 236-244.