UNIVERSITÁ DEGLI STUDI DI TRIESTE
Sede Amministrativa del Dottorato di Ricerca

UNIVERSITÁ DEGLI STUDI DI PADOVA
Sede Convenzionata

XXIV CICLO DEL DOTTORATO DI RICERCA IN
INGEGNERIA CIVILE E AMBIENTALE
Indirizzo Infrastrutture, Strutture e Sistemi di Trasporto

TRANSPORTATION DATA ANALYSIS. ADVANCES IN DATA MINING AND UNCERTAINTY TREATMENT

(Settore Scientifico-disciplinare ICAR/05)

DOTTORANDO
GREGORIO GECCHELE

DIRETTORE DELLA SCUOLA DI DOTTORATO
PROF. IGINIO MARSON

TUTORI
PROF. ROMEO VESCOVI
PROF. RICCARDO ROSSI

RELATORI
PROF. RICCARDO ROSSI
PROF. MASSIMILIANO GASTALDI

Anno Accademico 2011/2012

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the study of transportation systems, the collection and use of correct information representing the state of the system represent a central point for the development of reliable and proper analyses. Unfortunately in many application fields information is generally obtained using limited, scarce and low-quality data and their use produces results affected by high uncertainty and in some cases low validity.

Technological evolution processes which interest different fields, including Information Technology, electronics and telecommunications make easier and less expensive the collection of large amount of data which can be used in transportation analyses. These data include traditional information gathered in transportation studies (e.g. traffic volumes in a given road section) and new kind of data, not directly connected to transportation needs (e.g. Bluetooth and GPS data from mobile phones).

However in many cases this large amount of data cannot be directly applied to transportation problems. Generally there are low-quality, non-homogeneous data, which need time consuming verification and validation process to be used. Data Mining techniques can represent an effective solution to treat data in these particular contexts since are designed to manage large amount of data producing results whose quality increases as the amount of data increases.

Based on these facts, this thesis analyses the capabilities offered by the implementation of Data Mining techniques in transportation field, developing a new approach for the estimation of Annual Average Daily Traffic from traffic monitoring data.

In the first part of the thesis the most well-established Data Mining techniques are reviewed, identifying application contexts in transportation field for which they can represent useful analysis tools. Chapter 2 introduces the basic concepts of Data Mining techniques and presents a review of the most commonly applied techniques. Classification, Clustering and Association Rules are presented giving some details about the main characteristics of

well-established algorithms. In Chapter 3 a literature review concerning Data Mining applications in the transportation field is presented; a deeper analysis has been done with reference to some research topics which have extensively applied Data Mining techniques.

The second part of the thesis focuses on a deep critical review of the U.S. Federal Highway Administration (FHWA) traffic monitoring approach for the estimation of AADT, which represents the main applicative topic of this research. In Chapter 4 the FHWA factor approach is presented in its original form, while Chapter 5 reports a detailed summary of the modifications proposed in recent years. From the analysis of the review of FHWA approach, a new approach is proposed in Chapter 6, based on the use of Data Mining techniques (Fuzzy clustering and Artificial Neural Networks) and measures of uncertainty from Dempster-Shafter Theory.

The third part of the thesis (Chapter 7) presents the validation study of the proposed approach, reporting the results obtained in the case study context and discussing the main findings.

Finally conclusions and further developments of the research are reported in Chapter 8.

Part I

Data Mining

Chapter 2

Data Mining Concepts

2.1 Introduction

Data Mining (DM) is a general concept which is used to consider a wide number of models, techniques and methods, extremely different one to each other. Many authors have tried to define this concept, providing some definitions, such as:

> Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. (Hand, Manilla, and Smith 2001)

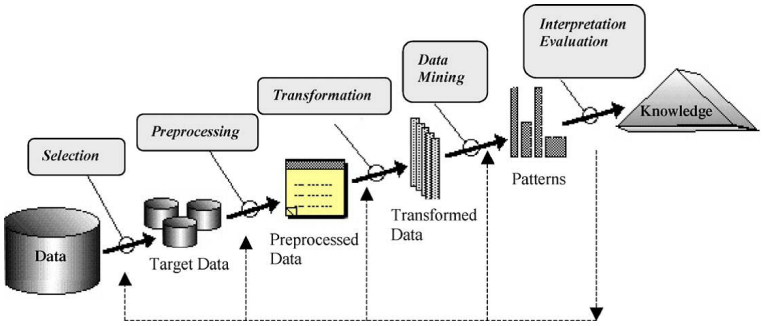
> Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases. (Simoudis 1998)

> Data mining is the exploration and the analysis of large quantities of data in order to discover meaningful patterns and rules. (Berry and Linoff 2004)

Following the framework developed by Fayyad et al. 1996, which incorporates the basic ideas of these definitions, DM could be considered a passage of the Knowledge Discovery in Databases (KDD) Process (Figure 2.1). Data obtained from a certain source are selected, pre-processed and transformed in order to be elaborated by Data Mining techniques. This preprocess is particularly important since in many cases data under analysis are:

- secondary data (i.e. data stored for reasons different from the analysis);
- observational data (i.e. data not obtained from a precise experimental design);

Figure 2.1: *KDD Process. Source: Fayyad et al. 1996*



- large amount of data (i.e. data for which it is difficult to define adequate research hypotheses).

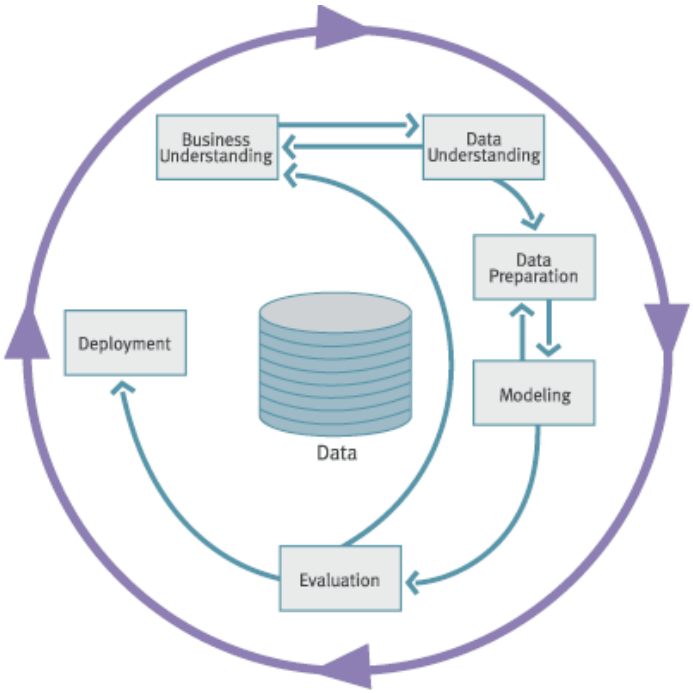
The results obtained by DM need to be interpreted and evaluated to come out with a better comprehension of the problem under analysis. This point is particularly important to understand the capabilities and the limits of DM. The application of DM techniques does not guarantee the solution of a problem, but gives some useful indications that the decision-maker must interpret to solve his/her problem. The more precise is the problem definition, the simpler would be the choice of the most effective technique (or set of techniques) and the better would be the final result of the process.

This concept could become more clear considering the CRISP-DM framework (Chapman et al. 2000), which provides a non proprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or a research unit.

As can be observed in figure 2.2, CRISP-DM is an iterative, adaptive process, consisting of 6 Phases:

1. Business/Research understanding Phase.
   (a) Enunciate the project objectives and requirements clearly in terms of the business or the research unit as a whole
   (b) Translate these goals and restrictions into the formulation of a data mining problem definition
   (c) Prepare a preliminary strategy for achieving these objectives
2. Data understanding Phase.
   (a) Collect the data
   (b) Use exploratory data analysis to familiarize yourself with the data and discover initial insights

Figure 2.2: *CRISP-DM. Source: Chapman et al. 2000*



(c) Evaluate the quality of the data

(d) If desired, select interesting subsets that may contain actionable patterns

3. Data preparation Phase.

   (a) Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive

   (b) Select the cases and variables you want to analyze and that are appropriate for your analysis

   (c) Perform transformations on certain variables, if needed

   (d) Clean the raw data so that is ready for the modeling tools

4. Modeling Phase.

   (a) Select and apply appropriate modeling techniques

   (b) Calibrate model settings to optimize results

   (c) Remember that often, several different techniques may be used for the same data mining problem

(d) If necessary, loop back to the data preparation phase to bring the form of data into line with the specific requirements of a particular data mining technique

5. Evaluation Phase.

   (a) Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field

   (b) Determine whether the model in fact achieves the objectives set for in the first phase

   (c) Establish whether some important facet of the business or research problem has not been accounted for sufficiently

   (d) Come to a decision regarding use of the data mining results

6. Deployment Phase.

   (a) Make use of the model created: model creation does not signify the completion of a project

   (b) Example of a simple deployment: Generate a report

   (c) Example of a more complex deployment: Implement a parallel data mining process in another department

   (d) For businesses, the customer often carries out the deployment based on your model

In practice the analyst can choose the technique to adopt from a large number of alternatives. Traditionally DM techniques have been divided in categories, which are related to the objective one would achieve from the analysis (Berry and Linoff 2004):

Classification. The main objective is assigning the data in pre-defined classes, usually described by qualitative variables;

Estimation. The main objective is producing estimates of a quantitative variable, generally continuous;

Prediction. The main objective is producing estimates of future values that can be assumed by a quantitative variable of interest;

Clustering. The main objective is subdividing observations (data) in groups not already defined (clusters), being maximized the similarity among observations belonging to the same group and minimized the similarities with observations in other groups;

Affinity Grouping (Association). The main objective is defining rules which describe existing patterns in data, connecting the variables of interest one to each other;

Profiling. The main objective is providing a description of the observations.

Another common classification of DM techniques is between *supervised* and *unsupervised* techniques. In the first case the dataset under analysis has clearly defined the solution that the algorithm has to learn and reproduce on new data (e.g. Classification task); in the second case the dataset do not have a pre-defined solution and the DM techniques try to identify patterns/relationships among data (e.g. Clusterin or Association rules).

In any case, the choice of the analyst will consider techniques which come from two different fields: machine learning and statistics. In this thesis more attention has been given to the analysis of machine learning approaches, since they represent more typically Data Mining techniques and are less explored tools for the research compared to statistical techniques.

For these reasons in the following sections the most important DM techniques will be introduced and described, following in particular the book of Witten and Frank (2005). Major details will be given for the categories commonly employed in transportation systems analysis, which have been considered and applied in this thesis: Classification, Clustering and Association Rules techniques.

2.2 Classification

Classification techniques are probably the most commonly applied DM techniques and have as a main objective the insertion of observations in pre-defined classes.

Given a dataset of elements (observations) $D = \{t_1, t_2, \ldots, t_n\}$ and a set of classes $C = \{C_1, C_2, \ldots, C_m\}$, a *classification problem* is defining a map $f : D \rightarrow C$ for which t_i is assigned to just one class. A class C_j has the elements mapped by f, that is: $C_j = t_i | f(t_i) = C_j, 1 \leq i \leq n, t_i \in D$.

Classes must be pre-defined, non overlapping and such that they partition completely the dataset. Generally a classification problem is solved in two steps, following a *supervised learning* approach:

1. *Training step*. A classification model is defined based on classified data.

2. *Application step*. Elements belonging to the dataset D are classified by the model developed in the training step.

It must be observe that a large number of techniques employed to solve classification problems can be applied to estimation or prediction problems. In these cases, models do not refer to pre-defined classes, but produce as output a quantitative response variable, generally continuous.

Four broad categories of techniques are generally employed to solve classification problem, based on the approach they adopt: techniques based

on measures of distance, Decision Trees, Artificial Neural Networks and Statistical approaches.

Excluding statistical approaches, the most important techniques belonging to each category will be presented in the following sections.

2.2.1 Techniques Based on Measures of Distance

The concept of distance (and similarity) can be successfully applied in classification problems considering that observations in the same class should be more similar one to each other than observations in other classes. The main difficulty in applying this approach is the choice of similarity measures adequate to the variables adopted.

Formally it is correct to distinguish between measures of similarity and measures of distance, depending on the type of variables adopted.

In case of **qualitative variables** similarity among elements in the dataset must be used.

The similarity $sim(t_i, t_j)$ between two observations t_i e t_j in a dataset D, is defined as a map from $D \times D$ to interval $[0, 1]$, such that $sim(t_i, t_j) \in [0, 1]$.

Similarity measures have some properties:

1. Non negativity: $\forall t_i, t_j \in D, sim(t_i, t_j) \geq 0$;
2. Normalization: $\forall t_i \in D, sim(t_i, t_i) = 1$;
3. Symmetry: $\forall t_i, t_j \in D, sim(t_i, t_j) = sim(t_j, t_i)$;

In practice qualitative variables are re-codified using binary variables (dummy variables) and similarity indices for these variables are used. Considering 2 observations codified using p binary variables, absolute frequencies are calculated for 4 situations:

- CP = co-presences. Number of variables for which both observations have value 1;
- CA = co-absences. Number of variables for which both observations have value 0;
- AP (and PA) = absences-presences (and presences-absences). Number of variables for which the first (second) observation has value 1 and the second (first) has value 0.

Different indices of similarity have been proposed, combining the aforementioned four values in different ways:

- Russel and Rao's index of similarity.

$$sim(t_i, t_j) = \frac{CP}{p} \tag{2.1}$$

- Jaccard's index of similarity.

$$sim(t_i, t_j) = \frac{CP}{CP + PA + AP} \tag{2.2}$$

  If there is a complete dissimilarity between observations ($CA = p$), the index is undetermined.

- Sokal e Michener's index of similarity. (simple matching coefficient)

$$sim(t_i, t_j) = \frac{CP + CA}{p} \tag{2.3}$$

If **quantitative variables** are adopted, measures of distance are calculated. For quantitative variables some properties are satisfied:

1. Non negativity: $\forall t_i, t_j \in D, dis(t_i, t_j) \geq 0$;
2. Identity: $\forall t_i, t_j \in D, dis(t_i, t_j) = 0 \Leftrightarrow t_i = t_j$;
3. Symmetry: $\forall t_i, t_j \in D, dis(t_i, t_j) = dis(t_j, t_i)$;
4. Triangular inequality: $\forall t_i, t_j, t_k \in D, dis(t_i, t_j) \leq dis(t_i, t_k) + dis(t_j, t_k)$;

In a k-dimensions space, a large number of measures can be used; two measures commonly adopted are:

1. Euclidean

$$dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2} \tag{2.4}$$

2. Manhattan

$$dis(t_i, t_j) = \sum_{h=1}^{k} |(t_{ih} - t_{jh})| \tag{2.5}$$

In particular Euclidan distance is the most adopted measure of distance, even if it can be highly influenced by the presence of extreme values. These effects are usually due to variables measured on different scales. Normalization (or standardization) can be sufficient to reduce or solve this problem.

Two extremely simple techniques based on the measure of similarity or distance are presented: the simplified approach and the K Nearest Neighbour.

Simplified Approach

This approach has been derived from Information Retrieval(IR) field. It assumes that each observation t_i in the dataset is defined as a vector of numerical values $\{t_{i1}, t_{i2}, \ldots, t_{ik}\}$ and that each class C_j is defined as a vector of numerical values $\{C_{j1}, C_{j2}, \ldots, C_{jk}\}$. Each observation is simply assigned to the class to which the measure of similarity is larger. The vector representative of each class is generally calculated using the center of the region which subdivides the training set observations.

K Nearest Neighbour

K Nearest Neighbour (KNN) is a very simple algorithm commonly adopted for classification. When a new observation is presented to the algorithm, it calculates the distance between this observation and each element in the training set. Then only the K nearest elements (the "nearest neighbours") are considered and the new observation is assigned to the class which contains the larger number of elements. Due to its simplicity, *KNN* algorithm is extremely sensitive to the choice of K value. A simple rule of thumb suggests to chose $K = \sqrt{T}$, where T is the number of elements belonging to the training set.

2.2.2 Decision Trees

Decision trees algorithms represent classification techniques which divide the instances on the basis of a hierarchy of the attribute space, first considering the most important attribute (the root of the tree) and progressively using the other attributes till the reaching the attribution of a certain class (the leaves of the tree).

Decision trees can be considered non-parametric predictive models, since they do not make specific hypotheses on the probability distribution of the dependent variable. This fact generally requires more computational resources and could produce some dependences from the observed data, limiting the generalization of the results to other datasets (overfitting problem).

Formally the problem can be expressed as:

Being $D = \{(t_1), \ldots, (t_n)\}$ a set of observations $t_i = \{t_{i1}, t_{i2}, \ldots, t_{ik}\}$ defined by the attributes $\{(A_1), \ldots, (A_h)\}$ and a set of classes $C = \{(C_1), \ldots, (C_m)\}$, a *decision tree* (*DT*) is a tree associated to D, with the following properties:

- each node in the tree is identified by an attribute, A_i;
- each branch is identified by a predicate applied to the attribute associated to the parent node;
- each leaf node is identified with a class, C_j.

Decision trees can be divided in two main types:

1. *Classification trees*, if the output is the assignment of an instance to one of the predetermined classes;

2. *Regression trees*, if the output is the estimate of a quantitative variable.

From the applicative point of view the differences are minimal, since changes are needed only in the definition of the measurements and in the interpretation of the results, not in the algorithmic structure.

For each output variable y_i, a regression tree produces an estimated value $\hat{y}_i$ that is equal to the mean of the response variable of the leaf m which the observation i belongs to, that is:

$$\hat{y}_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m} \tag{2.6}$$

In the case of a classification tree, the estimated values are the probabilities of belonging to a certain group π_i. For binary classification (i.e. classification with only two classes) the estimated probability of success is:

$$\pi_i = \frac{\sum_{l=1}^{n_m} y_{lm}}{n_m} \tag{2.7}$$

where the observations y_{lm} can assume the values 0 or 1, and the probability of belonging to a group is the observed proportion of successes in the group m.

In both cases $\hat{y}_i$ and π_i are constant for all the values of the group.

For each leaf of the tree a rule can be derived: usually it is chosen the one which correspond to the class with the majority of instances (majority rule). By this way each path in the tree represents a classification rule which divides the instances on the classes.

The learning algorithm has a top-down recursive structure. At each level the *best splitting attribute* is chosen as the one which induces the best segmentation of the instances and the algorithm creates as many branches as the number of predicates of the splitting attribute (*splitting predicates*). In case of binary trees the branches are two. The algorithm recursively analyses the remaining attributes, till the reaching of a certain stopping criterion which determines the definitive structure of the tree.

The main differences between decision trees regard:

1. the criterion function for the choice of the best splitting attribute;

2. the stopping criterion for the creation of the branches.

Criterion Functions

Considering the criterion function, at each step of the procedure (at each level of the tree structure) a function $\Phi(t)$, which gives a measure of the diversity between the values of the response variable in the children groups generated by the splitting ($s = 2$ for binary trees) and the ones in the parent parent group t, is used as index.

Being t_r $r = 1, \ldots, s$ the children groups generated by the splitting and p_r the proportion of observations in t allocated to each children node, with $\sum p_r = 1$, the criterion function is generally expressed as:

$$\Phi(s, t) = I(t) - \sum_{r=1}^{s} I(t_r) p_r \tag{2.8}$$

where $I(t)$ is an impurity function.

For regression trees the output variable are quantitative, therefore the use of the variance measure is a logical choice. More precisely for a regression tree the impurity of a node t_r can be defined as:

$$I_V(t_r) = \frac{\sum_{l=1}^{n_{t_r}} (y_{lt_r} - \hat{y}_{t_r})^2}{n_{t_r}} \tag{2.9}$$

where $\hat{y}_{t_r}$ is the mean estimated value for the node t_r, which has n_{t_r} instances.

For classification trees the most common impurity measures adopted are:

1. *Misclassification impurity*

$$I_M(t_r) = \frac{\sum_{i=1}^{n_{t_r}} \mathbf{1}(y_{lt_r}, y_k)}{n_{t_r}} = 1 - \pi_k \tag{2.10}$$

   where y_k is the class with estimated probability π_k; the notation $\mathbf{1}$ indicates the indication function, which assumes the value 1 if $y_{lt_r} = y_k$ and 0 otherwise.

2. *Gini impurity*

$$I_G(t_r) = 1 - \sum_{i=1}^{m} \pi_i^2 \tag{2.11}$$

   where π_i are the estimated probability of the m classes in the node t_r.

3. *Entropy impurity*

$$I_E(t_r) = -\sum_{i=1}^{m} \pi_i \log \pi_i \tag{2.12}$$

The use of the impurity functions could be extended in order to globally evaluate a decision tree. Being $N(T)$ the number of leaves of a tree T, the total impurity of the tree can be calculated as:

$$I_T = \sum_{m=1}^{N(T)} I(t_m) p_m \tag{2.13}$$

where p_m are the proportions of the instances in the final classification.

Stopping Criteria

Theoretically the stopping criteria would activate when all the instances of the training set were correctly classified. However in practice it's better to cut the latest branches added to the tree (pruning), to prevent the creation of excessively long tree and over-fitting problems. This means that the tree must give a classification parsimonious and discriminatory at the same time. The first property leads to decision trees with a small number of leaves, with decision rules easy to interpret. Conversely the second property leads to a large number of leaves, extremely different one to each other.

Other relevant factors

Other factors are relevant for a correct definition of a decision tree. They include:

- An accurate Exploratory Data Analysis process, which excludes outlier data and limits the number of classes;
- An adequate number of observations included in the training set;
- A balanced structure of the tree, which has the same length for each path from the root node to the leaves;
- The pruning process, which improve classification performances, removing sub-trees resulting from over-fitting phenomena.

Before introducing the more common algorithms adopted for building decision trees, it can be useful summarize the main advantages and limits of decision trees. The main advantages of decision trees are the ease of use and the efficiency, the availability of rules that facilitate the interpretation of the results, the capability of handling a large amount of attributes and the computational scalability.

The main limits are the difficulty in using continuous data or missing data, a tendency to over-fitting, that can be counterbalance by the pruning technique, the fact that correlation among the attributes are ignored and that the solution space is divided in rectangular regions and this is not good for all classification problems.

ID3

ID3 algorithm (Quinlan 1986) is based on the Information Theory principles. The rationale of the technique is minimizing the expected number of comparison choosing splitting attributes that give the maximum increment of information.

If one considers that decision trees divide the research space in rectangular regions, the division attribute which produces the maximum increment of information is that one which divides the research space in two subspaces with similar dimensions. In fact this attribute has attribute values with the same amount of information.

Therefore, at each level of the tree, ID3 calculates Entropy impurity associated to each attribute and chooses the attribute which produces the largest increment in the criterion function $\Phi(t)$, called information gain. Consequently information gain is calculated as the difference between the entropy before and after the splitting, that is:

$$\Phi(s,t) = Gain(s,t) = I_E(t) - \sum_{r=1}^{s} I_E(t_r)p_r \tag{2.14}$$

where:

- t_r $r = 1, \ldots, s$ are the child branches generated by the splitting;
- p_r is the proportion of observations in t allocated to each child branch, with $\sum p_r = 1$.

C4.5 and C5.0

C4.5 algorithm (Quinlan 1993) represents an improvement of the ID3 algorithm from different points of view. The main one is the substitution of the Information Gain with the *GainRatio* as a criterion function, defined as:

$$\Phi(s,t) = GainRatio(s,t) = \frac{Gain(s,t)}{\sum_{r=1}^{s} p_r \log p_r} \tag{2.15}$$

which it is more stable and less influenced by the number of values of each attribute. Other improvement are summarized in the following paragraphs.

Numerical Attributes C4.5 algorithm produces binary trees, restricting the possibilities to a two-way split at each node of the tree. The gain ratio is calculated for each possible breakpoint in the domain of each attribute.

The main issue about using binary trees with numerical variables is that successive splits may continue to give new information and to create trees complex and particularly difficult to understand, because the test on a single numerical attribute are not locate together.

To solve this problem is possible to test against several constants at a single node of the tree or, in a simpler but less powerful solution, to prediscretize the attributes.

Pruning A clear distinction can be made between two different types of pruning strategies:

1. Prepruning (Forward pruning). This strategy involves the choice of when stopping the development of the sub-trees during the tree-building process.

2. Postprunign (Backward pruning). In this case the process of pruning is made after the tree was built.

Even if the first strategy seems to be more attractive, since it can limit the development of sub-trees, the second one allows the presence, in some cases, of synergies between the attributes, that the first one can eliminate in the building phase. Analysing in more depth postpruning strategy (since it the most implemented in learning algorithm), two different operations can be considered:

1. Subtree Replacement

2. Subtree Raising

At each node a learning scheme can decide which one of the two techniques adopting, even both or none of them.

Subtree Replacement refers to the operation of taking some sub-trees and replace them with single leaves. This operation certainly decreases the accuracy on the training set, but can give an opposite effect on test set. When implemented, this procedure starts from the leaves and goes back to the root node.

Subtree Raising refers to the operation of replacing an internal node with one of the node below it. It is a more complex and time-consuming procedure and it's not always necessary to implement it; for this reason it is usually restricted to the subtrees of the most popular branch.

A relevant point is how to decide when to perform the two operations.

Estimating Errors To do this is necessary to estimate the error rate that would be expected at a particular node given an independently chosen dataset, both for leaves and internal nodes. C4.5 algorithm uses the training set to do this, but it is more statistically sound using an independent dataset (different from the training and the test sets), performing the so-called *reduced-error* pruning.

C4.5 algorithm analyses what happens on a certain node of the tree, considering that the majority class could be used to represent the node. From the total number of instances N, a certain error E, represented by the minority classes, is made.

At this point it is assumed that the true probability of errors at the node is q and that the N instances are generated by a Bernoulli process with parameter q, of which E represent the errors. Since the values of E and N are measured on the training data, and not on an independent dataset, a pessimistic estimate of the error is made, using the upper limit of the confidence limit.

This means that, given a confidence level c (the default value used by C4.5 is 25%), a confidence limit z is such that:

$$Pr\left[\frac{f-q}{q(1-q)/N} > z\right] = c \tag{2.16}$$

where N is the number of instances, $f = E/N$ is the observed error rate, and q is the true error rate. This upper confidence limit can be used as a pessimistic estimate of the error rate e at the node considered:

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{2.17}$$

Fixing the value of the confidence level c to 25% gives a value of $z = 0.69$, even if a higher level can be chosen.

In practice the errors is calculated at each node, considering the combined error estimate for the children and the estimate error for the parent node. If the error of the parent is less the the error for the children nodes, they are pruned away.

The estimated errors obtained with this calculation must be considered with particular attention, since they are based on particular strong assumptions; however the method seems to work reasonably well in the practice.

Rules The final decision tree can be used to create a set of rules describing the classification of the instances; the use of the estimated error allows the selection of the smallest set of rules, even if this process can lead to long computational efforts.

'**C5.0** algorithm is a commercial version of C4.5, implemented in many DM packages, modified in order to be used efficiently in datasets with large amount of data. One of the main improvements is the implementation of boosting, which allows to create multiple training datasets; specific classification trees are created using these subsets and merged together in a final tree with the boosting operation.

CART

The *Classification and Regression Tree* (CART) is a technique that produces binary trees (regression or classification ones) on the basis of the entropy impurity I_E. Differently from the ID3 at each node it creates only two branches, selecting the best splitting following the criterion function:

$$\Phi(s,t) = 2p_L p_R \sum_{j=1}^{m} |P(C_j|t_L) - P(C_j|t_R)| \tag{2.18}$$

The function is calculated with reference to the node t for each of the two possible splitting s:

1. L and R represent the two children created with the split;
2. p_L and p_R are the probability that the instances of the training set are on the right or the left part of the tree; they were estimated as the ratio between the instances of each child branch and the total number of instances of the training set;
3. $P(C_j|t_L)$ and $P(C_j|t_R)$ are the probability that an instance belongs to the class C_j and to the left or right child branch; they were estimated as the ratio between the instances of the class C_j in each child branch and number of instances in the parent node.

Missing values are not considered in the learning phase, which iteratively goes on until the splitting of the tree does not increments the performances of the tree.

The stopping criterion of the splitting process is related to the global performance index of the tree and to the pruning strategy. Being T_0 the biggest tree and T a general smaller tree, the pruning process determines an optimal tree starting from T_0, minimizing the loss function:

$$C_\alpha(T) = I(T) + \alpha N(T) \tag{2.19}$$

where, for a given tree T, $I(T)$ is the total impurity function, $N(T)$ is the number of leaves in the tree and α is a constant value which linearly penalizes the complexity of the tree.

The pruning process should be operated in combination with accurate treatments of available data, distinguishing between training set, adopted for the building of the tree, and testing set, adopted for a correct evaluation of the model, including the calculation of loss function and the pruning process. In this sense cross-validation, which separates one set of observations (learning sample) to another completely independent set of observations (testing sample), can be an effective solution

CHAID

The CHAID (Chi-squared Automatic Interaction Detector) tree classification method was originally proposed by Kass 1980. CHAID is a recursive partitioning method that builds non-binary trees, based on an algorithm particularly well suited for the analysis of larger datasets, to solve both regression-type or classification-type problems.

The basic algorithm differs in case of classification or regression problems. In the first case, when the dependent variable is categorical, relies on the Chi-square test to determine the best next split at each step, while for regression-type problems the program will actually compute F-tests. Specifically, the algorithm proceeds as follows:

Preparing predictors. The first step is to create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations (prediscretization). For categorical predictors, the categories (classes) are "naturally" defined.

Merging categories. The next step is to cycle through the predictors to determine for each predictor the pair of predictor categories that is least significantly different with respect to the dependent variable, computing a Chi-square test for classification problems and F tests for regression problems.

If the respective test for a given pair of predictor categories is not statistically significant as defined by an alpha-to-merge value, then it will merge the respective predictor categories and repeat this step (i.e., find the next pair of categories, which now may include previously merged categories).

If the statistical significance for the respective pair of predictor categories is significant (less than the respective alpha-to-merge value), then (optionally) it will compute a Bonferroni adjusted p-value for the set of categories for the respective predictor.

Selecting the split variable. The next step is to choose the predictor variable that will yield the most significant split, i.e. having the smallest

adjusted p-value; if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

Continue this process until no further splits can be performed (given the alpha-to-merge and alpha-to-split values).

A general issue of CHAID, is that the final trees can become very large, diminishing the ease of understanding characteristic of decision tree methods.

2.2.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are highly connected systems of basic computational elements, created with the objective of imitate the neurophysiology of human brain.

A basic neural network is formed by a set of computational elements (called nodes, neurons or units), connected one to each other by weighted connections. Neurons are organised in layers and each node is connected with neurons belonging to previous or following layer. Each node operates independently by the others and its activity is determined by the input values received simultaneously by the neurons belonging to the previous layer. The neuron is activated if input signals overcome a pre-fixed threshold (bias) and produces a (generally unique) output signal.

Being j a generic neuron with 'bias' θ_j, it receives n input signals $\mathbf{x} = (x_{1j}, x_{2j}, \ldots, x_{nj})$ from the neurons of the previous layer, with associated weights $\mathbf{w} = (w_{1j}, w_{2j}, \ldots, w_{nj})$. The neuron elaborates input signals $\mathbf{x}$ using a combination function and the result (potential P_j) is transferred by a transfer function, producing the final output y_j.

Generally the combination function is a linear combination of input signals $\mathbf{x}$ and bias θ_j, which can be represented as a further input with signal $x_0 = 1$ and weight $w_{0j} = -\theta_j$:

$$P_j = \sum_{i=0}^{n} x_{ij} w_{ij} \tag{2.20}$$

The output y_j of the j-th neuron is given by the application of a transfer function to the potential P_j:

$$y_j = f(\mathbf{x}, \mathbf{w}) = f(P_j) = f\left(\sum_{i=0}^{n} x_{ij} w_{ij}\right) \tag{2.21}$$

The functional form of transfer function $f(P_j)$ is defined during the ANN model specification. The most common transfer functions used are:

1. Linear transfer function

$$f(P_j) = \beta P_j + \alpha \tag{2.22}$$

where α and β are constant values.

2. Step-wise transfer function

$$f(P_j) = \begin{cases} \alpha & P_j > \theta_j \\ \beta & P_j \leq \theta_j \end{cases} \tag{2.23}$$

When $\alpha = 1$, $\beta = 0$ and $\theta_j = 0$ $f(P_j)$ is the sign function, which takes values 0 if the input is negative and +1 if it is negative.

3. Sigmoidal transfer function

$$f(P_j) = \frac{1}{1 + e^{-\alpha P_j}} \tag{2.24}$$

where α is a positive parameter.

4. Hyperbolic tangent transfer function

$$f(P_j) = \frac{1 - e^{-\alpha P_j}}{1 + e^{-\alpha P_j}} \tag{2.25}$$

5. Gaussian transfer function

$$f(P_j) = e^{\frac{-P_j^2}{V}} \tag{2.26}$$

where P_j is the mean and V is the variance of the function.

6. Softmax transfer function

$$softmax(v_j) = \frac{e^{v_j}}{\sum_{h=i}^{g} e^{v_h}} \quad j = 1, \ldots, n \tag{2.27}$$

Each ANN has its own structure, which organizes the neurons in three types of layers: input layer, output layer and hidden layers. The input layer neurons accept input from the external environment, and generally each input neuron represents an explicative variable. The output layer neurons send data produced by the neural network to the external environment. Hidden layer neurons connect input layer with output layer, without any relationship with external environment. One or more hidden layers can be introduce in the neural network structure. Their main role is elaborating the information obtained from the input layer.

Neural networks can be classified on the basis of four characteristics of their architecture (or topology):

Differentiation between input and output layers. ANNs can have separate input and output layers (the majority of cases) or having layers which work at the same time like input and output layers;

Number of layers.

Direction followed by the information flow in the network.

- Feed-forward networks: the information is elaborated from one layer to the next one, following one specific direction.
- Feed-back networks: the information can be elaborated from one layer to the other connected layers, in any direction.

Type of connections.

- Fully connected networks: each neuron in a layer is connected only to all the neurons in the next layer;
- Fully interconnected networks: each neuron in a layer is connected to all the neurons in the other layers.

The choice of the architecture is made considering the objective of the analysis to be made with the ANN and data characteristics. As an example, some commonly applied neural networks are:

- Auto-associative Neural Networks. The architecture has one layer of fully interconnected neurons which behave like input and output layers;
- Single Layer Percetrons (SLP). These networks have a feed-forward architecture, with n input neurons $x_1, x_2, \ldots, x_n$ and p output neurons $y_1, y_2, \ldots, y_n$ fully connected.
- Multi-Layer Perceptrons (MLP), These networks have a feed-forward architecture with n input neurons, p output neurons and h_i neurons in hidden layer i. The number of hidden layers i can vary depending on the needs.

A final parameter to classify ANN is given by the type of training process they follow:

- in case of **supervised learning** the values of explanatory variable and dependent variable are given to ANN for each observation in the training dataset. The objective of the ANN is modifying its structure (i.e. weights) such that the sum of distances between observed and estimated values of the response variables is minimum. The ANN obtained at the end of the learning can be applied for classification and estimation tasks;

- in case of **unsupervised learning** only the values of explanatory variable are given to ANN for each observation in the training dataset. The objective of the ANN is modifying its structure (i.e. weights) such that the observations are clustered in an effective way. The ANN obtained at the end of the learning can be applied for clustering task.

In the following sections major details will be given about Multi-Layer Perceptrons (MLP), which are the ANNs most applied for classification and estimation tasks and Adaptive Resonance Theory Neural Networks, as representative of another type of Neural Networks. Details concerning unsupervised neural networks are given in section 2.3.3, since they can be applied in clustering problems.

Multi-Layer Perceptrons

Multi-Layer Perceptrons (MLP) are feed-forward architecture with fully connected neurons, organised in one input layer, one output layer and one or more hidden layers. The simplest MLP is composed by one input layer with n neurons $x_1, x_2, \ldots, x_n$, one output layer p with neurons $y_1, y_2, \ldots, y_p$ and h neurons in the hidden layer.

The layers are connected one to each other by two sets of weights: the first set is composed by weights w_{ik} $(i = 1, \ldots, n; k = 1, \ldots, h)$, which connect the neurons in the input layer with the neurons in the hidden layer, the second one by weights z_{kj} $(k = 1, \ldots, h; j = 1, \ldots, p)$, which connect the neurons in the hidden layer with the neurons in the output layer.

Each node in the input layer elaborates the information from the input layer producing the output:

$$h_k = f(\mathbf{x}, \mathbf{w}_k) \tag{2.28}$$

The output from hidden layer neurons is fed to the output layer neurons which produce the final output:

$$y_j = g(\mathbf{h}, \mathbf{z}_j) \tag{2.29}$$

Therefore the final output produced by the j-th neuron of the output layer is given by:

$$y_j = g_j\Big(\sum_k h_k z_{kj}\Big) = g_j\Big(\sum_k z_{kj} f_k\Big(\sum_i x_i w_{ik}\Big)\Big) \tag{2.30}$$

Some aspects are important for the correct design of a MLP:

- Coding of variables. Quantitative variables are usually described by a single neuron. For categorical variables one neuron is needed for each mode of the variable (as done with dummy variables);

- Variable transformation. In some cases it can be useful transform original explanatory variables. In particular standardization can be a good choice when variables are measured using different scales;
- Choice of the architecture. This choice is relevant for the quality of final results, however it is difficult to define specific guidelines. The analysis of performances calculated with techniques such as cross validation can give useful indication to compare alternative architectures;
- Learning process. The adaptation of weights in the training phase must be carefully analysed. In particular attention should be given to two aspects:
  - The **choice of the error function** between observed values and values determined by the MLP;
  - The **choice of the optimization algorithm**.

Choice of the error function Given a training set of observations $D = \{(x_1, t_1), \ldots, (x_n, t_n)\}$, the definition of error function is based on the principle of maximum likelihood, which leads to the minimization of function:

$$E(\mathbf{w}) = -\sum_{i=1}^{n} \log[p(\mathbf{t}_i|\mathbf{x}_i; \mathbf{w})] \tag{2.31}$$

where $p(\mathbf{t}_i|\mathbf{x}_i; \mathbf{w})$ is the distribution of response variable, conditioned by values of input variables and by weights of the neural network.

If the MLP is adopted for the estimation of a continuous response variable (**regression case**) each component $t_{i,k}$ of the response vector t_i is defined by:

$$t_{i,k} = y_{i,k} + e_{i,k} \quad k = 1, \ldots, q \tag{2.32}$$

where $y_{i,k} = \mathbf{y}(\mathbf{x}_i, \mathbf{w})$ is the k-th component of the output vector $\mathbf{y}_i$ and $e_{i,k}$ is a random error component. Random errors are assumed to be normally distributed, therefore the error function is:

$$E(\mathbf{w}) = \sum_{i=1}^{n} \sum_{k=1}^{q} (t_{i,k} - y_{i,k})^2 \tag{2.33}$$

that is similar to a least-squares function.

Otherwise if the MLP is adopted for the estimation of a categorical response variable (**classification case**) the output is the estimated probability that an observation belongs to the different classes.

Each class is represented by a neuron, and the activation of the neuron is the conditional probability $P(C_k|\mathbf{x})$ where C_k is the k-th class and $\mathbf{x}$ is the

input vector. Therefore the output value represent the estimated probability that the i-th observation belongs to the k-th class C_k.

The error function becomes:

$$E(\mathbf{w}) = -\sum_{i=1}^{n}\sum_{k=1}^{q} t_{i,k}\log(y_{i,k}) + (1 - t_{i,k})\log(1 - y_{i,k}) \tag{2.34}$$

Choice of the optimization algorithm The value of function error $E(\mathbf{w})$ is highly non-linear with reference to the weights, therefore different minima, which satisfy $\nabla E = 0$, could exist. The search of the optimum point w^* is done iteratively, with algorithms that from an initial estimate $\mathbf{w}^{(0)}$ generate a series of solution $w^{(s)}, s = 1, 2, \ldots$ which converges to the value $\hat{\mathbf{w}}$. The algorithms follow some basic steps:

1. identify a research direction $d^{(s)}$:
2. choose the value $\alpha^{(s)}$ and set $\mathbf{w}^{(s+1)} = \mathbf{w}^{(s)} + \alpha(s)d(s) = \mathbf{w}^{(s)} + \Delta\mathbf{w}^{(s)}$;
3. if the convergence (or stopping) criterion is verified, set $\hat{\mathbf{w}} = \mathbf{w}^{(s+1)}$, otherwise $s = s + 1$ and algorithms go back to step 1.

These algorithms must be used carefully, since different elements can influence the convergence to the optimal solution, including:

- the choice of initial weights $\mathbf{w}^{(0)}$;
- the choice of convergence criterion, which can be set as a function of the number of iterations, the computing time or the value of error function;
- the *learning rule*, that is the way the increment of weights $\Delta\mathbf{w}^{(s)}$ is calculated.

In particular the choice of the *learning rule* is important for the convergence of the algorithm. Different learning rules have been proposed and are commonly used. Here only the most important are reported.

One considers the j-th neuron of the neural network, which produces the output y_j, connected to x_{ij} input neurons by weights w_{ij}. Some learning rule are:

- **Hebb's rule** (Hebb 1949)

$$\Delta w_{ij} = c x_{ij} y_j \tag{2.35}$$

  where c is the learning rate. As a rule of thumb it can be assumed that $c = \frac{1}{N}$, where N is the number of elements in the training dataset.

- **Delta rule** (Widrow and Hoff 1960)

$$\Delta w_{ij} = c x_{ij}(d_j - y_j) \tag{2.36}$$

  where d_j is the value assumed by the output in the training dataset and c is again the learning rate.

- **Generalized Delta rule**

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{2.37}$$

  where η is a learning parameter, usually included in the interval $[0, 1]$, and $\frac{\partial E}{\partial w_{ij}}$ is the gradient of the error for the weight w_{ij}.

The Generalized Delta rule is applied in the **backpropagation algorithm**, which is probably the most applied algorithm for the optimization of neural networks. The structure of this algorithm is similar to the general one; the name is due to the characteristic that the weights are updated at each iteration 'going back' from the output to the input layer. The analytical complexity of the formulation depends on different factors, such as the type of function error, the type of transfer function and the type of neurons to update.

One considers the simple MLP structure introduced at the beginning of this section and the notation adopted. In addition to this the neurons have a continuous non-linear transfer function (e.g. sigmoidal) and t_j^m is the correct output for input pattern μ. The backpropagation algorithm follows some steps:

- Given an input pattern $\mathbf{x}^\mu$, calculate the output of hidden and output neurons

$$h_k^\mu = \Phi\left(\sum_{i=0} w_{ik} x_i^\mu\right) \tag{2.38}$$

$$y_j^\mu = \Phi\left(\sum_{k=0} z_{kj} h_k^\mu\right) \tag{2.39}$$

- Calculate the error function for the

$$E_w = \frac{1}{2} \sum_\mu \sum_j \left(t_j^\mu - y_j^\mu\right)^2 \tag{2.40}$$

  that can be expanded in the form

$$E_w = \frac{1}{2} \sum_{\mu} \sum_{j} \left(t_j^{\mu} - \Phi \left[\sum_{k} z_{kj} \Phi \left(\sum_{i} w_{ik} x_i^{\mu} \right) \right] \right)^2 \tag{2.41}$$

- Apply the generalized delta rule to calculate the variation of weights z_{kj}

$$\Delta z_{kj} = -\eta \frac{\partial E}{\partial z_{kj}} = \eta \sum_{\mu} \left(t_j^{\mu} - y_j^{\mu} \right) \Phi' \left(P_j^{\mu} \right) h_k^{\mu} \tag{2.42}$$

  where Φ' is the derivative of Φ and P_j^{μ} is the activation potential of neuron i for pattern μ. If $\delta_j^{\mu} = (t_j^{\mu} - y_j^{\mu}) \Phi'(P_j^{\mu})$ is introduced, the formulation can be simplified as

$$\Delta z_{kj} = \eta \sum_{\mu} \delta_j^{\mu} h_k^{\mu} \tag{2.43}$$

- Calculate the variation Δw_{ik} for weights w_{ik}

$$\Delta w_{ik} = -\eta \frac{\partial E}{\partial w_{ik}} = -\eta \sum_{\mu} \frac{\partial E}{\partial h_k^{\mu}} \frac{\partial h_k^{\mu}}{\partial w_{ik}} \tag{2.44}$$

  that can be expanded in the form

$$\Delta w_{ik} = \eta \sum_{\mu} \sum_{j} \left(t_j^{\mu} - y_j^{\mu} \right) \Phi' \left(P_j^{\mu} \right) z_{kj} \Phi' \left(P_j^{\mu} \right) x_i^{\mu} \tag{2.45}$$

  or

$$\Delta w_{ik} = \eta \sum_{\mu} \sum_{j} \delta_j^{\mu} z_{kj} \Phi' \left(P_j^{\mu} \right) x_i^{\mu} \tag{2.46}$$

The algorithm can be applied following two alternative modes.

- **batch** (*offline*) approach. The changes of weights values are applied after that all the observations in the training dataset have been evaluated and a global error has been calculated.
- **incremental** (*online*) approach. The error is calculated for each observation in the training dataset and the weights are modified consequently. This is generally preferred since it allows to examine a wider number of possible solutions.

In some applications the error function E_w is particularly complex. The main effect is that the finding of the optimal solution can be difficult (a local minimum is found) or the algorithm can be very slow. Some changes to the original algorithm have been proposed, including:

- the adoption of a modified version of generalized delta rules:

$$\Delta w_{ij}(t+1) = -\eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(t) \tag{2.47}$$

  where the modifications to weights at time (iteration) $t+1$ are not only based on the gradient, but also on the latest modification of weights, multiplied by a coefficient α (the *momentum*), which tends to prevent possible oscillations of the algorithm around the solution;

- the introduction of adaptive parameters (η and α), which vary their values during the learning process to accelerate the convergence to the solution or to improve the search of the solution;

- specific controls in the choice of initial weights;

- addition of 'noise' during the learning process to avoid the finding of local minima.

Adaptive Resonance Theory

Adaptive Resonance Theory (ART) Neural Networks have one layer of processing units, which are fully connected with input buffer with types of weights (Figure 2.3). This structure is interesting since:

1. ART implements an analogic classification that is considered more "natural" compared with cluster analysis in classifying patterns based on the numerical computational results;

2. ART is able to overcome the so-called stability-plasticity dilemma. When an arbitrary pattern comes into the network, the previously stored memory is not damaged; instead, a new class is automatically set up for it.

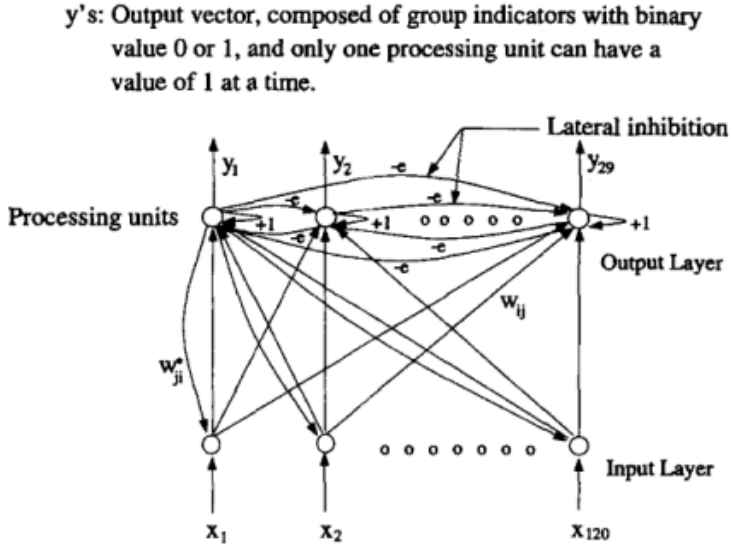
3. The ART processes on-line training so that no target patterns are needed.

Analysing in details Figure 2.3:

- n = number of inputs of the network;

- x_i = i-th component of the input vector (0 or 1);

Figure 2.3: *Schematic of ART1 Network. Source: Faghri and Hua 1995*



- y_j = j-th output;
- w_{ij} = the weight for the connection from the j-th output to the i-th input;
- w^*_{ji} = the weight for the connection from the i-th input to the j-th output
- ρ = the so-called 'vigilance parameter', a constant having values between 0 and 1;
- k = index that denotes the winning output element (largest value among output)

The relationship between the 2 kinds of weight vectors is given by the equation:

$$w^*_{ji} = \frac{w_{ij}}{1 + \sum_{k=1}^{n} w_{kj}} \tag{2.48}$$

Initially all weights w_{ij} are set to 1 and w^*_{ji} are equal to $\frac{1}{(1+n)}$.

The ART1 develops using these steps:

1. Compute the outputs using the formula

$$x_j = \sum_{k=1}^{n} w_{kj} \tag{2.49}$$

2. Determine the largest output with a 'winner takes all' strategy, and let the winner be X_k;

3. Rate the input pattern match with the following formula:

$$r = \frac{\sum_{i=1}^{n} w_{ik} x_i}{\sum_{i=1}^{n} x_x} \tag{2.50}$$

4. If $r < \rho$, set $X = 0$ and go to step 2.

5. If $r > \rho$, for all i, if $x_i = 0$ and $w_{ik} = 0$, then, recompute w^*_{ji} for all i, if any weights have been changed.

ART1 stored vectors, and checked the committed processing units in order according to how well the vectors $[w^*_{j1}, \ldots, w^*_{jn}]$ being stored match the input pattern. If none of the committed processing units match well enough, an uncommitted unit will be chosen. The network sets up certain categories for the input vector, and classifies the input pattern into a proper category. If the input pattern does not match any of those categories, the network creates a new category for it.

2.3 Clustering

Clustering analysis techniques are the most applied DM descriptive techniques. The main objective of this kind of analysis is subdividing observations (data) in groups not already defined (clusters), being maximized the similarity among observations belonging to the same group (internal cohesion) and minimized the similarities with observations in other groups (external separation).

Formally the problem can be defined as:

Being $D = \{t_1, t_2, \ldots, t_n\}$ a dataset of elements (observations), a certain measure of similarity $sim(t_i, t_l)$ defined between two observations $t_i, t_l \in D$ and an integer value K, the *clustering problem* is defining a map $f : D \rightarrow \{1, \ldots, K\}$ where every t_i is assigned to a cluster K_j , $1 \leq j \leq K$. Given a certain cluster K_j, $\forall t_{jl}, t_{jm} \in K_j$ and $t_i \notin K_j$, $sim(t_{jl}, t_{jm}) > sim(t_{jl}, t_i)$.

From the definition of the clustering problem, some relevant points can be highlighted:

- Clustering analysis follows an unsupervised approach, that is a reference condition or situation does not exist, or is not available to the analyst;
- Differently from classes in classification problem, the meaning of each cluster is not known *a priori*, therefore the analyst must interpret the results and define them;
- It is not known *a priori* the optimal number k of groups to identify, therefore there is not a unique solution;
- A satisfactory solution is highly influenced by the inclusion (exclusion) of not significant (significant) variables; Exploratory Data Analysis is a relevant preliminary step to consider in order to avoid these situations;
- The presence of outlier data could negatively affect the identification of significant clusters; however in some cases the identification of outliers is the main objective of cluster analysis and outliers represent relevant data.

Clustering methods can be subdivided in different ways, but traditionally they have been divided in two broad categories, *hierarchical* and *non hierarchical* methods.

- *Hierarchical methods* produce a hierarchy of partitions. At one side each object is assigned to a cluster (n clusters, having n observations) and at the other side all the observations are included in the same group (1 cluster). Hierarchical method are called *agglomerative* if at each step the two most similar clusters are merged in a new cluster, or *divisive* if at each step one cluster is splitted in new clusters;
- *Partitioning methods* create only one partition of data, placing each observation in one of the k clusters.

More recently other types of clustering have been developed and are nowadays applied in different contexts, in particular:

- *Model-based clustering* (Fraley and Raftery 2002), which 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;

Furthermore new algorithms have been developed to deal with specific needs (Dunham 2003), including the analysis of non-quantitative data (qualitative data, images, ..) or the real-time treatment of large amount of data (streaming data).

Clustering models can also be classified in other ways, considering other characteristics they present:

- Connectivity models: models based on distance connectivity (hierarchical clustering);
- Distribution models: models which use statistic distributions, such as multivariate normal distributions, to model clusters (model-based clustering);
- Centroid models: models which represent each cluster by a single mean vector (many partitioning models)
- Density models: models which defines clusters as connected dense regions in the data space (models applied in large databases such as DBSCAN and OPTICS)

Another important distinction that is useful to introduce here is between hard clustering and soft clustering.

- In hard clustering, data is divided into distinct clusters, where each data element belongs to exactly one cluster.
- In soft clustering (also referred to as fuzzy clustering), data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster.

In the following sections some details will be given about the most important clustering methods for quantitative data commonly applied in practice: hierarchical, partitioning and model-based methods. Considering the increasing attention given to large database, basic indications concerning algorithms for this type of datasets are also presented. For all these clustering methods hard clustering situations will be considered.

Finally Fuzzy C-means Clustering will be introduced, as a representative of soft clustering approach.

2.3.1 Distances Between Clusters

The most common measures of similarity and distance have been presented introducing classification techniques (2.2.1). The same measures are applied in the clustering analysis, but some details are given in this section since the distance between clusters can be evaluated in different ways, given a pre-fixed measure of distance.

Given a cluster K_m with n observations $t_{m1}, t_{m2}, \ldots, t_{mn}$ one can define:

$$Centroid = C_m = \frac{\sum_{i=1}^{n} t_{mi}}{n} \tag{2.51}$$

$$Radius = R_m = \sqrt{\frac{\sum_{i=1}^{n} (t_{mi} - C_m)^2}{n}} \tag{2.52}$$

$$Diameter = D_m = \sqrt{\frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (t_{mi} - t_{mj})^2}{(n)(n-1)}} \tag{2.53}$$

The *centroid* C_m could be used as a representative point of a cluster, even if it could not belong to it; alternatively one can use the *medoid* M_m which is an observation of the cluster positioned in proximity of its center.

Being K_i and K_j two clusters and $dis(t_{il}, t_{jm})$ the distance between two observations $t_{il} \in K_i$ and $t_{jm} \in K_j$, the distance between K_i and K_j can be calculated in different ways:

- *Single link method*

$$dis(K_i, K_j) = \min[dis(t_{il}, t_{jm})] \; \forall t_{il} \in K_i \notin K_j \text{ e } \forall t_{jm} \in K_j \notin K_i \tag{2.54}$$

- *Complete link method*

$$dis(K_i, K_j) = \max[dis(t_{il}, t_{jm})] \; \forall t_{il} \in K_i \notin K_j \text{ e } \forall t_{jm} \in K_j \notin K_i \tag{2.55}$$

- *Average link method*

$$dis(K_i, K_j) = \text{mean}[dis(t_{il}, t_{jm})] \; \forall t_{il} \in K_i \notin K_j \text{ e } \forall t_{jm} \in K_j \notin K_i \tag{2.56}$$

- *Centroid method*

$$dis(K_i, K_j) = dis(C_i, C_j) \tag{2.57}$$

  where C_i and C_j are the centroids of K_i and K_j, respectively.

- *Medoid method*

$$dis(K_i, K_j) = dis(M_i, M_j) \tag{2.58}$$

  where M_i and M_j are the medoids of K_i and K_j, respectively.

- *Ward method*

  Ward's minimum variance method (Ward 1963) minimizes an objective function to produce clusters with the maximum internal cohesion and the maximum external separation.

  The *Total Deviance* (T) of the p variables is divided in two parts: "*Within Groups Deviance*"(W) and "*Between Groups Deviance*"(B):

  $$T = W + B \tag{2.59}$$

  In practice, having n observations grouped in k clusters, the deviance terms of equation 2.59 are calculated by:

  – *Total Deviance* (T), being x_{is} is the value of variable s taken by observation i and $\bar{x}_s$ is the overall mean for variable s.

  $$T = \sum_{s=1}^{p} \sum_{i=1}^{n} (x_{is} - \bar{x}_s)^2 \tag{2.60}$$

  – *Within Groups Deviance* (W)

  $$W = \sum_{k=i}^{g} W_k \tag{2.61}$$

  where W_k is the deviance of the p variables of k-th cluster, which has n_k observations and centroid $\bar{x}_k = [\bar{x}_{1k}, \ldots, \bar{x}_{pk}]$, defined as:

  $$W_k = \sum_{s=i}^{p} \sum_{i=1}^{n_k} (x_{is} - \bar{x}_{sk})^2 \tag{2.62}$$

  – *Between Groups Deviance* (B)

  $$B = \sum_{s=i}^{p} \sum_{k=1}^{g} n_k (\bar{x}_{sk} - \bar{x}_s)^2 \tag{2.63}$$

At each step the Ward's method aggregates the groups which produce the minimum increments of the Within Groups Deviance W and consequently the higher increment of Between Groups Deviance B.

2.3.2 Hierarchical Methods

Clustering hierarchical methods produce a series of partitions of data, associated to successive levels of grouping, following on ordering that can be graphically represented by a tree structure.

At the root of the tree all the observations belong to the same cluster, while at the top level (leaves of the tree), each observation belongs to a separate cluster. Between these two extreme situations, $n - 2$ levels exist and for each of them only one partition of the n observations in k groups.

As already written, hierarchical methods are called *agglomerative* if the clusters are merged going from the top to the bottom of the tree, or *divisive* if the clusters are splitted in a bottom-up way. Since in practical applications divisive methods are scarcely applied in this review only agglomerative methods are described.

Agglomerative Methods

Given a dataset D with n observations $\{t_1, t_2, \ldots, t_n\}$, $A = \{n \times n\}$ is the relative Distance Matrix, which has in each cell the distances between couples of observations $A[i, j] = dis(t_i, t_j)$ and $DE = \{d, k, K\}$ is the tree structure, where d is the threshold distance, k is the number of cluster created and K is the set of clusters.

The algorithm, common to different agglomerative hierarchical methods, has some basic steps:

1. Creation of the tree structure $DE = \{d, k, K\}$ with one cluster for each observation;

2. Determination, at each step, of the couple of clusters separated by the minimum distance d. The clusters identified are merged in a new cluster of the upper level;

3. The Distance Matrix A and the threshold distance d are updated to consider the new cluster added to the tree structure;

4. The process terminated when all the observations belong to the same cluster.

Differences among agglomerative methods regard the choice of the method used to compute the distances between clusters 2.3.1, which can differentiate the final segmentation of the dataset.

However, the main difficulty associated to hierarchical is the choice of the most correct number of clusters k, which guarantees the maximum internal cohesion and the minimum external separation. To solve this problem, some performance indices have been defined and can help the analyst to determine the optimal number of clusters. Usually the combined adoption of more indices is the most reasonable choice.

The most important criteria commonly adopted are:

- Pseudo F Statistic (Calinski and Harabasz 1974; Milligan and Cooper 1985). This statistic analyses the hierarchy at each level and a peak

value of PSF reveals the optimal number of clusters. It is calculated by:

$$PSF = \frac{\frac{T-P_G}{G-1}}{\frac{P_G}{n-G}} \tag{2.64}$$

where:

- $T = \sum_{i=1}^{n} \|x_i - \bar{x}\|^2$;
- $P_G = \sum W_j$ calculated for the G groups at the G-th level of the hierarchy;
- G = the number of clusters at a given level of the hierarchy;
- n = the number of observations;
- x_i = i-th observation;
- $\bar{x}$ = sample mean vector;
- $W_k = \sum_{i \in C_k}^{n} \|x_i - \bar{x_k}\|^2$;
- $\bar{x_k}$ = mean vector for group C_k.

- 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:

$$DB = \frac{1}{n} \sum_{i=1,1 \neq j}^{n} \max \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \tag{2.65}$$

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.

- Dunn Index (Dunn 1974). This index aims to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart, as compared to the within cluster variance. The Dunn index is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance, therefore for a given assignment of clusters, a higher Dunn index indicates better clustering. It can be calculated by:

$$DI_m = \min_{1 \leqslant i \leqslant m} \left\{ \min_{1 \leqslant j \leqslant m, j \neq i} \left\{ \frac{\delta(C_i, C_j)}{\max_{1 \leqslant k \leqslant m} \Delta_k} \right\} \right\} \forall i, j, k \tag{2.66}$$

where:

- $\delta(C_i, C_j)$ is the inter-cluster distance between clusters C_i and C_j;
- Δ_k measures the intra-cluster distance of cluster k.

- Silhouette measure (Rousseeuw 1987). The silhouette width $S(i)$ takes values in the interval $[-1; +1]$. Observations with large values of $S(i)$ are very well clustered, small $S(i)$ (around 0) means that the observation lies between two clusters, and observations with a negative $S(i)$ are probably placed in the wrong cluster. The overall average silhouette width for the entire plot is simply the average of the $S(i)$ for all items in the whole dataset.

  For each observation i, the silhouette width $S(i)$ is defined as:

$$s(i) = \frac{(b(i) - a(i))}{\max(a(i), b(i))} \tag{2.67}$$

  where:

  - $a(i)$ = average dissimilarity between i and all other points of the cluster to which i belongs. If i is the only element in its cluster, $S(i) = 0$;
  - $b(i) = \min_C d(i, C)$, being $d(i, C)$ the average dissimilarity of i to all observations of other clusters C. $b(i)$ represents the dissimilarity between i and its "neighbor" cluster, that is the nearest one to which it does not belong to.

- Goodman and Kruskal's index G2 (Goodman and Kruskal 1954). This index considers all possible quadruples (q, r, s, t) of input parameters and determines if this quadruple is concordant or discordant, based on the relative distances d(x,y) between the samples x and y.

  A quadruple is called concordant if one of the following two conditions is true:

  - $d(q, r) < d(s, t)$, being q and r in the same cluster, and s and t in different clusters;
  - $d(q, r) > d(s, t)$, being q and r in different clusters, and s and t in the same cluster.

  Conversely, a quadruple is called discordant if one of following two conditions is true:

  - $d(q, r) < d(s, t)$, q and r are in different clusters, and s and t are in the same cluster;
  - $d(q, r) > d(s, t)$, q and r are in the same cluster, and s and t are in different clusters.

A good clustering is one with many concordant and few discordant quadruples. Let S_c and S_d denote the number of concordant and discordant quadruples, respectively, then the Goodman-Kruskal index is defined as:

$$GK = \frac{S_c - S_d}{S_c + S_d} \tag{2.68}$$

Large values of GK indicate a good clustering.

- Decomposition of Deviance Indices.

  The implementation of the Ward's method is based on the decomposition of the Total Deviance of the p variables($T = W + B$) such that a good clustering of data is characterised by a small Deviance Within Groups W and a large Deviance Between Groups B.

  Synthetic indices can be calculated from the values of T, W and B.

  - R^2 index. This index takes values in the interval $[0, 1]$: tte closer to 1 is R^2, the better is the quality of the clusters identified ($W_k \approx 0$ e $B \approx T$). The main shortcoming of this index is that the best value of R^2 is obtained when each observation has its own cluster. In fact at the same time the maximum homogeneity and the maximum separation are obtained, but this result is not useful for applications.

$$R^2 = 1 - \frac{W}{T} = \frac{B}{T} \tag{2.69}$$

  - *Root-Mean-Squared Standard Deviation* (*RMSSTD*). This index derives from R^2 index and considers only the portion of deviance within groups that is added at each step. At the h-th step ($h = 2, \ldots, n-2$) the index is defined by:

$$RMSSTD = \sqrt{\frac{W_h}{p(n_h - 1)}} \tag{2.70}$$

    where W_h is the deviance within the cluster created at the h-th, n_h is the number of observations in the cluster and p are the variables considered. If two very dissimilar clusters are merged ,*RMSSTD* has a strong increment, which suggests to stop the procedure at the previous step.

  - *Semi Partial* R^2 (*RMSSTD*). Also this index evaluates the local contribution given by the h-th step of the agglomerative algorithm. It is defined by:

$$SPRSQ = \frac{W_h - W_r - W_s}{T} \tag{2.71}$$

where h is the group obtained at the h-th step from the merging of clusters r and s, T is the total deviance of observations, W_h, W_r and W_s are the deviance within groups h, r and s, respectively.

SPRSQ measures the increment of the deviance within groups obtained from the merging of clusters r and s; as observed for *RMSSTD* a relevant increment between two successive steps can be associated to the merging of two dissimilar clusters, and the procedure can be stopped at the previous step.

2.3.3 Partitioning Methods

Differently form hierarchical methods, partitioning methods create only one partition of data, placing each of the n observations in one of the k clusters, where k is chosen by the analyst.

Partitioning methods produce this partition satisfying criteria of optimality, generally expressed by the maximization of an objective function. This approach is generally more efficient and more robust than that one of hierarchical methods. In fact, many methods do not need t store the distance matrix, as happens for hierarchical methods, with relevant computational advantages. Therefore they can be better implemented in large databases.

However the large number of possible solutions lead to constrained results, which often correspond to local maxima. In addition to this, the main difficulty of these methods is related to the choice of the value k by the analyst. Since this is often difficult, the algorithms are applied varying the value of k and evaluating the results on the basis of the indices introduced for agglomerative clustering methods.

Algorithms based on squared error

Algorithms based on squared error determine the partition of data minimizing the squared error.

Given a cluster K_i, with observations $t_{i1}, t_{i2}, \ldots, t_{im}$ and centroid C_k, the squared error is defined by:

$$se_{K_i} = \sum_{j=1}^{m} \|t_{ij} - C_{Ki}\|^2 \tag{2.72}$$

Considering a set of clusters $K = \{K_1, K_2, \ldots, K_k\}$, the squared error for is defined by:

$$se_K = \sum_{j=1}^{k} se_{K_j} \tag{2.73}$$

At the beginning of the procedure each observation is randomly assigned to a cluster. Then, at each step, each observation t_i is assigned to the cluster whose centroid is the closest to the observation. The centroids of the new clusters are re-calculated and the squared error is calculated considering the new partition of data. The procedure is stopped when the decrement of successive squared error is lower than a pre-fixed threshold.

K-means algorithm

K-means algorithm is probably the most famous partitioning method.

The algorithm chooses initial seeds as initial values for the *K*-means, which are representative of the centroids of the clusters in the *p*-dimensional space. Seeds must be sufficiently dispersed in the variable space to guarantee an adequate convergence of the algorithm. Specific sub-algorithms, which impose a minimum distance among seeds, have been developed to accomplish this task.

Once the initial seeds have been selected, the iterative structure of the algorithm begins:

- Assignment of the observation to the closest mean;
- Calculation of the mean for each cluster.

The algorithm ends when the maximum number of iterations is reached or when a certain convergence criterion (such as a minimum value of the squared error) is satisfied.

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.

Different modifications of K-means have been proposed in recent years. Just citing some of the most recent development:

- *X-means* (Pelleg and Moore 2000) which allows the identification of the optimal number of clusters using the Bayesian Information Criteria (BIC);
- *K-modes* (Chaturvedi, Green, and Caroll 2001), which is a nonparametric approach to derive clusters from categorical data, following an approach similar to K-means;

- *K-means++* (Arthur and Vassilvitskii 2007), which was designed to choose the seeds for the k-means trying to avoid the sometimes poor clusterings found by the standard k-means algorithm;

PAM Algorithm

Partitioning Around Medoids(*PAM*, also called *K-medoids*) algorithm is a clustering method which adopts medoids instead of centroids, obtaining a relevant advantage in the treatment of missing data.

Initially k observations belonging to D are randomly chosen as medoids and the others are associated to the cluster with the closest medoid (Build Step). At each iteration the non-medoid observations are analysed, testing if they can become new medoids, improving the quality of the partition, that is minimizing the sum of the dissimilarities of the observations to their closest medoid (Swap Step).

Considering a cluster K_i represented by medoid k_i, the algorithm evaluates if any other observation t_h of the cluster can be changed with t_i, becoming the new medoid. C_{ijh} is the changing of cost for the observation t_j associated to the change of medoid from t_i to t_h. Repeating this process for all the observations of cluster K_i, the total changing of cost is equal to the change of the sum of distances of observations to their medoids.

As a consequence of the medoid change four different conditions could happen:

1. $t_j \in K_i$ but $\exists$ another medoid t_m such that $dis(t_j, t_m) \leq dis(t_j, t_h)$
2. $t_j \in K_i$ but $dis(t_j, t_h) \leq dis(t_j, t_m) \forall$ other medoid t_m
3. $t_j \in K_m, \notin K_i$ and $dis(t_j, t_m) \leq dis(t_j, t_h)$
4. $t_j \in K_m, \notin K_i$ but $dis(t_j, t_h) \leq dis(t_j, t_m)$

Therefore the total cost associated to the medoid change becomes:

$$TC_{ih} = \sum_{j=1}^{k} C_{jih} \tag{2.74}$$

Compared to K-means the main improvement provided by PAM is a main robust structure; however its use is not suggested for large datasets, because is highly penalized by its complexity.

Self-Organizing Neural Networks

Artificial Neural Networks can be used to solve clustering problems adopting unsupervised learning process. In this case ANNs are called *Self-Organizing Neural Networks*, since the only parameters are the weights of

the neural network, which self-organizes to detect significant groupings in data. Unsupervised learning can be competitive and non-competitive.

In *non competitive* learning the weight of connection between two nodes of the network is proportional to the values of both nodes. Hebb rule is used to update the values of weights. Given the j-th neuron of the neural network connected to x_{ij} input neurons with weights w_{ij}, Hebb rule is defined by:

$$\Delta w_{ij} = c x_{ij} y_j \tag{2.75}$$

where y_j is the output of the j-th neuron and c is the learning rate.

In *competitive* learning neurons compete one with each other and the winner can update its weights. Usually the network has two layers (input and output); in the input layer there are p neurons, representative of the p explanatory variables which describe the observations, connected with the neurons of output layer.

When an observation is fed to the neural network each node in the output gives an output value, based on the values of the connection weights. The neuron whose weights are the most similar to the input values is the winner. Following the *"Winner Takes All"* rule, the output is set to 1 for the winner and 0 for the other neurons, and weights are updated.

At the end of the learning process some relations are detected between observations and output nodes. These relations mean that some clusters have been identified in the dataset: the values of the weights of nodes grouped in a cluster are the mean values of the observations included in this cluster.

The most famous neural network which adopt competitive learning are *Self-Organizing Maps* (*SOM*), or Kohonen Networks.

Self-Organizing Maps

Self-Organizing Maps (SOM) are Artificial Neural Networks based on unsupervised competitive learning. They are also known as *Self-Organizing Feature Maps* (*SOFM*) or Kohonen Networks form the name of the mathematician who first proposed them (Kohonen 1982).

Kohonen Networks map each p-dimensional observation into a 1 or 2-dimensional space. In the latter case the output space is represented by a grid of output neurons (*competitive layer*), which guarantees the spatial correlation of clusters in the output space. This contiguity of similar clusters is due to the fact that the update of neurons is done for the winner neuron and a group of neuron in its neighbourhood. Using this approach at the end of the learning spatial partitions of neurons are obtained, which graphically represent the presence of clusters.

The algorithm is composed by some steps:

1. The weights w_{ij} between the i-th input neuron and the j-th output neuron are defined at iteration t as $w_{ij}(t)$, $0 \leq i \leq n-1$, where n is the number of input. Initial values of weights are randomly chosen in the interval $[0, 1]$ and the values $N_j(0)$ s set, where $N_j()$ is the number of neurons in the neighbourhood of the j-th neuron at the iteration $t = 0$.

2. Observation $\mathbf{X} = x_0(t), x_1(t), \ldots, x_{n-1}(t)$ is fed to the neural network, where $x_i(t)$ is the $'i$-th input.

3. Distances d_j between input neuron and each output neuron j are calculated. If Euclidean distance is chosen:

$$d_j^2 = \sum_{i=1}^{n} (x_i(t) - w_{ij}(t))^2 \tag{2.76}$$

4. Neuron which has the minimum distance value is selected and called j^*.

5. Weights of node j^* and nodes included in the neighbourhood defined by $N_{j^*}(t)$ are updated. The new weights are calculated by:

$$w_{ij}(t+i) = w_{ij}(t) + \eta(t)(x_i(t) - w_{ij}(t)) \quad \text{for } j = j^* \text{ e } 0 \leq i \leq n-1 \tag{2.77}$$

   where $\eta(t)$, $0 \leq \eta(t) \leq n-1$ is the learning rate, which decreases with t. In this manner the adaptation of weights is progressively slowed down. In a similar manner dimensions of $N_{j^*}(t)$ decreases, stabilizing the learning process.

6. Algorithm goes bask to step 1.

The learning rate η is initially set to values greater than 0.5 and decreases during the learning process, which usually needs 100 to 1000 iterations. Kohonen suggested the adoption of a linear decrease as a function of the number of iteration. SOM are effective and useful clustering techniques, in particular in those cases when it is important to maintain the spatial order of input and output vectors.

2.3.4 Model-based Clustering

Model-based clustering has a completely different approach compared to non-parametric methods and has a particular attractiveness for its capability of determining the optimal number of groups.

Model-based clustering 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 2002). The probability density function for the k-th group can be written as:

$$f_x(\mathbf{x}_i|\mu_k, \mathbf{\Sigma_k}) = \frac{\exp -\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_i - \mu_k)}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} \tag{2.78}$$

where:

- μ_k is the p dimensional mean vector;
- Σ_k is the $p \times p$ covariance matrix;
- p is the number of attributes used for the clustering.

Therefore each cluster forms an ellipsoid centered at its means μ_k with geometric characteristics determined by the covariance matrix Σ_k, which can be decomposed as:

$$\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^T \tag{2.79}$$

where:

- λ_k is the first eigenvalue of Σ_k, which specifies the volume of the k-th cluster;
- $\mathbf{D}_k$ is the orthogonal matrix of eigenvectors, which determines the orientation of the k-th cluster;
- $\mathbf{A}_k = \left[\alpha_{1k}, \ldots, \alpha_{pk}\right]$, which determines the shape of the k-th in the p dimensional space.

$\mathbf{D}_k$, λ_k and $\mathbf{A}_k$ can be considered as separate and independent parameters and different models can be built and analyzed (Banfield and Raftery 1993). Identifiers 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 parallel the coordinate axes.

The identification of clusters can be done following two alternative approaches: the classification likelihood approach and the mixture likelihood approach.

In the *classification likelihood approach*, the following likelihood function is maximized, with the objective of identifying parameters θ and class labels $\gamma = (\gamma_1, \ldots, \gamma_n)^T$ adopted for the classification:

$$L_C(\theta_1, \ldots, \theta_G; \gamma_1, \ldots, \gamma_n|\mathbf{x} = \prod_{i=1}^{n} f_{\gamma i}(\mathbf{x}_i|\theta_{\gamma i}) \tag{2.80}$$

In practice exact maximization is impractical for combinatorial aspects introduced by the presence of labels (Fraley and Raftery 2002), therefore models-based hierarchical methods are commonly implemented as alternative choice. Adopting a hierarchical approach pairs of clusters which produce the largest increase in maximum likelihood are merged; the sub-optimal solution found is generally a good approximation of the optimal grouping obtained with the exact maximization.

In the *mixture likelihood approach*, the objective is identifying parameters θ and τ which maximize the likelihood function:

$$L_M(\theta_1, \ldots, \theta_G; \tau_1, \ldots, \tau_G | \mathbf{x} = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(\mathbf{x}_i | \theta_k) \tag{2.81}$$

where τ_k is the probability that an element belongs to k-th cluster, which meets the following constraints:

$$\tau_k \geq 0 \tag{2.82}$$

$$\sum_{k=1}^{G} \tau_k = 1 \tag{2.83}$$

To obtain the maximum-likelihood estimation an equivalent log-likelihood function is derived and the Expectation-Maximization (EM) algorithm is adopted (Fraley and Raftery 1998). Bayesian Information Criteria (BIC) is finally applied to find the maximum mixture likelihood and consequently the optimal number of clusters:

$$BIC = 2L - r \log n \tag{2.84}$$

where:

- L is the log-likelihood of the model;
- r is the total number of parameters to be estimated in the model;
- n is the number of elements.

BIC can represent a valid measure of quality since a term is added to the log-likelihood to penalize the complexity of the model. Otherwise the fit of the model naturally increases adding more term to the model.

The main difference between classification and mixture likelihood approaches is that in the former each element is assigned to a unique cluster, while in the latter each object is assigned with a certain probability to each cluster.

2.3.5 Methods for Large Databases

The clustering methods already introduced represent a selection of the most applied traditional techniques, which are limited when applied to large datasets, in particular dynamic datasets. This particular type of data, which is becoming quite common, requires that:

1. data must be read only one time;
2. the algorithm is *online*, i.e. the 'best' solution of the algorithm is found when it is executed;
3. the algorithm can be easily suspended, stopped and restarted;
4. results must be updated incrementally, when data are added or subtracted to the database;
5. memory resources are limited;
6. algorithm can make a scan of the database;
7. each observation is processed only one time.

BIRCH Algorithm

BIRCH algorithm (*Balanced Iterative Reducing and Clustering Using Hierarchies*) (Zhang, Ramakrishnan, and Livny 1996) is an incremental hierarchical algorithm, which can be succesfully used in large database since it needs limited memory resources and read the data only one time.

Its structure is based on the concepts of clustering feature *CF* and *CF* Tree:

- A *clustering feature* (*CF*) is the triplet $(N, \vec{LS}, SS)$, where N is the number of elements in a cluster, $\vec{LS}$ is the sum of elements in a cluster and SS is the sum of the square of the elements in a cluster;
- A *CF Tree* is a balanced tree with branching factor B, which is the maximum number of children that can be generated by a node. Each node contains a triplet *CF* for each of its children. If the node is a leaf, it is representative of a cluster and has a *CF* for each sub-cluster, which cannot have a diameter larger than the threshold T.

Therefore *CF* Tree is a tree which is built adding observations and respecting the maximum diameter T allowed for each leaf, the maximum number of children B that can be generated by a node and memory limits. The diameter is calculated as the mean of the distances calculated between all the couples of observations which belong to the cluster. A larger value

of T produces a smaller tree, which is a good clustering solution in case of limited memory resources.

In the meanwhile clustering features, associated to each node of the tree, summarise the characteristics of the clusters, speeding up the update of the tree and reducing the access to the data to only one time.

The building of the CF Tree is a dynamic and incremental process. Defined by parameters B and T the limits of the tree, each observation is considered and the distance between the centroid of the clusters is determined using the information available from the clustering feature. If the parameters B and T are respected, the new observation is added to the closest cluster, the clustering feature of the cluster is updated and the same is done for the clustering features from the cluster to the root of the tree.

If the conditions are violated the new observations is added to the node as a new cluster and the interested clustering features are calculated and updated.

BIRCH algorithm is very efficient if the threshold value T has been correctly identified, otherwise the tree must be rebuilt. Furthermore BIRCH is adapt in case of spherical clusters, since it is strongly related to the maximum diameter T for the definition of the boundaries of clusters.

DBSCAN Algorithm

DBSCAN algorithm (Density-Based Spatial Clustering of Applications with Noise) (Ester et al. 1996) is a partitioning algorithm based on the measure of density, which is particularly interesting for the possibility of identifying clusters of arbitrary shape.

The algorithm is guided by parameters *MinPts*, which defines the minimum number of elements in a cluster and *Eps*, which is the maximum distance between two distinct elements in the same cluster. Some preliminary definitions must be given to have a correct comprehension of the algorithm:

- The $Eps - neigborhood$ of the element p is the set of elements that are within a circle of radius *Eps* centered in p;
- If the $Eps-neigborhood$ of p has a minimum number of elements *MinPts*, then p is a *core point*;
- Given the values *MinPts* and *Eps*, the element p is "directly density-reachable" by q if:
  1. $dis(p,q) \leq Eps$
  2. $\mid r \mid dis(r,q) \leq Eps \mid \geq MinPts$

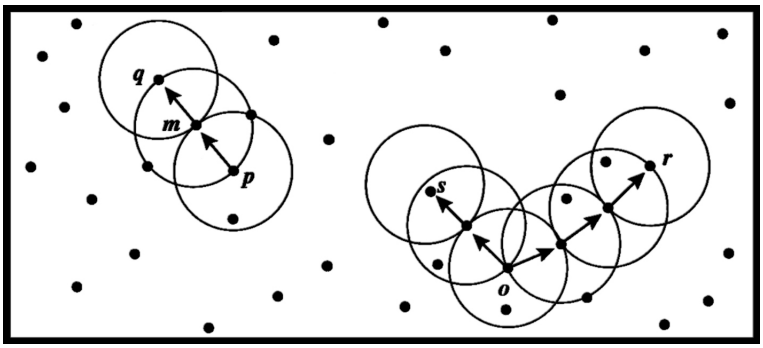
  i.e. if p is within the $Eps - neigborhood$ of q and q is a *core point*.

- Given the values *MinPts* and *Eps*, the element p is "**density-reachable**" by q if exists a chain of elements $p_1, \ldots, p_n$, for which $p_1 = q$ and $p_n = p$, such that p_{i+1} is directly density-reachable by p_i, with $1 < i < n$.
- Given the values *MinPts* and *Eps*, the element p is "**density-connected**" to q if exists an element o, such that p and q are density-reachable by o.

In Figure 2.4 are reported some examples of these concepts, where the circles have radii equal to *Eps* and *MinPts* = 3:

- the elements r and s are "density-connected" to each other by o;
- the element q is "density-reachable" by p.

Figure 2.4: *DBSCAN Distances. Source: Ester et al. 1996*



Using these concepts the density of a cluster is the criterion which determines the belonging of an element to a cluster: each cluster has a central set of elements directly density-reachable, very close one to each other (the core points), rounded by a series of other elements in the border of the cluster, sufficiently closed to the central points (border points). Finally elements not belonging to any cluster are defined as 'noise' and considered as outliers.

OPTICS Algorithm

Algorithm DBSCAN is a powerful method to detect cluster with arbitrary shape. However the quality of results is influenced by the choice of the correct values of parameters *MinPts* and *Eps*. This is a common issue for many clustering methods, but this fact is more relevant in case of multi-dimensional data, especially if data distributions are distorted with respect of some dimensions.

OPTICS algorithm (Ordering Points to Identify the Clustering Structure) (Ankerst et al. 1999) has been developed to solve this problem, giving as

a result an ordering of clusters that can be automatically analysed. This ordering is equivalent to the clustering obtained from DBSCAN algorithm through a wide range of parameters' values.

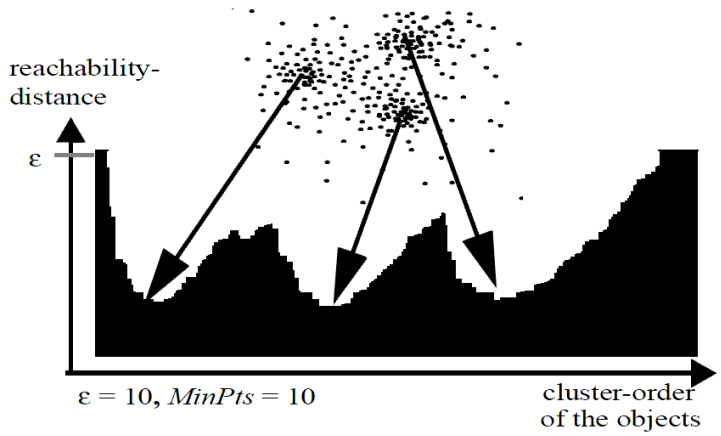
In fact for a constant value of *MinPts*, the clusters with a higher density (with smaller values of *Eps*) are completely included in density-connected sets with lower density. Executing DBSCAN algorithm with a progressive variation of the parameter *Eps*, it is possible to obtain an ordering of clusters starting from the ones with higher density (Figure 2.5).

Therefore two values are calculated or each of the elements analysed: the *core distance* and the *reachability distance*:

- The *core distance* of an element p is the smallest *Eps*' which makes p a *core object*. If p is not a core items, this distance is indefinite.
- The *reachability distance* of an element q with respect to another element p is the maximum value between the *core distance* of p and the Euclidean distance between p and q. If p is not a core items, this distance is indefinite.

Analysing this couple of values, associated to each element of the dataset, is possible to establish alternative clustering solutions, evaluating the influence of the choice of the values of distance *Eps*'.

Figure 2.5: *OPTICS. Illustration of the Cluster Ordering. Source: Ankerst et al. 1999*



2.3.6 Fuzzy Clustering

Differently from hard clustering, fuzzy clustering allows data elements to belong to more than one cluster, assigning elements using membership

levels to each cluster. One of the most widely used fuzzy clustering algorithm is the Fuzzy C-Means (FCM) Algorithm, developed by Dunn in 1973 (Dunn 1973) and improved by Bezdek [(Bezdek 1981),(Bezdek, Ehrlich, and Full 1984)].

The FCM algorithm attempts to partition a finite set of N elements $X = \{x_1, \ldots, x_N\}$ into C fuzzy clusters minimizing an objective function, as done by the k-means algorithm. The standard objective function is:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \|x_i - c_j\|^2 \quad 1 \leq m \leq \infty \tag{2.85}$$

where:

- m is the fuzzifier, which can assume any real number greater than 1;
- u_{ij} is the degree of membership of element x_i to the j-th cluster;
- x_i is the i-th element of the dataset, measured on d variables;
- c_j is the d-dimension center of the j-th cluster;
- $\|*\|$ is any norm expressing the similarity between any measured data and the center.

The fuzzifier m determines the level of cluster fuzziness: large m result in smaller memberships u_{ij} and hence, fuzzier clusters. When $m = 1$ the memberships u_{ij} converge to 0 or 1, which represents a crisp partitioning. Usually, in the absence of experimentation or domain knowledge, m is commonly set to 2.

Fuzzy partitioning is carried out through an iterative optimization of the objective function, with the update of membership u_{ij} and the cluster centers c_j calculated by:

$$u_{ij} = \frac{1}{\sum_k^C \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{2/(m-1)}} \tag{2.86}$$

$$c_j = \frac{\sum_{j=1}^{C} u_{ij}^m \cdot x_i}{\sum_{j=1}^{C} u_{ij}^m} \tag{2.87}$$

The structure of FCM algorithm is very similar to the k-means algorithm:

- Choose the number c of clusters to be determined;
- Assign randomly to each point coefficients for being in the clusters;
- Repeat the following steps until the algorithm has converged:

– Compute the centroid for each cluster, using formula 2.87.
– For each point, compute the membership to the clusters, using formula 2.86.

This convergence of the algorithm is controlled by

$$\max_{ij} \left\{ |u_{ij}^{(k+1)} - u_{ij}^{(k)}| \right\} < \epsilon \tag{2.88}$$

where ϵ is a termination criterion between 0 and 1 and k are the iteration steps.

The final output of the algorithm are a list of c cluster centres $C = \{c_1, \ldots, c_c\}$ and a partition matrix $U = u_{i,j} \in [0, 1],\ i = 1, \ldots, n,\ j = 1, \ldots, c$.

The algorithm minimizes intra-cluster variance as well, but has the same problems observed for k-means algorithm, in particular the minimum is a local minimum, and the results depend on the initial choice of weights. To reduce these limitations the algorithm should be run different times, considering stable results.

2.4 Association Rules

Association rules represent widely applied Data Mining methods. However in this sections only the basic concepts are given, introducing the well-known Apriori algorithm, since they have not be implemented in the following sections of this thesis.

Differently from Classification and Clustering methods, which are *global* approaches, Association rules are *local* methods. This means that the algorithm concentrates mainly on a part of the complete dataset, in terms of both records and variables of interest, instead of considering the complete dataset(records or variables of interest) as done by global approaches.

The main objective of Association rules is identify relationships among the ”items”, grouping them based on common ”purchase behavior”.

The generic term ”items” is used since the type of elements included in the database depends on the type of analysis considered. As an example, for a Market Basket Analysis they are the purchase done by clients in a supermarket, for a Web clickstream analysis they are the website pages visited, for an Accident Analysis the accident happened in a given road network.

The term ”purchase behavior” refers to how data have been collected. Generally data are organised in transactional databases, where each record includes all the purchases made by a client in each transaction (i.e. the list of products bought by the client or the pages visited in the website).

The information acquired at each transaction may change depending on the objective of the analysis. In the simplest cases only type of product

are analysed; in more detailed analyses quantity, cost and other elements of interest could be added.

In the following paragraphs the basic concepts of Association rules will be given considering the simple case where only types of product are analysed. Each product is described by a binary variable, which takes the value 1 if the product was purchased and 0 if not.

Given a set of *items* $I = \{I_1, I_2, \ldots, I_m\}$ and a transitional database $D = \{t_1, t_2, \ldots, t_n\}$, where $t_i = \{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ con $I_{ij} \in I$, an *association rule* is the implication $X \Rightarrow Y$, where $X, Y \subset I$ are *itemsets* for which $X \cap Y = \emptyset$.

This means that the association rule $X \Rightarrow Y$ describes a significant relationship between two sets of items of interest (*itemsets*) from transactions collected in a database. The fact that a significant relationship $X \Rightarrow Y$ exists does not mean that a causal relationship exist between X and Y. Association rules only states that the two terms of the rule occur together in the database. Other types of analyses (such as the analysis of the temporal sequences of purchases) must be adopted to identify causal relationships among items. Since association rule is a *local* model, it selects only two itemsets among the total number of items in the database: the itemset X, which is the premise of the rule, and itemset Y, which is the consequence of the rule.

To identify significant relationships and define significant association rule, some indices of significance are adopted. Given an association rule $X \Rightarrow Y$ three indices can be defined: *support*, *confidence* and *lift*.

The *support* s of the association rule $X \Rightarrow Y$ is the percentage of transactions in the database which contains $X \cup Y$, that is the probability that both events X and Y occurred simultaneously in the dataset:

$$s = \text{support}\{X \Rightarrow Y\} = \frac{N_{X \cup Y}}{N} = P(X \cap Y) \tag{2.89}$$

where $N_{X \cup Y}$ and N are the number of transactions which contains $X \cup Y$ and the total number of transactions, respectively.

The *confidence* α of the association rule $X \Rightarrow Y$ is the ratio between the number of transactions which contains $X \cup Y$ and the number of transactions which contains X, that is the conditional probability that event Y, given that event Y has occurred:

$$\alpha = \text{confidence}\{X \Rightarrow Y\} = \frac{N_{X \cup Y}}{N_X} = \frac{\text{support}\{X \Rightarrow Y\}}{\text{support}\{X\}} \tag{2.90}$$

$$= \frac{P(X \cap Y)}{P(X)} \tag{2.91}$$

where $N_{X \cup Y}$ is the number of transactions which contains $X \cup Y$ and N_X is the total number of transactions which contains X, respectively.

The *lift* l of the association rule $X \Rightarrow Y$ is the ratio between the probability that Y will occur when X occurs to the general probability that Y will occur,

that is:

$$\text{lift}\{X \Rightarrow Y\} = \frac{\text{confidence}\{X \Rightarrow Y\}}{\text{support}\{Y\}} = \frac{\text{support}\{X \Rightarrow Y\}}{\text{support}\{X\}\text{support}\{Y\}} \tag{2.92}$$

$$= \frac{P(X \cap Y)}{P(X) \cdot P(Y)} \tag{2.93}$$

Lift values are important to define the strength of the relationships among itemsets:

- **Lift > 1**: There exists a positive association between event X and event Y of the rule. In practice, if Lift = 2, it is twice as likely that event Y will occur when event X occurs than the likelihood that event Y will occur.
- **Lift = 1**: There is no association between occurrence of events X and Y. In practice, if Lift = 1, it is neither more likely nor more unlikely that event Y will occur when event X occurs, than the likelihood that event Y will occur. In these cases, X and Y are considered independent.
- **Lift < 1**: There exists a negative association between event X and event Y of the rule. In practice, if Lift < 1 it is less unlikely that event Y will occur upon occurrence of event X, than the likelihood that event Y will occur. If Lift = 0, then event Y will never occur simultaneously with event X (X and Y are mutually exclusive events).

2.4.1 Basic Algorithm

To identify significant rules, the database must be screened in an efficient rule by the algorithm. Traditional algorithms are based on some basic concepts:

- A *frequent itemset* l is an itemset which occurs a minimum number of times in the dataset, that is it has a minimum support s;
- L is the complete set of frequent itemsets in a database.

Traditional algorithm are based on two steps:

Step 1. Identification of frequent itemsets in the database;

Step 2. Generation of association rules based on the frequent itemsets.

The definition of set L is necessary to the definition of association rules, since for each association rule $X \Rightarrow Y$ must be $X \cup Y \in L$. The search can become time-consuming if specific algorithm are not implemented, since, given m records in the dataset, the total number of subsets is equal to 2^m.

Traditional algorithms differ for the type of algorithm they implement for the identification of "candidate" itemsets c, which are grouped in the "candidate set" C.

2.4.2 Apriori Algorithm

The Apriori algorithm is probably the most applied algorithm for the definition of association rules. The main characteristic of this algorithm is using the property that frequent itemsets are "*downward closed*" to optimize the search in the database.

One considers the dataset in Figure 2.6, which includes the elements $\{A, B, C, D\}$ with their subsets and the links representing the relationships among items.

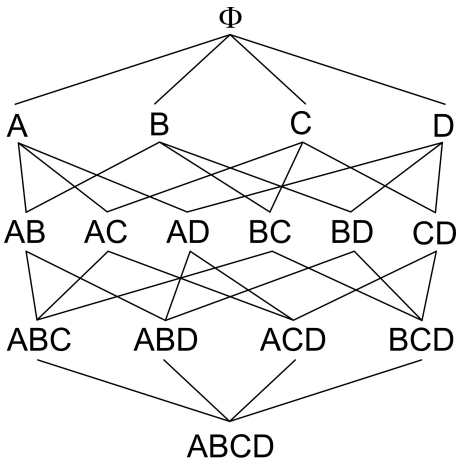


Figure 2.6: *APriori. Net of* $\{A, B, C, D\}$

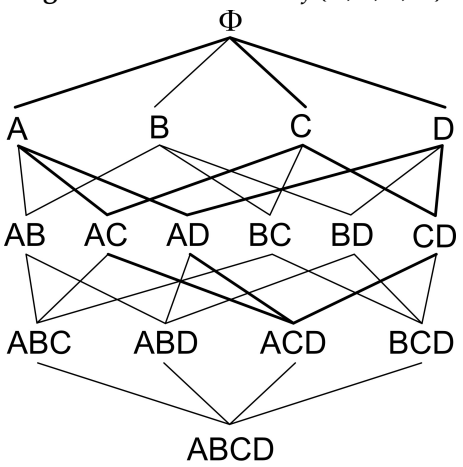


Figure 2.7: *APriori. Subsets of* $\{A, C, D\}$

If $ABCD$ is a frequent itemset, all the sets in the path from node $ABCD$ to the top of the net are frequent itemsets, since itemsets are downward closed. In Figure 2.7 the paths which refers to set ABC are highlighted and they include nodes $\{AC, AD, CD, A, C, D\}$.

The procedure followed by the algorithm is:

Step 1. Scan the entire database, generate candidate itemsets of dimension i and determine how many of them (C_i) are frequent;

Step 2. Only frequent itemsets L_i are used to generate candidates of the next level.

At the first level the generation of candidate itemsets is done considering all the couples in the database, while in the next levels the *Apriori-Gen* sub-algorithm is adopted. Apriori-Gen sub-algorithm combined couples of L_i frequent itemsets, which have $i - 1$ common items, generating C_{i+1} new candidate itemsets. To determine the frequency of C_{i+1} itemsets a new scan of the database is needed.

Chapter 3

Data Mining in Transportation Engineering

Data Mining techniques have been adopted with increasing attention in different research fields, including transportation engineering. This chapter summarizes some of the main applications reported in literature in the latest years (Barai 2003). The objective is not producing an exhaustive review (which is not feasible in practice given the amount of techniques and applications to be considered), but giving an idea of the research topics which could benefit of these innovative tools.

It must be noted that other transportation engineering topics not included in this review could potentially take advantage of using these techniques. Whenever basic objectives must be achieved (e.g. classification, clustering, estimation), DM techniques could be an efficient set of tools to be used by the analyst.

Moreover the analysis of the wide literature concerning traffic monitoring is reported separately in Chapter 5.

3.1 Knowledge Discovery Process

As observed in Section 2.4, DM can be considered a part of a more general process of knowledge discovery. Some paper dealt with the implementation of this process for transportation applications.

In their paper Zhaohui et al. (2003) identified the importance and the necessity of applying knowledge management (KM) in Intelligent Transportation System (ITS). The paper discussed possible targeted market, the application ranges of KM in ITS, the main contents and the primary methods of KM in ITS, finally proposing a typical model of KM in ITS.

Haluzová (2008) described the application of data mining methods (KDD framework) in the database of the DORIS transportation information system, used by the Prague Public Transit Company. Descriptive sta-

tistical methods and association rules were successfully applied to create knowledge about the behaviour of objects within the information system.

Similarly Maciejewski and Lipnicki (2008) applied data exploration and data mining techniques for the analysis of monitoring databases of a large transport system. The analyses focused on discovering relationships between key metrics of a transport system as such availability/usage profiles of the fleet and various factors on which they apparently depend, such as age. Based on real data from the transport system of the Polish Post, data exploration, data mining and efficient data summarization / visualization tools were found to be able to find bottlenecks, atypical patterns or event fraud-related events in the mail transport / mail delivery process.

Yang and Tong (2008) implemented KDD framework in a Information-Service-Oriented Experimental Transportation System (IETS) in China, with the aim of deducing useful traffic information from real-time traffic data. The example given by the authors showed that this method can be useful to discovery unknown information in traffic data.

Zhang, Huang, and Zong (2010) applied KDD framework (Knowledge Discovery in Databases) to decision making of railway operation safety in China. The application of KDD framework to accident treatment showed important significance to assist the decision making for train running safety.

Another application in railway context was proposed by **MELO007** who presented the so-called MPDF-DM methodology for the prediction of railroad demand. This methodology integrated existing methodologies (including for example CRISP-DM) and was successfully applied in a customer transport request database of Brasil.

Recently Rahman, Desa, and Wibowo (2011) reviewed data mining (DM) applications in logistics, considering their benefits and real-world expectations. The authors noted that very little is known about the usefulness of applying data mining in transport related research. Since frameworks for carrying out knowledge discovery and data mining (e.g. KDD or CRISP-DM) have been revised over the years to meet the business expectations, the authors proposed a framework to be tested within the context of transportation industry.

3.2 Pavement Management Systems

From the 1960s, Pavement Management Databases (PMS) have been designed and implemented by many transportation agencies. The main objective of these systems is supporting the efficient management of the infrastructure, in terms of planning, design, construction, maintenance, evaluation, and rehabilitation of pavements. With rapid increase of advanced information technology, many investigators have successfully integrated the Geographic Information System (GIS) into PMS for storing, retrieving,

analysing, and reporting information needed to support pavement-related decision making (G-PMS). G-PMS are enhanced with features and functionality by using a geographic information system (GIS) to perform pavement management operations, create maps of pavement condition, provide cost analysis for the recommended maintenance strategies, and long-term pavement budget programming.

With the increasing amount of pavement data collected, updated and exchanged due to deteriorating road conditions, increasing traffic loading, and shrinking funds, DM and KDD process have been implemented as a further tool for decision-makers.

Soibelman and Kim (2000) applied Knowledge Discovery in Databases (KDD) and Data Mining (DM) as tools to identify the novel patterns in construction fields. Leu, Chen, and Chang (2001) investigated the applicability of data mining in the prediction of tunnel support stability using an artificial neural network (ANN) algorithm. The case data of a railway tunnel were used to establish the model. The validation performed showed that the ANN outperformed the discriminant analysis and the multiple non-linear regression method in predicting tunnel support stability status. Attoh-Okine (Attoh-Okine 2002; Attoh-Okine 1997) presented applications of Rough Set Theory (RST) to enhance the decision support of the pavement rehabilitation and maintenance. Sarasua and Jia (1995) explored an integration of GIS Technology with knowledge discovery and expert system for pavement management.

More recently Zhou et al. (2010) explored the applicability of data mining and knowledge discovery (DMKD) in combination with Geographical Information Systems (GIS) technology to pavement management in order to better decide maintenance strategies, set rehabilitation priorities, and make investment decisions. Different data mining techniques, including decision trees and association rules, were used to find pertinent information hidden within a pavement database covering four counties within the State of North Carolina. The knowledge discovery process identifies seven rules, which guided the creation of maps (using GIS tools) of several pavement rehabilitation strategies. Analysing the results obtained for the pilot experiment in the project, the authors concluded that:

- the use of the DMKD method for the decision of road maintenance and rehabilitation can greatly increase the speed of decision-making, thus largely saving time and money, and shortening the project period;

- the DMKD technology can make consistent decisions about road maintenance and rehabilitation if the road conditions are similar, i.e., interference from human factors is less significant;

- integration of the DMKD and GIS technologies provides a Pavement

Management System with the capabilities to graphically display treatment decisions.

3.3 Accident Analysis

Analysis of accidents was found to be one of the topic with the largest number of DM applications. Analyses were developed for different transportation modes (e.g. for aviation incidents see Gürbüz, Özbakir, and Yapici (2009)), but the majority of studies have considered road accidents. Different research areas of interest can be identified, including Accident Detection (where prediction methods such as ANNs were adopted (Hu et al. 2003)), Black-spots and Risk factors identification.

In this review more details are given about the identification and quantification of risk factors, a rich research area that can be helpful to understand future developments of DM in transportation studies. Analysing the large amount of papers with DM techniques, two types of approaches can be identified, depending on how the spatial characteristics of accidents were treated.

In the first case traditional DM techniques (section 2) are applied, considering the spatial element as a variable like the others. In the second case DM techniques are extended, integrating some important concepts from Geographical Information Systems (GIS) and geographical sciences.

Some applications of the traditional approaches are the following.

Kalyoncuoglu and Tigdemir (2004) adopted an artificial neural networks (ANN) approach to simulate the effects of driver characteristics (i.e. gender, age, education, driving years, Kms drive per year) into the traffic accidents (Percentage of involvement in traffic accidents). The flexible and assumption-free ANN approach produced predictions considered by the authors highly satisfactory.

Chang (2005) compared a negative binomial regression model and an ANN model to analyse vehicle accident frequency in the National Freeway 1 in Taiwan. The results reported in the paper demonstrates that ANN is a consistent alternative method for analyzing freeway accident frequency.

Analysing another accident dataset, the same author (Chang and Chen 2005) found that also Classification and Regression Tree (CART) is a good alternative method for analyzing freeway accident frequencies and establishing empirical relationships between traffic accidents, highway geometric variables, traffic characteristics and environmental factors.

In 2006, Pande and Abdel-Aty (2006) used data mining process to relate surrogate measure of traffic conditions (data from freeway loop detectors) with occurrence of rear-end crashes on freeways. Freeway traffic surveillance data collected through underground loop detectors is a "observational" database maintained for various ITS (Intelligent Transportation Sys-

tems) applications such as travel time prediction. The results highlighted that a classification tree model with chi square test as splitting criterion was better than any of the individual or combined neural network models tested. The decision tree also provided simple interpretable rules to classify the data in a real-time application.

Finally, the same authors (Pande and Abdel-Aty 2009) applied association rules analysis (market basket analysis) to detect interdependence among crash characteristics using non-intersection crash data from the state of Florida for the year 2004. Based on the association rules discovered from the analysis, they found significant correlation of accidents with some conditions (e.g. lack of illumination or rainy conditions) consistent with the understanding of crash characteristics. The authors stated that the the potential of this technique might be realized in the form of a decision support tool for the traffic safety administrators.

In the more innovative approach specific attention is given to concept of *spatial autocorrelation* (Getis 2008), which explains the correlations among variables in a georeferenced space. This means that for certain dataset the correlations among variables can vary depending on the spatial localization. Without giving further details, it can be observed that some studies (Flahaut 2004; Flahaut et al. 2003; Khan, Qin, and Noyce 2008) have demonstrated that spatial autocorrelation can be a relevant element when analysing the spatial characteristics of events (such accidents), with relevant changes in the final results of analyses.

DM techniques are evolving in the sense of integrating all these aspects in common complex frameworks (Geurts, Thomas, and Wets 2005). In the future the adoption of these multidisciplinary approaches will probably represent a relevant topic for Transportation Engineering to be considered with interest (Spatial Data Mining (Ladner, Petry, and Cobb 2003; Miller and Han 2009) and Spatio-Temporal Data Mining (Cheng and Wang 2008; Mennis and Liu 2005)).

3.4 Traffic forecasting

In recent years Intelligent Transportation Systems (ITS) have been implemented widely throughout the world, giving the access to large amounts of data. In this particular case DM techniques are appropriate tools to acquire useful traffic patterns, in particular when "real-time" data are available. This fact has moved traffic operating systems from passive to proactive control and management through the adoption of traffic forecasting data. From the review of traffic-flow forecasting models done by Han and Song (2003) DM techniques represent one of the basic models currently adopted in this field.

In particular ANNs have been commonly used for the problem since

1990's, given the fact the appropriate neural network structure can approximate a given real-valued, continuous multi-variate function to any desired degree of accuracy (Hornik 1989). Different structures have been applied, including multi-layer perceptron, radial-based function and hybrid methods (Dougherty and Cobbett 1997; Van der Voort, Dougherty, and Watson 1996).

However, also other Data Mining techniques have been implemented with interesting results. Gong and Liu (2003) proposed an algorithm based on association rules mining and association analysis.

Wang et al. (2006) presented a dynamic traffic prediction model. The model deals with traffic flow data to convert them into traffic status. In this paper clustering analysis and classification analysis are used to develop the model and the classification model can be used to predict traffic status in real time. The experiment shows the prediction model can be used efficiently in the dynamic traffic prediction for the urban traffic flow guidance.

In their paper Gong and Lu 2008 proposed a Data Mining based Traffic Direction Control Algorithm (DMTDCA) to adjust the traffic direction of Direction-Changeable Lanes (DCLs) in a tunnel in Shanghai. Current traffic flow and short-term forecasted traffic flow of two tunnel entrances were analysed and the direction change is decided automatically and timely. Field tests showed an increase of average traffic capacity and a decrease of average queue length.

Interesting applications of DM traffic forecasting have been done also for air traffic flows and railway passenger flows.

Cheng, Cui, and Cheng (2003) employed a combination of neural networks and statistical analysis of historical data to forecast the inter-regional air traffic flow in China. Models for different prediction conditions were derived from the analysis of large collection of data radar. The accuracy of predictions was found satisfactory.

In railway case Xu, Qin, and Huang 2004 proposed an approach to forecast railway passenger flow based on spatio-temporal data mining. The approach first forecasts time sequence of target object using statistical principles, then figures out the spatial influence of neighbour objects using a neural network, and finally combines the two forecasting results using linear regression. Comparing with previous approaches, that did not consider the spatial influence, the approach resulted in higher forecasting accuracy.

3.5 Other Studies

In this section, other interesting applications of Data Mining in transportation field are reported.

Jiang and Huang (2009) addressed the problem of calibrating speed-density relationship parameters used by mesoscopic traffic simulators with

data mining techniques. The authors combining K-means with agglomerative hierarchical clustering, being able to reduce early-stage errors inherent in agglomerative hierarchical clustering resulted in improved clustering performance. The results from the case study in Beijing showed that the performance of the new algorithm was better than previous solutions.

Chen et al. (2004) applied DM techniques to online vehicle tracking system data, which are generally an enormous quantity of records. Vehicle behaviors could be mined to understand the status of every vehicle or driver (e.g. deviations from routes, driving against traffic regulations) and alert to abnormal conditions. The implementation of these techniques could give reduction of the operation costs, greater flexibly of dispatching vehicles, and therefore competitive advantages for the transportation industry.

Hayashi et al. (2005) presented a detection method of driver's drowsiness with focus on analysing individual differences in biological signals (pulse wave) and performance data (steering data). Biological signals of different drivers were analysed by neural networks, which successfully adapted to the differences observed among drivers. The correct detection of driver's drowsiness is a need for realization of safer traffic environment, contributing to prevent traffic accidents caused by human errors in a drowse.

Part II

Road Traffic Monitoring

Chapter 4

Traffic Monitoring Guide

Based on the literature review of Data Mining applications (chapter 3), road traffic monitoring was found to be a topic of interest for the application of these techniques.

Road traffic monitoring represent a relevant aspect of highway planning activities for many transportation agencies. Since data collection activities produce relevant management costs, both for equipment and personnel, monitoring programs need to be designed with attention to obtain the maximum quality of results.

While in Italy there is little guidance on how monitoring program must be implemented in practice (*Sistemi di Monitoraggio del Traffico. Linee guida per la progettazione*), in the U.S.A. the Federal Highway Administration (FHWA) provides the guidelines for the implementation of statewide data collection programs by way of the Traffic Monitoring Guide (TMG) (FHWA 2001).

TMG describes the analytical logic followed in a monitoring program and provides the information highway agencies need to optimize their frameworks. In practice each State or local highway agency has its own traffic counting needs, priorities, budgets, geographic and organizational constraints. However the general framework proposed by TMG can be applied in different contexts selecting different equipment for data collection, using different collection plans for obtaining traffic data, and emphasizing different data reporting outputs.

For this completeness TMG represents a reference for many other countries (Italy included), therefore in this thesis it has been analysed in depth, identifying the main issues still unsolved (Chapter 5).

4.1 Data Collection Design

Collecting data and predicting the traffic patterns for both short-term and long-term planning are the basic responsibilities of transportation agencies. One of the most important traffic parameter is the Annual Average

Daily Traffic (AADT), which should be known for each section in the road network. AADT is defined as the bi-direction traffic count representing an average 24-hour day in a year for a given road section. It is an essential information for pavement design, fuel-tax revenue projection, and highway planning.

Although it is ideal to monitor traffic volume for the entire year, data collection incurs cost and manpower; hence, AADT is often estimated with minimum data collection effort with monitoring activities. To provide accurate estimates of AADT traffic monitoring program must measure and account for the variability of traffic patterns as they occur over time and space. The variability in traffic conditions is related to different time scales, including time of day, day of week, season (month) of the year and different space levels, from the direction of traffic in the same section to geographical variations on a certain area, such as a Province or a Region.

Moreover differences in traffic variation also exist by type of vehicle, since truck volumes vary over time and space differently than automobile volumes (Hallenbeck et al. 1997). For this reason the latest version of TMG (FHWA 2001) suggests to differentiate the data collection using three or four simple categories, obtained aggregating existing FHWA classes, such as passenger vehicles (motorcycles, cars, and light trucks), single-unit trucks (including buses), single-unit combination trucks (tractor-trailers), and multi-trailer combination trucks.

To monitor traffic flows in a network, the basic program design recommended by TMG consists of two types of counts:

1. Permanent Traffic Counts (PTCs)
2. Short Period Traffic Counts (SPTCs)

which are combined using the Factor approach.

4.1.1 Permanent Traffic Counts

Permanent Traffic Counts are counts taken on a road section all year long using continuous traffic monitoring data collection equipments. PTCs provide knowledge of time-of-day, day-of-week, and seasonal travel patterns and very precise measurements of changes in travel volumes and characteristics at a limited number of locations.

The most common continuous data collection equipments include *Automatic Traffic Recorders* (ATR) and *Automatic Continuous Vehicle Classifiers* (often abbreviated AVC or CVC). AVCs should be preferred to ATRs, since they can provide information about volumes of different vehicle classes, while ATRs provides only total volume values. This difference is relevant for many traffic analyses, which are more dependent on truck volumes than they are on total traffic volumes, such as pavement design. Both types of

counters generally adopt inductance loops technology because it allows for reliable, long lasting installation of the vehicle detector. In case of AVCs dual loop systems are installed and estimates the total length of vehicles crossing the loops can be obtained.

Other equipments often included in traffic monitoring program are continuously operating weigh-in-motion (WIM) scales, placed to monitor vehicle weights, and volume and speed monitoring stations, that provide facility performance data to traffic management systems. However they are not described here since they are less important in the implementation of the monitoring program and less common in the Italian application context.

4.1.2 Short Period Traffic Counts

Short Period Traffic Counts are taken for short periods (from 24 hours to some weeks) using portable traffic counters based on various technologies (e.g. infrared, microwave, radar). SPTCs provide the geographic coverage needed to understand traffic characteristics on individual roadways. Because permanent counters are expensive to install, operate, and maintain, SPTCs must be taken on the road network to provide accurate measurements of traffic conditions.

The TMG recommends that SPTCs data collection consists of a periodic comprehensive coverage program over the entire system on a 6-year cycle (reduced to a 3-year cycle on the main roads of the network). However roadway sections that are of major interest (locations of pavement design projects, corridors for major investment studies, etc.) are counted often (every year or every two or three years), while other roadway sections with little activity or with stable traffic volumes may be uncounted for many years.

Depending on the the number and duration, SPTCs taken on a road section can be further classified as:

- Seasonal Traffic Counts (STCs) or Coverage Counts, when the duration is from 24 hours to some weeks and the number is from 2 to 12 times in a year;
- Short Period Traffic Counts (SPTCs) strictly meaning, when the duration is less then a week and they are taken only once in a year.

In the following parts of this thesis this distinction between STCs and SPTCs will be maintained.

4.2 TMG Factor Approach

TMG factor approach (also known as FHWA or traditional factor approach) indicates how SPTCs and PTCs are combined to estimate average

traffic conditions in a given section of the road network.

PTCs give detailed information about traffic conditions all the year long, but they are obtained only in a limited number of road sections. Conversely SPTCs can measure the traffic conditions in different road sections, but only for the limited amount of time during which the counts are taken.

TMG factor approach assumes that temporal characteristics affect all roads in the network and since continuous temporal data (PTCs) exist at several points to describe the temporal variation, it is possible to transfer this knowledge by developing a factoring mechanism.

The factoring process defines a set of roads as a 'road group' and all roads within that group are assumed to behave similarly. Then a sample of locations on roads from within that group is taken and data are collected (PTCs). The mean condition for that sample is computed and that mean value used as the best measure of how all roads in the group behave. Adjustments factors (seasonal adjustment factors) are calculated for each road group to account for temporal variability (time-of-day, day-of-week and seasonal) of the traffic stream. Road monitored with SPTCs are assigned to one of the 'road group' and road group seasonal adjustment factors are used to correct data obtained from SPTCs and estimate annual average conditions.

The procedure proposed in the TMG can be summarized in four steps:

Step 1: Define groups of roads which are assumed to behave similarly. A sample of locations on roads from within each group is taken and PTCs are collected, generally using an AVC;

Step 2: Compute the mean conditions for the sample of each road group and assume those mean values as the seasonal adjustment factors for the road group;

Step 3: Assign the road section in question, that is monitored with a SPTC, to one of the groups defined in step 1;

Step 4: Apply the appropriate seasonal adjustment factor to the SPTC of the road group to produce the AADT estimates for the road section in question.

Assuming that a State has decided to consider weekly and monthly variations of traffic volumes and has identified the road groups, the factor approach leads to the following computation (further details about the creation of road groups and the type of factors will be given in section 4.3).

The seasonal adjustment factor for an AVC site k for the i-th day of the week of the j-th month is calculated by:

$$f_{ij}^{k} = \frac{AADT_k}{ADT_{ijk}} \tag{4.1}$$

where $AADT_k$ is the AADT for the k-th AVC site, ADT_{ijk} is the average daily traffic recorded in the i-th day of week of the j-th month in the k-th AVC site.

Since AVC sites grouped together are supposed to have similar traffic patterns, the seasonal adjustment factors that correspond to (i, j) combinations are calculated for each road group. If n AVC sites are in road group c, the seasonal adjustment factor for the i-th day of week of the j-th month is calculated by:

$$f_{ijc} = \frac{1}{n} \sum_{k=1}^{n} \frac{AADT_k}{ADT_{ijk}} = \frac{1}{n} \sum_{k=1}^{n} f_{ij}^{k} \tag{4.2}$$

where $AADT_k$ is the AADT for the k-th AVC site in group c and ADT_{ijk} is the average daily traffic recorded in the i-th day of week of the j-th month in the same AVC site.

Once a road section is assigned to a group c, AADT can be estimated by multiplying the daily traffic count DT_{ij}, obtained for the i-th day of week of the j-th month, by the corresponding seasonal adjustment factor f_{ijc}:

$$AADT_{Estimate} = DT_{ij} \cdot f_{ijc} \tag{4.3}$$

DT is the 24 hour volume obtained from SPTC; if SPTC is for more than 24 hours, then DT is the average 24 hour volumes for the duration of SPTC.

To calculate the AADT and the ADT needed in equations 4.1 and 4.2, TMG suggests to use the AASHTO method (AASHTO 1992), which is more accurate than the simple average of all daily counts in case of missing data. Therefore the AADT (and similarly the ADT) for the k-th AVC site in the group is calculated by:

$$AADT_k = \frac{1}{7} \sum_{i=1}^{n} \left[\frac{1}{12} \sum_{j=1}^{12} \left(\frac{1}{n} \sum_{l=1}^{n} DT_{ijlk} \right) \right] \tag{4.4}$$

where DT_{ijlk} is the daily traffic count taken the i-th day of the week of the j-th month in the k AVC site. The index l represents the occurrence of the i-th day of the week in the j-th month, and n is the number of times that the i-th day of the week during month j occurs (usually between 1 and 5, depending on the calendar and the number of missing days).

For many years TMG factor approach has been applied using total volume values. This fact was mainly due to technological limits of vehicle monitoring equipments. Since substantial differences were observed between truck and passengers car travel patterns, TMG now suggests to apply the factoring approach in a separate way for different vehicle classes.

In the following sections the analysis will consider the case of a single vehicle class (that is the same of considering total volume). Further details

about specific characteristics of truck vehicle class will be given in section 4.5.

4.2.1 Basic Issues of TMG Factor Approach

TMG highlights some relevant issues that must be considered when applying the factor approach, since they are the source of many of the errors in annual traffic estimates. Different techniques used to create and apply traffic correction factors allow the user to control the errors associated with any given step of the procedure. Unfortunately, none of the available techniques can control for all of the limitations.

Variability and similarity. It is difficult to define groups of roads that are similar with respect to traffic variation, and the more mathematically alike the factoring groups created from the data, the more difficult it is to define the attributes that determine which roads belong to a given group.

Definition of road groups. The appropriate definition of a road group changes depending on the characteristic being measured.

Sample selection. It is difficult to select a representative sample of roads from which to collect data for calculating the mean values used as factors.

Incomplete datasets. The datasets used to compute adjustment factors are not complete.

Variability and similarity This issue relates to the fact that it is quite easy to define groups of roads with a high level of precision based on variability. However, the groups that can be easily defined based on variability usually do not have clear characteristics to identify the group. Therefore, the creation of factor groups usually involves balancing the need to easily define a group of roads against the desire to ensure that all roads within a given group have similar travel patterns.

This same trade-off occurs in the type and magnitude of errors in the factoring process. For groups that are easy to define but include wider ranges of travel patterns within the group, errors occur because the mean factor computed for the group may not be a good estimate of the ”correct” factor for a specific road segment. For groups that have very ”tight” factors but for which it is difficult to define the roads that fit, the error occurs in defining to which factor group a specific road segment belongs.

Definition of road groups This fact can be easily observed considering that trucks have different travel patterns than passenger cars. Factor groups that work extremely well for computing and applying total volume adjustments (dominated by car volume patterns) often do not work well for truck volume adjustments.

In general, the 'best factor groups' are those that can be readily defined and at the same time contain similar traffic patterns.

Sample selection The best alternative for selecting these sites is to first define the factor group and then perform a random selection of data collection sites. Normally, neither of these events takes place. Consequently, the "mean" value computed is often not the "true" mean value for the group, in particular if one "unusual" location is included within a group.

Data collection points are usually not "perfect" for two reasons.

- Permanent data collection site locations are often selected for a number of reasons, only one of which is factor computation. Some of these reasons are:
  - the need for data from a specific site;
  - the desire to track trends over time at sites that have been historically monitored;
  - the need for specific physical conditions (the availability of power or communications lines, the need for smooth, flat pavement).
- Because factor groups are often determined on the basis of data from existing data collection sites, actual site locations often exist before the grouping process and cost considerations tend to prevent their being moved. Thus, the data drive the grouping process, rather than the grouping process driving the selection of data collection points.

Incomplete datasets No data collection device is perfect, therefore within any given road network, a certain number of ATR or AVC devices will fail each year, from a few hours to several months. In some cases, so few data are available that the site may not be included in the factor computation at all.

A number of procedures, most notably the AASHTO process for computing AADT (AASHTO 1992), have been designed to limit the effects of missing data. However, because of the holes in the data, errors are introduced into the factors being computed and, in general, the more data that are missing, the more error that may be associated with the mean factors applied to any given location.

The best way to decrease the chance of these errors occurring is to monitor and repair permanent data collection equipment as quickly as

possible. It is also helpful to maintain more than the minimum number of counters within any given factor group, so that if one counter experiences a long data outage, the data from that counter can be removed from the computation process without adversely affecting the factor computation.

4.3 Details About Factor Approach

To be applied in practice, the basic structure of TMG factor approach needs to be further explained, giving some details about:

- the type of factors that can be used;
- the definition of road groups;
- the number of road sections to be monitored.

4.3.1 Types of Factor

Different procedures have been developed for calculating and applying factors. Cambridge Systematics Corporation (1994)'s study showed that different factoring techniques can result in reasonably similar levels of accuracy in the estimate of average annual conditions (Average Percentage of Errors between 7.0% and 7.6%).

The key point is that the factoring technique must account for all types of variation present in the data. This means that the level of aggregation that exists in each factor and the definition of 'seasonal' can change case by case. In some cases, day-of-week and seasonal adjustments are combined in a single factor; in other cases, these two components are treated as separate factors.

For seasonal adjustments, some techniques use monthly factors, whereas others use weekly factors. Both of these techniques can be successful. Seasonality does not necessarily vary smoothly from month to month. Consequently, some States find that weekly factors work better than monthly adjustment factors. However, others find that the monthly factors provide equally good annual adjustments and require considerably less effort to compute and apply.

Similarly for day-of-week factors, some States use day-of-week adjustments for each day. Others combine some weekdays (traditionally Tuesday to Thursday or Monday to Thursday). Both techniques can produce acceptable results if they are applied appropriately, that is if they can be representative of traffic pattern variations.

4.3.2 Road Groups Identification

The TMG suggests three alternative techniques for computing factor groups from existing PTCs data. Therefore the first step is to compute the adjustment factors that will be used in the group selection process for each site for which data are available. The analyst should pay particular attention to the quality of the data produced by each counting device.

As explained later strengths and weaknesses exist for each alternative. In many cases the combination of approaches could be better than following any one technique exclusively. The three techniques here introduced are:

1. cluster analysis;
2. geographic/functional assignment of roads to groups;
3. same road factor application.

Cluster Analysis

With the term "cluster analysis" the TMG refers to the application of the Ward's Minimum Variance method for hierarchical clustering (see section 2.3.2), which determines which AVCs or ATRs are most similar based on the computed factors.

The analyst's decision of determine at what point to stop the analysis should be done in one of two ways:

- The first way is to analyse the mathematical distance between the clusters formed, as usually done in clustering analysis;
- The second approach is to choose a predetermined number of groups to be formed.

In practice both options require that the analyst is able to define the group of roads a given cluster of continuous counters actually represents. This definition (which is spatial and functional) is necessary to assign arbitrary roadway segments to the newly created factor groups, since the analysts must understand the rules for assigning short counts to the factor groups.

This task is extremely difficult, therefore the cluster process is often modified by the use of secondary procedures (a combination of statistical analysis and analyst knowledge and expertise) to develop the final factor groups.

Geographic/Functional Classification of Roads Factor Groups

This approach is based on the available knowledge about traffic patterns of the roads in the network. The analyst allocates roads into alternative

factor groups on the basis of available knowledge, which is usually obtained from a combination of existing data summaries and professional experience with traffic patterns.

First, factor groups are selected based on a combination of functional roadway classification and geographic location. The use of functional classes for group selection makes it easy to assign individual road sections to factor groups and also allows the creation of factor groups that are intuitively logical. Sub-State geographical differences could be considered when different levels and types of economic activity exist and differences in traffic characteristics can occur.

Once the initial factor groups have been identified, PTCs data are examined for each group. For each factor and each factor group, the mean factor for the group and the standard deviation of that factor are computed. The standard deviation tells the analyst the size of the expected error of the average group factor. Since it is assumed that the continuous counters for which data are available are representative of a random sample of roads of that group, the errors should be roughly normally distributed about the factor group mean. If the standard deviation is too high (i.e., the error associated with factors computed for that group of roads is too large), the definition of roads included in that group may have to be changed. This can mean the creation of new factor groups (for example, splitting some roads based on geographical characteristics), or the redefinition of those groups. Changing factor group definitions effectively moves continuous counters from one factor group to another and allows the variation within a given group to be decreased.

Performing this analysis, particular attention must be devoted to the most important factors, which represent the most significant period of the year for that roads. For example, if the majority of traffic counting takes place from the middle of Spring to the middle of Fall, the factor group variation in January and December is less important (because these factors may never be used) than the variation in the key May through September time period, when most short duration traffic counts are taken. Similarly attention is needed to the presence of "outlier" sections, which are continuous counters that do not really fit within the basic pattern that is assumed to exist. For example, if a counter does not fit within a factor group, having a plot of that counter's data will allow the analyst to determine whether the data for that site contain potential errors that could affect the grouping process, indicate the need to create a recreational factor group, or provide the insight to place the stations in another group.

Same Road Application of Factors

The third approach proposed for the creation of factor groups can be applied only if a dense network of continuous counters exist. This process

assigns the factor from a single continuous counter to all road segments within an "influence area" of that counter site. The boundary of that influence zone is defined as a road junction that causes the nature of the traffic volume to change significantly. This approach has a particular interest since it avoids the application of a mean value that does not accurately describe traffic variation on that given road section and the problem of associating a specific road section with a vaguely defined factor group.

Difficulties in the application of this technique occur when the short duration count is not near the continuous counter and traffic patterns at the count location may be different than those found at the continuous counter. Application of factors from individual locations in this fashion creates considerable potential for bias in the factoring process.

Combining Techniques As noted at the beginning of this subsection, most States develop and apply factors by using some combination of the above techniques. For example, on road sections where continuous counters exist nearby, factors from specific counters can be applied to short duration counts on those roads. For all other road sections, group factors can be computed and applied. Factor groups can be initially identified by starting with the cluster analysis process followed by the use of common sense and professional judgement. In this way minor adjustments can be made to the cluster results in order to define the final factor groups in such a way that they can be easily identified for factor application.

4.3.3 ATR Sample Dimension

FHWA has fixed the precision levels requested by AADT estimates taken form different road groups, considering the geographical context, the functional level of roads and the approximate expected values of AADT. Appendix C of HPMS field manual (FHWA 2005) presents these levels, which have a confidence level included in the interval $80-90\%$ and a level of error in the range $5-10\%$.

HPMS standards for the precision of AADT estimates consider road groups corresponding to specific functional classes. Otherwise TMG is more focused on differences among roads in terms of temporal variations of traffic patterns, which are shown to be not related to functional classification, in particular for truck movements (Hallenbeck et al. 1997). In this sense TMG recommends to achieve a 10% level of error with a 95% confidence interval for each road group, excluding recreational road groups which are characterised by much more complex traffic patterns.

Assuming permanent count stations in a given seasonal group being randomly selected, the level of precision of AADT estimate for road group i and a seasonal factor k can be calculated as:

$$d_{ik} = t_{\frac{\alpha}{2},n-1} \frac{CoV_{ik}}{\sqrt{n_{ik}}} \tag{4.5}$$

where:

- d = precision as a percentage to the mean factor for road group i and a seasonal factor k;
- CoV = coefficient of variation of the seasonal factor k in road group i;
- n = required number of ATRs to be used in road group i for seasonal factor k;
- $t_{\frac{\alpha}{2},n-1}$ = value of Student's t statistic at $100(1-\alpha)$ percent confidence interval and $n-1$ degrees of freedom.

From equation 4.5 the number of ATRs needed to estimate AADT for road group i and a seasonal factor k can be calculated by :

$$n = \frac{\frac{t^2_{(1-\alpha/2),n-1} CoV^2}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2_{(1-\alpha/2),n-1} CoV^2}{d^2} - 1\right)} \tag{4.6}$$

where:

- d = precision as a percentage to the mean factor for road group i and a seasonal factor k;
- CoV = coefficient of variation of the seasonal factor k in road group i;
- n = required number of ATRs to be used in road group i for seasonal factor k;
- $t_{\frac{\alpha}{2},n-1}$ = value of Student's t statistic at $100(1-\alpha)$ percent confidence interval and $n-1$ degrees of freedom.
- N number of road sections belonging to road group i

Student t statistic $t_{\frac{\alpha}{2},n-1}$ can be substituted by Z statistic in case of sample dimensions larger than 30 sections.

If counts are routinely taken over a nine-month period, the one month with the most variable monthly adjustment factor (among those nine months) should be used to determine the variability of the adjustment factors and should thus be used to determine the total sample size desired. In that way, factors computed for any other month have higher precision. For most factor groups, at least six continuous counters should be included within each factor group. This is an initial estimation based on AADT factor groups. If it is assumed that some counters will fail each year because of equipment, communications, or other problems, a margin of safety may be achieved by adding additional counters.

4.4 Alternatives to Factor Approach

An alternative to factoring exists. This technique is not commonly used, but it is appropriate where factor groups are not readily known and the annual traffic estimate must be very accurate. Work done showed that for volume counts by vehicle classification, it was possible to achieve accurate annual estimates by taking 4 week-long STCs per year at the same location (Hallenbeck and O'Brien 1994). This approach provides sufficient data to overcome the primary sources of variation in the data collection process:

- Taking week-long counts removes the day-of-week variation;
- Counting at the same location four times at equally spaced intervals removes the majority of seasonal bias.

4.5 Specificities of Truck Vehicles

As already exposed, traditional factor approach can be applied to truck traffic patterns and, in general, to more vehicle classes. However, the characteristics that need to be accounted for can be very different and in practice some points must be considered:

- Functional class of roadway has been shown to have a very inconsistent relationship to truck travel patterns (Hallenbeck et al. 1997). Local truck traffic can be generated by a single facility such as a factory, or by a wider activity such as agriculture or commercial and industrial centers. These 'point' or 'area' truck trip generators create specific seasonal and day-of-week patterns much like recreational activity creates specific passenger car patterns.
- Geographic stratification and functional classification can be used to create truck volume factor groups that capture the temporal patterns and are reasonably easy to apply. Also Clustering analysis can be appropriate to identify the natural patterns of variation and to place the continuous counters in the road groups.
- Independently of the approach adopted for the identification of road groups, the information on variability must be reviewed to determine whether the roads grouped together actually have similar truck travel patterns. In practice no designed group will be optimal for all purposes or apply perfectly to all sites. At the same time, by changing the road groups, it may be possible to classify roads so that all roads have similar travel patterns for one vehicle class, but for another class patterns become highly variable. At some point, the analyst will need to determine the proper balance between the precision of the group

factors developed for these two classes, or they will have to accept the fact that different factor groups are needed for different vehicle classes. Then each road may end up in multiple factor groups depending on what vehicle classification volume is being factored. Use of multiple groups may result in a more accurate factor process but will certainly result in a more complicated and confusing procedure.

- If very precise adjustment factors are desired, it is possible that the factor process will require different factor groups for each vehicle class. In such a case, each class volumes may need to be adjusted using a specific factor process and the volume estimates independently obtained need to be added to produce the total AADT estimate.

Chapter 5

Review of Traffic Monitoring Guide

Given its importance in the U.S., the FHWA factor approach has been analysed and revised by many during the years. In this section the main findings of these efforts are reported, with the aim of identify the state of the art in this specific topic.

5.1 Introduction

The estimation of AADT based on factor approach and the combined use of limited observation of traffic flow and road groups has been practiced for nearly 40 years. Bodle 1967 classified the sources of possible errors of the factor approach into three categories:

1. Error due to the day-to-day variations in traffic volumes;
2. Error in grouping of road segments and the use of wrong adjustment factors;
3. Error in assigning the road segment where SPTC is obtained to the road group.

The fact that traffic volumes fluctuate constantly presents a problem when estimating AADT. This is an issue common among any estimation problem in the transportation field. The FHWA factor approach tries to reduce this effect, describing temporal traffic variations with seasonal adjustment factors for different period of the year, road groups and vehicle classes.

The other sources of possible errors identified by Bodle deal more with the correct implementation of FHWA factor approach in practice. Before discussing them, some details about the different importance of these source of errors is given in the next section.

5.2 Bias and Precision in MDT Estimation

Following the indications given by Davis (1997), Annual Average Daily Traffic (AADT) can be considered an estimate of the Mean Daily Traffic (MDT), which has to be considered as the expected daily traffic volume on a "typical day".

One considers the simple case of determining the value of MDT using a 24h SPTC, denoted by z. The value of z can vary day-to-day both systematically, due to seasonal and day-of-week quite predictable trends, both randomly, due to the unpredictable decisions of drivers.

FHWA factor approach can be represented by a multiplicative model:

$$E[z] = M_i W_j z_0 \tag{5.1}$$

where:

- z = the daily traffic count made during month $i, i = 1, \ldots, 12$ and day-of-week $j, j = 1, \ldots, 7$;
- $E[.]$ = the expected value of the random variable inserted in the square brackets,
- z_0 = Mean Daily Traffic,
- M_i = adjustment factor for month i,
- W_j = adjustment factor for day-of-week j.

In statistic an estimator of a parameter is said to be *unbiased* if the expected value of that estimator equals the parameter's true value; the difference between these two values is called the *bias* of the estimator. The same agreement does not exist speaking of *precision*; one point of view is referring to the variance of an estimator about its expected value.

Equation 5.1 indicates that a single day count will be an unbiased estimator of MDT z_0 only if the product $M_i W_j$ of the monthly and day-of-week factor equals 1. Otherwise, if the adjustment factors M_i and W_j are already known, an unbiased estimator of z_0 can be written as:

$$E\left[\frac{z}{M_i W_j}\right] = z_0 \tag{5.2}$$

In other terms, Equation 5.2 states that the factor approach helps to reduce bias since the group adjustment factors calculated from continuous recorders well represent the sample site's true ones. From this point of view the correct assignment of sample site to the correct group (i.e. the use of correct adjustment factors) is crucial to obtain reliable estimates.

Being z an estimator of z_0, the percentage error of estimation (PE) can be defined as:

$$PE = \frac{\hat{z} - z_0}{z_0} \times 100 \tag{5.3}$$

Since $\hat{z}$ is a random outcome that depends on the sample of traffic counts, some sample will give large values of PE, while others small ones. The measure of the theoretical tendency of an estimator $\hat{z}$ to produce estimates close to z_0 is the root mean square percentage error (RMSPE), defined as:

$$RMSPE = \sqrt[2]{E[PE]^2} \tag{5.4}$$

Letting $\overline{z} = E[\hat{z}]$ denote the expected value of estimator $\hat{z}$, the RMSPE of $\hat{z}$ can be calculated as:

$$RMSPE = \sqrt[2]{E[PE]^2} \tag{5.5}$$
$$= \sqrt[2]{Var[PE] + (E[PE])^2} \tag{5.6}$$
$$= \sqrt[2]{E(PE - E[PE])^2 + (E[PE])^2} \tag{5.7}$$

Since

$$E[PE] = E\left[\frac{\hat{z} - z_0}{z_0} \times 100\right] = \frac{\overline{z} - z_0}{z_0} \times 100 \tag{5.8}$$

therefore

$$RMSPE = \sqrt[2]{E\left(\frac{\hat{z} - z_0}{z_0} - \frac{\overline{z} - z_0}{z_0}\right)^2 + \left(\frac{\overline{z} - z_0}{z_0}\right)^2} \times 100 \tag{5.9}$$
$$= \sqrt[2]{E\left(\frac{\hat{z} - \overline{z}}{z_0}\right)^2 + \left(\frac{\overline{z} - z_0}{z_0}\right)^2} \times 100 \tag{5.10}$$
$$= \sqrt[2]{\frac{E\left(\hat{z} - \overline{z}\right)^2 + \left(\overline{z} - z_0\right)^2}{z_0}} \times 100 \tag{5.11}$$

Finally:

$$RMSPE = \sqrt[2]{\frac{E\left(\hat{z} - \overline{z}\right)^2 + \left(\overline{z} - z_0\right)^2}{z_0}} \times 100 \tag{5.12}$$

The result obtained highlights that the RMSPE can be divided in two distinct terms,each of them contributing to the estimator's mean distance to z_0.:

1. The variance of the estimator $\hat{z}$ (the left-hand term under the radical sign);

2. The square of the bias of $\hat{z}$ (the right-hand term under the radical sign);

If the estimator $\hat{z}$ is unbiased, meaning that $\bar{z} = z_0$, the bias term in equation 5.12 equals zero and RMSPE is equal to the coefficient of variation CoV of $\hat{z}$. It is important to notice that variance and bias have two distinct origins:

- The *variance* of $\hat{z}$ depends on the random day-to-day variability in traffic counts and usually decreases as the sample size increases;
- The *bias* of $\hat{z}$ can be unrelated to sample size, but to an incorrect accounting of seasonal and day-of-week trend variability.

This analysis of *variance* and *bias* is important for the practical consequences in AADT estimation:

- The use of SPTCs is expected to provide an accurate estimate of AADT (within the 15% percent of the true value) only if the seasonal adjustment factors are close to the site's true adjustment factors;
- The assignment of a section monitored with SPTCs to the incorrect road group could led to a tripling of the estimation error (Davis 1996).

Therefore is necessary to consider the importance of grouping road segments and assigning them to the correct road group. In the following sections each aspect is considered in detail and some of the most important improvements proposed in recent years are introduced.

5.3 Grouping of Road Segments

As introduced in section 4.3.2, the TMG suggests three ways to establish road groups based on the data obtained at AVC sites: geographical/functional classification, "same road" application of the factors, and clustering analysis. The choice of the "best" method depends on the availability of data and to the analyst's knowledge of the roadway network.

Since functional classification has been proved to be not effective in many situations (in particular for truck vehicles (Hallenbeck et al. 1997)) and "same road" application of the factors requires a dense network of AVC sites with high costs, the most popular approach is the clustering analysis.

However, the application of clustering analysis could have some drawbacks:

1. The road groups formed could not have at the same time a clear mathematical definition and a linguistic definition in terms of geographical/functional characteristics (FHWA 2001);

2. It is often difficult to establish an "optimal" number of groups, both using functional or cluster-based definition of road groups (FHWA 2001);

3. The clusters could not be the same over the years; that is, the group to which an ATR site belongs can change over time (Faghri, Glaubitz, and Parameswaran 1996; Ritchie 1986).

Different authors have tried to improve the accuracy of grouping step adopting different methods, including Genetic Algorithms (GAs), Artificial Neural Networks (ANNs), Regression Analysis and a large number of Clustering techniques.

Genetic Algorithms (GAs) are machine learning techniques, which reproduce the processes of evolution in nature. The origin of GAs is attributed to Holland's work (*Adaptation in Natural and Artificial Systems*) on cellular automata and there has been significant interest in GAs (Buckles and Petry 1994) in different fields, including job shop scheduling, training neural nets, image feature extraction, and image feature identification.

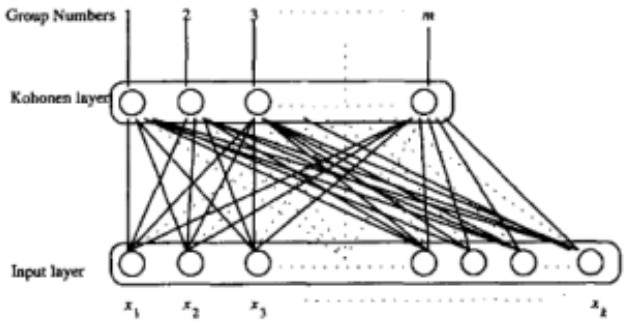
Lingras (2001) applied GAs to the definition of road groups from 264 monthly traffic patterns recorded between 1987 and 1991 on Alberta highways and compared the results with traditional hierarchical grouping approach (Sharma and Werner 1981). Groupings obtained from both the hierarchical and the GA approaches were analyzed for different numbers of groups, considering the values of the within-group errors as a quality measure. The results reported seems to prove that for a smaller number of groups GAs will be able to provide a better classification scheme. Similarly, if a larger number of groups is desired, hierarchical grouping may be a more suitable choice. In particular the hierarchical grouping error was found to be as high as 40% above the GAs error for the three-group classification, but when the number of groups were larger than 14, hierarchical grouping outperformed GAs.

ANNs have been implemented by many authors to define road groups of permanent counters (AVCs).

Lingras (1995) adopted an unsupervised learning model to deal with the traffic pattern classification. In his study he implemented a Kohonen network (see section 2.3.3) to group patterns represented by k-dimensional vectors into m groups (Figure 5.1).

In competitive learning neural networks, the output neurons compete with each other: the winner output neuron has the output of 1, the rest of the output neurons have outputs of 0. This means that a given pattern is classified into exactly one of the mutually exclusive classes (hard clustering). Lingras compared the Kohonen Neural Network approach with the hierarchical grouping approach to classify seasonal traffic patterns from 72 permanent traffic counters in Alberta.

Figure 5.1: *Example of Kohonen Neural Network. Source: Lingras 1995*



The comparative analysis was based on the similarity of grouping obtained with the two approaches. The results highlighted that:

- Road Groups determined were very similar and the Kohonen Neural Network approach could be used to approximate the hierarchical grouping technique;
- The classification was better using hierarchical grouping when a small number of groups is considered, using Kohonen Neural Networks in case of more input factors.
- Kohonen Networks could be successfully applied when further data need to be added or when incomplete patterns (for example due to equipment failures) were available.
- The use of Kohonen Networks should be extended to cope with other types of inputs, such as day-to-day and hour-to-hour variations, having a greater flexibility of use compared to hierarchical approach.

The results proposed were interesting, but no details were given by the authors on the final estimate of AADT.

Faghri and Hua (1995) similarly applied Artificial Neural Networks to determine seasonal factors and AADT estimation, comparing the results obtained with existing approaches. In this case Adaptive Resonance Theory (ART) was proposed by the authors (see section 2.2.3). Data from 29 ATR sites were provided by Delaware Department of Transportation and, based on these data, 12 monthly adjustment factors were calculated for each ATR site. ATR1 results were compared with the one obtained using clustering analysis (Ward's method, the average linkage method, and centroid method) and regression analysis.

The linear model adopted in the case study was:

$$f_{sm} = \alpha_{0m} + \alpha_{1m}X_1 + \alpha_{2m}X_2 + \alpha_{3m}X_3 \tag{5.13}$$

where $X_1 = 1$ if rural and 0 if urban, $X_2 = 1$ if recreational and 0 otherwise, $X_3 = 1$ if recreational and arterial, 0 otherwise.

The comparison was based on the deviation of the estimated seasonal factors sf obtained from clustering analysis *clus*, regression *regr* and ART1 $ART1$, from the actual *act*, calculated for different ATR i and month j:

$$err_{clus}(j) = \sum_i \left[sf_{clus}(i,j) - sf_{act}(i,j)\right]^2 \tag{5.14}$$

$$err_{regr}(j) = \sum_i \left[sf_{regr}(i,j) - sf_{act}(i,j)\right]^2 \tag{5.15}$$

$$err_{ART1}(j) = \sum_i \left[sf_{ART1}(i,j) - sf_{act}(i,j)\right]^2 \tag{5.16}$$

The average error obtained for different methods were calculated by:

$$avgerr_{clus}(j) = \frac{1}{12}\sum_i err_{clus}(j) \tag{5.17}$$

$$avgerr_{regr}(j) = \frac{1}{12}\sum_i err_{regr}(j) \tag{5.18}$$

$$avgerr_{ART1}(j) = \frac{1}{12}\sum_i err_{ART1}(j) \tag{5.19}$$

ATR1 was found to produces substantial improvements compared to the results of clustering and regression analysis. In fact the average error obtained by ATR1 was an 81% improvement on the regression analysis results and an 85% improvement on the cluster analysis. Based on these authors indicated that the neural network approach was significantly better for estimating the seasonal factors and provides more accurate results. However, as observed for Lingras (1995), no details were given by the author on the final estimates of AADT.

5.3.1 Clustering techniques

As written in section 4.3.2 TMG suggests the implementation of Ward's agglomerative hierarchical method, which minimizes the increase in total within-cluster sum of squared error.

Probably this choice was based on the study of Sharma and Werner (1981), who applied the Ward's method to group 45 permanent traffic counters (PTCs) in Alberta, Canada, based on monthly adjustment factors. They

combined the Ward's method with the Scheffe's S-method of multiple comparisons of group means to determine the optimal number of groups. The approach was successfully implemented in other contexts (Sharma and Allipuram 1993; Sharma et al. 1986) and became a reference method for other researchers (e.g. (Faghri and Hua 1995; Lingras 1995)).

However, other clustering techniques have been applied to road grouping of ATR or AVC sites.

Flaherty (1993) used the hierarchical clustering method and the k-means method to analyze the monthly factor data collected over a 5-year period from 28 PTCs in Arizona. The comparative analysis highlighted that functional classification of grouped roads was not relevant for the similarity of monthly factor patterns. Conversely geography and topography were more important to group similar roads.

Schneider and Tsapakis (2009) and Tsapakis et al. (2011) combined geographical and *k*-means method to identify road groups. The Ohio Highway network considered in the analysis was divided in areas due to different geographical characteristics. For each area k-means method was implemented analysing the effect of different vehicle classes (automobiles and trucks) and direction of flow at each site.

Recently some authors have compared the results obtained applying different clustering techniques to the same dataset, with the aim of identifying the "best" method to be applied for road grouping.

Li et al. (2003) compared eight agglomerative clustering methods based on data collected in Florida. Data from 21 ATR were used to calculate monthly adjustment factors to be used as attributes for the clustering analysis. As done by Faghri and Hua (1995), the quality of clusters was tested analysing the average pooled variance of cluster groups and the temporal stability of results. The study found that average linkage, centroid, and single linkage methods were more robust to outliers than the other methods. McQuitty's method performed better than the other methods on grouping ATR sites after outliers were eliminated. However the compositions of seasonal groups were not stable over time and the authors suggested that other variables should be included in the grouping to cope with the change in the spatially clustering.

Zhao, Li, and Chow (2004) extended the previous research applying model-based clustering analysis on monthly adjustment factors from 129 ATR sites located in Florida rural areas. To evaluate the quality of clustering the analysis process included:

1. The implementation of model-based strategy for clustering 2 to 100 groups using monthly adjustment factors as input variables;

2. Perform the same implementation adding the coordinates of ATR sites to the input variable dataset.

Also in this case the quality of results was evaluated analysing the homogeneity of road groups. An ATR site was considered correctly classified if its adjustment factors did not exceed the threshold defined as 10% of the mean adjustment factors obtained for the road group. The results showed that model-based clustering methods, such as the EEV model, could produce classifications with negligible grouping error (2.08%) when adjustment factors data were used in the analysis. However ATR sites belonging to the same road group were over-dispersed in space. By incorporating coordinates of the ATR sites in the clustering it was found that the EEI model was the best one since it produced the least grouping error.

Finally Gecchele et al. (2011) compared the implementation of hierarchical, partitioning and model-based clustering techniques for the creation of road groups based on adjustment factors. 54 AVC sites located on the rural road network of the Province of Venice in 2005 were analysed, considering direction traffic variability and a 2-class scheme which differentiates between passengers vehicles (PV) and truck vehicles (TV) with reference to a 5 m-length threshold. Seasonal adjustment factors were calculated for both classes and used together as input of clustering techniques.

Differently from previous studies the quality of road groups obtained with the different methods was determined analysing the accuracy of AADT estimates. Following a procedure adopted by other authors (Sharma et al. 1999, 2000), 24hr sample counts were generated from the dataset and, for each classification made by clustering methods, were used as follows:

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

For each AADT estimate $AADT_{i,Estimate}$ the absolute percent error was calculated by:

$$\Delta = \left| \frac{AADT_{Estimate} - AADT_{Actual}}{AADT_{Actual}} \right| \times 100 \tag{5.20}$$

The analysis of the errors obtained with the various methods was developed considering the Mean of the Absolute percent error (MAE) for

different tests: by vehicle type (passenger and truck vehicles), by different day-types and by different periods of the year. MAE was calculated by:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{AADT_{i,Estimate} - AADT_{i,Actual}}{AADT_{i,Actual}}\right| \times 100\right) \tag{5.21}$$

The results were that:

- Clustering methods identified a common basic structure of AVC groups with functional and geographical significance. Model-based clustering methods showed slightly better performances compared to the other methods and had a robust mathematical structure;
- Error patterns showed worse results in AADT estimation in some specific cases:
  - Analyzing day-type, Saturdays and Sundays had higher errors than Weekdays;
  - Analyzing the period of the year, summer period showed higher errors than winter period;
  - Analyzing the vehicle type, the errors were higher for truck vehicles than passenger vehicles.

Since a large difference was observed between AADT estimation errors for passenger and truck vehicles, the authors developed a similar analysis creating separated road groups for passenger and truck vehicles, with the aim of better reproducing the temporal traffic patterns of both vehicle classes (Rossi, Gastaldi, and Gecchele 2011).

In this case AADT estimates for passenger and truck vehicles were calculated in a separate manner using 48-hr sample counts generated from the main dataset. Mean Absolute Errors (MAEs) were calculated for passenger and truck vehicles, considering the day of the week (weekdays or weekends) and the period of the year in which the short counts were taken. The results were analysed using the Friedman Test (a non-parametric alternative to the one-way ANOVA with repeated measures), and the most relevant findings were that:

- MAE patterns were similar to the results obtained by previous studies, considering when SPTCs are taken and vehicle classes;
- The use of different seasonal factors for each vehicle class affected the number and the characteristics of the road groups identified. The use of road groups which consider the specificity of traffic patterns for passenger and truck vehicles has a positive effect on the accuracy of AADT estimates;

- Clustering methods show common error patterns and give comparable results in term of accuracy of AADT estimates.

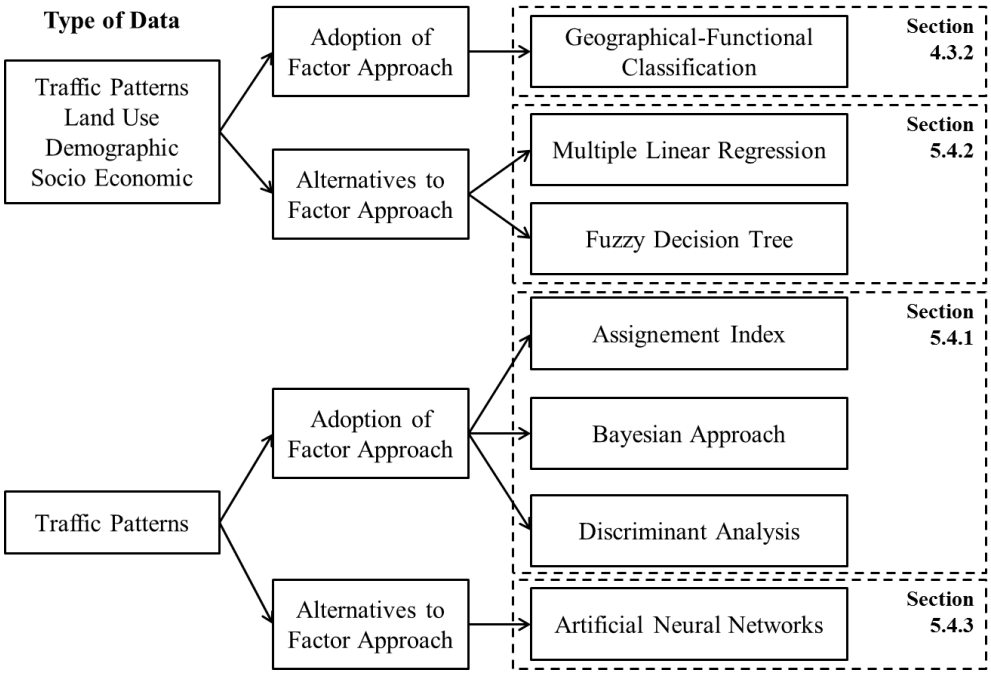
5.4 Assignment of Road Segments

As introduced in section 5.2 the assignment of a road segment monitored with SPTCs to the correct road group represents the most critical aspect of FHWA factor approach. In particular it was found that two-week counts taken in different months represent the minimum seasonal count to guarantee the correct assignment [(Davis and Guan 1996),(Davis 1997)].

To minimize the risk of large errors, the TMG suggests the use of weekly counts repeated in different periods of the year (STCs) in order to capture the seasonal variability of the monitored road sections. However there are no further specifications about the assignment process.

This lack of information has lead to the growth of a large number of different methods, in some cases alternative to the factor approach, based on different types of data. Following the scheme of Figure 5.2 they have been reviewed and the major finding are reported in the next sections.

Figure 5.2: *Scheme of Assignment Strategies*



5.4.1 Assignment Using Traffic Patterns

The use of traffic patterns of short counts can be considered the most important information for the assignment of a road section to road groups. Different methods have been proposed to accomplish this task, given a certain number of SPTCs or STCs at the specific road section.

One of the first proposal was made by Sharma and Allipuram (1993), who developed an Index of Assignment Effectiveness to correctly assign STCs to the road groups. In their paper they analysed the use of Seasonal Traffic Counts (STCs) and developed a method for the assignment of STC to road groups, rationalizing the length and frequency of STCs for various types and road facilities. Finally they suggested STC schedules for traffic monitoring. Their dataset included 61 PTC Sites in Alberta (Canada) monitored for year 1989. Based on monthly adjustment factors seven road groups were identified using the approach of Sharma and Werner (1981).

The study considered the appropriateness of a schedule in terms of its ability to correctly identify the pattern of seasonal variation at the location of the traffic count. Different STC schedules $S(L, F)$ were defined, being L the length of a count in weeks and F the frequency of counts in a year. The data from permanent counters allowed the generation of sample STC data: 12 STC samples were built for each schedule $S(L, F)$. The method of assignment proposed had two parts, the first concerning the PTC and the second one the STC.

The first part of the assignment considers *PTC* and can be subdivided in different steps:

- Choice of a Schedule $S(L, F)$;
- Selection of a sample PTC and generation of sample STCs according to the schedule chosen [12 samples];
- Computation of an array of potential errors associated to the assignment of STCs to each road group:

$$MSES_i = \frac{1}{12} \sum_{j=1}^{12} (f_{sj} - f_{ij})^2 \tag{5.22}$$

  where $MSES_i$ is the Mean Squared Error associated with the assignment of STC to i-th road group, f_{sj} is the monthly traffic factor of the STC for month j and f_{ij} is the average monthly factor of road group i for month j.

- Scaling of performances, giving a 100% effectiveness for the correct assignment (to the group the sample PTC belonged to) and 0% for the

worst one. A linear scale was used for intermediate values:

$$AE_i = \frac{\max MSE - MSE_i}{\max MSE - \min MSE} \times 100 \tag{5.23}$$

where AE_i is the effectiveness of the assignment to road group i.

Each PTC was chosen each at a time as a sample PTC, STCs were genrated and the Assignment Effectiveness to the road groups were calculated.

However when *STCs* are taken, the actual values of monthly factors f_{sj} are not available, due to the limited amount of data. Therefore a specific analysis must be conducted:

- Computation of the Sample Average Daily Traffic (SADT) at the sample site combining sample values monitored:

$$SADT = \frac{\text{Total Volume counted during schedule } S(L,F)}{\text{Total Number of days counted during schedule } S(L,F)} \tag{5.24}$$

- Computation of the Average Daily Traffic ADT_k and the seasonal traffic factors f_{sk} for the k-th count of the schedule $S(L,F)$, using the relationships:

$$ADT_k = \frac{\text{Total Volume during the } k\text{-th visit}}{\text{Number of days counted in the visit}} \tag{5.25}$$

$$f_{sk} = \frac{ADT_k}{SADT} \tag{5.26}$$

- Use of the average monthly factors of road groups f_{ik} to describe the seasonal traffic variation in the months in witch STC counts were undertaken, where i represents a road group and k a season (month).

- Comparison of the sample f_{sk} values with the f_{ik} values for each road group, using as error relationship:

$$MSES_i = \frac{1}{F} \sum_{(j=1)}^{F} (f_{sk} - f_{ik})^2 \tag{5.27}$$

where $MSES_i$ is the Mean Squared Error associated with the assignment of SPTC to i-th road group and F is the frequency or number of counts in the schedule $S(L,F)$. The value of correspondent AE can be found considering the STC belonging to the road group with the minimum error of assignment.

- Computation of the Index of Assignment Effectiveness (IAE), taking off repeated samples of a giving schedule $S(L, F)$ and moving from step 1 to 4. The resulting AE values can be used to compute the required IAE by:

$$IAE = \frac{1}{N} \sum_{(i=1)}^{I} n_i(AE_i) \tag{5.28}$$

where n_i is the number of times the sample site is assigned to road group i, AE_i is the AE value for road group i, I is the total number of road group and N is the total number of samples of a given $S(L, F)$ taken at the sample site.

The Sample Average Daily Traffic (SADT) was used also as a measure of schedule evaluation. In fact, the closer is SADT to AADT actual value, the more the chosen schedule is capable of describing temporal variations of traffic patterns. The absolute value of percent difference (PDA) between SADT and the actual AADT was used to evaluate the quality of schedules:

$$PDA = \left| \frac{SADT - AADT}{AADT} \right| \times 100 \tag{5.29}$$

The authors produced different tables with IAE and PDA values obtained for different schedule combinations. The results highlighted that an increment in frequency or duration of STCs increased IAE and reduced PDA. These trends were different among the road groups, as a consequence of specific characteristics of traffic patterns observed in the road groups.

Given these findings, the choice made by a highway authority would be the result of the trade-off between the accuracy of results and the cost of alternative schedules. The authors suggested to adopt values of 95% or higher for the IAE and 2% or lower for the PDA. Using these values, which ensure a fairly accurate assignment, the authors determined the minimum requirements for schedules $S(L, F)$ for different road groups.

Table 5.1: *Suggested Frequency of Counts of Different Durations. Source: Sharma and Allipuram 1993*

Type of route	$L = 1$	$L = 2$	$L = 3$	$L = 4$
Commuter	4	3	3	3
Average rural	6	4	3	3
Recreational	-	6	6	5

For rural highways in Alberta, the AADT volume error margins or precision at a 95% confidence level as found in that study were:

1. ±8.00 for two noncontinuous month counts during a year;

2. ±5.60% for three noncontinuous month counts during a year;

3. ±3.61% for four noncontinuous month counts during a year.

Sharma, Gulati, and Rizak (1996) extended the findings of previous paper considering the importance of accuracy of AADT estimates, based on data from 63 ATR sites located on Minnesota rural roads. Since in Minnesota 48hr SPTCs were adopted as standard sample counts, starting at 12.00 noon on Monday, Tuesday and Wednesday from May to September. Authors defined road groups (ATRG) using Sharma and Werner 1981's approach, based on 15 *AF* factors, which accounted for the starting day and the month of SPTCs (3 days by 5 months), that is:

$$AF(n,d,m) = \frac{AADT}{(\text{Average } n-h \text{ volume}_{d,m)}} \times \frac{n}{24} \tag{5.30}$$

where $AF(n,d,m)$ is the adjustment factor (AF) for n (e.g., 48) hours of counting starting at noon on day d (e.g., Monday) in the month m (e.g., July) and $AADT$ is the average annual daily traffic at the ATR site.

The authors analysed the effects of different SPTC's duration (n = 24, 48, 72 hours) extracting from each ATR site 6 counts/month for each duration. Totally $30 \times 63 = 1890$ SPTCs were generated for each duration. The Estimate of AADT ($EAADT$) was calculated by:

$$EAADT = \frac{24 \times SV(n,d,m)}{n} \times AF(n,d,m) \tag{5.31}$$

where $SV(n,d,m)$ was the sample volume.

The estimation error E for a sample was calculated by:

$$E = \frac{EAADT - AADT}{AADT} \times 100 \tag{5.32}$$

The errors were analysed for various ATRGs, checking for the normality of E. In the case study histograms of error distribution developed for various ATRGs and all the ATRGs combined together showed that the variable E followed a normal distribution and that the mean values of estimation errors for all groups were statistically equal to zero at a confidence level of 95%.

Assuming equality of variances for errors in the sample sites, the following relationship could be written;

$$S_e^2 = \frac{1}{N-1} \sum_{i=1}^{n_i} \sum_{k=1}^{n_s} (E_{ik} - \bar{E}_{ik})^2 \tag{5.33}$$

where:

- S_e^2 is the standard deviation of errors resulting from SPTCs from an ATRG;
- E_{ik} is the error at the i-th sample site for the k-th sample;
- $\bar{E}_{ik}$ is the mean value of E_{ik};
- n_i is the total number of sample sites included in the group;
- n_s is the number of samples taken at the sample site i;
- N is the total number of samples resulting from the multiplication of n_i and n_s.

The normal distribution of errors with mean 0 and standard deviation equal to S_e allowed computation of the confidence interval:

$$\pm Z_{(\frac{\alpha}{2})} \times S_e \tag{5.34}$$

where α is the confidence coefficient and $Z_{(\frac{\alpha}{2})}$ is the standard normal statistic corresponding to α. Using a confidence level equal to 95%, the absolute value of the 'lower bound' and the 'upper bound' were denoted as:

$$PB95 = |\pm Z_{0.025}| \times S_e = 1.96 S_e \tag{5.35}$$

The authors first investigated the AADT estimation errors in case of correct assignment of SPTCs to the ATR group to which they actually belong (Figures 5.2 and 5.3).

Table 5.2: *Effect of Road Type on AADT Estimation Errors from 48hr SPTCS. Source: Sharma, Gulati, and Rizak 1996*

ATR Group	Road Class	Number of Samples	Mean Error [%]	Standard deviation Se [%]	Error limit PB95 [%]
All ATRs	All rural highways	1890	-0.21	7.39	14.48
ATRG 1	Regional commuter 1	660	-0.17	5.39	10.56
ATRG 2	Regional commuter 2	510	0.07	5.78	11.33
ATRG 3	Average rural	300	-0.77	7.41	14.52
ATRG 4	Regional/recreational	270	-0.16	12.19	23.89
ATRG 5	Tourist/recreational	150	-0.74	8.20	16.07

However improper assignment of sample sites to ATR groups resulted in the use of incorrect adjustment factors and possibly large estimation errors. The standard deviation (Se) of the errors resulting from the 24hr, 48hr, and 72hr SPTCs were calculated for different AE categories: $> 50\%, 51-55\%, 56-60\%, ..., 91-95\%, > 95\%$. Figure 5.3 shows curves plotted between the Se values for the counts of various durations and the AE categories.

One can observe that:

Table 5.3: *Effect of Count Duration on AADT Estimation Errors. Source: Sharma, Gulati, and Rizak 1996*

ATR Group	Error limit PB95 for 24hr counts [%]	Error limit PB95 for 48hr counts [%]	Error limit PB95 for 72hr counts [%]
All ATRs	16.50	14.48	13.13
ATRG 1	12.89	10.76	10.76
ATRG 2	15.23	11.66	11.66
ATRG 3	17.01	14.62	14.62
ATRG 4	27.15	23.89	19.36
ATRG 5	16.50	16.07	14.56

- the relationships between the estimation errors *Se* and assignment effectiveness AE values indicate that the degree of correctness of sample site assignment to an ATR group has a large influence on the AADT estimation errors. Small decreases in assignment effectiveness can result in large increases in the magnitude of AADT errors;
- the AADT estimation errors are more sensitive to the correctness of sample site assignment to a proper ATR group than to the duration of counts investigated. A 24hr SPTC at an appropriately assigned site would result in better AADT estimates than a 72hr SPTC at the same site when it has been incorrectly assigned.

Davis and Guan (1996) changed completely the approach to the problem, employing the Bayesian theorem to assign a given road section to the road group with the higher posterior probability, defined by:

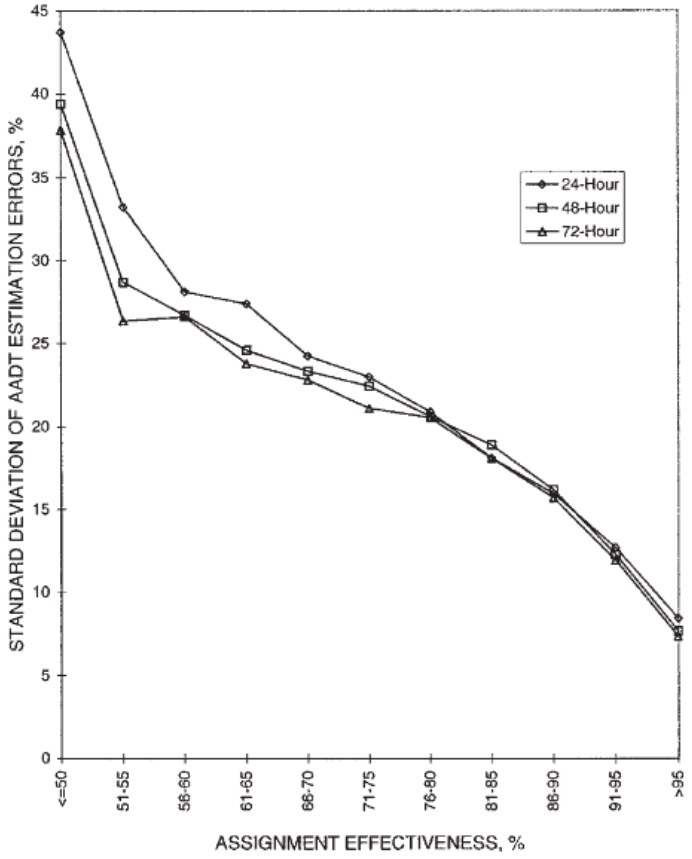
$$Prob[site \in G_k \mid z_1, \ldots, z_N] = \frac{f(z_1, \ldots, z_N \mid G_k)\alpha_k}{\sum_{i=1}^{n} f(z_1, \ldots, z_N \mid G_i)\alpha_i} \tag{5.36}$$

where:

- $f(z_1, \ldots, z_N \mid G_k)$ = a likelihood function measuring the probability of obtaining the count sample had the site actually belonged to a given road group G_k;
- $z_1, \ldots, z_N$ = a sequence of N daily traffic counts at a SPTC site;
- $G_1, \ldots, G_n$ = a total of n different road groups;
- α_k = probability that the given site belong to G_k (prior classification probability).

The prior classification probability was set equal to $\frac{1}{n}$, to represent the complete uncertainty regarding which road group a SPTC belonged. As a likelihood function in the posterior classification probability was used the linear regression model:

Figure 5.3: *Effect of AE on AADT Estimation Errors. Source: Sharma, Gulati, and Rizak 1996*



$$y_t = \mu + \sum_{i=1}^{12} \Delta_{t,i} m_{k,i} + \sum_{j=1}^{7} \delta_{t,j} w_{k,j} + \epsilon_t \tag{5.37}$$

where:

- y_t = natural logarithm of the SPTC z_t taken of day t;
- μ = expected log traffic count on a typical day;
- $\Delta_{t,i}$ = 1 if the count z_t was made during month $i, i = 1, \ldots, 12$ and 0 otherwise;
- $m_{k,i}$ = correction term for month i, characteristic of road group k;

- $\delta_{t,j} = 1$ if the count z_t was made during day $j, j = 1, \ldots, 7$ and 0 otherwise;

- $w_{k,j}$ = correction term for day-of-week j, characteristic of road group k;

- ϵ_t is the random error.

It was further assumed that the random errors $\epsilon_1, \ldots, \epsilon_N$ were randomly distributed with mean value 0 and covariance matrix $\sigma^2\mathbf{V}$, where σ^2 was the common unconditional variance of y_t and $\mathbf{V}$ is a $N \times N$ matrix of correlation coefficients such that the element $V_{s,t}$ in row s and column t was the correlation coefficient for y_s and y_t.

The model, validated using data from ATR sites in Minnesota, was shown to be able to produce mean daily traffic estimates (AADT estimates) that were near ±20% of actual values based on 14 well selected sampling days of the year. The main advantage introduced by this method was not an improvement of precision, but the reduction of the importance of subjective judgements in the attribution process of SPTC sites.

Recently the use of Linear Discriminant Analysis (Schneider and Tsapakis 2009; Tsapakis et al. 2011) was recently tested to determine the group assignment of SPTCs (24hr) with good results. In their latest paper Tsapakis et al. (2011) applied Linear Discriminant Analysis (LDA) to assign SPTCs to road groups using in their analysis continuous traffic volume data from 51 ATR sites, obtained in the state of Ohio during 2005 and 2006. The ATRs were grouped together by using a combination of geographical classification and cluster analysis. The *k*-means algorithm was used to further subdivide the geographical factor groups by using the 12 monthly factors of each site as input in cluster analysis. Seasonal adjustment factors were developed considering the total traffic volume at the site as well as the individual volume in each direction of travel. The data set was used both as a training set to estimate seasonal adjustment factors and create factor groupings and as a test set from which sample SPTCs are generated. Specifically, every daily count produced from each ATR was used as a sample 24hr count that is assigned to one of the previously defined factor groupings.

Two traffic parameters were selected to develop Discriminant Analysis (DA) assignment models:

- The set of 24 time-of-day factors estimated for weekdays:

$$F_{hh} = \frac{ADT}{HV_{hh}} \quad HV_{hh} \neq 0 \tag{5.38}$$

  where HV_{hh} is the hourly volume that corresponds to hh-th hour of a day and $hh = 1, 2, \ldots, 24$;

- the ADT, which represent the total traffic volume per day on a roadway section:

The methodology of the study relied on the assumption that the two variables may be used interchangeably or in combination to identify sites with similar traffic behavior, since the two variables capture both the variability and the magnitude of the daily traffic at a specific location. In fact, two stations may exhibit similar traffic patterns but carry significantly different traffic volumes, and viceversa.

Given the set of independent variables, LDA attempted to find linear combinations of those variables that best separate the groups of cases. These combinations are called discriminant functions and have the form displayed in the equation.

$$d_{ik} = b_{0k} + b_{1k}x_{i1} + \ldots + b_{pk}x_{ip} \tag{5.39}$$

where:

- d_{ik} is the value of the kth discriminant function for the ith case (standardized score);
- p is the number of predictors;
- b_{jk} is the value of the jth coefficient of the kth function;
- x_{ij} is the value of the ith case of the jth predictor.

The estimation of the discriminant functions is made using the means of the p predictors and the pooled within-group variance-covariance matrix $\mathbf{W}$, which is generated by dividing each element of the cross-products matrix by the within group degree of freedom:

$$\mathbf{D}_c = \mathbf{W}^{-1} \times \mathbf{M}_c \tag{5.40}$$

$$d_{c0} = -0.5 \times \mathbf{D}_c \times \mathbf{M}_c \tag{5.41}$$

where:

- $\mathbf{D}_c = (d_{c1}, \ldots, d_{cp})$;
- $\mathbf{W}^{-1}$ is the inverse of the within-group variance-covariance matrix;
- $\mathbf{M}_c = (X_{c1}, \ldots, X_{cp})$, means for function c on the p variables;
- d_{c0} is the constant for function c.

Two full discriminant models composed the basis for the model development. The first model, Equation 5.42, includes both the ADT and the time-of-day factors, and the second full model, Equation 5.43, considered the hourly factors only. The general form of the two full models was:

$$D_c^1 = d_{c,0} + d_{c,1}ADT + d_{c,2}F_1 + d_{c,3}F_2 + \ldots + d_{c,25}F_{24} \tag{5.42}$$

$$D_c^2 = d_{c,0} + d_{c,1}F_1 + d_{c,2}F_2 + \ldots + d_{c,24}F_{24} \tag{5.43}$$

where:

- D_c = standardized score of discriminant function c;
- F_{hh} = hourly factor that corresponds to *hh*-th hour of the day, and
- d_c = discriminant function coefficient.

12 assignment models were specified, using different variable-selection methods and algorithms, and tested on a large dataset, which consisted of 35,100 test SPTCs for each DA model. The statistical evaluation of the traditional method (based on functional classification of road sections) and the 12 DA models are based on three statistical measurements:

- the absolute error of the AADT estimate,

$$AE_{v,dd} = \frac{\mid AADT_{v,Actual} - AADT_{v,dd,Estimated} \mid}{AADT_{v,Actual}} \times 100 \tag{5.44}$$

- the mean absolute error (MAE);

$$MAE = \frac{1}{w}\sum_{s=1}^{w}\left(\frac{\mid AADT_{v,Actual} - AADT_{v,dd,Estimated} \mid}{AADT_{v,Actual}} \times 100\right) \tag{5.45}$$

- the standard deviation of the absolute error (SDAE):

$$SDAE = \sqrt{\frac{\sum_{s=1}^{w}\left(AE_{v,dd} - MAE\right)^2}{w-1}} \tag{5.46}$$

where:

- AE = absolute error;
- w = number of test SPTCs;
- s = SPTC $(1, \ldots, w)$;

- v = ATR index;
- dd = day of year;
- $AADT_{v,Actual}$ = actual AADT;
- $AADT_{v,dd,Estimated}$ = estimated AADT.

The analysis of the results, conducted using the ANOVA test, revealed that:

- the best-performing directional volume-based model, which employs the Rao's V algorithm, produced a mean absolute error (MAE) of 4.2%, which could be compared with errors reported in previous studies;
- an average decline in the MAE by 58% and in the SDAE by 70% is estimated over the traditional roadway functional classification;
- when directional-specific factors are used instead of total volume-based seasonal adjustment factors, the improvement in the average MAE is approximately 41% and 39% in the average SDAE.

5.4.2 Multiple Linear Regression

The use of Multiple Linear Regression to estimate AADT has been intensely investigated in recent years. Interesting results have been obtained in terms of the identification of factors (regressors) which affect the AADT values on a certain road section, however the results in terms of AADT estimate accuracy appears still limited.

The Multiple Linear Regression considers the model:

$$AADT_i = \beta_0 + \beta_1 X_{i1} + + \beta_j X_{ij} + \epsilon_i \tag{5.47}$$

where:

- $AADT_i$ = the value of the dependent variable for the i-th observation, $i = 1, \ldots, n$.
- X_{ij} = the value of the j-th independent variable in the observation, $j = 1, \ldots, m$
- β_0 = constant term;
- β_j = regression coefficient for the j-th independent variable;
- ϵ_i = error term;
- n = observation number;

- m = number of independent variables.

Mohamad et al. (1998) estimated AADT on low-volume county roads with a multiple linear regression model which incorporated quantitative and qualitative predictor variables, including relevant demographic variables. Field traffic data collected from 40 counties in Indiana were used for the calibration of the model and additional field traffic data collected from 8 counties to its validation.

The final model adopted had four predictor variables:

$$\log_{10} AADT = 4.82 + 0.82X_1 + 0.84X_2 + 0.24 \log(X_4) - 0.46 \log(X_{10}) \quad (5.48)$$

where:

- X_1 = type of road, 1 if urban and 0 if rural;
- X_2 = type of access, 1 easy access or close to state highway and 0 otherwise;
- X_4 = county population;
- X_10 = total arterial mileage of a county.

The model was considered acceptable, since had $R^2 = 0.77$ and $MSE = 0.1606$, and the predictors resulted significant (form t-statistic values) with signs as could be expected. The final model was applied on the validation set, composed by data from additional Indiana counties. The percentage difference between the observed and predicted values of the AADT ranged from 1.56% to 34.18%, with the average difference of 16.78%.

Xia et al. (1999) applied the Multiple linear Regression model to estimate AADT for Nonstate Roads in Broward County in Florida, an urban area with 1 million people living. The dataset adopted in the analysis was particularly large, with 450 count stations randomly divided in calibration (90% = 399) and validation (10% = 44) subsamples. The model was developed considering a large number of response variables, determined from the elaboration of roadway, socio-economic and accessibility data. In particular GIS tools were used to identify the influence of socio-economic and accessibility characteristics on the monitoring stations. Buffers with different values of radius were considered to define land-use characteristics around a counting station.

The final model was defined through a detailed variable selection process, which excluded the presence of outliers and avoided phenomena of multicollinearity among predictors. Variable selected were: the number of lanes on a roadway, the functional classification of roadway, the land use type; the automobile ownership, the accessibility to nonstate roads, the service employment.

The model was found acceptable (R^2 = 0.6061), with significant values of t-statics for the variables and signs coherent with the expectations, excepted the variable "Service employment". Also the validation of the model, applied on 40 data points, leads to acceptable results. The percent difference between observed and predicted AADT values ranged from 1.31% to 57%, with an average difference of 22.7%. The model also underestimated the AADT for about 5% of the entire test data set. Moreover, from the analysis of error distribution and cumulative percent of testing points, 50% of the test points had an error smaller than 20%, whereas 85% of the test points had an error smaller than 40%.

In a later paper, Zhao and Chung 2001 applied the same approach on a larger dataset that included all the AADTs for state roads, a new road function classification system, and a more extensive analysis of land-use and accessibility variables.

Four different models were calibrated: coefficients of determination (R^2) were acceptable and the signs of the coefficients were as expected. 82 different data points (9.1% ot the total dataset) were used to examine the models' predictive capability, comparing the AADTs for the roads estimated by the models to the actual AADTs observed. The testing results included the mean square of errors (MSEs) and the total error in percentage for the entire testing data set, which is less than 3% for all models. More interesting was the analysis of the cumulative percent errors and the analysis of the maximum error for each model. Comparison of the models shows that Model 1 had the best performance in terms of prediction errors; about 37% of the testing points had an error smaller than 10% and about 73% of the testing points had an error smaller than 30%, respectively.

Considering percentage error, the authors stated that the models in their current forms may not be adequate to meet the need of engineering design or the calibration of travel demand models. They may be used for tasks that do not require a high level of accuracy at individual sites such as estimating systemwide vehicle miles traveled.

In order to better estimate AADT, Zhao and Park (2004) applied the Geographically Weighted Regression (GWR) methods, which allow model parameters to be estimated locally instead of globally, as in the case of ordinary linear regression (OLR). The study, which used the same data and model variables of Zhao and Chung (2001), investigated the spatial variation of the parameter estimates, the local R^2 from the GWR model and the AADT estimation.

The basic idea of GWR is to allow different relationships between the dependent and independent variables at different points in space: the locally linear regression parameters at a point are affected more by observations near to that point than observations further away. The mathematical formalization of the model is similar to OLS models:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \epsilon i \tag{5.49}$$

where:

- y_i = dependent variable at location i ($i = 1, \ldots, n$) where n is the number of observations;
- x_{ik} = independent variable of the k-th parameter at location i;
- β_{ik} = estimated k-th parameter at location i;
- ϵi = error term at location i;
- p = number of parameters.

The model parameters β_{ik}, $(k = 0, \ldots, p)$ are estimated for each observation of y_i and x_{ik}, $(k = 0, \ldots, p)$. Therefore a total of $n \times (p + 1)$ parameters are estimated for n observations, whereas the number of parameters estimated for OLR model is $p + 1$. Error terms ϵi are assumed to be independent and identically distributed (i.i.d.) with zero means and constant variance σ^2. Parameters at a location i are estimated using the weighted least-squares approach, which uses weights to allow the influence of observations from different locations to vary according to their distance to location i. Weights can be determined using different weighting functions. In the study two popular weighting functions were tested: the bi-square function and the Gaussian function.

The results produced with the validation dataset (82 count stations) showed that both GWR models tested outperformed the OLR model in the accuracy of AADT estimates (Figure 5.4). In particular the 63.41% of the testing points had an error of less than 20% and 89% had an error smaller than 32% for GWR_{Bi} model. More detailed analysis of AADT estimates at specific sites were not provided, however the results appeared still not satisfactory, compared to factor approach.

More recently Wang and Kockelman (2009) proposed the use of Kriging-based methods for mining network and count data, over time and space. Using Texas highway count data (AADT estimates for 27738 sites from 1999 to 2005), the method forecasted future AADT values at locations where no traffic detectors were present.

The method of Kriging, first developed by Matheron (1963), relies on the notion that unobserved factors are autocorrelated over space, and the levels of autocorrelation decline with distance. The values to be predicted may depend on several observable causal factors (e.g., number of lanes, posted speed limit, and facility type) which create a "trend" estimate $\mu(s)$. In general, spatial variables can be defined as follows:

Figure 5.4: *Cumulative Percentage of Testing Points. Source: Zhao and Park 2004*

Error (%) ≤	Cumulative Percentage of Testing Points		
	OLR	GWR_Bi	GWR_Gau
10	34.15	34.15	30.49
20	51.22	63.41	54.88
30	68.29	89.02	78.05
40	76.83	95.12	87.80
50	84.15	96.34	95.12
60	89.02	97.56	97.56
70	93.90	97.56	97.56
80	95.12	97.56	97.56
90	95.12	97.56	97.56
100	96.34	98.78	98.78
200	98.78	100.00	100.00
300	100.00	100.00	100.00
Max. Error (%)	284.29	136.23	141.52
# with Error > 100%	3	1	1
Total Error[a] (%)	0.0004	-0.1904	0.0430

$$Z_i(s) = \mu_i(s) + \epsilon_i(s) \tag{5.50}$$

where:

- $Z_i(s)$ is the variable of interest (in the specific case the actual AADT);
- s is the location of site i, determined by coordinates (x, y);
- $\mu_i(s)$ is the deterministic trend;
- $\epsilon_i(s)$ is the random error component.

Based on the characteristics of $Z_i(s)$ three types of Kriging exist:

- *Ordinary Kriging*, if $\mu(s)$ is constant across locations (or explanatory information is lacking);
- *Simple Kriging*, if the trend is known but varies across locations;
- *Universal Kriging*, if trends depend on explanatory variables and unknown regression coefficients.

For all types of Kriging weak stationarity is assumed, therefore the correlation between $Z(s)$ and $Z(s+h)$ does not depend on the actual locations, but only on the distance h between the two sites, and the variance of $Z(s + h) - Z(s) = 2\gamma(h)$ for any s and h. This fact can be expressed by the formula:

$$\gamma(h) = \frac{1}{2} var[Z(s + h) - Z(s)] \tag{5.51}$$

where $var[Z(s + h) - Z(s)]$ is the variance (over all sites) between counts taken at sites s and $s + h$.

In Universal Kriging $\mu_(s)$ can be described using any deterministic function, such as the linear function $\mu_(s) = X\beta$, where X contains explanatory variables (e.g. number of lanes and facility type). In contrast, $\epsilon_i(s)$ reflects unobserved variation (e.g. local land use patterns and transit service levels).

To estimate the random component $\epsilon_i(s)$, first an appropriate "curve" or semivariogram model to best fit the relationship $\gamma(h)$ for a given dataset is chose. There are several commonly used models, including exponential, spherical and Gaussian models. These models all rely on three parameters that describe their shape while quantifying the level of spatial autocorrelation in the data:

- c_0, the "nugget effect", which reflects discontinuity at the variogram's origin, as caused by factors such as sampling error and short scale variability.
- a is the "range", a scale factor which determines the threshold distance at which $\gamma(h)$ stabilizes;
- $c_0 + c_1$ is the maximum $\gamma(h)$ value, called the "sill", with c_1 referred to as the "partial sill".

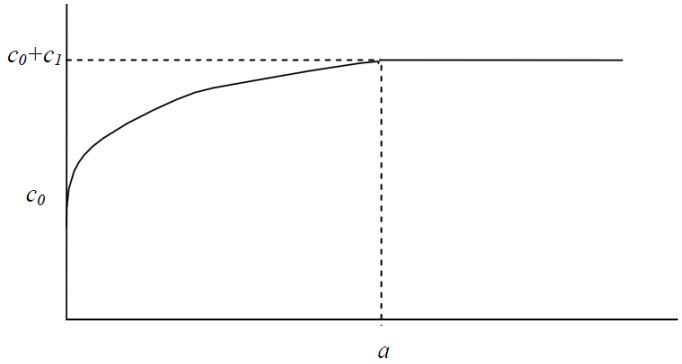
Figure 5.5 illustrates these parameters.

While using Ordinary Kriging it is simple to estimate the three shape parameters, in Universal Kriging, the vector of parameters β needs to be estimated simultaneously or iteratively (in synch with c_0, c_1 and a). This can be done using a series of feasible general least square (FGLS) regressions or to use restricted maximum likelihood estimation (REML) by assuming that errors follow a normal distribution.

The real challenge is the calculation of distance: currently, all available packages use Euclidean distances, which can be easily derived based on locations of the sites. The computational burden can increase dramatically if non-Euclidean distances are used and sample size is large.

The authors applied Kriging methods on their large dataset, spatially interpolating traffic counts on segments of the same functional class (interstate highways and principal arterials). To ensure computational tractability they rely on Euclidean distances and exponential semi-variogram specification thanks to its better fit.

Figure 5.5: *Illustration of Semivariogram. Source: Wang and Kockelman 2009*

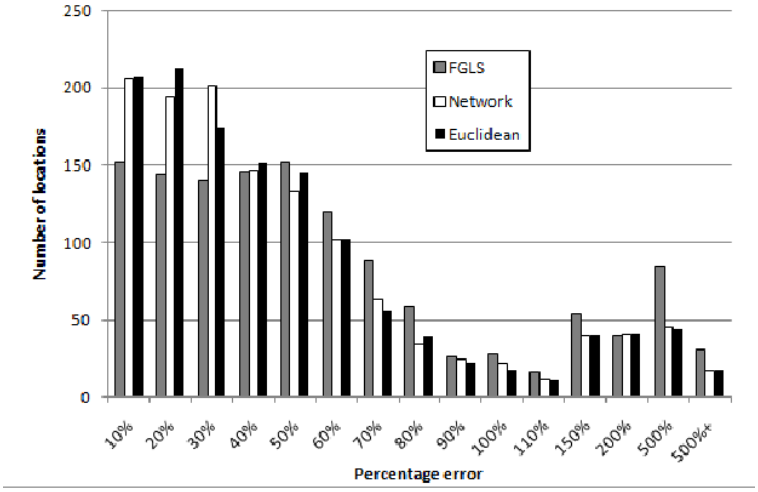


The error ratio obtained exhibited no clear spatial trends, however spatial interpolation using ordinary Kriging methods suggested that traffic volumes on different classes of roadways had rather different patterns of spatial autocorrelation.

Furthermore model validation suggested that, given the limited number of influential factors considered, the spatial interpolation method yielded fairly reliable estimation results, especially for road sections with moderate to high traffic volumes (overall AADT-weighted median prediction error was 31% across all Texas network sites). Compared to a previous studies, where AADT were estimated by assigning the AADT from its nearest sampling site (the standard technique for no-information AADT estimation), the authors concluded that the Kriging method is capable of yielding more reliable estimations than simply relying on the count of a location's nearest count location.

In a more recent paper Selby and Kockelman (2011) implemented Universal Kriging to the Texas network, adding further site attributes, including functional classification, lane numbers and speed limits to predict AADT in count sites. Compared to previous paper the authors considered more detailed data treatment options (i.e. AADT data transformation, network distances and Euclidean distances) and estimation methodologies. Universal Kriging was found to reduce errors (in practically and statistically significant ways) over non-spatial regression techniques (between 16% and 79% in the case study, depending on the data set and model specification used). Dividing up the roadways into groups by class improved the prediction by either method: lower errors were found on the urban roadways subset of data points, and, in particular, on the interstate subset, while at some sites errors remain quite high, particularly in less dense areas and on small roads near major highways (Figure 5.6).

Figure 5.6: *Comparison of Percentage Prediction Errors (Urban roads). Source: Selby and Kockelman 2011*



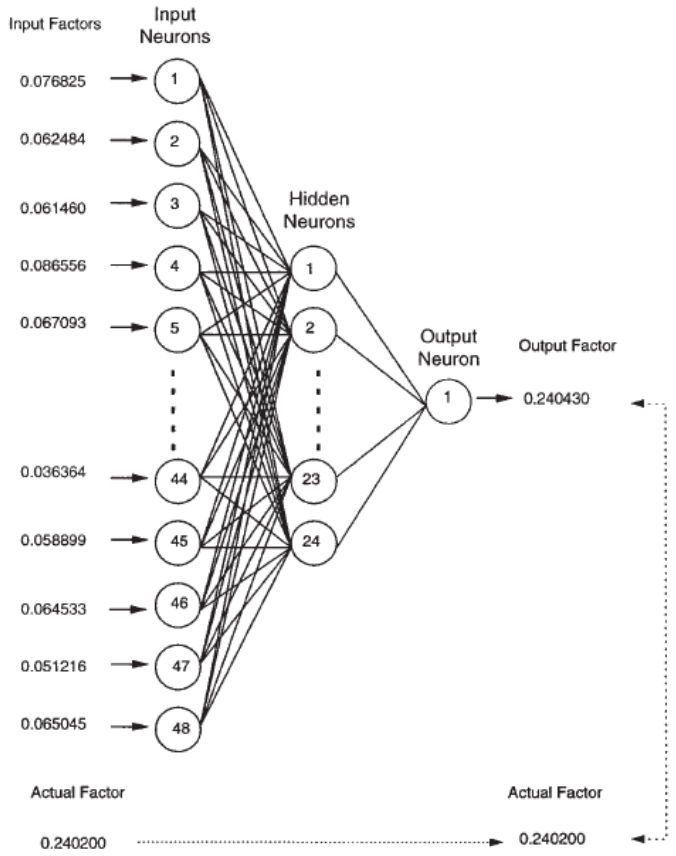
Concluding this review of methods which used land-use attributes in the AADT estimation process, Li, Zhao, and Chow (2006) considered this type of information to obtain the assignment of SPTCs to the road groups using a fuzzy-decision tree. The decision tree had the advantage of being conceptually easy to understand and did not require extensive data to operate. However, its implementation three counties in Florida was not satisfactory. The authors attributed the results to the fact that the four land use variables did not completely explain the traffic variations, that the sample size was limited, that ATR sites might not have reflected all representative land use patterns.

5.4.3 Artificial Neural Networks

The use of Artificial Neural Network was introduced in section 5.3 as a tool to create road groups. A different use of ANNs was proposed by a group of researchers from the University of Regina (Canada) to directly estimate AADT from SPTCs, avoiding the creation of road groups and, as a main consequence, the risk of incorrect assignment.

The first paper introducing the use of ANNs to estimate AADT from SPTCs was done by Lingras and Adamo (1996), but a more detailed description of the ANN was provided by (Sharma et al. 1999). In this paper the authors presented the ANN as a way to estimate AADT from 48hr SPTCs, comparing the results obtained with traditional factor approach (FACT model).

Figure 5.7: *ANN Model for AADT Estimation. Source: Sharma et al. 1999*



The ANN structure adopted had a multilayered, feedforward, back-propagation design (Figure 5.7), and neurons adopted the sigmoid transfer function. Data from 63 ATRs in Minnesota were used to test the quality of this approach. Since Minnesota's routine short-term counts are of 48hr duration starting at 12:00 p.m. on Monday, Tuesday, or Wednesday from April to October, the analysis was limited to three 48hr counts per week for each ATR during the counting period of 7 months, April to October.

Traditional factor approach was implemented using Sharma and Werner (1981)'s approach for ATRs grouping. On the basis of 21 composite expansion factors (3 days of week $\times$ 7 months of counting), 5 road groups were identified: ATRG1 and ATRG2 included commuter roads, roads of ATRG3 were average rural routes and ATRG4 and ATRG5 included recreational roads, with more traffic in summer and winter.

To evaluate the quality of AADT estimates, 48hr SPTCs were generated

from the dataset and were used as follows (see section 5.3.1):

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

The AADT estimated from a sample count was calculated by:

$$Estimated\,AADT = \frac{48 - hrsamplevolume}{2} \times adjustment\,factor \tag{5.52}$$

and the AADT estimation errors (Δ) were calculated by:

$$\Delta = \left| \frac{AADT_{Estimate} - AADT_{Actual}}{AADT_{Actual}} \right| \times 100 \tag{5.53}$$

The Neural Network approach was tested in different conditions:

- ANN1 (Model without classification): A single artificial neural network model was developed for all ATR sites without consideration of the day or month of counting. 48 neurons were included in the input layer, 24 in the hidden layer and and 1 in the output layer.
- ANN2 (Model with classification): A separate neural network model was developed for each ATRG developed for the traditional method, taking into account also the effect of the day and month of counting. The structure of the ANN is similar to ANN1 case.

For the training of both types of ANN, the input data were 48 input factors, defined as:

$$Input\,factor = \frac{\text{Hourly volume}}{SADT} \tag{5.54}$$

where SADT was the total 48hr sample volume divided by 2. Since supervised learning was adopted, also the neural network output was given in the learning phase, using the actual AADT factor, defined as:

$$ActualAADT\,factor = 0.25 \times \frac{AADT}{SADT} \tag{5.55}$$

In the testing phase the trained ANNs were fed with input factors for testing sample counts and the ANNs gave output factors to be used in AADT estimation as follows:

$$EstimatedAADT = 4(SADT \times Output\,factor) \tag{5.56}$$

The results obtained with traditional and ANN approaches were compared considering frequency and cumulative frequency distribution and average errors for each of the models (Figure 5.8). The neural network model ANN1 developed from unclassified data resulted in higher AADT estimation errors than those of the other two models. However the authors made a couple of relevant observations.

Figure 5.8: *AADT Estimation Errors Using Various Models. Source: Sharma et al. 1999*

(a) 95th Percentile Errors (%)

	ATR Group					
Model	All ATRs Combined	ATRG1	ATRG2	ATRG3	ATRG4	ATRG5
ANN1	22.91					
ANN2	19.83	18.29	19.96	17.89	20.79	22.91
FACT	15.28	12.17	12.94	16.89	16.94	22.83

(b) 85th Percentile Errors (%)

	ATR Group					
Model	All ATRs Combined	ATRG1	ATRG2	ATRG3	ATRG4	ATRG5
ANN1	16.64					
ANN2	14.26	13.42	14.69	11.64	15.37	19.15
FACT	10.04	7.61	8.39	11.70	12.82	15.28

(c) Average Errors (%)

	ATR Group					
Model	All ATRs Combined	ATRG1	ATRG2	ATRG3	ATRG4	ATRG5
ANN1	9.47					
ANN2	7.91	7.34	8.21	6.91	8.22	10.58
FACT	5.66	4.41	4.74	7.05	7.26	8.67

First, they considered that factor group classification improved results slightly when this kind of information is added in the neural network structure. Second, the experimental sample sites generated for this model

were correctly assigned to the various factor groups. In practice, in the absence of appropriate duration and frequency of seasonal counts covering all potential sites on the road network, it is unlikely that all short-period (e.g., 48hr) count sites would be assigned 100% correctly to one of the available factor groups. It is therefore reasonable that, in actual practice, the FACT model would not be able to provide comparable estimates of AADT, as shown in Table 1.

Therefore the authors conducted a preliminary investigation of the accuracy of two 48hr counts taken in two different months as compared with a single 48hr count, developing ANN1 model structure.

The first type of model was the ANN(1,48) model: in this case different models were developed for various months as done for ANN1 models. The second type of model used two 48hr counts and it was called ANN(2,48). It had 96 neurons in the input layer, 24 neurons in the hidden layer and a single neuron in the output layer. The network was fed with 96 hourly volumes in the form of input factors and with the desired output factor. No factors reflecting daily or monthly volume characteristics were provided as input to the model.

As can be observed from the results reported in Figure 5.8 the use of two 48hr counts taken in different months produced a notable reduction in the errors. The authors suggested that this fact could be related to the ANN's capability to distinguish seasonal variation in traffic and produce better AADT estimates by classifying patterns internally.

The authors implemented the ANN approach to another case study, focusing their attention to the estimation of AADT on low-volume roads (Sharma et al. 2000). Data from 55 ATR sites located on the provincial highway network of Alberta, Canada, were considered for the investigation. ATR sites were representative of different functional classes and trip purposes and they were divided in three groups, based on the AADT observed values: Group 1 ($AADT < 500$), Group 2 ($501 < AADT < 750$), Group 3 ($751 < AADT < 1000$).

Traditional factor approach (Sharma and Werner 1981) was implemented on the basis of 15 composite adjustment factors (3 days of week $\times$ 5 months of counting, May to September). 5 road groups were identified: ATRG1 (21 sites), ATRG2 (26 sites), ATRG3 (3 sites), ATRG4 (2 sites) and ATRG5 (3 sites). Also a subjective classification was made by traffic monitoring branch of Alberta Infrastructure, based on limited sample counts taken in the past and other factors, including geographic location and land-use characteristics. The grouping obtained consisted of four groups of low-volume roads: agriculture/resource (40 sites), resource (10 sites), tourist/resource (3 sites), tourist (2 sites).

The quality of AADT estimates obtained with the traditional factor approach was analysed considering three road classification schemes: the hierarchical grouping by Sharma and Werner 1981, the subjective classification

Figure 5.9: *AADT Estimation Errors Using Neural Network Models. Source: Sharma et al. 1999*

(a) ANN(1, 48) Model

Month	85th Percentile Errors (%)	95th Percentile Errors (%)
May	13.93	19.20
June	16.37	21.88
July	16.35	24.89
August	13.82	16.77

(b) ANN(2, 48) Model

Month (Day)	85th Percentile Errors (%)	95th Percentile Errors (%)
May-July (Monday)	12.02	16.68
May-July (Tuesday)	13.07	16.54
May-July (Wednesday)	9.83	14.14
June-August (Monday)	12.26	15.15
June-August (Tuesday)	11.63	15.92
June-August (Wednesday)	12.71	15.42

and a scheme in which all ATR sites belonged to the same group. 48hr SPTCs were generated from the dataset and were used as done in Sharma et al. 1999's paper, calculating the AADT estimation error Δ for each sample count (Equation 5.53).

The ANN approach used in this paper adopted a similar structure of the previous paper SLXL99. In this case the number of neurons in the input layer was equal to the total number of hourly volumes included in the sample counting problem: 96 for two 48hr counts, 48 for one 48hr counts. The hidden layer had a number of neuron equal to half of the number in the input layer and one neuron was included in the output layer. The learning process of the ANN is the same done by Sharma et al. (1999).

The results obtained using the different approaches were reported in Figures 5.10 and 5.11.

Traditional factor approach based on hierarchical grouping produced more accurate AADT estimates compared to the other classification schemes. However the implementation of the ANN approach was a valuable option if ANN(2,48) models were adopted. These ANNS produced AADT estimates comparable with factor approach results, with the advantage of not being

Figure 5.10: *AADT Estimation Errors Using Factor Approach. Source: Sharma et al. 2000*

	Percent Errors (%)		
Group	95th Percentile	85th Percentile	Average
Single Group	40.14	24.22	14.07

(**b**) For the subjective classification of study sites

	Percent Errors (%)		
Group	95th Percentile	85th Percentile	Average
Agriculture/Resource	33.82	23.30	12.68
Resource	31.53	20.44	11.30
Tourist/Resource	24.37	14.10	8.59
Tourist	23.03	11.75	9.24

(**c**) For hierarchical groups of study sites

	Percent Errors (%)		
Group	95th Percentile	85th Percentile	Average
ATRG1	24.73	17.40	9.71
ATRG2	28.99	20.01	11.21
ATRG3	41.51	29.61	16.32
ATRG4	20.64	15.74	8.76
ATRG5	45.14	29.88	17.90

influenced by the assignment process of SPTC to road groups, as stated by Sharma et al. 1999. In addition to this, the study highlighted that the effect of volume on AADT estimation errors were negligible, since there were no appreciable differences among percent error values for the various groups of low-volume roads.

These results were confirmed by the authors in a third study (Sharma et al. 2001), which analysed the implementation of similar ANN models, characterised by different number (1, 2 and 3) and duration (24, 48 and 72 hours) of short counts. The observed results, in terms of percent error of AADT estimates, highlighted in particular that:

- From a traffic monitoring point of view, if the AADT observed were less than 1000 vehicles they could be considered low-volume roads, without further AADT distinctions;
- The use of two 48hr short counts could be a reasonable choice, since there was a little gain in the accuracy of estimates when three 48hr counts were used instead of two.

Figure 5.11: *Comparison of AADT Estimation Errors. Source: Sharma et al. 2000*

(a) Summary of errors for the factor approach

	Percent Errors (%)		
Grouping Scheme	95th Percentile	85th Percentile	Average
Single Group	40.14	24.22	14.07
Subjective Groups	33.69	23.07	12.55
Hierarchical Groups	27.59	19.03	10.64

(b) Errors resulting from various ANN models

	Percent Errors (%)		
Model	95th Percentile	85th Percentile	Average
ANN(1, 48): May	32.2	24.14	15.57
ANN(1, 48): Jun	28.09	20.85	12.72
ANN(1, 48): Jul	35.16	26.33	15.63
ANN(1, 48): Aug	37.70	26.31	16.59
ANN(2, 48): May-Jul (Mon)	22.02	16.18	8.89
ANN(2, 48): May-Jul (Tue)	21.84	14.07	9.16
ANN(2, 48): May-Jul (Wed)	23.11	14.56	9.38
ANN(2, 48): Jul-Aug (Mon)	23.78	17.37	9.23
ANN(2, 48): Jul-Aug (Tue)	25.48	18.54	10.20
ANN(2, 48): Jul-Aug (Wed)	27.66	20.09	11.87

- The ANN using two 48hr short counts produce better estimations than those using 24-h counts, but there were not further advantages in using two 72hr counts;

5.5 Final Considerations

This review concerning AADT estimation based of FHWA factor approach has lead to the identification of two relevant issues, not well addressed by past literature:

- With reference to the grouping of road segments, it is difficult to identify the correct number and characteristics of road groups for a given road network. An AVC site could belong to more than one road group, and the groups cannot be easily defined in language (e.g., commuter road, recreational road);
- Considering the assignment problem of a given road section to a

road group, it was assumed that each road segment belongs to one group only, but it is apparent that a road section could exhibit the characteristics of more than one group.

Based on these considerations, a new approach has been developed and proposed to solve these issues, introducing specific theoretical frameworks.

Chapter 6

Proposed Approach

Based on the review of TMG factor approach, two points resulted critical and not well addressed by previous researches:

- the identification of the correct number and characteristics of road groups for a given road network;
- the treatment of situations in which a particular road segment belong to more than one group and how to measure the degree of similarity to each group.

The proposed approach (Gecchele et al. 2012) allows the analyst to deal with the situation when a road segment appears to belong to more than one group and to provide the degree of belonging to each group, while preserving the framework of the FHWA procedure. Fuzzy set theory is introduced to represent the vagueness of boundaries between individual road groups. To deal with the difficulties of identifying the group that matches the given road section, the measures of uncertainty (non-specificity and discord) are introduced. These measures help to interpret the quality of the estimates in an objective manner and also indicate the need for additional data collection.

As shown in Figure 6.1, the proposed approach has four steps:

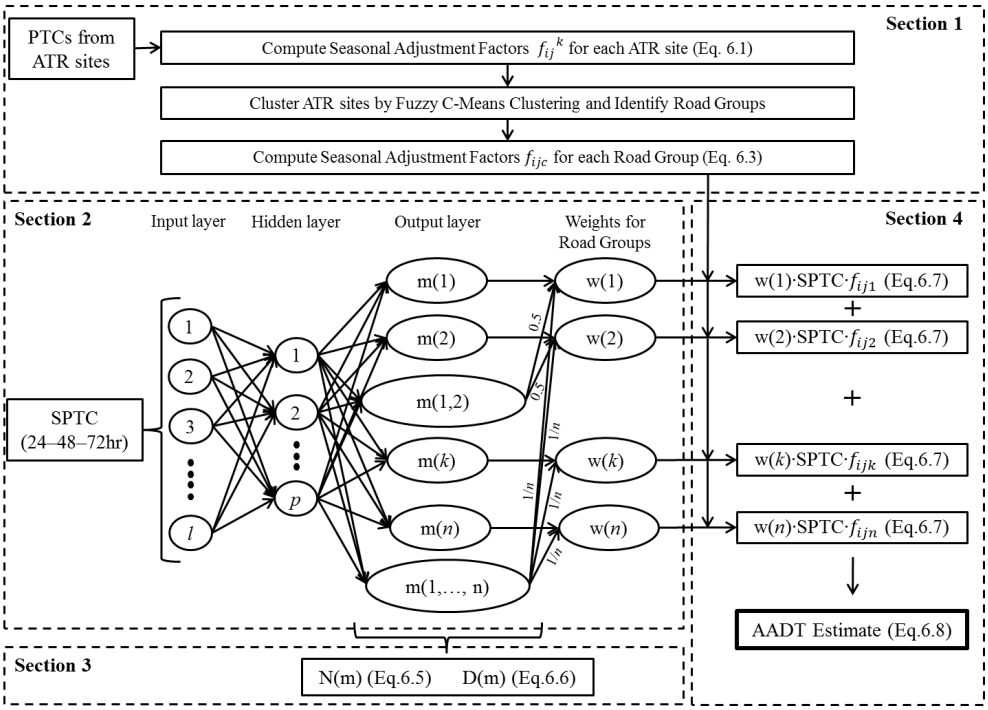
Step 1: Group the AVC sites using the fuzzy C-means algorithm on the basis of the seasonal adjustment factors of individual AVC (See section 1 of Figure 6.1) ;

Step 2: Assign the road segment for which SPTC is available to one or more of predefined road groups, using neural networks (See section 2 of Figure 6.1);

Step 3: Calculate the measures of uncertainty associated with the assignment to specific road groups (See section 3 of 6.1);

Step 4: Estimate AADT as the weighted by the average of SPTC daily volumes adjusted by seasonal adjustment factor of the assigned road group(s) (See section 4 of Figure 6.1).

Figure 6.1: *Scheme of the Proposed Approach*



More details about these steps are given in the sections below.

6.1 Grouping Step

When AVC is given for site *i*, the TMG's clustering approach assigns it to only one of the groups, assuming that the boundaries between road groups are well defined. This is not always the case in reality. The boundaries of the groups are often fuzzy; thus, AVC site *i* could belong to more than one group with different degrees, between 0 and 1, where 0 indicates no belonging to a group, and 1 represents complete belonging to a group (Klir and Yuan 1995).

The Fuzzy *C*-means (FCM) (Bezdek 1981; Bezdek, Ehrlich, and Full 1984) is a well-established algorithm to implement clustering when the boundaries of groups are fuzzy (see section 2.3.6). Given the number of groups *C*, the algorithm provides the degree that an AVC site belongs to each group used the seasonal adjustment factors as input variables.

Maintaining the structure of TMG factor approach, the seasonal adjustment factor for an AVC site *k* for the *i*-th day of the week of the *j*-th month is calculated by:

$$f_{ij}^{k} = \frac{AADT_k}{ADT_{ijk}} \tag{6.1}$$

where $AADT_k$ is the AADT for the k-th AVC site, ADT_{ijk} is the average daily traffic recorded in the i-th day of week of the j-th month in the k-th AVC site.

FCM requires that the number of groups (C) be specified in advance; however, the appropriate number is not known a priori. To solve this issue, the following procedure can be adopted:

- The value C is set to 2, that is the minimum number of groups (clusters) that can be identified;
- The FCM algorithm is ran for different time periods, changing the starting point and verifying the stability of results. The final output is the set of membership grades u_{ij} of each AVC site i to each road group j;
- These steps are repeated by incrementing the values of C to a maximum value C_{max}, which depends on the road network dimension and complexity;
- The optimal number of road groups C^* is chosen by comparing the values of some performance indices adopted for hierarchical clustering (see section 2.3.2), such as Dunn Index, Silhouette measure or Pseudo F Statistic.
- The set of membership grades corresponding to the optimal number of clusters C^* is assumed to be the final output.

Once the optimal number of groups C^* is fixed, the set of membership grades must be interpreted to identify the characteristics of the clustering obtained.

This analysis allows to distinguish among "*well defined*" AVC, which clearly belong to a road group, and "*I don't know*" cases, where the AVC could belong to different road groups. This distinction is based on the membership grades u_{ij} of each AVC site i to each road group j , following these rules (Hanesh, Sholger, and Dekkers 1984):

1. The highest membership grades u_{ij} is identified and set to u_{ijmax};
2. The ratios of all the membership grades u_{ij} to the highest one are calculated, that is:

$$ru_{ijk} = \frac{u_{ij}}{u_{ijmax}} \tag{6.2}$$

where j is the index for the j road group considered and k is the index for the road group with the highest membership grade. ru_{ijk} assumes the value 1 for the road group with the highest membership grade, and decreasing values as the membership grade to the road group decreases.

3. If the $u_{ijmax} > 0.5$ and the second highest $ru_{ijk} < 0.75$, then the AVC is considered as *clearly belonging* to the road group, otherwise the AVC is considered an "*I don't know*" case;

4. The AVC classified as an "*I don't know*" case could belong to the group with the highest membership grade or to the groups with $ru_{ijk} > 0.75$.

To better explain how the procedure is applied, one considers the different results obtained for a *clearly belonging* AVC (AVC 1) and for an "*I don't know*" AVC (AVC 2), whose membership grades to five road groups are reported in table 6.1.

Table 6.1: *Example of clearly belonging and "I don't know" AVCs*

	Group 1	Group 2	Group 3	Group 4	Group 5	Highest
AVC 1	0.7	0.2	0.1	0.0	0.0	0.7
AVC 2	0.05	0.0	0.35	0.3	0.3	0.35
AVC 1 Ratio	1.0	0.29	0.14	0.0	0.0	-
AVC 2 Ratio	0.14	0.0	1.0	0.85	0.85	-

- AVC 1 case: The highest membership grade u_{ijmax} is higher than 0.5 ($u_{11} = 0.7$ for Group 1). At the same time the ratio between the second highest value of membership grade ($u_{12} = 0.2$ for Group 2) and the highest one ($u_{11} = 0.7$ for Group 1) is lower than 0.75 ($ru_{121} = 0.29$). The result is that AVC 1 clearly belongs to Group 1.

- AVC 2 case: The highest membership grade is lower than 0.5 ($u_{23} = 0.35$ for Group 3). At the same time the ratio between the second highest value of the membership grade ($u_{24} = u_{25} = 0.3$ for Group 4 and 5) and the highest one ($u_{23} = 0.35$ for Group 3) is higher than 0.75 ($ru_{243} = ru_{253} = 0.85$). Therefore AVC 2 is considered an "*I don't know*" case. Since the membership grade to Group 3 is the highest one ($u_{23} = 0.35$) and only for Group 4 and 5 the ratio is higher than 0.75 ($ru_{243} = ru_{253} = 0.85$) it is assumed that AVC 2 could belong to "Group 3 or 4 or 5".

As done in the TMG factor approach (section 4.2), the seasonal adjustment factors that correspond to (i, j) combinations are calculated for the $C*$ road groups. If n AVC sites clearly belong to road group c, the seasonal adjustment factor for the i-th day of week of the j-th month is calculated by:

$$f_{ijc} = \frac{1}{n}\sum_{k=1}^{n}\frac{AADT_k}{ADT_{ijk}} = \frac{1}{n}\sum_{k=1}^{n} f_{ij}^{k} \tag{6.3}$$

where $AADT_k$ is the AADT for the k-th AVC site *clearly belonging* to group c and ADT_{ijk} is the average daily traffic recorded in the i-th day of week of the j-th month in the same AVC site.

6.2 Assignment Step

Artificial Neural Networks (ANNs) are an effective technique to assign an input to an output when the causal relation is not well-understood.

In this study, a multi-layered, feed-forward, back-propagation design is constructed, adopting the sigmoid transfer function for the neurons of the network. The ANN has three layers: one input layer, one output layer and one hidden layer, characterised as follows:

- the input layer includes a node describing the period of the year in which the counts are taken (generally the month) and a node for each hourly factor taken from the SPTC. The hourly factor, h_l, is defined as:

$$h_l = \frac{HT_l}{DT} \tag{6.4}$$

  where $l = 0, \ldots, 23$ is the starting time for the hourly count, HT_l is the hourly traffic volume for hour l and DT is the daily traffic for the specific SPTC. Since the number of hourly factors can vary depending on the duration of SPTCs (e.g. 24hr, 48hr), the number of neurons in the input layer will vary consequently (i.e. 24 or 48).

- the output layer has a node for each road group and nodes for all the power sets, such as Groups (1 or 2), Groups (2 or 3 or 5), including Groups$(1, \ldots, C^*)$, which is the case of total ignorance or "*I don't know at all*".

- the hidden layer has a variable number of nodes depending on the type of SPTCs considered. A rule of thumb that can be adopted as a first approximation is h = (attribute + groups)/2, where h is the number of hidden layers, attribute is the number of input variables and groups the number of groups.

6.3 Measures of Uncertainty in Assignment

A new aspect of this research is to investigate the uncertainty associated with assigning a road section to road groups. This section develops two measures of uncertainty for this purpose.

One case is when one cannot say either this group or that group because the traffic pattern fluctuation is not consistent enough to specify to just one group. The other case is when traffic patterns are conflicting so that a certain pattern at a particular time points to one group and at other times points to another group. The uncertainty related to the former is called *Non-Specificity*, and that of the latter is called *Discord*. These two measures are developed in the Dempster-Shafer theory. While details of this theory can be found in (Klir and Wierman 1999), the expressions for these measures are provided here.

Consider that the output nodes of the neural network represents all possible combinations of road groups; that is, not only road groups 1 to $C*$, but also all the power set of the road groups, e.g, (1 or 2), (2 or 4), (1 or 2 or 3), etc., a total of 2^n. Consider also that the weight associated with each final node is the degree that the traffic pattern supports individual power sets of road groups.

Let the weights associated with final node as $m(x)$, where x is a road group or more than one road group, and $m(x) = 1$.

- When $m(A) = 1$, it is certain that the road section in question belongs to road group A.
- When $m(A \text{ or } B) = 1$, the road section in question belongs either A or B, but the specific is uncertain.
- When $m(X) = 1$, where X is all road groups, the road section in question can belong to any of the group, in other words, the situation of "*I don't know*".

Given this probability distribution, $m(x)$, the measure of non-specificity, $N(m)$, and the measure of conflict $D(m)$, are developed.

$N(m)$ provides the measure of uncertainty that the available traffic pattern has no specific information about which road group the road section belongs to, or *Non-Specificity*. It can be calculated by:

$$N(m) = \sum_{A \in F} m(A) \cdot \log_2 | A | \tag{6.5}$$

where $| A |$ is the number of road groups in power set A.

The value of $N(m)$ is within $[0, \log_2 | X |]$. X is the universal set (all road groups) and $| X |$ is the number of these groups. The minimum of $N(m)$ is obtained when $m(A) = 1$, or the probability of belonging to a particular road group is one. The maximum of $N(m)$ corresponds to the case of "not able to assign to any *specific group*".

$D(m)$ provides the measure of uncertainty that the available traffic pattern contains conflicting information; it can belongs to one group and also another group, information of *Discord*. It can be calculated by:

$$D(m) = -\sum_{A \in F} m(A) \cdot \log_2 \left(\sum_{B \in F} m(B) \frac{| A \cap B |}{| B |} \right) \tag{6.6}$$

where $| B |$ and $| A \cap B |$ are the numbers of power sets associated with group B and for the intersections among subsets A and B, respectively. These measures are used to characterize the traffic pattern data collected, SPTC, at the road section which is to be classified to one or more of the predefined road groups.

6.4 AADT Estimation

The final step is to estimate AADT for the road section in question, where SPTC are available.

The degree that a road section belongs to each group, which is found in the output of the neural network, is used to calculate AADT.

For example, if the degree of belonging to road group 1 and (1 or 2) are $m(1) = 0.4$ and $m(1,2) = 0.6$, respectively, the final weights adopted for the estimation are calculated as $w(1) = 0.4 + 0.6/2 = 0.7$ and $w(2) = 0.6/2 = 0.3$. Therefore, the AADT estimate for a given SPTC is calculated by:

$$AADT = w(1) \cdot SPTC \cdot f_{ij1} + w(2) \cdot SPTC \cdot f_{ij1} \tag{6.7}$$

where f_{ij1} and f_{ij2} are found in Equation 6.3.

In the case of d-days SPTCs, the estimation of AADT is repeated d times using the 24hr volume data. The final AADT estimate is the average of these AADT estimates. For example, for a 3-days (72hr) SPTC the final AADT estimate is:

$$AADT_{Final} = \frac{AADT_1 + AADT_2 + AADT_3}{3} \tag{6.8}$$

where each AADT is the AADT estimated using the 24hr volume data for each of the d monitoring days.

Part III

Case Study Analysis

Chapter 7

Case Study

The approach presented in Chapter 6 has been implemented in a real world situation, testing its validity using traffic data. Different combination of SPCTs were tested, evaluating the accuracy of AADT estimates obtained in each condition.

Given the road network and the AVC sites, the Fuzzy C-means algorithm was used to create road groups. SPTCs were extracted from the AVC sites, and AADTs were estimated from the SPTCs, based on the assignment to road groups given by the Artificial Neural Network. The estimated AADT value were compared with the actual AADTs of the AVC sites.

Results have been interpreted considering measures of uncertainty (discord and non-specificity) and compared with those obtained by two approaches proposed in previous studies.

7.1 Data Source

The case study traffic data were obtained from the SITRA (TRAnsportation Information System or "Sistema Informativo TRAsporti") monitoring database of the Province of Venice.

Since 2000 the local Transportation Office is responsible of the monitoring activities taken on the rural road network of the Province of Venice. The Department of Structural and Transportation Engineering of the University of Padova (Della Lucia 2000) collaborated to the design of the original program and is still involved in data management and validation processes.

The design of the system was inspired by the procedure proposed by Traffic Monitoring Guide. Automatic Vehicle Classifiers (AVCs) were installed in a small number of road sections to permanently collect traffic data (PTCs). Road groups were identified based on similarity of seasonal adjustment factors calculated from PTCs. SPTCs (48hr to 2 weeks) and STCs (2 to 12 weeks) were periodically taken in the other road sections, following a detailed monitoring program. These road sections were assigned

to road groups and AADT could be estimated adjusting STPCs with the corresponding seasonal adjustment factors.

Variations of traffic patterns in different period of the year are accounted using 18 seasonal adjustment factors. These seasonal adjustment factors are calculated based on the combinations of:

- 3 day-types (Weekdays, Saturdays, Sundays);
- 6 two-month periods (January-February, March-April, May-June, July-August, September-October, November-December).

Monitoring road sections are generally equipped dual loop systems. Since the network links are mainly two-lane roads, two AVCs were installed in each site, one for each direction of traffic flow. AVCs collect hourly traffic volumes for different vehicle and speed classes (Tables 7.1,7.2). Hourly data are aggregated to obtain some relevant traffic parameters for each road section, including:

- Annual Average Daily Traffic (AADT);
- Annual Average Daytime Traffic;
- Seasonal Adjustment Factors;
- 30th highest annual hourly volume;
- Peak hour factors;
- Traffic growth trends.

Table 7.1: *Length Classes Adopted by SITRA Monitoring Program*

Class	**Length [m]**	**Level of Aggregation**		
		0	**1**	**2**
I	$L < 5.0$	LU01	LU11	LU21
II	$5.0 < L < 7.5$	LU02	LU12	LU22
III	$7.5 < L < 10.0$	LU03	LU13	
IV	$10.0 < L < 12.5$	LU04		
V	$12.5 < L < 16.5$	LU05		LU23
VI	$16.5 < L < 18.0$	LU06	LU14	
VII	$L \geq 18$	LU07		

Periodically the Transportation Office issues reports with traffic data summaries for each road section, that can be used by pratictioners and other public agencies (e.g. Figure A.1 reported in the Appendix). More detailed data can be also obtained from the website of the monitoring program (`http://trasporti.provincia.venezia.it/pianif_trasp/osservat.html`).

Table 7.2: *Speed Classes Adopted by SITRA Monitoring Program*

Class	Speed [km/h]	Level of Aggregation	
		0	**1**
I	$V < 30$	LU01	LU11
II	$30 < V < 50$	LU02	
III	$50 < V < 70$	LU03	LU12
IV	$70 < V < 90$	LU04	
V	$90 < V < 110$	LU05	
VI	$110 < V < 130$	LU06	LU13
VII	$V \geq 130$	LU07	

Since 2000 the number of AVCs has increased, reaching a good coverage of different types of road in the network. In 2012 the number of working AVC sites is 39, that is 78 AVCs which monitor directional traffic volumes (see Figure A.2 reported in the Appendix).

The traffic data for the study are the volumes obtained for the year 2005 at 42 AVCs (see Figure A.3 reported in the Appendix). The remaining AVCs have been excluded from the analysis since they were affected by considerable amounts of missing data in some periods of the year due to vandalisms and equipment failures.

Some assumptions have been made in the analysis:

- Data from monitored road sections have been analysed separately for each direction, based on the findings of Tsapakis et al. 2011;

- Estimation of AADT has been done for passenger vehicles only. Passengers vehicles data were divided by truck vehicles data, with reference to a 5 m-length threshold. This choice was made following the indications given by the FHWA concerning the specificities of truck traffic pattens, as reported in section 4.5.

7.2 Data Treatment

The total amount of available data at the AVC sites was 12,695 days of counts, which corresponded to 304,680 hours of counts. Hourly volumes of each AVC were sampled to form SPTCs of different durations (24 to 72 hours). Different SPTCs combinations have been tested, to simulate alternative SPTCs campaigns performed in the road network. For each combination a specific dataset has been created, as reported in Table 7.3.

Data included in each dataset were divided into 2 groups using the stratified holdout method (Witten and Frank 2005):

Table 7.3: *SPTCs Datasets Used for Simulations*

Number	Duration	Day-type	Starting Days
1	24hr	Weekdays	Mondays to Fridays
2	24hr	Saturdays	-
3	24hr	Sundays	-
4	48hr	Weekdays	Mondays to Thursdays
5	48hr	Week-ends	Saturdays
6	72hr	Weekdays	Mondays to Wednesdays
7	72hr	Week-ends	Fridays

1. Training dataset (70% of samples), used for the learning process of the ANN adopted for the Assignment of SPTCs;

2. Testing dataset (30% of samples), used to evaluate the accuracy of the AADT estimates.

Therefore SPTCs included in the testing dataset were used as the input to the ANN which has been developed from the training dataset.

Analysing the number of samples included for each AVC (Table A.1 reported in Appendix), it must be observed that:

- AVCs have different sample dimensions. AVC with less data have been included in the analysis only if all the seasonal adjustment factors could be calculated. This means that missing data did not concentrate in specific periods of the year but were distributed throughout the year;

- Counts taken during Italian holidays (e.g. Christmas, Easter, Celebration of Italian Republic) have been excluded from the analysis. In fact these holidays can occur in weekdays, but usually they show traffic patterns more similar to week-ends. Their exclusion from the analysis preserved the integrity and the quality of seasonal adjustment factors.

7.3 Model Implementation

Three tasks were conducted for the implementation of the proposed model: identifying road groups, developing and executing the artificial neural networks, and calculating AADT.

7.3.1 Establishing Road Groups Using Fuzzy C-Means

PTCs data from the 42 AVCs were used to establish the road groups (see section 6.1). Seasonal adjustment factors f_{ij}^{k} were calculated for each AVC

k, following Eq. 6.1. Then the Fuzzy C-means algorithm was applied using the adjustment factors as inputs. The algorithm was tested by changing the values of C from 2 to 20, running the algorithm for different time periods (10), changing the starting point and verifying the stability of results. The best number of groups was chosen by comparing the values of four indices:

- the Dunn Index;
- the Silhouette measure;
- the Pseudo F Statistic;
- Goodman and Kruskal's index G2.

Based on these criteria, the best number of groups C^* was found to be 8 for the case study. For each AVC k the membership grades to each road group u_{ij} (Table 7.4) were analysed with the procedure presented in section 6.1 and the belonging to a Road Group was determined.

Table 7.4: *Membership Grades of the AVCs to Different Road Groups*

AVC Number	Group1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group8	Road Group
1	0.01	0.01	0.01	0.03	**0.91**	0.02	0.00	0.01	5
2	0.01	0.01	0.01	0.04	0.05	**0.79**	0.01	0.08	6
3	0.05	0.02	0.04	**0.78**	0.04	0.05	0.01	0.01	4
4	0.03	0.01	0.02	**0.87**	0.03	0.02	0.01	0.01	4
5	0.19	0.17	**0.60**	0.02	0.01	0.01	0.00	0.00	3
6	0.07	0.03	**0.90**	0.00	0.00	0.00	0.00	0.00	3
7	0.02	0.01	0.01	0.06	**0.81**	0.05	0.01	0.03	5
8	0.01	0.01	0.01	0.02	0.02	**0.91**	0.00	0.02	6
9	**0.61**	0.09	0.26	0.02	0.01	0.01	0.00	0.00	1
10	**0.62**	0.08	0.26	0.02	0.01	0.01	0.00	0.00	1
11	**0.66**	0.06	0.25	0.02	0.01	0.00	0.00	0.00	1
12	**0.68**	0.07	0.21	0.02	0.01	0.01	0.00	0.00	1
13	**0.85**	0.04	0.10	0.01	0.00	0.00	0.00	0.00	1
14	0.33	0.09	0.26	0.25	0.04	0.02	0.00	0.01	(1, 3, 4)
15	0.32	0.09	0.25	0.26	0.03	0.03	0.01	0.01	(1, 3, 4)
16	0.47	0.09	0.40	0.03	0.01	0.00	0.00	0.00	(1, 3)
17	0.40	0.10	0.45	0.03	0.01	0.01	0.00	0.00	(1, 3)
18	0.32	0.10	0.26	0.25	0.03	0.03	0.00	0.01	(1, 3, 4)
19	0.08	**0.74**	0.14	0.02	0.01	0.01	0.00	0.00	2
20	0.08	**0.76**	0.13	0.02	0.01	0.00	0.00	0.00	2
21	0.12	0.08	**0.79**	0.01	0.00	0.00	0.00	0.00	3
22	0.14	0.13	**0.69**	0.02	0.01	0.01	0.00	0.00	3
23	0.08	**0.77**	0.12	0.02	0.01	0.00	0.00	0.00	2
24	0.10	**0.72**	0.13	0.02	0.01	0.01	0.00	0.01	2
25	0.13	0.27	**0.56**	0.02	0.01	0.01	0.00	0.00	3
26	0.14	0.22	**0.60**	0.02	0.01	0.01	0.00	0.00	3
27	0.08	0.05	**0.86**	0.01	0.00	0.00	0.00	0.00	3
28	0.12	0.16	**0.71**	0.01	0.00	0.00	0.00	0.00	3
29	0.26	0.16	**0.54**	0.03	0.01	0.00	0.00	0.00	3
30	0.15	0.06	**0.76**	0.01	0.01	0.01	0.00	0.00	3
31	0.36	0.28	0.28	0.05	0.01	0.01	0.00	0.01	(1, 2, 3)
32	0.27	0.27	0.36	0.05	0.02	0.02	0.00	0.01	(1, 2, 3)
33	0.01	0.01	0.01	0.04	**0.89**	0.03	0.00	0.01	5
34	0.01	0.01	0.01	0.04	0.04	**0.86**	0.00	0.03	6
35	0.09	0.05	0.07	**0.63**	0.09	0.05	0.00	0.02	4
36	0.00	0.00	0.00	0.04	0.00	0.00	**0.89**	0.07	7
37	0.01	0.01	0.01	0.02	0.02	0.05	0.00	**0.88**	8
38	0.06	**0.78**	0.15	0.01	0.00	0.00	0.00	0.00	2
39	0.00	0.00	0.00	0.00	0.01	0.03	**0.90**	0.06	7
40	0.01	0.00	0.01	0.01	0.02	0.00	0.05	**0.90**	8
41	**0.78**	0.05	0.14	0.02	0.01	0.00	0.00	0.00	1
42	**0.80**	0.04	0.13	0.02	0.01	0.00	0.00	0.00	1

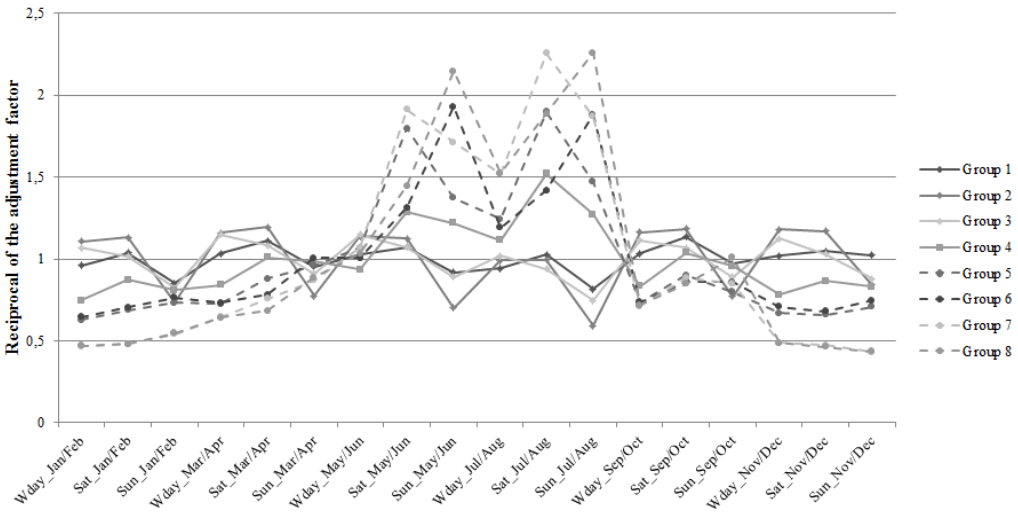
AVCs "clearly belonging" to a group are marked in bold

The seasonal adjustment factors that correspond to (i, j) combinations were calculated for the C^* road groups with Equation 6.3. Reciprocals of the seasonal adjustment factors rf_{ijc} were defined by:

$$rf_{ijc} = \frac{1}{f_{ijc}} \tag{7.1}$$

and they could represent the characteristics of fluctuations better than the seasonal adjustment factor, f_{ijc}. Average seasonal adjustment factors of the road groups and their reciprocals are reported in Tables A.2 and A.3 in Appendix. In Figure 7.1 the average reciprocals of the seasonal adjustment factors rf_{ijc} for different days and periods of the year are plotted for each road group.

Figure 7.1: *Reciprocals of the Seasonal Adjustment Factors for the Road Groups*



Analysing the plot some clearly distinguishable traffic patterns can be observed:

- Groups 1 (7 AVCs), 2 (5 AVCs), and 3 (10 AVCs) could be characterized as *commuter road* groups. These groups show stable traffic patterns, with seasonal adjustment factors close to one, in particular for weekdays. Weekly traffic patterns occur in similar manner during the year, but some differences exist among groups:

  - Group 1 passenger car volumes increase from weekdays to Saturdays and decrease from Saturdays to Sundays;

  - Group 2 shows a pattern similar to Group 1, but the decrease observed during Sundays is more relevant, in particular in summer period (seasonal adjustment factors $Sun_{May/Jun}$ and $Sun_{Jul/Aug}$);
  - Group 3 pattern is characterised by a decrease of traffic volumes during week-ends, in particular on Sundays.

- Groups 5 (3 AVCs), 6 (3 AVCs), 7 (2 AVCs) and 8 (2 AVCs) could be characterized as *recreational road* groups. Traffic patterns are characterised by a strong variability during the year: very small volumes in winter period and high peaks in summer time, with higher variations observed for Groups 7 and 8.

  A deeper analysis of the composition of the groups shows that the AVCs of the same site belong to different groups: Group 5 - Group 6 or Group 7 - Group 8. This fact is due to different traffic patterns observed in each direction during summer week-ends (May/Jun and Jul/Aug): Groups 5 and 7 have peaks in traffic volumes during Saturdays, while Groups 6 and 8 have their peaks during Sundays.

  These differences can be explained considering the holiday trips made by people during week-ends: the vacationers reach the holiday resorts on Saturdays and go back home in Sundays driving in the opposite direction.

- Group 4 includes 3 ATRs with intermediate characteristics.

- Seven AVCs were classified as "I don't know" cases. They belonged to Group "1 or 2 or 3" (2 AVCs), Group "1 or 3"(2 AVCs), Group "1 or 3 or 4"(3 AVCs), that is they had traffic patterns similar to commuter roads.

Furthermore spatial distribution of road groups were analysed, to evaluate if the differences among road groups could also be interpreted based on the knowledge of land-use characteristics.

As can be observed in Figure A.4 reported in the Appendix, AVCs were grouped following clear spatial patterns which confirmed previous observations:

- AVCs belonging to commuter road groups (1, 2, 3) are located in the inland parts of the province, where tourist activities are limited and traffic patterns are supposed to be quite stable during the year;

- AVCs belonging to recreational road groups (5, 6, 7, 8) are located in the coastal part of the Province of Venice (groups 7 and 8) or to roads which give access to the tourist facilities (groups 5 and 6). This means that in summer period these roads are supposed to be characterised by very high passenger car volumes compared to winter period;

- AVCs belonging to road group 4 are representative of intermediate characteristics between commuter and recreational roads. Their location in the road network (between the inland and the coastal line) confirms these characteristics;

- AVCs classified as "I don't know cases", which are characterised by commuter-type patterns, are in the inland parts of the Province.

7.3.2 Developing the Artificial Neural Networks

Multi-layered, feed-forward artificial neural networks (ANNs) were developed in order to assign the SPTCs to the road groups (see section 6.2).

Different structures of ANN are adopted, corresponding to the different SPTCs combinations analysed, as reported in Table 7.5. Applying the proposed approach to the case study, the number of output nodes was reduced to 11, since in the training dataset 8 road groups and 3 "I don't know" situations were found. Moreover different ANNs were trained for each datasets, maintaining the structure corresponding to the specific duration of SPTCs used. That is 24hr SPTCs taken on weekdays were used to train a network, while 24hr SPTCs taken on Saturdays were used for another network with the same structure.

Table 7.5: *Characteristics of ANNs Used for the Assignment*

Datasets	SPTCs Duration	Input Nodes	Hidden Nodes	Output Nodes
1,2,3	24hr	25	30	11
4,5	48hr	49	60	11
6,7	72hr	73	84	11

Some further details about the training process, repeated in the different training datasets, are:

- Learning cycles = 25,000;

- Momentum α = 0.2;

- Learning rate η = 0.3.

7.3.3 Calculation of AADT

In this process each SPTC in the test dataset is assigned by the corresponding ANN, obtaining the probabilities of belonging to each road group. The SPTC volume is used to estimate the AADT following only Equation 6.7 for 24hr SPTCs and also Equation 6.8 in case of 48hr and 72hr SPTCs.

7.4 Results and Discussion

7.4.1 Examination of the Estimated AADTs

The actual AADT for an AVC and the estimated AADT from SPTC for the same site were compared. The percent absolute estimation error was used as a measure of goodness of estimate:

$$\Delta = \left| \frac{AADT_{Estimate} - AADT_{Actual}}{AADT_{Actual}} \right| \times 100 \tag{7.2}$$

The Mean Absolute Error (MAE) and the Standard Deviation of Absolute Error (SDAE) (Equations 5.45 and 5.46) were used to analyse the accuracy of the AADT estimates made with the proposed approach.

Tables 7.6 and 7.7 summarize MAEs and SDAEs for the AVCs for different tests: by group, by different durations of SPTCs and by different day-types (weekdays, weekends, weekdays + weekends). Differences between the MAE and SDAE of road groups in the various conditions were tested using a T-test at a 95% confidence level.

Table 7.6: *Mean Absolute Error (MAE) of the Road Groups for Different Combination of SPTC and Time Periods*

SPTC	Group	Total				Weekdays		Weekends	
		Perfect	ANN	D <20%	% Samples	Perfect	ANN	Perfect	ANN
24	1	5.57	8.28	7.20	79.27	4.96	7.03	7.41	11.15
24	2	6.26	9.05	7.75	78.21	5.90	7.71	7.82	12.44
24	3	5.99	7.11	6.52	85.07	5.73	6.59	6.96	8.43
24	4	8.46	10.68	10.19	83.65	8.06	9.08	11.25	14.72
24	5	11.24	16.29	15.39	68.54	10.74	15.50	13.01	17.80
24	6	9.28	15.55	15.32	73.28	8.92	12.64	10.59	21.12
24	7	16.16	20.23	19.11	83.51	14.85	17.32	19.28	27.29
24	8	14.88	19.15	16.67	81.22	13.55	15.35	19.73	28.54
24	Total	7.98	10.88	9.96	81.09	7.47	9.45	9.17	14.21
48	1	4.54	5.49	4.90	86.06	4.32	5.14	5.41	6.60
48	2	5.53	6.14	5.55	94.47	5.50	5.98	5.67	6.81
48	3	5.20	5.43	5.24	94.79	5.14	5.34	5.43	5.80
48	4	7.27	7.79	7.38	91.79	7.28	7.50	7.25	8.98
48	5	9.16	10.64	9.79	82.89	8.87	10.33	10.01	11.49
48	6	7.42	9.04	7.86	83.21	7.26	8.38	7.91	11.04
48	7	14.49	16.33	15.64	87.20	13.62	15.53	17.88	19.36
48	8	13.43	14.79	13.92	86.14	12.67	13.48	16.47	19.91
48	Total	6.83	7.66	7.11	90.63	6.63	7.31	7.56	8.89
72	1	4.22	4.76	4.24	85.41	3.97	4.49	5.01	5.55
72	2	5.33	5.50	5.31	96.66	5.36	5.56	5.21	5.33
72	3	4.91	4.96	4.87	97.20	4.80	4.79	5.22	5.46
72	4	6.72	6.86	6.73	92.39	6.88	6.90	6.24	6.72
72	5	8.60	9.44	9.12	83.89	8.13	8.90	9.83	10.89
72	6	7.22	8.07	7.70	86.11	7.04	8.00	7.71	8.32
72	7	13.72	14.29	14.19	88.55	12.69	13.28	16.67	17.02
72	8	12.90	13.83	12.87	88.28	12.18	13.02	15.06	16.16
72	Total	6.46	6.85	6.56	92.12	6.24	6.60	7.09	7.55

The MAEs and the SDAEs have been calculated for different cases:

Perfect case. The performances refer to the case of assigning all the samples coming from an AVC to the correct group. This is the ”perfect” case since in practice it is difficult that 100% of SPTCs are assigned to the correct group. However it has been introduced as a benchmark for the proposed approach;

Table 7.7: *Standard Deviation of Absolute Error (SDAE) of the Road Groups for Different Combination of SPTC and Time Periods*

SPTC	Group	Total				Weekdays		Weekends	
		Perfect	ANN	D <20%	% Samples	Perfect	ANN	Perfect	ANN
24	1	5.13	10.48	8.97	79.27	4.41	8.98	5.96	12.03
24	2	5.50	13.78	12.63	78.21	4.99	11.23	6.66	16.81
24	3	5.26	9.04	8.29	85.07	4.82	8.14	5.55	10.48
24	4	7.14	11.26	11.17	83.65	6.34	8.39	9.16	15.27
24	5	9.76	13.81	14.51	68.54	9.81	12.37	9.68	15.04
24	6	8.55	20.24	19.66	73.28	8.68	10.73	7.67	30.59
24	7	15.55	19.82	19.78	83.51	14.91	17.47	15.41	22.50
24	8	15.42	22.32	19.87	81.22	15.06	16.62	16.07	30.50
24	Total	7.27	12.94	12.12	81.09	6.81	10.37	7.93	16.05
48	1	4.12	6.50	5.64	86.06	3.94	6.23	4.51	7.03
48	2	4.52	7.60	4.90	94.47	4.30	6.22	5.20	10.57
48	3	4.53	5.03	4.74	94.79	4.39	4.94	4.92	5.28
48	4	5.54	6.61	5.97	91.79	5.20	5.45	6.55	9.34
48	5	7.75	8.93	8.29	82.89	7.40	8.85	8.21	8.68
48	6	6.45	8.96	7.52	83.21	6.55	8.17	6.04	10.74
48	7	13.31	15.14	15.14	87.20	12.66	14.79	15.14	16.15
48	8	13.55	15.45	15.30	86.14	12.92	13.92	15.54	19.61
48	Total	5.99	7.67	6.79	90.63	5.76	7.11	6.59	8.93
72	1	3.92	4.90	4.00	85.41	3.76	4.80	4.23	7.03
72	2	4.20	4.67	4.13	96.66	4.00	4.56	4.59	10.57
72	3	4.32	4.43	4.23	97.20	4.18	4.17	4.59	5.28
72	4	4.99	5.23	5.04	92.39	4.53	4.58	6.20	9.34
72	5	7.22	7.90	7.74	83.89	6.64	7.34	8.37	8.68
72	6	5.94	6.76	6.62	86.11	6.11	6.61	5.53	10.74
72	7	12.36	12.95	12.89	88.55	10.82	11.53	15.72	16.15
72	8	12.70	13.54	12.54	88.28	11.55	12.42	15.59	19.61
72	Total	5.61	6.13	5.72	92.12	5.28	5.76	6.32	6.87

ANN. These data represent the values obtained applying the proposed approach;

D < 20%. In this case only the samples with a discord less than the 20% of the maximum value have been used to calculate the statistics. The percentage of samples passing this threshold is indicated as %*Samples*.

The following observations can be made based on the values of MAE and SDAE reported in Tables 7.6 and 7.7:

- Recreational roads (Groups 5, 6, 7, 8) have larger values of MAE and SDAE compared to those of the commuter roads (Groups 1, 2, 3) in any condition tested, because of the effect of the higher variability in traffic patterns;

- SPTCs taken on weekdays give more precise AADT estimate compared to the ones taken during weekends. Differences between Perfect and ANN assignments are significant for 24hr SPTCs during weekdays and weekends for Groups 1, 2, 3, 6, 7 and 8. Adopting 48hr and 72hr SPCTS the difference is significant only for Group 1;

- the increase in the duration of the SPTCs has a positive influence on the AADT estimate. The MAE and the SDAE decrease both in the perfect assignment, both in the case of ANN assignment. Except for the Group 1 sites, the differences between the results obtained in the two cases for 48hr and 72hr SPTCs are not statistically significant;

- the use of SPTCs assigned to a road group with low values of discord (< 20% of maximum) gives better AADT estimates compared of the use of all the SPTCs available. The differences are statistically significant for 24hr and 48hr SPTCs excepting Groups 7 and 8, while there is no significant difference using 72hr SPTCs. The percentage of samples with low values of discord increases as the duration of SPTCs increase for the majority of AVCs.

To better understand the effect of discord measure on AADT estimates, further analyses were conducted on available data. Since the results presented similar schemes for the different day-types considered, only the "Total" case sample data are reported here for a subset of the AVCs under analysis. However the results obtained for each AVCs in the different conditions tested are reported in the Appendix.

Figure 7.2 reports the percentage of 48hr SPTCs with values of discord lower than different thresholds for 14 AVCs. Figure 7.3 shows the corresponding MAEs obtained from the selected SPTCs for the same AVCs. One can observe that:

- The number of SPCTs with low values of discord decreases as the chosen threshold value of discord decreases;

- The use of SPTCs with low values of discord produces lower MAEs. This effect increases as the chosen threshold value dicreases;

- The number of SPTCs with low values of discord and the corresponding reduction in the MAEs can vary depending on the AVC considered;

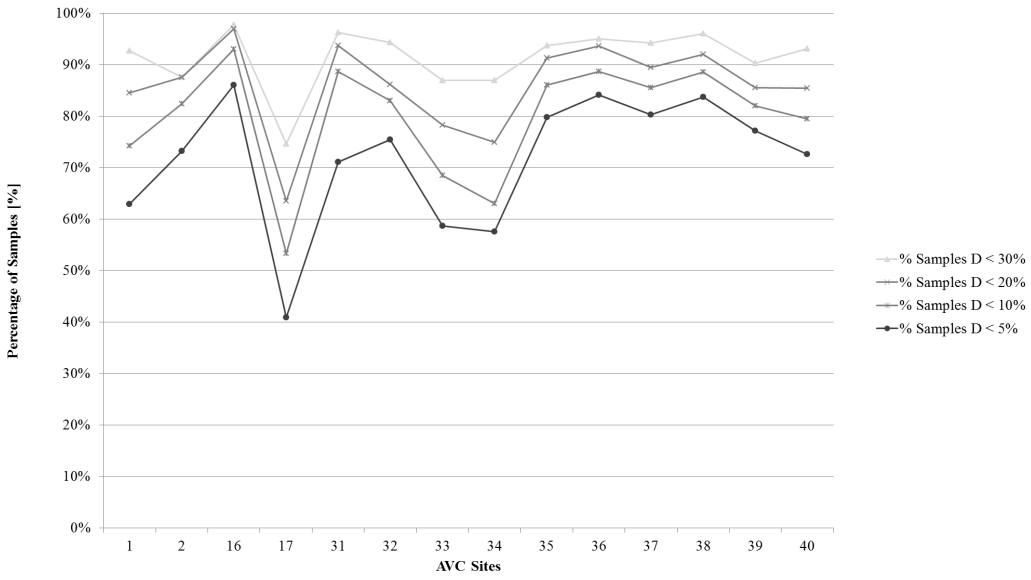
- The use of a discord threshold equal to the 20% of the maximum seems to be a compromise between the possibility of reducing the MAE obtained form SPTCs counts and the need of using as many SPTCs as possible.

The effect obtained on MAEs due to the choice of SPTCs with low values of discord (with a threshold set to 20% of maximum) was analysed for different durations at the same 14 AVCs (Figure 7.4). As can be observed:

- The use of SPTCs with low values of discord gives more accurate AADT estimates than the use of all SPTCs also for 24hr and 72hr SPTCs;

- The reductions of MAE are more relevant for 24hr SPTCs than 72hr SPTCs. Furthermore the increase of SPTCs' duration produces more relevant reduction compared to the selection of SPTCs based on values of discord measure;

Figure 7.2: *Percentage of 48hr SPTCs With Values of Discord Lower Than Different Thresholds for 14 AVC Sites*



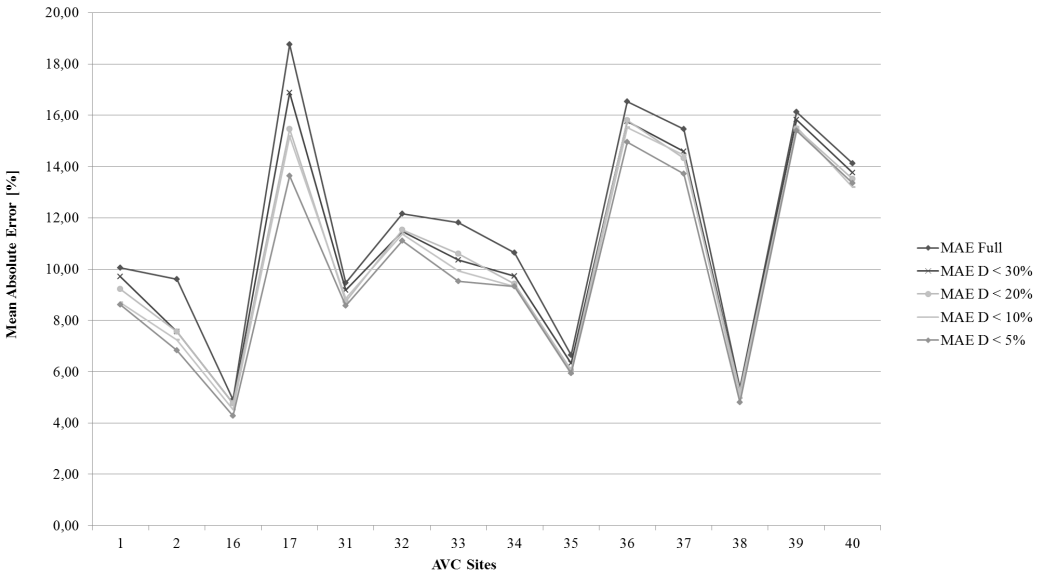
- Those SPTC < 20% maximums vary among the AVCs in the observed network, that is the difficulty in classifying to particular road group is different among the AVCs.

The results here presented suggest that consistency in the traffic pattern pointing to particular road pattern is important to obtain accurate AADT. When the traffic pattern observed from a given SPTC clearly points to a road group, the AADT accuracy increases. Discord measure can be used to quantify the consistency of the assignment. However the results highlights that the quality of assignment can change among the AVC sites.

AADT estimates were analysed also in terms of non-specificity measure. For the same 14 AVCs Figure 7.5 presents the percentage of SPTCs of different durations with low non-specificity values (< 15% and < 25% of maximum). One can observe that:

- Small differences in the percentage of SPTCs are observed for different threshold values of non-specificity;
- AVCs 16, 17, 31 and 32 were classified as ”I don’t know” cases and they showed a low percentage of samples with low values of non-specificity compared to other AVCs;
- An increase in the duration of counts reduces the SPTCs with low

Figure 7.3: *Mean Absolute Error for 14 ATR Sites Based on Different SPTC Durations and Discord Values*



values of non-specificity mainly for sections clearly belonging to a group, as for AVC sites 33 and 34;

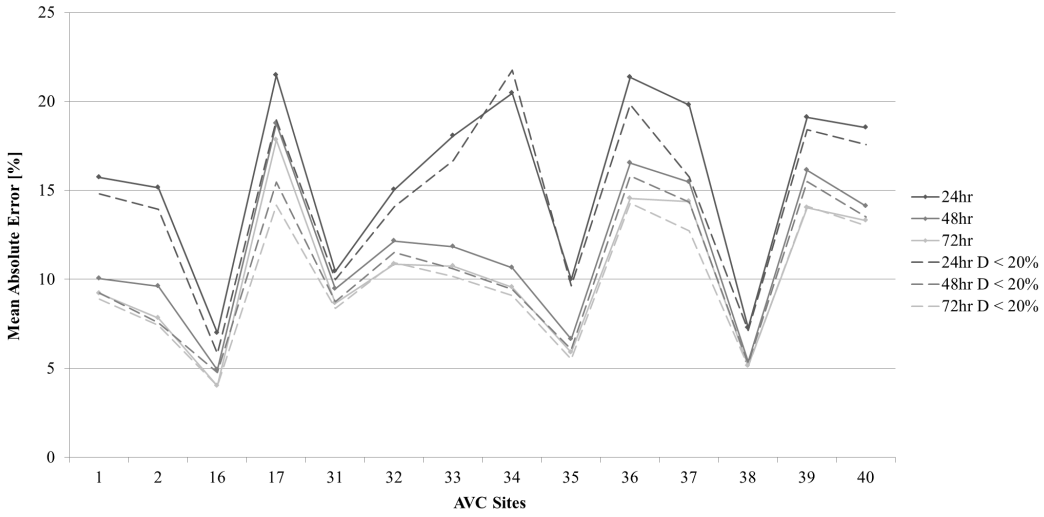
- AVCs clearly belonging to a road group and "I don't know" cases show different trends as the duration of SPTCs increases:

  - For AVCs clearly belonging to a road group, if the duration increases, the number of SPTCs with low values of non-specificity increases as an effect of a better classification done by the ANN;

  - For AVCs which could belong to more than one road group, if the duration increases, the number of SPTCs with low values of non-specificity decreases. This fact can be explained considering that these AVCs cannot be clearly assigned to a road group, that is they can belong to more than one group. The more they are assigned to more than one group (when duration increases), the more the assignment is "correct" and the number of SPTCs with low values of non-specificity decreases.

These findings suggested that the non-specificity measure is useful to highlight another aspect of the uncertainty in AADT estimates, that is the identification of SPTCs characterized by uncertain assignment. Based on

Figure 7.4: *Mean Absolute Error for 14 ATR Sites Based on Different SPTC Durations*



these results, the practical application of the proposed method should pay attention to the following:

- SPTCs, preferably 48hr-long, should be taken during weekdays to have sufficient information for a correct assignment of the road section to road groups;

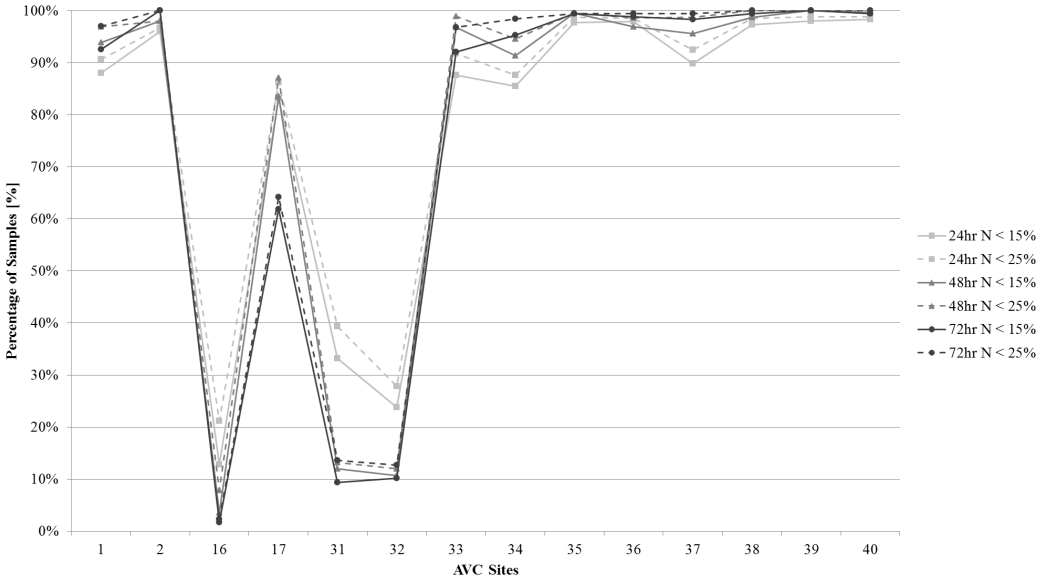
- Discord and non-specificity are important measure to evaluate the quality of estimates and improve their interpretability. Low values of discord are related to accurate AADT estimates and high values of non-specificity indicate uncertain assignments of SPTCs. SPCTs with high discord measure or high non-specificity suggest the need for additional data collection.

7.4.2 Comparison with Other Models

The accuracy of AADT estimates presented in the previous section in terms of MAE and SDAE were found close to previous studies (Cambridge Systematics Corporation 1994; Sharma et al. 1999). However a detailed comparison was not possible because large differences exist in terms of counts and network characteristics among different studies.

As an example Tsapakis et al. (2011) recently obtained from 24hr SPTCs assigned to road groups MAEs ranging from 4.4% to 4.2%. However no details were given by the authors about the number of road groups considered or the distribution of errors for different day-of-week or road groups.

Figure 7.5: *Percentage of Samples with Non-specificity Lower Than Different Thresholds for 14 AVC Sites*



Similarly Sharma et al. (1999) using 48hr SPTCs on weekdays from April to October obtained a MAE of 4.41% and 4.74% for commuter road groups, 7.26% and 8.67% for recreational road groups and 7.05% for average road group, assuming a perfect assignment. In this case it seems not correct assuming that the network of the Province of Venice has the same characteristics of the rural road network of Minnesota. Consequently the obtained AADT estimation accuracies cannot be directly compared.

For these reasons the results obtained with the proposed model were compared with those obtained by two approaches proposed in previous studies using the same dataset. 48hr SPTC data taken on weekdays were used since they were found to be the best solution for AADT estimate from SPTCs.

The two model tested were the methods proposed by Sharma et al. (1999) (see section 5.4.3 for details) (here called Sharma et al.) which is an ANN model with hourly factors as inputs without the definition of road groups, and the approach proposed by Tsapakis et al. (2011), based on the use Linear Discriminant Analysis (LDA)(see section 5.4.1 for details).

The MAEs of these approaches are presented in Table 7.8, according to the road groups from which the SPTCs are extracted. For Sharma et al. approach the road grouping step is not needed, but the MAEs obtained from the different AVCs are grouped to facilitate the comparison of results.

Moreover, since the proposed approach uses singular SPTCs, Sharma et al. approach has been adapted to the specific case, considering only one 48hr SPTC at each time.

Table 7.8: *Comparison of MAE of the Proposed Model with Previous Models*

SPTC [hrs]	Group	Proposed	Sharma et al.	LDA
48	1	5.1	8.7 **	7.6 *
48	2	6.0	8.8 *	8.0 **
48	3	5.3	7.2 **	7.6 **
48	4	7.5	9.2	9.6
48	5	10.3	16.1 **	13.2 **
48	6	8.4	14.1 **	14.4
48	7	15.5	15.8	18.8
48	8	13.5	15.5	18.3 **

** = statistically significant at a 95% confidence level (Paired T-Test)
* = statistically significant at a 90% confidence level

Table 7.8 shows that the AADT estimated by the proposed approach shows smaller MAE compared with Sharma et al. and LDA. The differences are significant for groups 1, 2, 3 (commuter roads) and group 5 (recreational roads). The difference is statistically significant also for groups 6 and 8.

Moreover these results highlight that:

- the use of procedures which follow FHWA factor approach (identification of road groups and assignment of SPTCs) can give accurate AADT estimate, better than a direct estimate of AADT, as done by Sharma et al.'s model;
- ANN model performs better than LDA model. This results might be due to the presence of non-linear relationships between short traffic patterns and seasonal traffic patterns, that the ANN model can better represent.

Chapter 8

Conclusions and Further Developments

In recent years the technological evolution is increasing the availability of data that can be used in transportation studies. However the quality and the amount of these data often represent a relevant issue when accurate analysis need to be performed. Since Data Mining techniques have been proved to be useful tools to be used in these situations, this thesis has analysed the capabilities offered by the implementation of Data Mining techniques in transportation systems analysis.

The first part of the thesis has presented a review of well-established Data Mining techniques, focusing on classification, clustering and associations techniques, which are the most applied in practice. It has been reported that their use has become quite common also in transportation studies, where they have been used in different contexts.

The second part of the thesis has analysed in details FHWA factor monitoring approach for the estimation of Annual Average Daily Traffic. The review of literature has identified in the definition of road groups and the assignment of sections monitored with short counts (SPTCs) the most relevant research issues. A new approach that estimates AADT based on limited traffic counts while preserving the basic structure of the current FHWA procedure has been proposed. In this approach fuzzy set theory is used to represent the vagueness of boundaries between individual road groups. The measures of uncertainty (non-specificity and discord) from Dempster-Shafer theory are also introduced to deal with the difficulties of identifying the group that matches the given road section. These measures help to increase the interpretability of results, giving the degree of uncertainty when assigning a given road section to a specific road group. This information is particularly important in practice when a given road section can match the patterns of more than one roads group.

The approach has been implemented using data from 42 Automatic Ve-

hicle Classifiers located on the rural road network of the Province of Venice. The results obtained in the estimation of AADT for passenger vehicles, also compared with those obtained by two approaches proposed in previous studies, highlighted that:

- the accuracy of estimates in terms of Mean Absolute Error and Standard Deviation of Absolute Error is found better than the previous studies;
- for the commuter roads, errors are smaller than the case of the recreational roads given the different variability of traffic patterns;
- sample counts taken on weekdays give more accurate AADT estimates than the ones taken during weekends;
- increase in the duration of the SPTCs results in a decrease in the errors in AADT estimates, particularly when the counting duration increases from 24 to 48 hours.

Concerning the use on uncertainty measures, it has been observed that:

- the measure of discord is useful to indicate the quality of the estimates made from SPTCs, in particular a low value of discord is related to an accurate AADT estimate. Conversely the non-specificity measure indicates more uncertainty in the assignment of the SPTCs to the road groups. When SPTCs have high discord or high non-specificity, then an additional data collection is needed

These findings suggest that when apply the proposed method SPTCs should be measured during weekdays, preferably for 48 hours. At the same time the discord and the non-specificity are important measures for evaluating the quality of the estimate and they can be used to identify the need for additional data collection.

Given the positive results obtained in the experimental phase of the research, the design of a software tool to be used in next future in real world applications has begun. However it the future, this work could be further extended to the following topics:

- Examine whether the proposed method can be applied to estimate the AADT for freight vehicles volumes, because the volume fluctuations of freight vehicles are different from the ones of passenger cars;
- Examine the influence of socio-economic and land-use characteristics of the environment of the road section when identifying the road group and assigning the SPTCs.

Chapter 9

Bibliography

Data Mining Techniques

Ankerst, M. et al. (1999). "OPTICS: Ordering Points To Identify the Clustering Structure". In: *ACM SIGMOD International Conference on Management of Data*. ACM Press, pp. 49–60.

Arthur, D. and S. Vassilvitskii (2007). "k-means++: the Advantages of Careful Seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035.

Banfield, J. and A. Raftery (1993). "Model-Based Gaussian and Non-Gaussian Clustering". In: *Biometrics* 49.3, pp. 803–821.

Berry, K. and G. Linoff (2004). *Data Mining Techniques for Marketing Sales and Customer Support*. New York: John Wiley Sons, Inc.

Bezdek, C.J. (1981). *Pattern Recognition with Fuzzy Objective Function Algoritms*. New York: Plenum Press.

Bezdek, C.J., R. Ehrlich, and W. Full (1984). "FCM: the Fuzzy C-Means Clustering Algorithm". In: *Computational Geosciences* 10.2-3, pp. 191–203.

Buckles, B.P. and F.E. Petry (1994). *Genetic Algorithms*. Los Alamitos, California: IEEE Computer Society Press.

Calinski, R. and J. Harabasz (1974). "A Dendrite Method for Cluster Analysis". In: *Communications in Statistic* 3.1, pp. 1–27.

Chapman, P. et al. (2000). *CRISP-DM Step-by-Step Data Mining Guide*. `http://www.crisp_dm.org`.

Chaturvedi, A., P.E. Green, and J.D. Caroll (2001). "K-modes Clustering". In: *Journal of Classification* 18.1, pp. 35–55.

Davies, D. and D. Bouldin (1979). "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1.4, pp. 224–227.

Dunham, M.H. (2003). *Data Mining. Introductory and Advanced Topics*. Upper Saddle River, New Jersey: Prentice Hall.

Dunn, J.C. (1973). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". In: *Journal of Cybernetics* 3, pp. 32–57.

— (1974). "Well Separated Clusters and Optimal Fuzzy Partitions". In: *Journal of Cybernetics* 4.1, pp. 95–104.

Ester, M. et al. (1996). "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. Ed. by U. M. Fayyad E. Simoudis J. Han. AAAI Press, pp. 226–231.

Fayyad, U. et al. (1996). *Advances in Knowledge Discovery and Data Mining*. pp.1-34. AAAI Press/The MIT Press.

Fraley, C. and A.E. Raftery (1998). "How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis". In: *The Computer Journal* 41.8, pp. 578–588.

— (2002). "Model-based Clustering, Discriminant Analysis and Density Estimation". In: *Journal of the American Statistical Association* 97.458, pp. 611–631.

Goodman, L. and W. Kruskal (1954). "Measures of Associations for Cross-Validations". In: *Journal of the American Statistical Association* 49, pp. 732–764.

Hand, D.J., H. Manilla, and P. Smith (2001). *Principles of Data Mining*. New York: MIT Press.

Hanesh, M., R. Sholger, and M.J Dekkers (1984). "The Application of Fuzzy C-Means Cluster Analysis and Non-Linear Mapping to a Soil Data Set for the Detection of Polluted Sites". In: *Physics and Chemistry of the Earth* 26.11-12, pp. 885–891.

Hebb, D.O. (1949). *The Organisation of Behavior*. New York: Wiley and Sons.

J.H., Holland. *Adaptation in Natural and Artificial Systems*. Ann Arbor, Michigan: University of Michigan Press.

Kass, G.V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data". In: *Applied Statistics* 29.2, pp. 119–127.

Kohonen, T. (1982). "Self-Organized Formation of Topologically Correct Feature Maps". In: *Biological Cybernetics* 43, pp. 59–69.

Milligan, G. and M. Cooper (1985). "An Examination of Procedures of Determining the Number of Cluster in a Data Set". In: *Psychometrika* 50.2, pp. 159–179.

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

Quinlan, J.R. (1986). "Induction of Decision Trees". In: *Machine Learning*, pp. 81–106.

— (1993). *C4.5: programs for machine learning*. San Matteo: Morgan Kaufman.

Rousseeuw, P.J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20.1, pp. 53–65.
Simoudis, E. (1998). "Discovering Data Mining: From Concept to Implementation". In: ed. by P. Cabena et al. Upper Saddle River, NJ: Prentice Hall.
Ward, J. (1963). "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301, pp. 236–244.
Widrow, B. and M.E. Hoff (1960). "Adaptive Switching Circuits". In: *IRE WESCON Convention Record*. Vol. IV, pp. 96–104.
Witten, I.W. and E. Frank (2005). *Data Mining. Practical Machine Learning Tools and Techniques*. Second Edition. San Francisco, CA: Morgan Kaufman Publisher.
Zhang, T., R. Ramakrishnan, and M. Livny (1996). "BIRCH: An Efficient Data Clusterig Method for Very Large Databases". In: *SIGMOD '96 Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*. ACM Press, pp. 103–114.

Transportation Applications of Data Mining

Attoh-Okine, N. (2002). "Combining Use of Rough Set and Artificial Neural Networks in Doweled-Pavement-Performance Modeling. A Hybrid Approach". In: *Journal of Transportation Engineering* 128.3, pp. 270–275.
Attoh-Okine, N.O. (1997). "Rough Set Application to Data Mining Principles in Pavement Management Database". In: *Journal of Computing in Civil Engineering* 11.4, pp. 231–237.
Barai, S.K. (2003). "Data Mining Application in Transportation Engineering". In: *Transport* 18.5, pp. 216–223.
Chang, L.-Y. (2005). "Analysis of Freeway Accident Frequencies: Negative Binomial Regression Versus Artificial Neural Network". In: *Safety Science* 43, pp. 541–557.
Chang, L.-Y. and W.-C. Chen (2005). "Data Mining of Tree-based Models to Analyze Freeway Accident Frequency". In: *Journal of Safety Research* 36, pp. 365–375.
Chen, C.-Y. et al. (2004). "Using Data Mining Techniques on Fleet Management System". In: *24th Annual ESRI International User Conference*. Ed. by ESRI.
Cheng, T., D. Cui, and P. Cheng (2003). "Data Mining for Air Traffic Flow Forecasting: A Hybrid Model of Neural-Network and Statistical Analysis". In: *2003 IEEE International Conference on Intelligent Transportation Systems*. Ed. by IEEE. Shanghai, China, pp. 211–215.
Cheng, T. and J. Wang (2008). "Integrated Spatio-temporal Data Mining for Forest Fire Prediction". In: *Transactions in GIS* 5.12, pp. 591–611.

Dougherty, M. and M. Cobbett (1997). "Short-term Inter-urban Traffic Forecasts Using Neural Networks". In: *International Journal of Forecasting* 13, pp. 21–31.

Flahaut, B. (2004). "Impact of Infrastructure and Local Environment on Road Unsafety. Logistic Modeling with Spatial Autocorrelation". In: *Accident Analysis and Prevention* 36, pp. 1055–1066.

Flahaut, B. et al. (2003). "The Local Spatial Autocorrelation and the Kernel Method for Identifying Black Zones. A Comparative Approach". In: *Accident Analysis and Prevention* 35, pp. 991–1004.

Getis, A. (2008). "A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective". In: *Geographical Analysis* 40, pp. 297–309.

Geurts, K., I. Thomas, and G. Wets (2005). "Understanding Spatial Concentrations of Road Accidents using Frequent Item Sets". In: *Accident Analysis and Prevention* 37, pp. 787–799.

Gong, X. and X. Liu (2003). "A Data Mining Based Algorithm for Traffic Network Flow Forecasting". In: *KIMAS 2003*. Ed. by IEEE. Boston, MA, USA, pp. 243–248.

Gong, X. and Y. Lu (2008). "Data Mining Based Research on Urban Tide Traffic Problem". In: *11th International IEEE Conference on Intelligent Transportation Systems*. Ed. by IEEE. Beijing, China.

Gürbüz, F., L. Özbakir, and H. Yapici (2009). "Classification Rule Discovery for the Aviation Incidents Resulted in Fatality". In: *Knowledge-Based Systems* 22.8, pp. 622–632.

Haluzová, P. (2008). "Effective Data Mining for a Transportation Information System". In: *Acta Polytechnica* 48.1. Czech Technical University Publishing House, pp. 24–29.

Han, C. and S. Song (2003). "A Review of Some Main Models for Traffic Flow Forecasting". In: *2003 IEEE International Conference on Intelligent Transportation Systems*. Ed. by IEEE. Shanghai, China, pp. 216–219.

Hayashi, K. et al. (2005). "Individualized Drowsiness Detection During Driving by Pulse Wave Analysis with Neural Network". In: *IEEE Intelligent Transportation Systems*. Ed. by IEEE, pp. 901–906.

Hornik, K. (1989). "Multilayer Feedforward Networks Are Universal Approximators". In: *Neural Networks* 2, pp. 359–366.

Hu, M. et al. (2003). "Development of the Real-time Evaluation and Decision Support System for Incident Management". In: *2003 IEEE International Conference on Intelligent Transportation Systems*. Shanghai, China: IEEE, pp. 426–431.

Jiang, Z. and Y.-X. Huang (2009). "Parametric Calibration of Speed-Density Relationships in Mesoscopic Traffic Simulator with Data Mining". In: *Information Sciences* 179, pp. 2002–2013.

Kalyoncuoglu, S. and M. Tigdemir (2004). "An Alternative Approach for Modelling and Simulation of Traffic Data: Artificial Neural Networks". In: *Simulation Modelling Practice and Theory* 12, pp. 351–362.

Khan, G., X. Qin, and D. Noyce (2008). "Spatial Analysis of Weather Crash Patterns". In: *Journal of Transportation Engineering* 134.5, pp. 191–202.

Ladner, R., F. Petry, and M. Cobb (2003). "Fuzzy Set Approaches to Spatial Data Mining of Association Rules". In: *Transactions in GIS* 1.7), pages = 123-138.

Leu, S.-S., C.-N. Chen, and S.-L. Chang (2001). "Data Mining for Tunnel Support Stability: Neural Network Approach". In: *Automation in Construction* 10.4, pp. 429–441.

Maciejewski, H. and T. Lipnicki (2008). "Data Exploration Methods for Transport System Dependability Analysis". In: *Third International Conference on Dependability of Computer Systems DepCoS-RELCOMEX 2008*. IEEE, pp. 398–405.

Mennis, J. and J. Liu (2005). "Mining Association Rules in Spatio-Temporal Data: An Analysis of Urban Socioeconomic and Land Cover Change". In: *Transactions in GIS* 1.9, pp. 5–17.

Miller, H. and J. Han (2009). *Geographic Data Mining and Knowledge Discovery*. Second Edition. Chapman Hall/CRC Data Mining and Knowledge Discovery Series.

Pande, A. and M. Abdel-Aty (2006). "Assessment of Freeway Traffic Parameters Leading to Lane-change Related Collisions". In: *Accident Analysis and Prevention* 38, pp. 936–948.

— (2009). "Market Basket Analysis of Crash Data From Large Jurisdictions and its Potential as a Decision Support Tool". In: *Safety Science* 47, pp. 145–154.

Rahman, F.A., M.I. Desa, and A. Wibowo (2011). "A Review of KDD-Data Mining Framework and its Application in Logistics and Transportation". In: *7th International Conference on Networked Computing and Advanced Information Management (NCM), 2011*, pp. 175 –180.

Sarasua, W.A. and X. Jia (1995). "Framework for Integrating GIS-T with KBES: a Pavement Management System Example". In: *Transportation Research Record: Journal of the Transportation Research Board* 1497. Transportation Research Board of the National Academies, Washington, D.C., pp. 153–163.

Soibelman, L. and H. Kim (2000). "Generating Construction Knowledge with Knowledge Discovery in Databases". In: *Computing in Civil and Building Engineering* 2, pp. 906–913.

Van der Voort, M., M.S. Dougherty, and S.M. Watson (1996). "Combining Kohonen Maps with ARIMA Time Series Models to Forecast Traffic Flow". In: *Transportation Research C* 4.5, pp. 307–318.

Wang, Y. et al. (2006). "Dynamic Traffic Prediction Based on Traffic Flow Mining". In: *6th World Congress on Intelligent Control and Automation*. Ed. by IEEE. Dalian, China, pp. 6078–6081.

Xu, W., Y. Qin, and H. Huang (2004). "A New Method of Railway Passenger Flow Forecasting Based on Spatio-Temporal Data Mining". In: *7th*

International IEEE Conference on Intelligent Transportation Systems. Ed. by IEEE. Washington, D.C., USA, pp. 402–405.

Yang, X. and M. Tong (2008). "Knowledge Discovery in Information-Service-Oriented Experimental Transportation System". In: *2008 International Conference on Intelligent Computation Technology and Automation*. IEEE, pp. 787–790.

Zhang, C., Y. Huang, and G. Zong (2010). "Study on the Application of Knowledge Discovery in Data Bases to the Decision Making of Railway Traffic Safety in China". In: *International Conference on Management and Service Science (MASS), 2010*, pp. 1 –10.

Zhaohui, W. et al. (2003). "Application of Knowledge Management in Intelligent Transportation Systems". In: *2003 IEEE International Conference on Intelligent Transportation Systems*. Shanghai, China: IEEE, pp. 1730–1734.

Zhou, G. et al. (2010). "Integration of GIS and Data Mining Technology to Enhance the Pavement Management Decision-Making". In: *Journal of Transportation Engineering* 136.4, pp. 332–341.

Road Traffic Monitoring

AASHTO, Join Task Force on Traffic Monitoring Standards (1992). *AASHTO Guidelines for Traffic Data Programs*. Tech. rep. AASHTO.

Bodle, R.R. (1967). "Evaluation of Rural Coverage Count Duration for Estimating Annual Average Daily Traffic". In: *Highway Research Record, HRB, National Research Council* 199. Washington, D.C., pp. 67–77.

Cambridge Systematics Corporation (1994). *Use of Data from Continuous Monitoring Sites*. Tech. rep. Prepared for FHWA, Volume I and II, Documentation, FHWA.

Davis, G.A. (1996). *Estimation Theory Approaches to Monitoring and Updating Average Daily Traffic*. Tech. rep. Rep.No. 97-05. Center of Transportation Studies, University of Minnesota, Minneapolis.

— (1997). "Accuracy of Estimates of Mean Daily Traffic: A Review". In: *Transportation Research Record: Journal of the Transportation Research Board* 1593. Transportation Research Board of the National Academies, Washington, D.C., pp. 12–16.

Davis, G.A. and Y. Guan (1996). "Bayesian Assignment of Coverage Count Locations to Factor Groups and Estimation of Mean Daily Traffic". In: *Transportation Research Record: Journal of the Transportation Research Board* 1542. Transportation Research Board of the National Academies, Washington, D.C., pp. 30–37.

Della Lucia, L. (2000). *Campagna di monitoraggio del traffico sulla rete di interesse regionale 1999-2000*. Tech. rep. In Italian. Padova: Dipartimento Costruzioni e Trasporti.

Faghri, A., M. Glaubitz, and J. Parameswaran (1996). "Development of Integrated Traffic Monitoring System for Delaware". In: *Transportation Research Record: Journal of the Transportation Research Board* 1536. Transportation Research Board of the National Academies, Washington, D.C., pp. 40–44.

Faghri, A. and J. Hua (1995). "Roadway Seasonal Classification Using Neural Networks". In: *Journal of Computing in Civil Engineering* 9.3, pp. 209–215.

FHWA (2001). *Traffic Monitoring Guide*. Tech. rep. U.S. Department of Transportation.

— (2005). *HPMS Field Manual*.

Flaherty, J. (1993). "Cluster Analysis of Arizona Automatic Traffic Record Data". In: *Transportation Research Record: Journal of the Transportation Research Board* 1410. Transportation Research Board of the National Academies, Washington, D.C., pp. 93–99.

Gecchele, G. et al. (2011). "Data Mining Methods for Traffic Monitoring Data Analysis: a Case Study". In: *Procedia Social and Behavioral Sciences* 20, pp. 455–464.

Gecchele, G. et al. (2012). "Advances in Uncertainty Treatment in the FHWA Procedure for Estimating Annual Average Daily Traffic Volume". In: *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, Washington, D.C. Accepted for Publication.

Hallenbeck et al. (1997). *Vehicle Volume Distributions by Classification*. Tech. rep. FHWA-PL-97-025. FHWA.

Hallenbeck, M. and A. O'Brien (1994). *Truck Flows and Loads for Pavement Management*. Tech. rep. FHWA.

Klir, G.J. and M.J. Wierman (1999). *Uncertainty-Based Information. Elements of Generalized Information Theory*. 2nd edition. Heidelberg, Germany: Physica-Verlag.

Klir, G.J. and B. Yuan (1995). *Fuzzy sets and fuzzy logic. Theory and applications*. Upper Saddle River, New Jersey, USA: Prentice Hall PTR.

Li, M.-T., F. Zhao, and L.F. Chow (2006). "Assignment of Seasonal Factor Categories to Urban Coverage Count Stations Using a Fuzzy Decision Tree". In: *Journal of Transportation Engineering* 132.8, pp. 654–662.

Li, M.-T. et al. (2003). "Evaluation of Agglomerative Hierarchical Clustering Methods". In: *Presented at the 82th Annual Meeting of the Transportation Research Board*. Transportation Research Board of the National Academies, Washington, D.C.

Lingras, P. (1995). "Hierarchical Grouping Versus Kohonen Neural Networks". In: *Journal of Transportation Engineering* 121.4, pp. 364–368.

— (2001). "Statistical and Genetic Algorithms Classification of Highways". In: *Journal of Transportation Engineering* 127.3, pp. 237–243.

Lingras, P. and M. Adamo (1996). "Average and Peak Traffic Volumes: Neural Nets, Regression, Factor Approaches". In: *Journal of Computing in Civil Engineering* 10.4, pp. 300–306.

Matheron, G. (1963). "Principles of Geostatistics". In: *Economic Geology* 58, pp. 1246–1266.

Ministero dei Lavori Pubblici. *Sistemi di Monitoraggio del Traffico. Linee guida per la progettazione*. Ispettorato Generale per la Circolazione e la Sicurezza Stradale.

Mohamad, D. et al. (1998). "Annual Average Daily Traffic Prediction Model for County Roads". In: *Transportation Research Record: Journal of the Transportation Research Board* 1617. Transportation Research Board of the National Academies, Washington, D.C., pp. 69–77.

Ritchie, S. (1986). "Statistical Approach to Statewide Traffic Counting". In: *Transportation Research Record: Journal of the Transportation Research Board* 1090. Transportation Research Board of the National Academies, Washington, D.C., pp. 14–21.

Rossi, R., M. Gastaldi, and G. Gecchele (2011). "FHWA Traffic Monitoring Approach. Innovative procedures in a study case". In: *SIDT Scientific Seminar*. Venice.

Schneider, W.H. and I. Tsapakis (2009). *Review of Traffic Monitoring Factor Groupings and the Determination of Seasonal Adjustment Factors for Cars and Trucks*. Tech. rep. FHWA/OH-2009/10A. Project performed in cooperation with the Ohio Department of Transportation and the Federal Highway Administration. Ohio Department of Transportation.

Selby, B. and K.M. Kockelman (2011). "Spatial Prediction of AADT at Unmeasured Locations by Universal Kriging". In: *Presented at the 90th Annual Meeting of the Transportation Research Board*. Transportation Research Board of the National Academies, Washington, D.C.

Sharma, S.C. and R. Allipuram (1993). "Duration and Frequency of Seasonal Traffic Counts". In: *Journal of Transportation Engineering* 119.3, pp. 344–359.

Sharma, S.C., B.M. Gulati, and S.N. Rizak (1996). "Statewide Traffic Volume Studies and Precision of AADT Estimates". In: *Journal of Transportation Engineering* 122.6, pp. 430–439.

Sharma, S.C. and A. Werner (1981). "Improved Method of Grouping Provincewide Permanent Traffic Counters". In: *Transportation Research Record: Journal of the Transportation Research Board* 815. Transportation Research Board of the National Academies, Washington, D.C., pp. 12–18.

Sharma, S.C. et al. (1986). "Road Classification According to Driver Population". In: *Transportation Research Record: Journal of the Transportation Research Board* 1090. Transportation Research Board of the National Academies, Washington, D.C., pp. 61–69.

Sharma, S.C. et al. (1999). "Neural Networks as Alternative to Traditional Factor Approach to Annual Average Daily Traffic Estimation from Traf-

fic Counts". In: *Transportation Research Record: Journal of the Transportation Research Board* 1660. Transportation Research Board of the National Academies, Washington, D.C., pp. 24–31.

Sharma, S.C. et al. (2000). "Estimation of Annual Average Daily Traffic on Low-Volume Roads". In: *Transportation Research Record: Journal of the Transportation Research Board* 1719. Transportation Research Board of the National Academies, Washington, D.C., pp. 103–111.

Sharma, S.C. et al. (2001). "Application of Neural Networks to Estimate AADT on Low-Volume Roads". In: *Journal of Transportation Engineering* 127.5, pp. 426–432.

Tsapakis, I. et al. (2011). "Discriminant Analysis for Assigning Short-Term Counts to Seasonal Adjustment Factor Groupings". In: *Transportation Research Record: Journal of the Transportation Research Board* 2256. Transportation Research Board of the National Academies, Washington, D.C., pp. 112–119.

Wang, X. and K.M. Kockelman (2009). "Forecasting Network Data: Spatial Interpolation of Traffic Counts Using Texas Data". In: *Transportation Research Record: Journal of the Transportation Research Board* 2105. Transportation Research Board of the National Academies, Washington, D.C., pp. 100–108.

Xia, Q. et al. (1999). "Estimation of Annual Average Daily Traffic for Nonstate Roads in a Florida County". In: *Transportation Research Record: Journal of the Transportation Research Board* 1660. Transportation Research Board of the National Academies, Washington, D.C., pp. 32–40.

Zhao, F. and S. Chung (2001). "Contributing Factors of Annual Average Daily Traffic in a Florida County: Exploration with Geographic Information System and Regression Models". In: *Transportation Research Record: Journal of the Transportation Research Board* 1769. Transportation Research Board of the National Academies, Washington, D.C., pp. 113–122.

Zhao, F., M.-T. Li, and L.F. Chow (2004). *Alternatives for Estimating Seasonal factors on Rural and Urban Roads in Florida*. Tech. rep. Final Report. Research Office. Florida Department of Transportation.

Zhao, F. and N. Park (2004). "Using Geographically Weighted regression Models to Estimate Annual Average Daily Traffic". In: *Transportation Research Record: Journal of the Transportation Research Board* 1879. Transportation Research Board of the National Academies, Washington, D.C., pp. 99–107.

Appendix A

Case Study Data

Figure A.1: *Road Section Data Summary (SITRA Monitoring Program)*

PROVINCIA DI VENEZIA - MONITORAGGIO TRAFFICO 2000-2005

SS 14 "della Venezia Giulia" a Portegrandi (km 16+850)

Strada	SS 14 della Venezia Giulia
Codice sezione	0168
Progressiva chilometrica	16+850
Località	Portegrandi
Comune	Quarto d'Altino
Direzione A	verso Portegrandi - San Donà di Piave
Direzione B	verso Mestre - Venezia
Limite di velocità	90 km/h
Larghezza carreggiata	n.d.

Parametri		Anno 2000	2001	2002	2003	2004	2005
Traffico Diurno Medio (bidirezionale)	$TDM_{feriale}$	-	-	12.544	13.265	13.248	12.943
	TDM_{sabato}	-	-	13.338	14.104	14.086	13.762
	$TDM_{festivo}$	-	-	11.284	11.933	11.917	11.643
	TDM	-	-	12.478	13.195	13.177	12.874
Traffico Giornaliero Medio (bidirezionale)	$TGM_{feriale}$	-	-	17.440	17.662	17.905	17.463
	TGM_{sabato}	-	-	20.490	20.750	21.036	20.517
	$TGM_{festivo}$	-	-	18.959	19.200	19.465	18.984
	TGM	-	-	18.093	18.323	18.575	18.116
Flusso 30° Ora	*Direzione A*	-	-	1.350	1.287	1.275	1.274
	Direzione B	-	-	1.193	1.365	1.349	1.427
	Direzione A+B	-	-	1.936	2.269	2.169	2.110
Ora di Punta 7.00 – 9.00	*Direzione A*	-	-	701	*967*	707	818
	Direzione B	-	-	1.381	*1.819*	1.437	1.449
	Direzione A+B	-	-	2.082	*2.786*	2.144	2.267
Ora di Punta 17.00 – 19.00	*Direzione A*	-	-	1.258	*1.608*	1.356	1.356
	Direzione B	-	-	868	*1.099*	945	1.083
	Direzione A+B	-	-	2.125	*2.707*	2.302	2.439
Velocità	*V10 (km/h)*	-	-	116	100	105	107
	V50 (km/h)	-	-	87	81	82	83
Composizione veicolare	*Autovetture*	-	-	82,10%	87,11%	84,75%	83,80%
	Comm. leggeri	-	-	10,05%	6,60%	8,25%	8,32%
	Comm. pesanti	-	-	7,85%	6,29%	7,00%	7,88%

N.B.: Per le spiegazioni dei parametri consultare pag. 20
I dati in corsivo sono stimati su un numero ridotto di giornate di rilievo

Figure A.2: *Province of Venice AVC Sites (year 2012)*

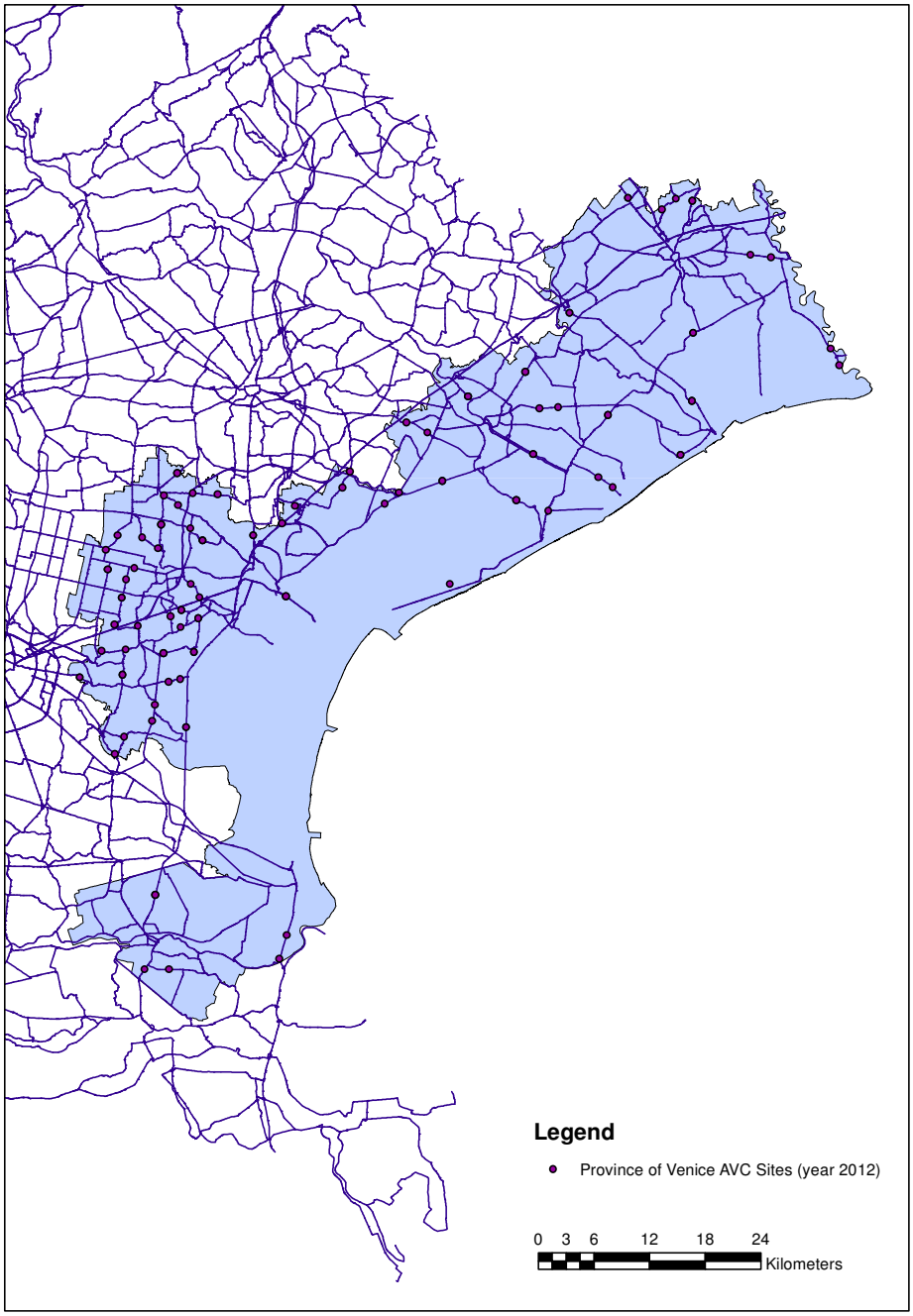


Figure A.3: *Case Study AVC Sites*

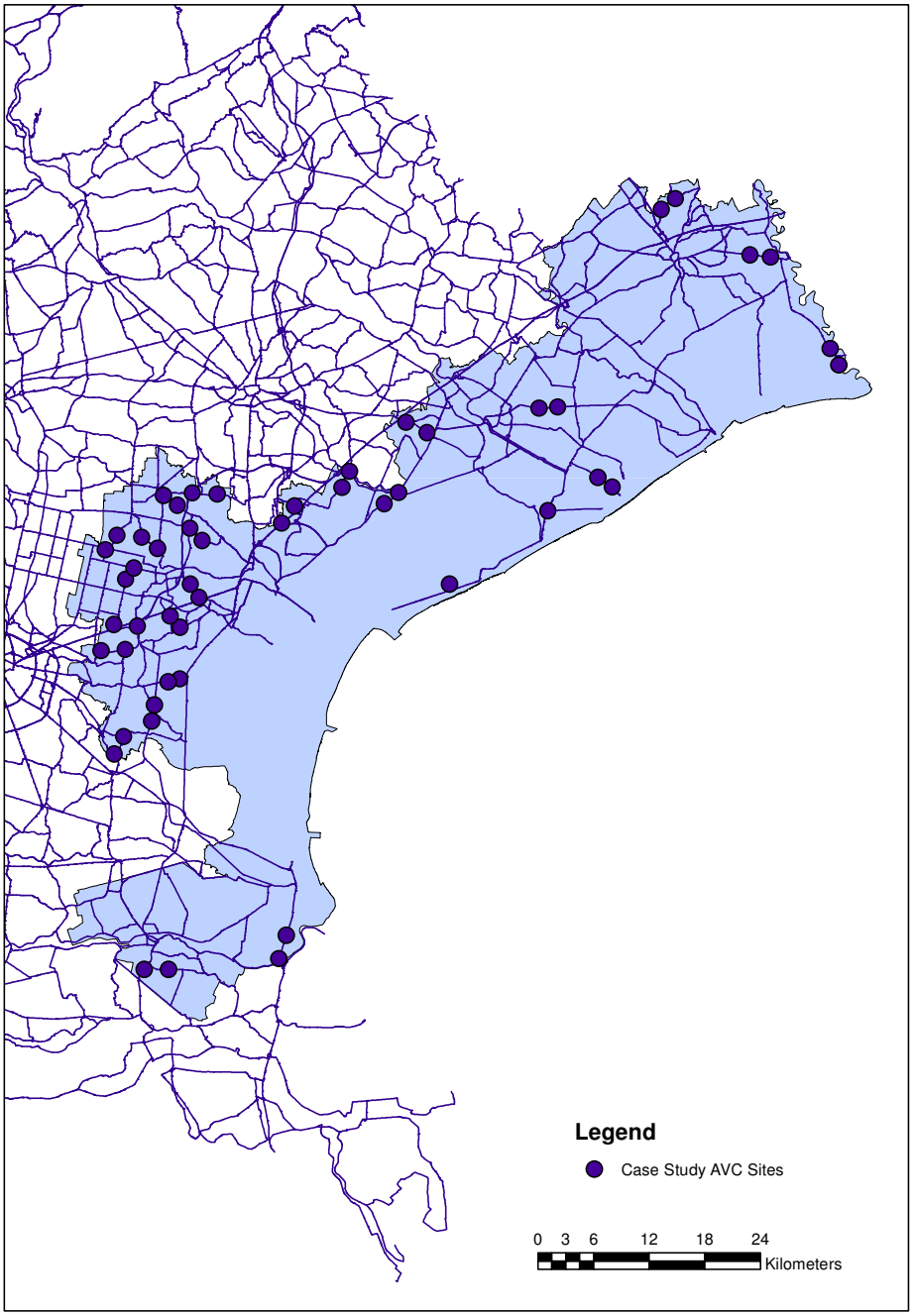


Table A.1: *AVC Sites Sample Size*

AVC Number	AVC Site		24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Road Section	Direction	Total	Weekdays	Week-ends	Total	Weekdays	Week-ends	Total	Weekdays	Week-ends
1	ANASS014h0168	A	150	99	51	97	73	24	67	49	18
2	ANASS014h0168	B	150	99	51	97	73	24	67	49	18
3	ANASS309h0800	A	349	250	99	241	194	47	187	142	45
4	ANASS309h0800	B	349	250	99	241	194	47	187	142	45
5	VNTSR011h4005	A	342	247	95	232	189	43	179	137	42
6	VNTSR011h4005	B	342	247	95	232	189	43	179	137	42
7	VNTSR089h0142	A	115	76	39	74	55	19	50	36	14
8	VNTSR089h0142	B	113	74	39	73	54	19	50	36	14
9	VNTSR245h0107	A	343	246	97	237	191	46	185	140	45
10	VNTSR245h0107	B	343	246	97	237	191	46	185	140	45
11	VNTSR515h0239	A	345	246	99	236	189	47	183	137	46
12	VNTSR515h0239	B	345	246	99	236	189	47	183	137	46
13	VNTSR516h0413	B	348	251	97	240	195	45	187	143	44
14	xVESP012h0126	A	343	244	99	232	185	47	178	133	45
15	xVESP012h0126	B	342	243	99	229	182	47	174	129	45
16	xVESP013h0075	A	335	240	95	230	185	45	178	134	44
17	xVESP013h0075	B	332	237	95	227	182	45	175	131	44
18	xVESP018h0025	B	334	239	95	230	185	45	179	135	44
19	xVESP026h0021	A	349	250	99	241	194	47	188	142	46
20	xVESP026h0021	B	349	250	99	241	194	47	188	142	46
21	xVESP027h0017	A	343	247	96	235	191	44	183	140	43
22	xVESP027h0017	B	344	248	96	236	192	44	183	140	43
23	xVESP032h0092	A	350	251	99	242	195	47	189	143	46
24	xVESP032h0092	B	350	251	99	242	195	47	189	143	46
25	xVESP035h0065	A	344	248	96	235	191	44	182	139	43
26	xVESP035h0065	B	344	248	96	235	191	44	182	139	43
27	xVESP036h0045	A	293	210	83	198	159	39	152	114	38
28	xVESP036h0045	B	293	210	83	198	159	39	152	114	38
29	xVESP039h0034	A	339	242	97	232	186	46	180	135	45
30	xVESP039h0034	B	338	241	97	232	186	46	180	135	45
31	xVESP040h0102	A	244	172	72	159	126	33	118	86	32
32	xVESP040h0102	B	244	172	72	159	126	33	118	86	32
33	xVESP041h0045	A	145	95	50	92	68	24	63	45	18
34	xVESP041h0045	B	145	95	50	92	68	24	63	45	18
35	xVESP054h0048	B	305	218	87	212	171	41	165	127	38
36	xVESP074h0117	A	338	239	99	229	182	47	178	131	47
37	xVESP074h0117	B	343	244	99	234	187	47	185	138	47
38	xVESP081h0022	B	332	237	95	227	182	45	175	132	43
39	xVESP090h0033	A	341	242	99	232	185	47	180	134	46
40	xVESP090h0033	B	344	245	99	235	188	47	182	137	45
41	xVESP093h0024	A	270	175	95	180	136	44	129	101	28
42	xVESP093h0024	B	288	193	95	191	147	44	141	107	34

Figure A.4: *Road Groups Identified with FCM Algorithm*

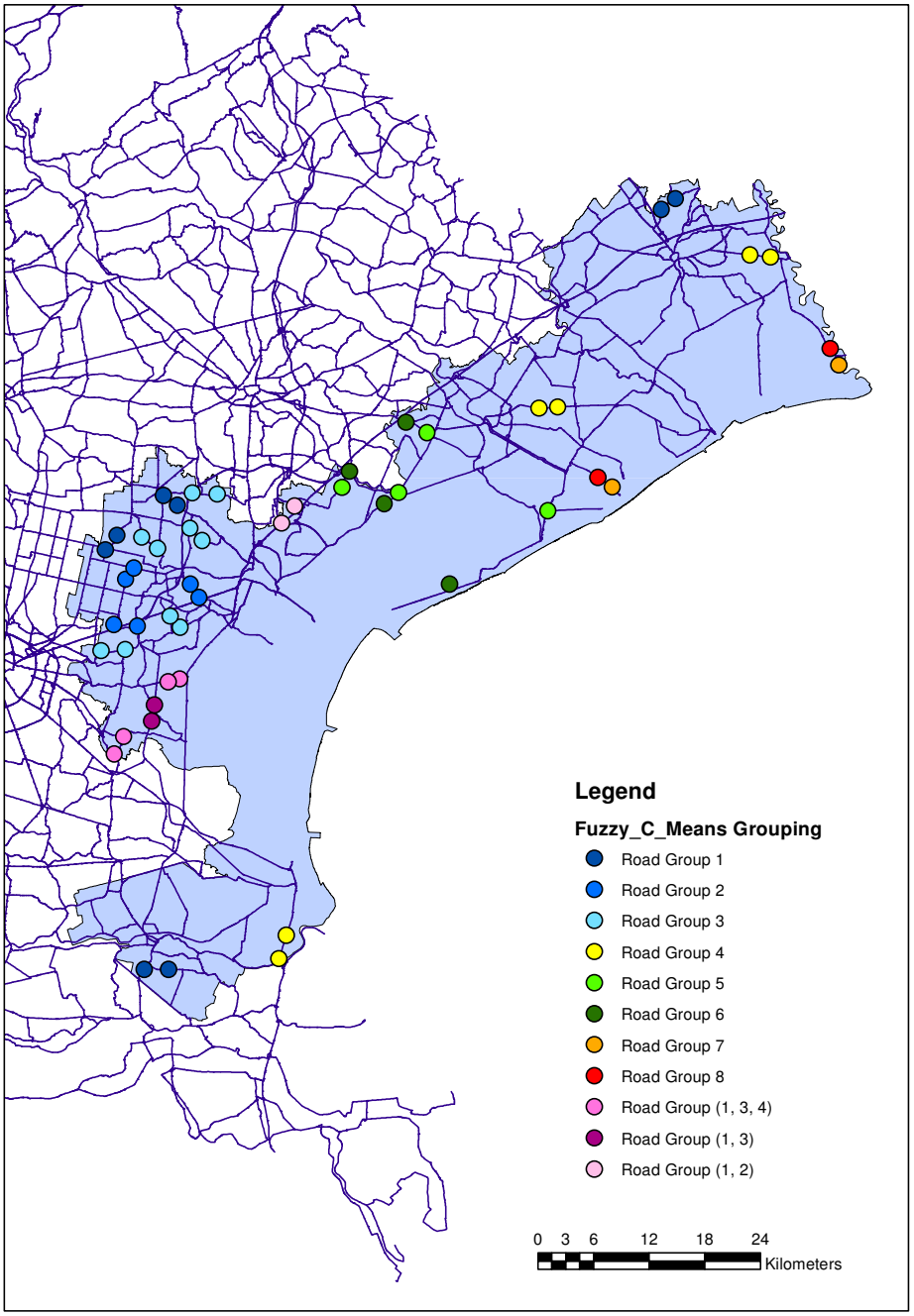


Table A.2: *Average Seasonal Adjustment Factors for Road Groups*

Road Group	AVC sites	Jan/Feb			Mar/Apr			May/Jun			Jul/Aug			Sep/Oct			Nov/Dec		
		Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun
Group 1	7	1.04	0.96	1.18	0.97	0.90	1.05	0.97	0.93	1.09	1.06	0.97	1.22	0.97	0.88	1.03	0.98	0.95	0.98
Group 2	5	0.90	0.88	1.36	0.86	0.84	1.29	0.88	0.89	1.43	1.01	1.00	1.69	0.86	0.84	1.30	0.85	0.85	1.19
Group 3	10	0.94	0.99	1.21	0.87	0.92	1.10	0.87	0.93	1.12	0.98	1.07	1.34	0.90	0.94	1.12	0.89	0.97	1.14
Group 4	3	1.34	1.14	1.23	1.19	0.99	1.02	1.07	0.78	0.82	0.90	0.66	0.79	1.20	0.96	1.04	1.28	1.16	1.20
Group 5	3	1.59	1.46	1.36	1.38	1.14	1.04	0.94	0.56	0.73	0.80	0.53	0.68	1.37	1.11	1.25	1.49	1.51	1.41
Group 6	3	1.55	1.42	1.31	1.37	1.28	0.99	1.00	0.76	0.52	0.84	0.70	0.53	1.36	1.16	1.16	1.41	1.47	1.34
Group 7	2	2.14	2.08	1.85	1.55	1.32	1.15	0.94	0.52	0.58	0.66	0.44	0.53	1.40	1.13	1.17	2.05	2.12	2.32
Group 8	2	2.13	2.09	1.83	1.56	1.46	1.13	0.97	0.69	0.47	0.66	0.53	0.44	1.39	1.18	0.99	2.05	2.16	2.29

"Jan/Feb" = "January/February". "Mar/Apr" = "March/April". "May/Jun" = "May/June". "Jul/Aug" = "July/August". 'Sep/Oct" = "September/October". "Nov/Dec"=November/December"
"Wday" = "Weekdays". "Sat" = "Saturdays". "Sun" = "Sundays"

Table A.3: *Average Reciprocals of the Seasonal Adjustment Factors for Road Groups*

Road Group	AVC sites	Jan/Feb			Mar/Apr			May/Jun			Jul/Aug			Sep/Oct			Nov/Dec		
		Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun	Wday	Sat	Sun
Group 1	7	0.96	1.04	0.85	1.03	1.11	0.95	1.03	1.07	0.92	0.94	1.03	0.82	1.03	1.14	0.97	1.02	1.05	1.02
Group 2	5	1.11	1.13	0.74	1.16	1.19	0.77	1.14	1.13	0.70	0.99	1.00	0.59	1.16	1.18	0.77	1.18	1.17	0.84
Group 3	10	1.07	1.01	0.82	1.15	1.08	0.91	1.15	1.07	0.89	1.02	0.94	0.75	1.11	1.07	0.89	1.12	1.03	0.88
Group 4	3	0.75	0.87	0.81	0.84	1.01	0.98	0.94	1.29	1.22	1.11	1.52	1.27	0.83	1.04	0.96	0.78	0.86	0.83
Group 5	3	0.63	0.69	0.73	0.72	0.88	0.96	1.07	1.80	1.37	1.24	1.90	1.47	0.73	0.90	0.80	0.67	0.66	0.71
Group 6	3	0.65	0.71	0.76	0.73	0.78	1.01	1.00	1.31	1.93	1.19	1.42	1.88	0.74	0.86	0.86	0.71	0.68	0.75
Group 7	2	0.47	0.48	0.54	0.65	0.76	0.87	1.07	1.91	1.71	1.52	2.26	1.87	0.71	0.88	0.85	0.49	0.47	0.43
Group 8	2	0.47	0.48	0.55	0.64	0.68	0.89	1.03	1.45	2.15	1.52	1.89	2.26	0.72	0.85	1.01	0.49	0.46	0.44

'Jan/Feb" = "January/February". "Mar/Apr" = "March/April". "May/Jun" = "May/June". "Jul/Aug" = "July/August". 'Sep/Oct" = "September/October". "Nov/Dec"=November/December"
"Wday" = "Weekdays". "Sat" = "Saturdays". "Sun" = "Sundays"

Table A.4: *MAE of AVC Sites. "Total" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	11.10	15.72	14.82	8.85	10.05	9.23	8.71	9.19	8.88
2	9.81	15.15	13.93	7.60	9.60	7.57	7.20	7.82	7.42
3	8.43	11.89	11.09	7.60	8.01	7.56	7.04	7.01	7.00
4	9.14	10.18	9.83	8.24	8.72	8.52	7.55	7.68	7.66
5	5.93	7.42	6.84	4.79	4.85	4.68	4.42	4.36	4.11
6	5.28	6.25	5.61	4.49	4.44	4.41	4.19	4.18	4.13
7	10.51	15.10	14.65	9.05	10.06	9.54	8.24	8.39	8.35
8	8.08	11.02	10.29	6.74	6.89	6.57	6.93	6.84	6.58
9	5.34	6.24	5.59	4.35	4.60	4.35	4.09	4.26	3.97
10	5.06	7.52	7.29	4.20	4.46	4.32	4.01	4.41	4.09
11	5.70	7.05	6.30	4.89	5.24	4.79	4.59	4.81	4.45
12	6.49	8.13	7.00	5.47	6.08	5.60	5.19	5.38	4.99
13	4.80	6.27	5.67	3.82	4.20	3.85	3.57	3.70	3.62
14	6.33	9.87	9.34	4.42	7.57	7.49	4.19	6.28	6.15
15	6.49	10.34	9.67	5.10	6.10	5.40	4.92	5.22	4.86
16	4.05	6.98	5.83	3.34	4.90	4.77	3.09	4.03	4.01
17	4.23	21.48	18.95	3.39	18.77	15.46	3.18	17.84	14.16
18	4.83	10.08	8.61	3.55	6.08	5.78	3.29	5.35	4.77
19	7.61	11.19	9.03	7.32	8.22	7.44	7.21	7.37	7.32
20	5.61	9.25	7.26	4.66	5.40	4.67	4.37	4.59	4.31
21	5.13	7.07	6.18	4.27	5.42	4.80	4.02	4.20	3.92
22	7.00	8.11	7.39	6.54	6.74	6.48	6.44	6.53	6.61
23	5.73	8.13	7.42	5.32	5.39	5.22	5.00	5.06	4.86
24	6.29	9.39	7.92	5.34	6.32	5.20	5.00	5.34	4.98
25	7.25	9.38	8.60	6.82	7.34	7.38	6.45	6.49	6.42
26	7.55	8.40	8.04	7.33	7.37	7.39	6.89	6.92	6.97
27	4.42	5.27	4.75	3.55	3.87	3.56	3.32	3.35	3.27
28	5.49	6.22	5.65	4.28	4.41	4.24	4.08	4.09	4.01
29	6.39	7.13	6.75	5.34	5.35	5.26	5.05	5.23	5.22
30	5.49	5.88	5.42	4.58	4.55	4.24	4.23	4.24	4.01
31	7.93	10.44	9.98	6.59	9.47	8.74	6.19	8.66	8.36
32	10.20	15.03	14.08	9.41	12.15	11.52	8.84	10.86	10.93
33	12.10	18.07	16.69	9.58	11.82	10.59	8.85	10.74	10.15
34	9.94	20.47	21.74	7.93	10.64	9.43	7.53	9.54	9.10
35	7.81	9.98	9.65	5.97	6.64	6.06	5.57	5.89	5.53
36	15.93	21.36	19.81	14.18	16.54	15.80	13.66	14.55	14.27
37	15.02	19.79	15.76	13.43	15.46	14.33	12.97	14.37	12.73
38	6.05	7.28	7.13	5.04	5.37	5.25	5.04	5.15	5.10
39	16.39	19.10	18.42	14.79	16.12	15.48	13.79	14.04	14.11
40	14.73	18.52	17.57	13.44	14.12	13.51	12.83	13.29	13.00
41	5.94	10.08	8.33	4.61	7.17	5.52	4.18	4.77	4.33
42	5.70	12.69	10.23	4.47	6.70	5.88	3.88	5.98	4.22

Table A.5: *SDAE of AVC Sites. "Total" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	10.83	17.09	16.91	8.04	8.80	8.27	7.54	8.28	8.00
2	10.26	22.16	19.61	7.49	11.06	8.40	7.55	8.92	9.07
3	6.67	14.37	13.95	5.44	5.91	5.11	4.85	4.66	4.53
4	6.87	7.53	7.12	5.77	6.09	6.03	5.19	5.36	5.24
5	5.47	12.63	11.35	4.25	4.85	4.71	4.03	3.94	3.72
6	4.00	9.04	9.10	3.35	3.22	3.21	3.22	3.26	3.20
7	8.34	11.28	12.52	6.91	7.62	7.29	6.75	6.88	7.18
8	6.87	14.31	14.80	5.22	6.05	5.05	4.71	4.82	4.72
9	4.39	6.29	5.07	3.87	4.39	3.98	3.66	3.97	3.51
10	4.41	9.71	9.80	3.60	4.23	3.82	3.47	3.93	3.71
11	5.71	7.70	7.07	5.53	6.00	5.49	5.35	5.80	4.99
12	6.31	9.01	6.99	5.11	6.88	5.84	4.91	5.62	4.75
13	4.16	7.74	7.34	3.15	4.18	3.55	2.92	3.40	3.30
14	6.08	9.89	9.89	3.95	10.39	10.56	3.85	6.34	6.31
15	5.41	11.66	11.55	4.20	7.40	6.03	4.07	5.80	5.68
16	3.73	10.06	8.10	3.00	6.43	6.38	2.82	3.50	3.49
17	3.77	26.10	25.01	3.03	23.16	22.43	2.87	26.37	25.22
18	4.41	12.62	8.88	2.92	6.76	6.57	2.56	4.84	3.42
19	5.67	16.65	12.26	4.49	9.70	4.66	4.07	4.22	4.21
20	5.77	13.83	11.48	4.50	6.57	5.00	4.54	4.63	4.25
21	4.93	12.06	11.75	4.39	7.73	6.71	4.25	4.76	4.00
22	4.95	6.73	5.22	4.41	4.49	4.26	4.21	4.20	4.19
23	5.17	15.12	15.62	4.53	4.76	4.44	4.08	4.28	3.71
24	5.16	14.96	14.49	4.55	11.10	4.43	4.09	5.95	4.15
25	5.51	11.59	9.97	5.19	5.87	5.70	4.91	4.91	4.75
26	6.46	7.76	7.51	6.18	6.11	6.18	6.02	5.97	6.02
27	4.30	6.91	6.09	3.57	4.07	3.51	3.26	3.18	3.07
28	4.89	7.75	6.85	3.95	4.17	3.85	3.62	3.65	3.53
29	6.36	9.37	8.97	5.25	5.06	5.00	5.11	5.42	5.44
30	5.74	6.53	6.05	4.79	4.73	4.27	4.58	4.99	4.40
31	6.00	10.15	9.64	4.94	9.84	8.11	4.48	6.78	6.21
32	9.16	21.35	20.96	7.88	10.43	9.40	7.02	8.72	8.79
33	10.12	13.07	14.10	8.30	10.36	9.30	7.36	8.54	8.03
34	8.52	24.24	24.57	6.64	9.77	9.12	5.55	6.52	6.07
35	7.88	11.88	12.43	5.41	7.83	6.79	4.95	5.66	5.35
36	17.25	23.43	23.42	15.13	17.48	17.48	14.13	14.22	14.26
37	16.89	24.02	19.30	15.00	18.31	17.80	14.20	15.68	13.67
38	5.72	8.34	9.30	4.54	5.86	5.98	4.24	4.27	4.31
39	13.86	16.22	16.13	11.49	12.81	12.80	10.60	11.67	11.52
40	13.96	20.63	20.44	12.09	12.59	12.79	11.21	11.39	11.42
41	5.34	13.52	9.95	3.96	12.33	9.77	3.92	4.55	4.23
42	5.59	19.38	16.60	3.65	7.50	7.03	3.20	7.06	3.54

Table A.6: *Maximum Error of AVC Sites. "Total" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	56.87	96.63	96.63	35.75	35.86	35.81	31.17	33.27	31.38
2	53.48	174.61	129.00	35.91	52.24	45.81	32.06	44.77	44.77
3	59.60	135.40	135.40	40.35	40.06	40.06	31.30	30.08	30.08
4	63.07	42.12	42.12	38.62	38.55	38.55	37.58	37.44	37.44
5	49.94	133.35	133.35	33.30	38.71	38.71	27.33	27.13	27.13
6	22.80	133.57	133.57	15.49	15.37	15.37	14.57	15.76	15.76
7	35.12	40.65	40.65	29.54	33.49	29.62	25.54	25.62	25.62
8	39.70	124.02	124.02	21.62	35.50	21.62	18.82	18.68	18.68
9	26.34	50.78	37.64	24.42	24.42	24.42	22.93	22.94	22.94
10	33.07	91.73	91.73	23.17	26.54	23.34	22.41	22.44	22.44
11	40.85	49.10	46.64	40.74	48.18	48.18	39.53	39.55	39.55
12	43.32	65.13	55.07	25.37	50.97	34.71	24.40	30.26	24.54
13	25.52	71.30	71.30	17.98	33.52	21.67	17.34	21.38	21.38
14	33.23	104.43	104.43	22.53	140.25	140.25	17.52	63.71	63.71
15	27.45	116.09	116.09	19.66	51.19	37.20	17.73	42.97	42.97
16	21.17	88.43	63.50	20.77	62.90	62.90	18.35	18.02	18.02
17	20.67	122.63	122.63	17.27	113.37	113.37	16.12	113.54	113.54
18	33.52	115.85	68.52	15.88	74.76	74.76	15.29	36.92	16.35
19	40.53	141.96	135.66	22.75	126.50	24.72	21.24	21.82	21.82
20	36.43	131.18	131.18	24.25	53.57	47.39	22.47	26.56	19.68
21	29.39	125.90	125.90	27.23	54.49	54.49	27.20	33.38	26.47
22	28.23	46.34	30.11	26.71	27.20	26.71	26.95	26.95	26.95
23	38.39	168.74	168.74	28.58	28.05	26.24	27.30	27.40	25.09
24	36.59	185.26	185.26	27.44	154.36	29.19	26.29	60.01	28.04
25	29.14	122.09	122.09	26.43	40.56	40.56	23.88	22.27	22.27
26	29.85	63.48	63.48	26.79	26.79	26.79	24.99	25.03	25.03
27	25.98	55.94	55.94	17.13	28.74	16.97	16.38	16.11	16.11
28	31.46	87.65	87.65	19.98	24.25	19.78	16.72	16.72	16.72
29	45.25	113.43	113.43	28.38	28.27	28.27	24.70	34.56	34.56
30	34.60	42.00	42.00	23.45	26.57	23.05	21.36	35.95	21.60
31	49.78	69.70	69.70	32.21	77.26	75.51	20.25	39.77	29.09
32	46.77	234.44	234.44	34.93	63.78	42.38	30.07	33.11	33.11
33	47.40	65.94	65.94	38.79	49.87	38.71	33.04	32.25	31.88
34	39.30	180.31	180.31	26.56	44.27	44.27	23.99	30.83	26.22
35	49.20	94.21	94.21	35.16	70.78	70.78	29.29	29.26	29.26
36	122.72	162.81	162.81	104.08	105.47	105.47	107.53	107.53	107.53
37	123.88	226.25	153.48	97.10	128.73	128.73	97.25	96.63	96.63
38	36.27	104.15	104.15	25.03	54.86	54.86	24.14	24.27	24.27
39	104.35	100.47	100.47	89.19	89.17	89.17	81.33	81.31	81.31
40	113.57	185.75	185.75	77.65	77.50	77.50	77.00	76.92	76.92
41	31.42	94.21	81.56	26.77	102.01	102.01	23.54	22.97	22.97
42	40.11	142.96	115.03	18.08	41.36	41.36	17.19	37.68	13.95

Table A.7: *Number of Samples of AVC Sites. "Total" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	150	150	111	97	97	82	67	67	58
2	150	150	111	97	97	85	67	67	59
3	349	349	295	241	241	223	187	187	173
4	349	349	294	241	241	224	187	187	178
5	342	342	283	232	232	223	179	179	172
6	342	342	297	232	232	227	179	179	178
7	115	115	72	74	74	64	50	50	44
8	113	113	88	73	73	64	50	50	48
9	343	343	296	237	237	222	185	185	180
10	343	343	227	237	237	209	185	185	165
11	345	345	284	236	236	219	183	183	173
12	345	345	283	236	236	220	183	183	175
13	348	348	296	240	240	222	187	187	179
14	343	343	269	232	231	218	178	177	171
15	342	342	266	229	229	208	174	174	158
16	335	335	295	230	230	223	178	178	173
17	332	332	236	227	225	143	175	173	125
18	334	334	299	230	229	218	179	178	171
19	349	349	292	241	241	232	188	188	185
20	349	349	264	241	241	226	188	188	177
21	343	343	297	235	233	218	183	180	175
22	344	344	294	236	236	222	183	183	179
23	350	350	295	242	242	235	189	189	184
24	350	350	274	242	242	225	189	189	183
25	344	344	277	235	235	223	182	182	174
26	344	344	281	235	235	225	182	182	177
27	293	293	253	198	198	187	152	152	148
28	293	293	274	198	198	188	152	152	151
29	339	339	280	232	231	219	180	178	177
30	338	338	290	232	230	215	180	178	172
31	244	244	210	159	159	149	118	118	114
32	244	244	219	159	159	137	118	118	113
33	145	145	98	92	92	72	63	63	49
34	145	145	100	92	92	69	63	63	48
35	305	305	250	212	208	190	165	159	147
36	338	338	292	229	221	207	178	169	160
37	343	343	273	234	228	204	185	175	162
38	332	332	228	227	227	209	175	175	169
39	341	341	275	232	228	195	180	176	157
40	344	344	285	235	234	200	182	181	162
41	270	270	225	180	141	126	129	83	74
42	288	288	198	191	155	122	141	94	73

Table A.8: *MAE of AVC Sites. "Weekdays" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	10.50	12.95	11.86	8.59	10.12	9.26	7.96	7.81	7.87
2	9.21	10.87	9.35	7.18	8.10	7.49	6.68	6.82	6.22
3	8.25	9.75	9.29	7.73	7.88	7.75	7.27	7.27	7.25
4	9.59	10.31	9.97	8.64	9.11	8.84	8.03	8.20	8.05
5	5.64	6.70	6.20	4.64	4.70	4.51	4.13	3.98	3.70
6	5.22	6.13	5.75	4.45	4.41	4.37	4.01	3.97	3.90
7	11.14	17.33	17.70	9.57	10.53	10.11	8.43	8.73	8.45
8	8.79	9.35	8.68	7.12	6.88	7.01	7.03	6.95	6.59
9	4.67	5.58	4.93	3.93	4.23	3.90	3.68	3.78	3.45
10	4.40	6.06	5.87	3.90	4.08	4.05	3.69	4.11	3.76
11	5.33	6.32	5.84	4.98	5.29	4.85	4.59	4.54	4.35
12	6.35	7.40	6.56	5.43	6.01	5.64	4.82	4.97	4.64
13	4.26	5.24	5.16	3.56	3.91	3.66	3.27	3.46	3.39
14	5.08	8.75	7.94	3.99	6.71	6.54	3.48	5.62	5.48
15	5.86	9.34	8.41	4.71	4.75	4.57	4.48	3.94	3.71
16	3.63	6.30	5.01	3.08	4.58	4.46	2.77	3.54	3.50
17	3.95	26.67	24.25	3.19	21.76	19.04	2.84	19.03	15.25
18	4.62	8.27	7.13	3.48	6.24	5.86	3.13	5.40	5.04
19	8.18	10.32	9.19	7.85	8.69	8.00	7.94	8.11	8.05
20	5.16	9.04	6.87	4.38	5.35	4.55	4.07	4.46	4.11
21	4.59	6.30	5.33	4.12	5.08	4.63	3.92	3.84	3.70
22	7.05	7.97	7.43	6.78	6.84	6.79	6.74	6.83	6.91
23	5.20	5.56	5.02	5.13	5.09	4.99	5.00	4.98	4.70
24	5.43	6.75	5.94	5.17	5.66	5.20	5.02	5.35	4.83
25	7.32	8.61	8.16	7.07	7.51	7.64	6.73	6.69	6.69
26	7.74	8.20	7.77	7.61	7.70	7.69	7.30	7.38	7.40
27	3.85	4.66	4.39	3.25	3.63	3.28	2.97	3.00	2.89
28	5.23	6.13	5.35	4.26	4.38	4.16	3.79	3.81	3.70
29	5.75	6.26	5.63	4.96	5.04	4.95	4.56	4.73	4.73
30	4.88	4.97	4.53	4.24	4.14	3.83	3.89	3.68	3.69
31	7.38	9.60	9.25	6.51	8.65	8.42	6.12	8.17	8.21
32	10.28	14.69	14.10	10.08	12.41	11.72	9.98	11.45	11.58
33	10.57	16.23	14.22	8.44	10.35	9.45	8.00	10.16	8.58
34	8.75	17.70	19.10	7.49	10.14	9.64	7.39	10.21	9.20
35	6.35	7.18	6.72	5.46	5.51	5.31	5.34	5.23	5.09
36	14.87	16.93	16.26	13.59	15.83	15.24	12.67	13.41	13.26
37	13.95	15.75	12.48	12.87	14.14	13.19	12.31	13.45	12.27
38	5.52	6.87	6.63	4.97	5.10	5.01	4.79	4.89	4.84
39	14.83	17.71	16.44	13.65	15.22	14.56	12.72	13.16	13.04
40	13.16	14.96	14.24	12.47	12.82	12.17	12.04	12.58	12.16
41	5.08	10.39	8.09	4.25	6.62	5.76	3.99	4.51	4.12
42	4.65	8.25	7.25	4.15	5.86	5.20	3.73	6.07	4.00

Table A.9: *SDAE of AVC Sites. "Weekdays" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	10.89	13.20	13.54	7.70	9.14	8.60	6.86	6.96	7.06
2	10.68	11.73	11.64	7.69	9.56	8.90	7.46	7.60	7.53
3	6.16	9.70	9.45	5.29	5.28	5.21	4.51	4.37	4.34
4	6.62	6.98	6.62	5.55	5.92	5.82	4.64	4.76	4.62
5	4.42	11.32	9.02	3.81	4.52	4.30	3.50	3.29	2.93
6	3.90	10.05	10.11	3.45	3.32	3.30	3.28	3.33	3.25
7	9.32	11.70	13.30	7.45	8.09	7.66	7.00	7.02	7.18
8	7.66	7.98	7.56	5.56	5.41	5.31	5.10	5.24	5.14
9	3.97	6.30	4.62	3.60	4.27	3.65	3.39	3.67	2.99
10	3.74	7.62	6.40	3.41	3.86	3.67	3.28	3.83	3.51
11	5.75	7.68	7.10	5.79	6.28	5.65	5.82	5.90	5.25
12	5.66	7.51	5.74	5.17	7.04	6.05	4.86	5.44	4.61
13	3.40	6.55	6.60	2.67	3.55	3.29	2.48	3.26	3.22
14	4.74	7.08	6.36	3.65	4.52	4.24	3.20	3.58	3.37
15	5.05	9.00	8.08	3.96	4.98	4.85	3.62	3.36	3.13
16	3.36	10.47	8.28	2.67	6.75	6.69	2.38	2.99	2.94
17	3.50	28.93	28.76	2.76	24.85	25.58	2.51	27.59	26.75
18	4.63	9.34	7.77	2.89	7.26	7.04	2.51	4.10	3.48
19	5.46	13.88	10.73	4.24	9.61	4.41	3.85	4.03	4.02
20	5.15	15.20	12.22	4.25	6.96	5.16	4.03	4.47	3.90
21	4.68	10.90	9.75	4.46	7.67	7.07	4.28	4.24	3.91
22	4.65	6.18	4.80	4.38	4.32	4.36	4.12	4.12	4.10
23	4.67	6.83	4.88	4.32	4.27	3.96	4.07	4.11	3.25
24	4.68	11.51	9.41	4.34	5.74	4.42	4.09	6.20	3.76
25	5.49	8.10	6.19	5.19	5.92	5.94	4.93	4.84	4.71
26	6.48	7.18	6.61	6.27	6.31	6.34	6.14	6.17	6.20
27	3.88	6.57	6.27	3.43	4.06	3.37	3.17	3.08	2.87
28	4.76	8.46	7.27	3.86	4.16	3.74	3.58	3.62	3.44
29	5.24	7.38	5.55	4.59	4.63	4.60	4.37	4.84	4.86
30	4.72	5.30	4.67	4.47	4.45	3.90	4.39	4.12	4.13
31	4.90	8.30	7.88	4.58	8.25	8.06	4.36	5.76	5.78
32	9.58	22.54	22.64	8.30	11.13	9.85	7.37	9.17	9.29
33	9.20	12.20	12.83	7.06	9.32	8.97	6.06	8.04	6.71
34	7.69	12.49	13.21	6.40	9.54	9.85	5.77	6.98	6.43
35	6.25	8.49	8.43	4.77	5.14	4.73	4.45	4.62	4.61
36	16.50	18.44	18.37	14.39	16.87	16.97	12.53	12.31	12.53
37	16.54	17.54	12.72	14.32	16.13	15.20	12.82	14.56	12.68
38	5.00	8.71	9.83	4.36	4.50	4.49	3.99	3.97	3.98
39	13.33	16.51	16.24	10.94	12.70	12.64	9.10	10.75	10.39
40	13.57	15.70	15.30	11.51	11.72	11.71	10.28	10.29	10.32
41	4.38	15.48	10.47	3.51	11.75	11.48	3.36	3.63	3.54
42	3.97	11.70	9.78	3.44	6.86	6.68	3.10	7.86	3.19

Table A.10: *Maximum Error of AVC Sites. "Weekdays" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	56.87	62.96	62.96	35.75	35.86	35.81	28.01	28.09	28.09
2	53.48	65.90	65.90	35.91	45.81	45.81	32.06	41.03	41.03
3	59.60	108.31	108.31	40.35	40.06	40.06	26.78	26.74	26.74
4	63.07	38.12	35.95	32.98	32.98	32.98	19.26	19.08	19.08
5	26.83	121.70	112.97	21.21	38.71	38.71	16.29	15.90	15.04
6	22.80	133.57	133.57	15.49	15.37	15.37	14.57	15.76	15.76
7	35.12	40.65	40.65	29.54	33.49	29.62	25.54	25.62	25.62
8	39.70	39.56	39.56	21.62	21.62	21.62	18.82	18.68	18.68
9	26.34	50.78	37.64	24.42	24.42	24.42	22.93	22.94	22.94
10	23.43	76.03	44.81	23.17	23.34	23.34	22.41	22.44	22.44
11	40.85	49.10	46.64	40.74	48.18	48.18	39.53	39.55	39.55
12	27.96	54.98	31.94	25.37	50.97	34.71	24.40	30.26	24.54
13	19.88	71.30	71.30	13.94	21.67	21.67	11.82	21.38	21.38
14	24.96	37.01	37.01	17.92	23.73	23.73	13.77	22.11	22.11
15	24.67	52.67	52.67	19.38	37.20	37.20	15.70	17.32	17.32
16	20.94	88.43	63.50	13.73	62.90	62.90	13.08	14.97	14.97
17	20.67	122.63	122.63	13.16	113.37	113.37	11.98	113.54	113.54
18	33.52	66.67	57.44	15.88	74.76	74.76	15.29	22.77	16.35
19	40.53	135.66	135.66	22.75	126.50	24.72	21.24	21.82	21.82
20	36.43	131.18	131.18	22.57	53.57	47.39	19.00	26.56	19.68
21	29.39	125.90	125.90	27.23	54.49	54.49	27.20	26.47	26.47
22	28.23	46.34	29.01	26.71	26.71	26.71	26.95	26.95	26.95
23	34.96	51.82	42.46	28.58	28.05	26.24	27.30	27.40	17.05
24	36.59	117.08	117.08	27.44	51.34	29.19	26.29	60.01	24.14
25	29.14	87.47	31.48	26.43	40.56	40.56	23.88	21.16	18.72
26	29.85	53.67	28.59	26.79	26.79	26.79	24.99	25.03	25.03
27	25.73	55.94	55.94	17.13	28.74	16.97	16.38	16.11	16.11
28	31.46	87.65	87.65	19.98	24.25	19.78	16.72	16.72	16.72
29	28.11	69.70	29.17	23.48	23.44	23.44	21.36	34.56	34.56
30	24.76	35.26	35.26	22.87	26.57	23.05	21.36	21.60	21.60
31	26.19	57.21	57.21	21.73	75.51	75.51	15.29	28.55	28.55
32	46.77	234.44	234.44	34.93	63.78	42.38	30.07	33.11	33.11
33	36.51	43.86	43.86	30.48	38.71	38.71	26.87	32.25	27.20
34	32.68	45.16	45.16	26.11	44.27	44.27	23.99	30.83	26.22
35	49.20	94.21	94.21	26.03	33.17	28.25	20.59	20.59	20.59
36	114.42	114.40	114.40	102.58	105.47	105.47	64.85	64.25	64.25
37	123.88	121.33	70.14	89.93	89.91	89.91	75.94	87.04	75.81
38	36.27	104.15	104.15	25.03	25.03	25.03	24.14	24.27	24.27
39	104.35	100.47	100.47	89.19	89.17	89.17	38.71	50.33	50.33
40	113.57	112.98	112.98	74.53	74.53	74.53	49.63	49.11	49.11
41	26.75	94.21	81.56	16.69	102.01	102.01	15.11	15.21	15.21
42	16.04	105.33	53.91	17.86	41.36	41.36	17.19	37.68	13.95

Table A.11: *Number of Samples of AVC Sites. "Weekdays" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	99	99	75	73	73	65	49	49	45
2	99	99	72	73	73	68	49	49	43
3	250	250	225	194	194	188	142	142	135
4	250	250	229	194	194	182	142	142	140
5	247	247	214	189	189	184	137	137	131
6	247	247	230	189	189	185	137	137	136
7	76	76	46	55	55	48	36	36	33
8	74	74	66	54	54	49	36	36	34
9	246	246	223	191	191	183	140	140	136
10	246	246	164	191	191	174	140	140	125
11	246	246	221	189	189	180	137	137	132
12	246	246	217	189	189	180	137	137	132
13	251	251	236	195	195	186	143	143	140
14	244	244	191	185	184	174	133	132	128
15	243	243	194	182	182	171	129	129	124
16	240	240	213	185	185	181	134	134	130
17	237	237	157	182	180	101	131	129	98
18	239	239	214	185	184	175	135	134	129
19	250	250	231	194	194	189	142	142	139
20	250	250	197	194	194	181	142	142	134
21	247	247	228	191	189	178	140	137	134
22	248	248	224	192	192	185	140	140	137
23	251	251	230	195	195	192	143	143	139
24	251	251	214	195	195	183	143	143	139
25	248	248	207	191	191	183	139	139	133
26	248	248	212	191	191	186	139	139	136
27	210	210	184	159	159	150	114	114	112
28	210	210	198	159	159	150	114	114	113
29	242	242	210	186	185	180	135	133	132
30	241	241	217	186	184	175	135	133	132
31	172	172	161	126	126	121	86	86	85
32	172	172	160	126	126	112	86	86	81
33	95	95	62	68	68	54	45	45	33
34	95	95	68	68	68	53	45	45	35
35	218	218	186	171	167	155	127	121	113
36	239	239	222	182	174	165	131	122	115
37	244	244	208	187	181	167	138	128	122
38	237	237	166	182	182	172	132	132	129
39	242	242	192	185	181	156	134	130	118
40	245	245	219	188	187	170	137	136	125
41	175	175	152	136	97	88	101	55	47
42	193	193	140	147	111	90	107	60	47

Table A.12: *MAE of AVC Sites. "Week-ends" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	12.25	21.09	20.98	9.64	9.81	9.12	10.76	12.95	12.38
2	10.98	23.45	22.39	8.90	14.17	7.87	8.60	10.53	10.64
3	8.89	17.31	16.87	7.08	8.55	6.57	6.33	6.17	6.12
4	7.98	9.87	9.33	6.60	7.13	7.13	6.05	6.02	6.22
5	6.69	9.30	8.82	5.47	5.50	5.48	5.35	5.59	5.42
6	5.44	6.55	5.10	4.65	4.57	4.58	4.80	4.88	4.88
7	9.28	10.76	9.27	7.55	8.67	7.85	7.74	7.53	8.05
8	6.73	14.19	15.11	5.65	6.90	5.15	6.66	6.56	6.56
9	7.03	7.92	7.62	6.07	6.15	6.50	5.37	5.76	5.58
10	6.73	11.22	10.98	5.45	6.06	5.63	5.01	5.35	5.10
11	6.62	8.86	7.89	4.54	5.04	4.53	4.60	5.61	4.80
12	6.82	9.94	8.43	5.60	6.33	5.45	6.32	6.59	6.07
13	6.18	8.93	7.66	4.97	5.46	4.83	4.56	4.47	4.43
14	9.39	12.62	12.76	6.11	10.92	11.27	6.30	8.21	8.17
15	8.04	12.82	13.06	6.59	11.32	9.27	6.17	8.90	9.06
16	5.12	8.69	7.96	4.40	6.19	6.10	4.06	5.49	5.55
17	4.94	8.53	8.41	4.22	6.80	6.86	4.20	14.33	10.19
18	5.36	14.64	12.33	3.84	5.43	5.46	3.78	5.18	3.92
19	6.16	13.39	8.42	5.10	6.31	4.97	4.98	5.11	5.11
20	6.73	9.79	8.39	5.81	5.61	5.17	5.30	5.00	4.92
21	6.51	9.05	8.98	4.92	6.86	5.57	4.37	5.36	4.65
22	6.88	8.48	7.27	5.48	6.27	4.94	5.46	5.56	5.64
23	7.05	14.66	15.89	6.07	6.60	6.24	5.03	5.29	5.35
24	8.47	16.08	14.97	6.01	9.02	5.21	4.94	5.29	5.43
25	7.06	11.36	9.90	5.75	6.63	6.19	5.53	5.84	5.56
26	7.07	8.93	8.87	6.13	5.95	5.97	5.58	5.45	5.55
27	5.88	6.81	5.71	4.78	4.88	4.70	4.36	4.39	4.46
28	6.16	6.46	6.44	4.36	4.52	4.55	4.94	4.93	4.93
29	7.98	9.28	10.14	6.85	6.60	6.70	6.51	6.68	6.68
30	7.00	8.13	8.06	5.94	6.19	5.99	5.26	5.88	5.05
31	9.22	12.43	12.35	6.89	12.58	10.12	6.40	9.97	8.79
32	10.00	15.84	14.03	6.86	11.14	10.61	5.77	9.28	9.28
33	15.02	21.56	20.94	12.84	15.99	13.99	10.99	12.19	13.38
34	12.20	25.72	27.37	9.18	12.05	8.73	7.88	7.86	8.82
35	11.48	16.98	18.17	8.07	11.27	9.38	6.32	7.97	6.99
36	18.50	32.07	31.07	16.47	19.14	18.02	16.41	17.53	16.85
37	17.66	29.74	26.26	15.66	20.53	19.48	14.91	16.88	14.12
38	7.40	8.29	8.46	5.34	6.49	6.37	5.80	5.95	5.92
39	20.22	22.51	22.99	19.30	19.58	19.16	16.92	16.50	17.37
40	18.62	27.33	28.62	17.29	19.28	21.09	15.21	15.43	15.86
41	7.53	9.50	8.82	5.73	8.37	4.97	4.86	5.27	4.70
42	7.84	21.69	17.42	5.52	8.80	7.78	4.35	5.82	4.62

Table A.13: *SDAE of AVC Sites. "Week-ends" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	10.71	22.01	21.29	9.15	7.84	7.13	9.05	10.43	10.20
2	9.37	32.98	27.32	6.83	13.96	6.19	7.82	11.64	12.02
3	7.83	21.29	22.26	6.04	8.04	4.46	5.80	5.42	5.10
4	7.37	8.80	8.70	6.42	6.55	6.75	6.46	6.74	6.97
5	7.51	15.44	16.58	5.80	6.08	6.29	5.38	5.43	5.38
6	4.26	5.68	4.09	2.90	2.82	2.85	2.97	2.96	2.96
7	5.92	9.07	8.93	4.87	6.06	5.93	6.27	6.69	7.52
8	4.84	21.58	26.43	4.01	7.78	3.91	3.68	3.68	3.68
9	4.92	6.00	5.83	4.46	4.60	4.75	4.17	4.53	4.43
10	5.46	12.97	14.93	4.12	5.25	4.31	3.90	4.13	4.14
11	5.52	7.49	6.77	4.33	4.75	4.74	3.63	5.47	4.06
12	7.70	11.82	9.98	4.94	6.28	4.86	4.96	6.02	5.06
13	5.47	9.73	9.55	4.54	6.10	4.61	3.91	3.76	3.49
14	7.75	14.37	14.96	4.63	21.08	21.73	4.79	10.84	11.01
15	5.95	16.25	17.44	4.78	11.78	8.90	4.98	8.98	9.68
16	4.39	8.77	7.21	3.96	4.71	4.63	3.72	4.46	4.50
17	4.29	7.84	7.84	3.87	6.07	6.26	3.61	22.31	18.55
18	3.76	17.72	10.35	3.05	4.14	4.22	2.69	6.67	3.12
19	5.97	22.09	16.96	4.84	9.94	4.94	3.96	4.03	4.03
20	7.01	9.57	8.93	5.29	4.71	4.29	5.79	5.13	5.21
21	5.29	14.50	16.54	4.06	7.90	4.77	4.16	6.04	4.25
22	5.67	8.00	6.41	4.45	5.19	3.38	4.41	4.36	4.38
23	6.08	25.20	30.69	5.30	6.30	6.07	4.14	4.84	4.88
24	5.68	19.91	24.24	5.34	22.30	4.51	4.12	5.15	5.21
25	5.61	17.56	16.75	5.14	5.66	4.32	4.79	5.12	4.83
26	6.40	9.12	9.79	5.63	5.00	5.19	5.47	5.10	5.20
27	4.96	7.54	5.51	3.90	3.98	3.89	3.32	3.30	3.38
28	5.18	5.58	5.58	4.35	4.27	4.32	3.64	3.66	3.66
29	8.37	12.89	14.72	7.19	6.42	6.40	6.72	6.71	6.71
30	7.52	8.50	8.49	5.78	5.46	5.30	5.00	6.76	5.10
31	7.93	13.44	13.77	6.18	14.10	8.31	4.86	8.96	7.43
32	8.11	18.29	15.70	5.42	7.20	7.13	4.85	7.26	7.26
33	11.19	14.04	15.33	10.61	12.13	9.70	9.78	9.77	9.70
34	9.59	37.20	38.77	7.27	10.47	6.31	5.09	4.98	5.20
35	10.08	15.74	17.41	7.20	13.44	11.86	6.34	7.84	7.17
36	18.78	29.97	32.61	17.69	19.54	19.43	17.71	18.08	17.86
37	17.52	33.32	30.22	17.43	24.58	26.25	17.66	18.33	16.44
38	7.06	7.27	7.64	5.24	9.59	10.45	4.92	5.04	5.23
39	14.44	15.02	14.99	12.58	12.76	12.95	13.73	13.79	14.08
40	14.20	27.67	29.64	13.65	14.64	15.88	13.51	14.13	14.33
41	6.49	8.91	8.81	4.99	13.59	3.47	5.52	6.01	5.29
42	7.52	27.29	25.39	4.16	8.65	7.73	3.51	5.46	4.14

Table A.14: *Maximum Error of AVC Sites. "Week-ends" Sample*

AVC Number	24hr SPTCs			48hr SPTCs			72hr SPTCs		
	Perfect	ANN	D<20%	Perfect	ANN	D<20%	Perfect	ANN	D<20%
1	51.37	96.63	96.63	33.21	31.07	19.40	31.17	33.27	31.38
2	50.58	174.61	129.00	26.74	52.24	19.78	30.37	44.77	44.77
3	34.56	135.40	135.40	32.69	34.70	21.73	31.30	30.08	30.08
4	42.50	42.12	42.12	38.62	38.55	38.55	37.58	37.44	37.44
5	49.94	133.35	133.35	33.30	31.60	31.60	27.33	27.13	27.13
6	19.95	30.41	17.67	12.34	12.50	12.50	11.53	11.42	11.42
7	23.89	39.64	39.64	15.54	19.71	19.09	19.07	21.18	21.18
8	16.95	124.02	124.02	14.09	35.50	14.05	12.40	12.48	12.48
9	24.31	36.36	36.36	21.07	21.15	21.15	20.70	20.99	20.99
10	33.07	91.73	91.73	21.89	26.54	21.75	17.32	17.57	17.57
11	28.66	37.62	28.67	27.25	27.20	27.20	24.99	26.44	26.44
12	43.32	65.13	55.07	19.95	30.34	22.12	21.00	29.21	22.03
13	25.52	61.25	61.25	17.98	33.52	17.87	17.34	15.43	13.46
14	33.23	104.43	104.43	22.53	140.25	140.25	17.52	63.71	63.71
15	27.45	116.09	116.09	19.66	51.19	35.12	17.73	42.97	42.97
16	21.17	54.41	46.15	20.77	18.89	18.89	18.35	18.02	18.02
17	19.09	47.78	47.78	17.27	24.98	24.98	16.12	99.71	99.71
18	15.74	115.85	68.52	12.12	17.73	17.73	9.85	36.92	14.31
19	28.97	141.96	115.20	19.83	64.34	20.69	18.78	18.97	18.97
20	33.56	46.70	41.61	24.25	17.01	17.01	22.47	17.34	17.34
21	25.37	118.36	118.36	20.82	44.00	20.80	19.46	33.38	19.46
22	27.93	43.06	30.11	23.51	27.20	16.31	20.92	20.99	20.99
23	38.39	168.74	168.74	21.67	24.65	24.65	20.17	25.09	25.09
24	28.85	185.26	185.26	21.87	154.36	18.47	16.90	28.04	28.04
25	26.37	122.09	122.09	25.21	31.35	15.78	22.23	22.27	22.27
26	27.02	63.48	63.48	26.29	19.24	19.24	22.85	21.10	21.10
27	25.98	47.93	34.30	15.77	15.65	15.65	12.80	12.81	12.81
28	22.50	24.14	24.14	16.70	15.93	15.93	16.06	16.03	16.03
29	45.25	113.43	113.43	28.38	28.27	28.27	24.70	24.69	24.69
30	34.60	42.00	42.00	23.45	20.04	18.44	18.81	35.95	18.78
31	49.78	69.70	69.70	32.21	77.26	31.94	20.25	39.77	29.09
32	36.00	103.75	103.75	22.53	25.37	24.47	20.57	24.42	24.42
33	47.40	65.94	65.94	38.79	49.87	37.03	33.04	31.88	31.88
34	39.30	180.31	180.31	26.56	41.92	17.66	21.78	21.86	21.86
35	44.22	82.57	82.57	35.16	70.78	70.78	29.29	29.26	29.26
36	122.72	162.81	162.81	104.08	103.89	103.89	107.53	107.53	107.53
37	104.91	226.25	153.48	97.10	128.73	128.73	97.25	96.63	96.63
38	29.87	30.28	30.28	23.59	54.86	54.86	20.50	21.36	21.36
39	80.54	80.40	80.40	69.82	69.82	69.82	81.33	81.31	81.31
40	86.36	185.75	185.75	77.65	77.50	77.50	77.00	76.92	76.92
41	31.42	48.00	48.00	26.77	86.48	13.18	23.54	22.97	22.97
42	40.11	142.96	115.03	18.08	33.39	29.83	12.01	24.38	12.00

Ringraziamenti

Desidero ringraziare la "Fondazione Ing. Aldo Gini" di Padova e il Collegio Docenti della Scuola di Dottorato in Ingegneria Civile e Ambientale dell'Università di Trieste per il prezioso supporto finanziario che ha reso possibile il mio periodo di ricerca presso il Virginia Tech.

Ringrazio il Prof. Romeo Vescovi e il Prof. Riccardo Rossi per avermi offerto questa posibilità di crescita professionale e personale e per la disponibilità dimostratami in questi anni.

Ringrazio inoltre il Prof. Massimiliano Gastaldi per il supporto, la disponibilità e lo stimolante scambio di idee che mi ha sempre saputo offrire.

Un ringraziamento speciale va al Prof. Shinya Kikuchi, mio "tutore" durante la mia permanenza al Virgina Tech, per la sua capacità di stimolare la continua ricerca di nuove soluzioni e la squisita accoglienza che mi ha regalato.

Un grazie alla mia famiglia per il supporto, l'aiuto e l'affetto che mi hanno dato in questi anni, grazie ai quali ho potuto superare anche i momenti più difficoltosi.

Infine un pensiero speciale a Giulia, la mia gioia più grande.