1 INTRODUCTION

1.1 OBJECTIVES

The implementation of Intelligent Transport Systems (ITS) has made vast quantities of real-time traffic data available, by making use of current road network infrastructure that enables information to be gathered on-line. Detectors that measure traffic flow, speed and occupancy are an example. How to use real-time traffic data, as well as historical data, to provide short-term traffic prediction, remains an open problem for researchers. The problem of short-term traffic prediction involves determining the evolution of traffic flows or, equivalently, of the network state. The ability to predict the network state dynamically is essential in traffic management and for traffic information centres particularly, since it enables them to apply traffic control and traffic management policies to prevent traffic congestion rather than dealing with traffic problems after congestion has already occurred.

Advanced traffic management systems (ATMS) and advanced traffic information systems (ATIS) must consider, in real time, short time intervals in which neither demand nor flows are constant and homogenous. Demand and flow behave dynamically, that is, they are both time-dependent. The concept of traffic management, as defined by Barceló (1991), is broader than the classic concept of traffic control, because it takes action over time, including control over space, such as, for instance, redistributing flows by rerouting, that is, by proposing alternatives routes. Therefore, traffic management applications require dynamic modelling that shows flow variation over time.

All proposals for advanced traffic management and control systems that are based on telematic technologies agree on the importance of short-term prediction of traffic flow evolution, which is equivalent to the short-term prediction of the network state, for correct decision-making in traffic management, information dissemination to users, etc. Several system architectures have been proposed and evaluated in European projects in recent years. Although the achievements of these projects cannot be applied or extrapolated to complex urban structures, other models that are more suited to complex networks have been developed, by Cascetta (1993) and Barceló (1997), for example. Unfortunately, these models do not appear to be appropriate for full dynamic applications, and so we had to look elsewhere in our search for a suitable prediction model. The promising features of neural networks, which make them suitable for use as predictive tools (Baldi and Hornik, 1995), encouraged us to explore this approach. The approach, which is based on real-time detector measurements combined with historical OD matrices, involves determining a short-term forecast of a sliced OD matrix. The forecast OD matrix could be used as input for a microscopic traffic simulator such as AIMSUN; thus the evolution of traffic flows and, as a consequence, the forecast network state could be obtained.

According to this dynamic vision of demand, we can consider each of the OD matrix's components as a time series. Therefore, forecasting an OD matrix consists in performing the forecast for each component in the matrix, that is, in simultaneously forecasting many multivariate time series. Solutions to this problem that are based on classic forecasting methods, such as Box-Jenkins or Kalman filtering, have been proposed by several authors (Davis, 1993; Davis et al., 1994; Van der Ziipp and Hamerslag, 1996). The approaches proposed provide relatively good results for linear infrastructures, such as motorways, although it remains unclear whether they would provide reliable results in the case of more complex networks, such as urban networks. In some of the most promising cases (Davis, 1994), however, the computational task required practically invalidates their use in real-time applications in large-scale networks and makes it advisable to look for other methods.

Neural networks appear to be natural candidates for forecasting models, particularly if their easily parallelisable structure is taken into account, and high computational speed is required to achieve a system's objectives. Further reasons to consider a neural network approach are the results reported by Chakraborty (1992) for multivariate time series analysis using neural networks and by Weigend (1992) in his evaluation of their predictive capabilities compared to other classic models.

The dynamic prediction of the network state in terms of the OD matrix by means of neural networks has one main drawback: the amount of data required for the proper training of the neural network. This thesis proposes solving this handicap by partitioning the neural network in terms of clusters of independent or almost independent OD pairs. This technique allows an original neural network of a large size to be split into a set of smaller neural networks that are easier to train. Before the clustering problem can be solved, however, the paths that are most likely to be used between each OD pair must be identified.

Short-term forecasting leads, in this way, to the critical problem of dynamic traffic assignment, which is solved in this thesis by a microsimulation-based heuristic. In the thesis, some of the most critical aspects of the dynamic simulation of road networks are discussed, namely heuristic dynamic assignment, implied route choice models and the validation methodology, a key issue in determining the degree of validity and significance of the simulation results. The work is divided into two parts: the first provides an overview of how the main features of microscopic simulation were implemented in the microscopic simulator AIMSUN (AIMSUN 2002) and the second is a detailed discussion of heuristic dynamic assignment and sets guidelines for calibrating and validating dynamic traffic assignment parameters. The calibrated and validated simulation model is then used to conduct a dynamic traffic assignment, whose output identifies the paths that are most likely to be used, which will be clustered in subsets that connect the OD pairs and will define the neural networks for the forecast.

1.2 THESIS OVERVIEW

After the introductory chapter, the remaining chapters of the thesis are organised as follows. In Chapter 2, a scheme for the architecture of decision-support systems for traffic management is proposed and discussed, and the importance of short-term forecasting of the evolution of traffic flows or the network state in management decisions is made explicit. The research carried out to explore the performance of neural networks in relation to the demand prediction problem is then described, in terms of the quality of the results provided and the computational requirements for real-time applications. The dynamic prediction of the network state in terms of the OD matrix by means of neural networks has two main drawbacks: the size of the neural network, and the amount of data required for the proper training of the neural network. This work proposes solving these handicaps by partitioning the neural network by considering independent or almost independent OD pairs, which results in a subset of independent neural networks of a smaller size and less data that requires training. A way of identifying independent or almost independent OD pairs is to identify the paths that are most used between each OD pair. Identifying the paths requires a dynamic traffic assignment process. In Chapter 3, the dynamic traffic assignment implemented in the microscopic AIMSUN simulator is described and the implied route choice models and the validation methodology, a key issue in determining the degree of validity and significance of the simulation results, is discussed in chapter 4. Figure 1.1 depicts the thesis outline in terms of requirements and proposals.

As shown in Figure 1.1, Chapter 2 proposes a scheme for the architecture of decision-support systems for traffic management and traffic information systems in which the main aim is the short-term forecasting of the evolution of traffic flows or the network state. The proposed scheme has been evaluated in European projects during the drafting of the thesis; of them, we might draw attention to ARTIS (ARTIS, 1994), PETRI (PETRI, 1996) and CAPITALS (CAPITALS, 1998). The requirements for the short-term prediction of the network state are formulated by determining the short-term prediction of traffic demand, expressed as an OD matrix, to which a dynamic component is introduced, and then the network state is obtained using the AIMSUN simulation output.

To address the problem of the OD structure we consider origins and destinations as pairs, I being the set of all OD pairs in the network. If origin r and destination s are the i-th OD pair, g_i denotes the corresponding entry of demand matrix G, which represents the total number of trips between origin r and destination s. Therefore,

$$O/D_{(r,s)} = g_i \,, \quad i = (r,s) \in I$$

where I denotes the set of all OD pairs in the network. The total number of trips between an origin r and a destination s is not a fixed value over time but a dynamic value (i.e. it is time-dependent), as depicted in Figure 1.2. According to this dynamic vision of demand, forecasting an OD matrix consists in simultaneously forecasting many multivariate time series.

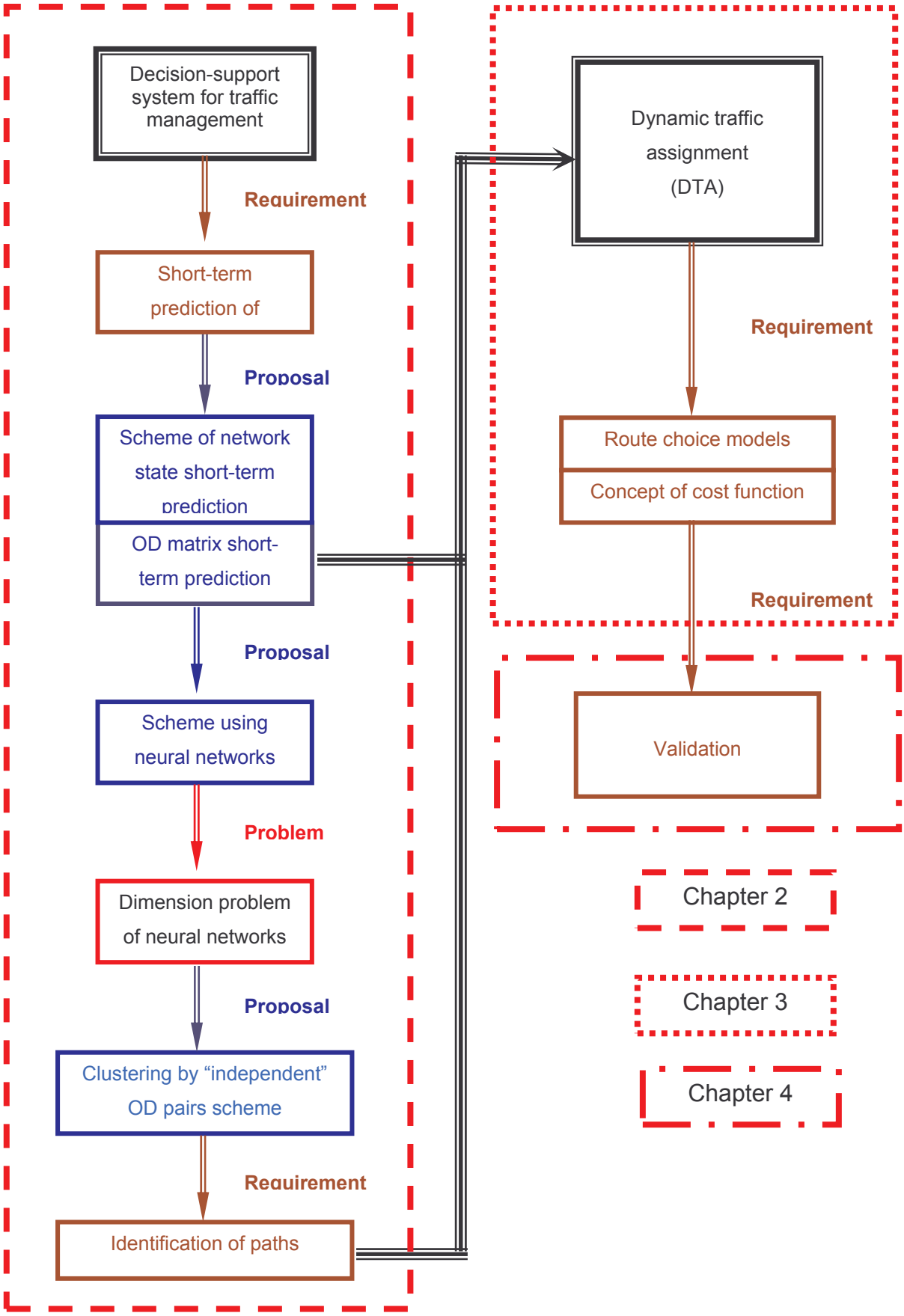

Figure 1.1. Thesis outline

The research explored in this work uses neural networks as a prediction model for achieving the system's objectives and considers the real-time requirement and the positive results reported when its predictive capabilities have been evaluated in comparison to other classic models.

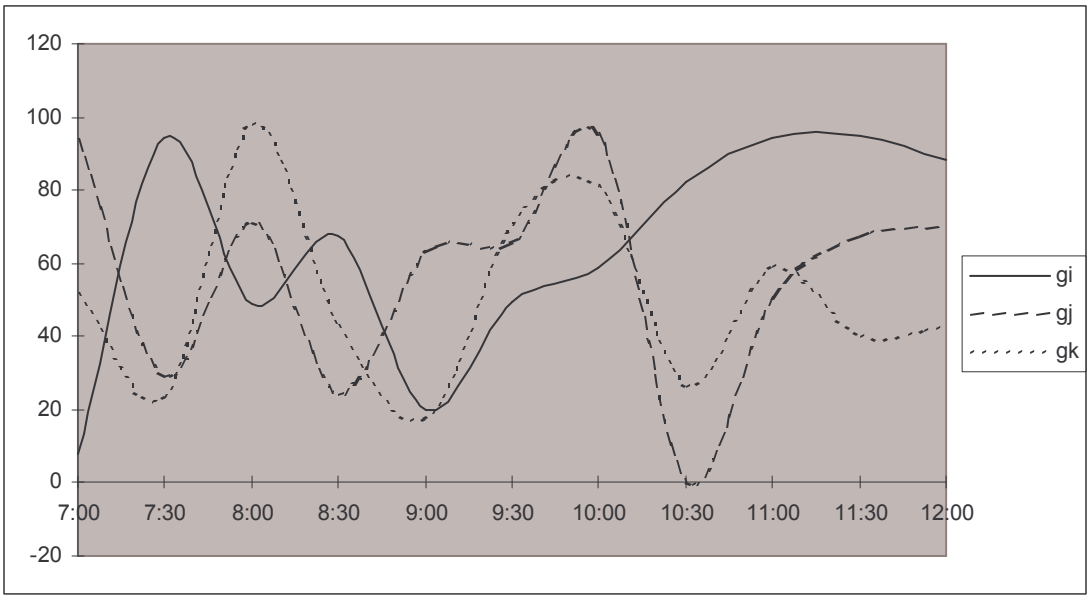

Figure 1.2. Dynamic demand

A neural network, as defined by Hecht-Nielsen (1989), consists of a set of interconnected computational units or neurons, each of which performs a computational process on a weighted sum of inputs according to a specific function, as shown in Figure 1.3. A neural network model is generally characterised by three elements: the topology of the neural network, the neurons' characteristics and the rules of the training or learning process.

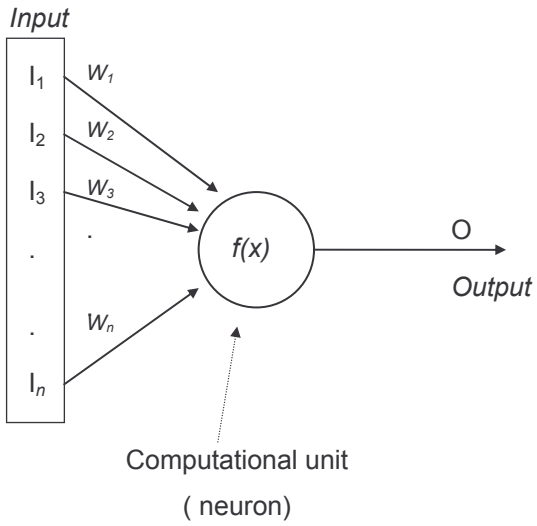

Figure 1.3. Neuron representation

The topology explored in this work, according to the requirements of analysing and predicting multivariate time series, is a multi-layer perceptron neural network. This topology corresponds to a feed-forward network in which the neurons are arranged in layers and every neuron on each layer is connected directly to all the neurons on the next layer.

The neuron is characterised by a nonlinear activation function and in our case the activation function selected is the following sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function (Hecht-Nielsen, 1989) is a bounded differentiable real function that is defined for all real input values. It rapidly approaches asymptotically fixed finite upper and lower limits as its argument gets larger or smaller respectively, and this limited dynamic range effectively implements noise suppression and cut-off as Másson (1990) shows. Given the nature of our problem, in which there are continuous inputs and outputs, the main reason for selecting the above function is the fact that it makes use of the sigmoid rule, one of the most frequently used nonlinear activation rules. The training or learning algorithm used is an ad hoc version of the back propagation algorithm described by Hecht-Nielsen (1989). It is a supervised learning process, given that the weights of the different neuron connections are iteratively changed with reference to a set of predefined patterns specified as a set of input-output pairs. At each step, the computational error is estimated to be

$$E = \frac{1}{\mathrm{P}\, N_L} \sum_{p=1}^{P} \left\| t^{(p)} - S^{(p)}(L) \right\|^2$$

$$E = \frac{1}{\mathrm{P}\, N_L} \sum_{p=1}^{P} \sum_{n=1}^{N_L} \left(tn^{(p)} - Sn^{(p)}(L) \right)^2$$

where $t^{(p)}$ is the p-th desired output and $S^{(p)}(L)$ is the p-th output produced by the neural network. Back propagation tries to minimise the total squared error E using a gradient algorithm.

The training process gauges the neural network, i.e. determines the different weights of the link connections, and this depends on a set of desired input and output pairs. The experiments were conducted with AIMSUN, a microscopic simulator that provides the detector measurements that correspond to the simulation of traffic flows obtained from an OD matrix as output. Then, from the historical OD matrix, small perturbations in this historical OD matrix, expressed as percentage variations, and the detector measurements generated by simulation, the necessary inputs for the training module can be simulated.

The training process requires the input of a historical time-sliced OD matrix, as well as the patterns to train the neural network that has to produce the forecast. Time-sliced OD matrices

are currently unavailable and their production is difficult and costly. Our proposal, evaluated in the European CAPITALS project, consists in generating the sliced OD matrix that is receiving information from detector flows and applying a matrix adjustment.

The prediction process forecasts the OD matrix in the next interval from the detector measures collected and the historical OD matrix using a multilayer perceptron neural network and applying a feed-forward algorithm.

The dynamic prediction of the network state in terms of the OD matrix using neural networks has the disadvantage of the amount of data required to properly train the neural network. Our proposal consists in reducing the size of the neural network whilst not diminishing its capacity for representing the road network. This reduction is based on determining, for each OD pair, the *k* current paths most likely to be used and partitioning them considering the sharing of links. The partitioning condition may be very strict in most cases; it would thus be desirable to admit a certain degree of overlapping if no significant errors are induced. Our proposal considers a cluster analysis, in which the degree of overlapping can be controlled as a function of the similarity level between clusters.

The microsimulation using traffic demand defined in terms of the OD matrix becomes the principal point due to its use in several steps in the scheme proposed, such as the simulation to determine the forecast network state, the generation of the input patterns in the training process and the path identification process to determine the OD pair clusters. Chapter 3 discusses some of the most critical aspects of the dynamic simulation of road networks, namely heuristic dynamic assignment; the implied route choice models; the implementation in AIMSUN and the validation methodology, an issue that is key in determining the degree of validity and significance of the simulation data, is presented in chapter 4.

The assessment by simulation of ITS applications requires a substantial change of traditional paradigms in microscopic simulation, in which vehicles are generated at the input sections in the model and perform turnings at intersections according to probability distributions. In such a model, vehicles have neither origins nor destinations and move randomly on the network. The required simulation approach must be based on a route-based microscopic simulation paradigm. In this approach, vehicles are input into the network according to the demand data defined as an OD matrix (that is preferably time-dependent) and they drive along the network following specific paths in order to reach their destination. In route-based simulation, new routes are to be calculated periodically during the simulation, and a route choice model is needed, if alternative routes are available, to determine how the trips are assigned to these routes.

The key question that this approach raises is whether this simulation can be interpreted in terms of heuristic dynamic traffic assignment or not. We therefore propose investigating the answer to this question for the case of the microscopic route-based simulator AIMSUN (Barceló et al., 1995, 1998) used in the thesis.

The route-based simulation process in AIMSUN can be interpreted in terms of a heuristic approach to dynamic traffic assignment similar to the one proposed by Florian et al. (2001), which consisted of the following:

1. A method to determine the path-dependent flow rates on the paths in the network, based on a route choice function.

2. A dynamic network loading method, which determines how these path flows give rise to time-dependent arc volumes, arc travel times and path travel times, heuristically implemented by microscopic simulation.

The simulation process implemented, which is based on time-dependent routes, consists of the procedure described below, a conceptual diagram of which is depicted in Figure 1.4.

Heuristic dynamic assignment procedure

Step 0 Calculate the initial shortest path(s) for each OD pair using the defined initial costs.

Step 1 Simulate for a time interval Δt, having assigned to the available path K_i the fraction of the trips between each OD pair i for that time interval according to the probabilities $P_{k,}$ $k \in K$ estimated by the selected route choice model.

Step 2 Update the link cost functions and recalculate the shortest paths, with the updated link costs.

Step 3 If there are guided vehicles or variable message panels that propose rerouting, provide the information calculated in Step 2 to the drivers that are dynamically allowed to reroute during trips.

Step 4 **Case A** (Preventive dynamic assignment)

If all the demand has been assigned, then stop. Otherwise, go to Step 1.

Case B (Reactive dynamic assignment)

If all the demand has been assigned and the convergence criteria hold, then stop.

Otherwise,

Go to Step 1 if all of the demand has not been assigned yet

Or

Go to step 0 and start a new major iteration.

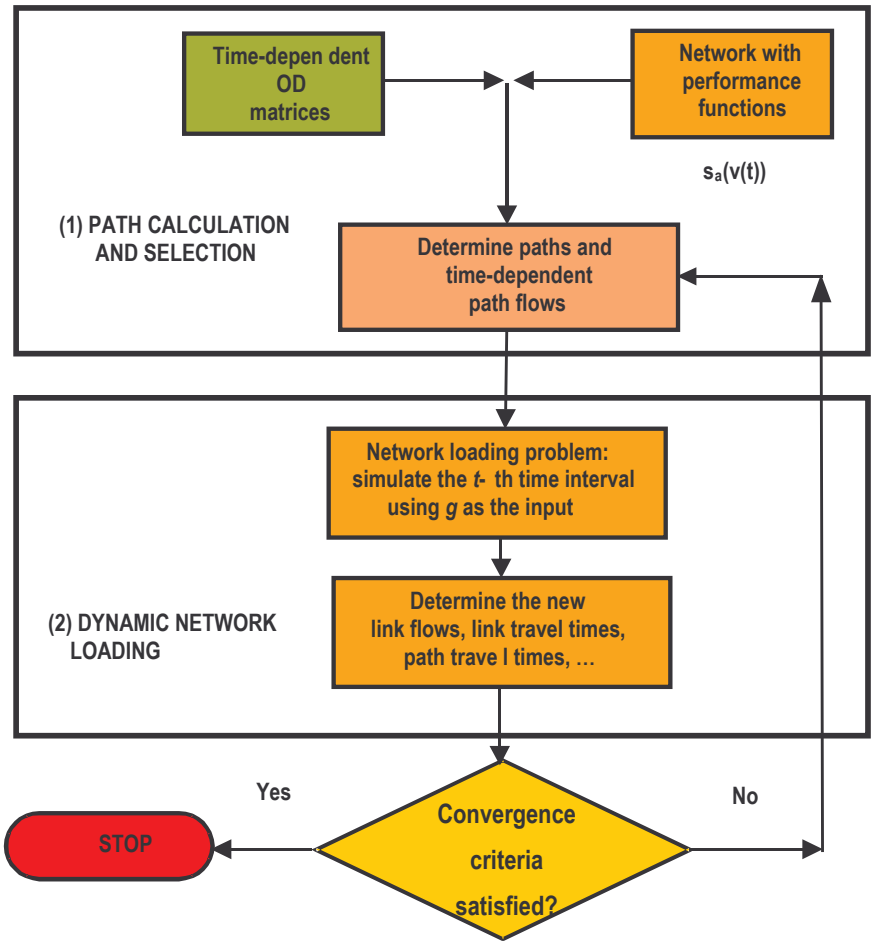

Figure 1.4. Conceptual diagram of the heuristic dynamic traffic assignment

Depending on how the link cost functions are defined, and whether the procedure is applied as one pass method completed when all the demand has been loaded or applied as part of an iterative scheme that is repeated until a given convergence criterion is satisfied, it corresponds either to a “preventive” or en-route dynamic traffic assignment, or to a “reactive” or heuristic equilibrium assignment. In the first case, route choice decisions are made for drivers entering the network at a time interval based on the travel times experienced, i.e. the travel times of the previous time interval, and the link cost function is defined in terms of the average link travel times in the previous interval. Alternatively, a heuristic approach to equilibrium can be based on repeating the simulation scheme a number of times and defining a link cost function including predictive terms, as proposed by Friesz et al. (1993) and Xu et al. (1999). This could be interpreted in terms of a day-to-day learning mechanism.

The simulation experiments carried out as part of this thesis were implemented in AIMSUN. The logit, C-logit and proportional route choice functions were selected from the default route choice functions available. The multinomial logit route choice model defines the choice probability P_k of

alternative path k, $k \in K_i$, as a function of the difference between the measured utilities of that path and all other alternative paths, such that

$$P_k = \frac{e^{\theta V_k}}{\sum_{l \in K_i} e^{\theta V_l}} = \frac{1}{1 + \sum_{l \neq k} e^{\theta (V_l - V_k)}}$$

where V_i is the perceived utility for alternative path i (i.e. the opposite of the path cost or path travel time), and θ is a scale factor that fulfils two roles: it makes the decision based on differences between utilities independent of measurement units, and it influences the standard error of the distribution of expected utilities, in that way determining a trend towards using many alternative routes or concentrating on only a very few. It is thus a critical parameter in calibrating whether the logit route choice model leads to a meaningful selection of routes or not.

One of the drawbacks of using the logit function is a tendency towards route oscillations in the routes used, whereby the corresponding instability leads to a kind of flip-flop process. Our experience shows that there are two main reasons for this behaviour: the properties of the logit function and the logit function's inability to distinguish between two alternative routes when there is a high degree of overlapping.

The instability of the routes used can be substantially improved when the network topology allows for alternative routes with little or no overlapping at all by playing with the shape factor of the logit function and frequently recomputing the routes. However, in large networks, where there are many alternative routes between origin and destination, some of which exhibit a certain degree of overlapping, the use of the logit function may still exhibit some weaknesses. To counter this drawback, the C-logit model (Cascetta et al., 1996 and Ben-Akiva and Bierlaire, 1999) was implemented.

In this model, the choice probability P_k, of each alternative path k belonging to the set K_i of available paths that connect the i-th OD pair is defined as

$$P_k = \frac{e^{\theta (V_k - CF_k)}}{\sum_{l \in K_i} e^{\theta (V_l - CF_l)}}$$

where V_i is the perceived utility for alternative path i, i.e. the opposite of the path cost, and θ is the scale factor, as in the case of the logit model. The 'commonality factor' of path k, denoted as CF_k, is directly proportional to the degree of overlapping of path k with other alternative paths. Thus, highly overlapped paths have a larger CF factor and their utility is therefore less than that of similar paths. CF_k is calculated as follows:

$$CF_k = \beta \cdot \ln \sum_{l \in I_{rs}} \left(\frac{L_{lk}}{L_l^{1/2} L_k^{1/2}} \right)^{\gamma}$$

where L_{lk} is the length of the arcs common to paths *i* and *k*, while L_l and L_k are the length of paths *l* and *k* respectively. Depending on the two factor parameters β and γ, the 'commonality factor' is more or less weighted. Larger values of β mean that the overlapping factor is of greater significance with respect to the utility V_i; γ is a positive parameter, whose influence is smaller than β, and it has the opposite effect. The utility V_i used in this model for path *i* is the opposite of the path travel time tt_i (or path cost if it has been thus defined by the user).

Another option is the estimation of the choice probability P_k of path *k* , $k \in K_i$, in terms of a generalisation of Kirchoff's laws given by the function

$$P_k = \frac{CP_k^{-\alpha}}{\sum_{l \in K_i} CP_l^{-\alpha}}$$

where CP_l is the cost of path *l* and α is, in this case, the parameter whose value has to be calibrated.

This option has been included in the study for the sake of completeness, although its foundations are unclear from the point of view of discrete choice theory in the context of modelling road users' route choice behaviours, at least insofar as it has been proposed by other authors and used in other studies (Fellendorf and Vortisch, 2000). It could be interpreted as an aggregated flow behaviour resulting from individual decisions in which the intensities of the flows are distributed among the available routes, similarly to the way in which electrical flows are distributed in electrical networks, assuming that link travel times or travel time-related cost functions may be taken to be similar to electrical resistance.

The statistical methods and techniques for validating a traffic simulation model that are based on the standard statistical comparison of the model's and system's outputs can consider global measurements and/or disaggregated measurements, but a critical aspect in the calibration/validation of a dynamic traffic assignment model is determining the values of the dynamic traffic assignment parameters that lead to a meaningful selection of paths. No formal convergence proof can be given for the dynamic traffic assignment proposed, since the heuristic network loading process based on microscopic simulation does not have an analytical form. The method proposed is based on the assumption that, insofar as the assignment described may be associated with a heuristic approach to a preventive dynamic equilibrium assignment (Xu et al., 1999), properly selecting the path should lead to such equilibrium. An assignment's progress towards equilibrium, and therefore the quality of the solution, may be measured using the relative gap function *RGap(t)* (Florian et al., 2001) and (Janson, 1991), which estimates, at time

interval *t*, the relative difference between the total travel time actually experienced and the total travel time that would have been experienced if the travel times for all vehicles had been equal to the current shortest path, such that

$$Rgap(t) = \frac{\sum_{i \in I} \sum_{k \in K_i} h_k(t)[s_k(t) - u_i(t)]}{\sum_{i \in I} g_i(t) u_i(t)}$$

where

t is the time interval used in the dynamic traffic assignment algorithm;

I is the set of all OD pairs;

$k \in K_i$ is the set of paths for *i*-th OD pair;

g_i *is* the traffic demand of OD pair *i*;

$h_k(t)$ is the path flow assigned to path $k \in K_i$ that connects OD pair *i* at interval *t*;

$s_k(t)$ is the total travel time experienced by all vehicles assigned to path $k \in K_i$ that connects OD pair *i* at interval *t*; and

$u_i(t)$ is the total travel time experienced by all vehicles assigned to the shortest path that connects OD pair *i* at interval *t.*

1.3 THESIS OUTLINE

The remaining chapters of this dissertation are organized as follows. Chapter 2 discusses the architecture for advanced traffic management and control system using the microscopic simulation models to support sound decision-making processes, proposing neural networks as dynamic mechanism for the short-term prediction of the network state and highlighting the problem of pattern generation for the neural network training process and the drawback related to the size of the neural network. Chapter 3 presents a heuristic for the dynamic traffic assignment implemented in the scope of this research with the microscopic simulator AIMSUN. Chapter 4 describes the methodology for validation of microscopic models. Chapter 5 presents the computations results for validating the heuristic dynamic assignment based on microsimulation. Finally, Chapter 6 concludes this dissertation by highlighting the main contributions and discussing directions for future research.