

NEURAL NETWORK FOR TRAVEL DEMAND FORECAST USING GIS AND REMOTE SENSING

André Dantas¹, Koshi Yamamoto¹, Marcus V. Lamar², Yaeko Yamashita³

¹Dept. of Civil Engineering -Nagoya Institute of Technology,
Showa, Gokiso, 466-8555, Nagoya, Japan

² Dept. of Electrical and Computer Engineering – Nagoya Institute of Technology
Showa, Gokiso, 466-8555, Nagoya, Japan

³Master Programme in Transportation, Dept. of Civil Engineering, University of Brasilia
Campus Universitario Darcy Ribeiro, Brasilia, Brazil

E-mail: andre@keik1.ace.nitech.ac.jp

Abstract

This paper describes an application of Neural Networks in the development of a travel forecast model for transportation planning. The model intends to quantify trips within the urban area through the representation of the land use-transportation system interaction. The data to express such a complex interaction is mainly obtained from Remote Sensing images that are processed in a Geographical Information System. We present, in this paper, model's basic formulation and the results of a case study conducted in Boston metropolitan area.

1. INTRODUCTION

Travel demand forecast is an essential element for transportation planning in order to evaluate future needs of an urban area. In a planning context, these forecasts are used to generate the necessary information for decision-makers to decide where, when and how transportation system has to be improved. This problem has been researched since the 60's and many models were developed following a standard formulation called *four-step process* (trip generation/attraction, trip distribution, modal split and allocation). In this formulation, the urban area is divided into homogeneous sectors (traffic zones) where socioeconomic indicators such as population, income, household and car ownership are collected. Through statistical processing, these data are used to calculate current trips and then forecast future scenarios by simulating socioeconomic changes and / or transportation system improvements [4].

Despite the extensive application, travel forecast models have been mostly criticized due to the non-incorporation of the land use-transportation system interaction. This complex interaction directly affects the travel process since the locational arrangement of activities generates the movement of people and goods as well as the creation and / or improvement of the transportation system leads to the generation of new activities within the urban area. According to Rodrigue [7], this dynamic phenomenon has been conceived under a static and deterministic approach that does not adapt mathematical methods to urban processes and their dynamical relationships. Additionally, Harris [4] shows that socioeconomic information does not represent realistically land use-transportation interaction and data's requirements in forecast models are so massive and costly to be applicable in real problems.

Just recently, it can be observed some efforts to overcome these limitations and they are concentrated on two distinct approaches. The first approach is devoted to the development of Neural Network (NN) based travel forecast models. NN are dedicated to represent the complex nature of land use-transportation system interaction through a self-adaptable model. In this direction, Rodrigue [7] presented an innovative and theoretical conception considering the economic structure, population, accessibility and spatial interactions towards the forecast of the land use-transportation system. However, this conception is mainly based on socioeconomic data and concerned about general urban planning. On the other hand, the second approach focused on the incorporation of the land use-transportation interaction by exploring the integration of Remote Sensing (RS) and Geographical Information Systems (GIS). Taco *et al.* [8] elaborated and applied a travel generation model contemplating land use patterns

extracted from aerial photograph images. Nevertheless, Taco's model did not evaluate dynamic changes nor the interactions between patterns and it performed just the modelling of the travel generation phase.

In this work, it is intended to gather these two distinct approaches in order to establish a travel forecast model integrating NN, RS and GIS. Following Openshaw's experience [6] on the integration of NN and GIS, the model explores RS images as the main source of information on the urban dynamic. These images are processed in GIS to obtain indicators related to land use-transportation system interactions such as areas of the patterns, extension of modes and spatial location. From these data, NN is used to forecast the trips within the urban area. In this paper, firstly we concentrate on the problem formulation describing the basic assumptions for travel forecast modelling. Next, it is briefly outlined a methodology for RS-GIS-NN integration. Finally, it is reported the case study that was conducted in Boston Metropolitan Area in order to verify model's efficiency.

2. PROBLEM FORMULATION FOR A NEURAL NETWORK APPROACH

The urban area can be considered as a locational arrangement of activities, where people move from one place to another to take part on socioeconomic events [4]. These activities are represented by land use patterns that occupy the earth surface and are separated by distance. Therefore, trips can be conceived as the disequilibrium of land use patterns in different locations inside the urban area. For instance, one area typically occupied by residential land use patterns generates trips that are attracted to another area formed by industrial, or commercial, or services, etc. Along the time, these interactions between patterns change according to the land use modifications and the accumulation of activities that gradually transform the urban environment.

On the other side of the travel process, the transportation system acts as a supply and inducting element. Through the use of transportation system (streets, roads, highways, bus lines, subway, train, etc), the interaction between land use patterns can be processed in the sense that the infrastructure provides the condition for movements within the urban area. In addition to its supply role, the transportation system can be used to induce the development. It means that a change in accessibility determines the change in the value of land and it may affect the way land is currently used. In a temporal perspective, if such change does occur the travel process will change [4].

From this background, it is assumed that occupied area and extension of each mode can quantify land use patterns and transportation systems, respectively. So, be an urban area that is divided in Homogenous Aggregated Sectors (*HAS*), which consists of a set of land use patterns. Trips (T_{ij}) between a pair of *HAS* pairs are verified, where i and j denote the origin and destination, respectively. Land use is represented by occupied area (meters square) of each pattern in each *HAS*, which is assigned to RLU_{HAS} (Residential Land use), CLU_{HAS} (Commercial Land use) and SLU_{HAS} (Service Land use). Transportation system is characterized by the extension of each mode (meters) in each *HAS*, which is associated to RTS_{HAS} (Road Transportation System), BTS_{HAS} (Bus Transportation System), STS_{HAS} (Subway Transportation System) and TTS_{HAS} (Train Transportation System). In order to represent the spatial distribution of a pair of *HAS*, it is defined the distance D_{ij} (meters).

In a feedforward Multilayer Perceptron (MLP), it is established the input and output vectors [9]. The input vector (X) is defined as

$$X_{ij}=(RLU_i, CLU_i, SLU_i, RTS_i, BTS_i, STS_i, TTS_i, RLU_j, CLU_j, SLU_j, RTS_j, BTS_j, STS_j, TTS_j, D_{ij}) \quad (1)$$

The definition of the output vector depends on the level of planning analysis. In transportation planning context, travel-demand forecasting attempts to quantify the amount of travel on system. This quantification can be performed in terms of trip values or levels of intensity of movements. We propose two formulations for output layer of the NN structure as described following:

2.1 Structure 1

This model assigns the NN as a function approximator. The network has to be able to model the relation between trips T_{ij} and input vector X_{ij}

$$\overline{T_{ij}} = \Psi(X_{ij}) \quad (2)$$

where $\overline{T_{ij}}$ is the forecasted trip value and Ψ is the non linear mapping function performed by the NN. The activation functions of the hidden and output units are set to logistic bipolar and identity function, respectively. For the training and testing data, the output vector Y assumes $Y = T_{ij}$.

The output of network generates an integer value as result that can be used in the analysis of small parcels of the urban area. The more exact the result, the more representative and useful it can be for microanalysis such as congestion and traffic lights.

2.2 Structure 2

Here, the NN is assigned as a pattern classifier. NN's output is used to forecast the levels of urban movements as main element to be evaluated in strategic planning. The model's final output classifies the forecasted movements in levels such as high, medium-high, medium, medium-low and low. The output vector of the training / testing data set has to be modified. So, T_{ij} is linearly quantized in z levels and represented by z output units y defined as

$$y_k = \begin{cases} 1.0, & \alpha_k \leq T_{ij} < \beta_k \\ 0.0, & \text{otherwise} \end{cases} \quad \text{where } k \in (1, 2, \dots, z) \quad (3)$$

where α_k and β_k , are the minimum and maximum T_{ij} values that define the k level. It is clear that the number of output units in the NN structure is directly related to number of z levels.

3. METHODOLOGY FOR THE INTEGRATION OF RS AND GIS AND NN

The description of this methodology is concentrated on the integration process and the role of its components. The integration process starts with data obtainment. In the sequence, satellite images and aerial photographs are stored in RS database and Origin/Destination matrix is organized in Trips database. Finally, maps containing transportation system information (road, subway, train, bus, etc) are transferred to Maps database. In this obtainment process, it is fundamental to define a common aggregation level once this definition will direct affect all modelling process.

After this step, RS data is used in GIS environment in order to process multispectral analysis and aerial photograph interpretation generating the land use patterns. After, they are incorporated into GIS database that previously stored transportation and trips data. Spatial-temporal queries are processed using this database and the results are conducted to NN processing. In this module, query results are pre-processed and then divided in training and test data sets. NN software performs the training process defining a function that is tested and revised until the expected level of forecast is reached. Finally, NN returns to GIS, where travel forecast function is applied to future scenarios.

4. CASE STUDY

This case study intended to verify model's efficiency and evaluate the variations in output layer. It was conducted in Boston Metropolitan Area (Massachusetts State – USA), which covers about 1400 square miles (3580 square kilometers) with nearly three million people live. It was selected a study area in Boston South area, near to Boston Medical Center that involves seventeen Traffic Zones (TZ). It was assumed that the TZ 's were equivalent to HAS definition. The study area is mainly occupied by residential land use and it is located very near to downtown.

RS and Transportation system data were obtained from *MassGIS* database. In this database, data was projected to Massachusetts State plane Mainland Zone (FIPZONE 2001) coordinate system, Datum NAD83, unit meters. It was used black and white digital orthophotos produced in 1992 in 1:5000 scale. Additionally, bus route maps from Metropolitan Boston Transportation Authority (MBTA) were incorporated to Transportation data. Finally, Central Transportation Planning Staff (CTPS) provided access to travel data related to 1990 survey that involves all purposes trips as well as TZ definition.

In the sequence, GIS database construction activities are described. Next, NN experiments and results are introduced.

4.1 GIS database

Using a GIS software, firstly, TZ 's were defined as shown in figure 1. Next, Land Use patterns were obtained following *United States Geological Service* (USGS) classification system [3] and Taco's methodology [8]. It was

analyzed up to Level II (Residential, Commercial and Services) according to model's requirements (see item 2). In the sequence, Transportation System was transformed in digital format inside the GIS database.

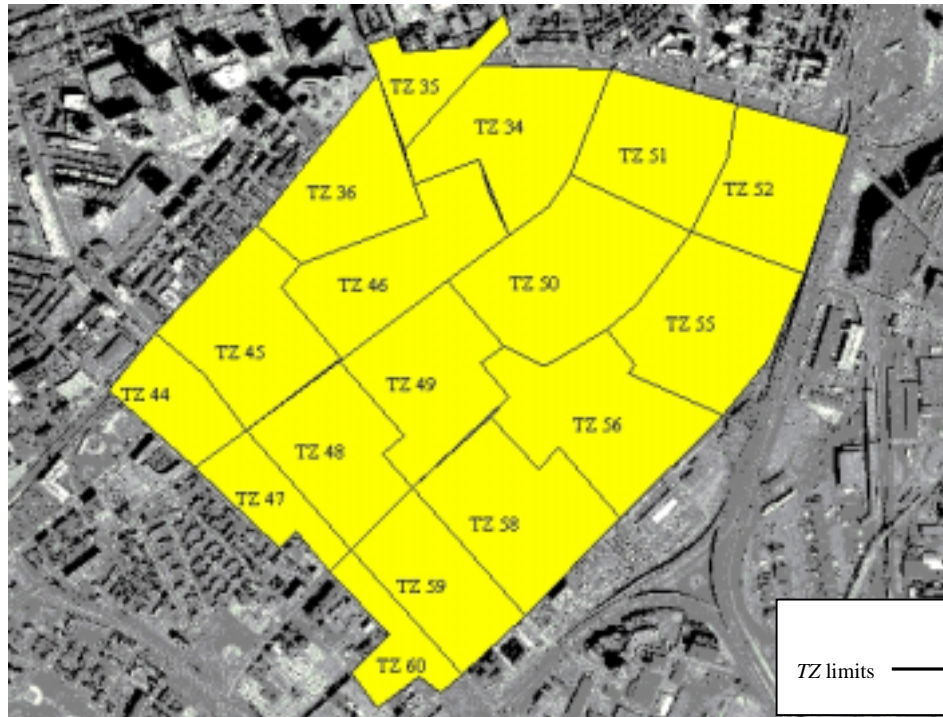


Figure 1: Study area with the selected TZ

4.2 NN experiments and results

From the spatial queries, it was generated a data set formed by 289 vectors. The input vector X was normalized using

$$xn_c = -1 + 2 \left[\frac{(x_c - x_c^{\min})}{(x_c^{\max} - x_c^{\min})} \right] \quad (4)$$

where x_c is the value from X input vector for characteristic c , according to Eq.(1), xn_c is the normalized value and x_c^{\min} and x_c^{\max} are the minimum and maximum values for the related characteristic.

Using this normalized data set, it was defined the training (75%) and testing (25%) data set randomly selected, so 216 and 73 vectors were assigned to training and test sets, respectively. Applying a backpropagation algorithm, the networks were trained until Minimum Square Error (MSE) error in the test set was reached. We simulated the Structures 1 and 2 varying the number of layers and units in the hidden layer. In both cases we used an input layer with 15 units. The best results for each structure are described and discussed next.

4.2.1 Structure 1

A four-layer structure composed by two hidden layers with 15 and 7 units in each one presented the best results. It was reached a $MSE=10.0$. Figure 2 shows expected (T) and forecasted (\hat{T}) trips and the absolute difference between them (error) for the test data set. The average error was 2.38 trips, which denotes a good result considering the complete data set. However, this error is high in some special cases where the network was not efficient to do the modelling.

In these cases, one or more characteristics of the input vector (X) were zero. It is resulted due to the adoption of the TZ 's aggregation level that provided data for a very limited area, where some input vector characteristics do not exist. It is expected that NN can achieve a better level of forecast, if more aggregated areas are used.

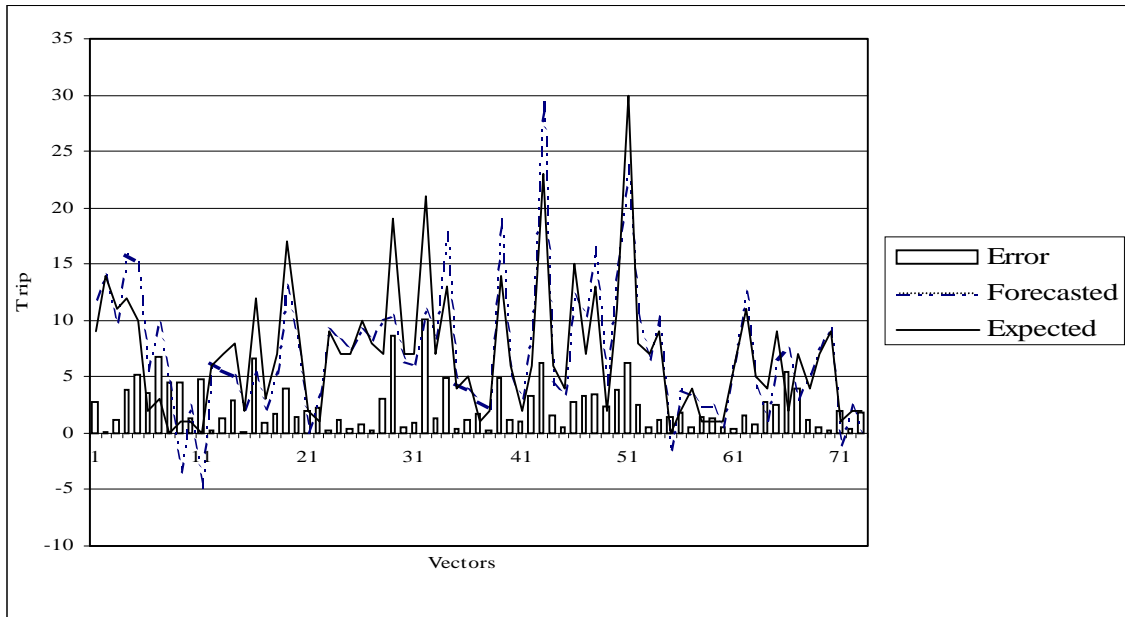


Figure 2: Forecasted and Expected Trips

4.2.2 Structure 2

First, the T matrix was quantized according to Eq.(3) and considering $z=5$ levels. For this structure, the best result ($MSE=0.15$) was also for a four-layer structure with 15 and 7 units in the hidden layers. It represents a recognition rate of 94.5%. Analyzing in detail the forecasted results, it is verified that the errors occurred in five testing vectors. These vectors were related to levels 2, 3 and 4. On the other hand, all the test vectors related to level 5 ($0 \leq T_{ij} \leq 14$) were correctly forecasted.

The composition of training set can explain this NN performance. It is observed that we are treating with a very imbalanced problem since there is a great concentration of data on level 5 (252 vectors – 87% of the total), becoming that the dominant level. Meanwhile, levels 1, 2, 3 and 4 have only 1, 3, 5 and 28 vectors, respectively. This composition led to the construction of training and testing data sets that resulted in NN efficiency just for those vectors in level 5.

Despite the high recognition rate, the final results were compromised by data composition. This problem has been observed in real-life data and some research has been developed in order to solve it. Some efforts have been dedicated to change the backpropagation algorithm to avoid such imbalanced training [1]. Alternatively, we propose in this work a solution focused on the treatment of training / test data in order to reach a smooth distribution of the vectors. In this sense, it is necessary to add some data in those levels, which are formed by few vectors (levels 1,2,3,4) and select some from level 5.

It was defined that 40 and 10 vectors of each level will form the training and testing data, respectively. First, new data was estimated for those vectors related to levels 1, 2, 3 and 4 by adding gaussian noise with average zero and variance 0.001 to those existing vectors, generating new noisy versions. This operation was performed in order to supply the NN training with vectors highly correlated with original ones. Next, the number of vectors in the level 5 was reduced creating a new set of more representative vectors using the *LBG* algorithm [5]. The *LBG* aims to design *quasi-optimum* codebooks in vector quantization based coding systems. It was applied to create 50 template vectors for the level 5 by minimizing the *MSE* of the representation for all vectors of this level.

Using this balanced data training set, the best result ($MSE=0.10$ and recognition rate 94.0%) was obtained for a four-layer NN with 15 and 7 units in the hidden layers. The balanced data provided a better modelling for those levels that had only few samples in the original imbalanced training data set. Table 1 shows the performance reached in the test set. In all categories, the NN was able to forecast correctly the most part of testing vectors. In the vectors that resulted in an incorrect forecast, it is noticed that one or more characteristics of the input vector (X) are zero. Specifically, the errors were related to trips that do not contain train and subway transportation systems. Physically,

these errors express how difficult it is for NN to reach the expected result without the information about the transportation system.

Table 1 – Recognition rate for levels

	Level 1	Level 2	Level 3	Level 4	Level 5	Final
Imbalanced data	0%	0%	0%	33%	100%	94.5%
Balanced data	100%	100%	100%	80%	90%	94.0

5. CONCLUSIONS

This paper presented the development of a NN based model for transportation planning that provides the representation of urban dynamic in travel forecast process. After the definition of NN's formulation for travel demand modelling, the case study for Boston Metropolitan Area proved model's efficiency. For both formulations (Linear and quantized output), it was reached a considerable recognition rate. The results showed that selected characteristics are suitable to represent the problem

The formulation for linear output showed to be less exact than the quantized one. It was expected since in the linear formulation (Structure 1) the NN is used to model a forecast function leading to a poor generalization capability. Meanwhile, the quantized formulation (Structure 2) establishes levels, which leads to a very limited variety of outputs to be forecasted. Consequently, it becomes easier for Structure 2 to reach better *MSE* results as observed in the simulations.

For both formulations, it was detected that the nature of the data has important influence in the modelling. In the simulation of the Structure 2, we developed a special treatment on the data that provided better results than using original imbalanced data. However, there is another possible direction to solve this limitation that is related to the aggregation level of the data. In these experiments, *TZ* definition was used considering that there was not any other suitable definition of *HAS*. In Boston's case, *TZ* definition was designed for microanalysis generating units with a very limited area and specific characteristics of land use and transportation system. It is expected that using a macro aggregation level these kinds of errors would be eliminated. These improvements will be incorporated to the spatial-temporal model that is expected to process travel forecast considering a multi-year evolution of land use patterns and transportation system.

Acknowledgments: The authors would like to express gratitude by the scholarship supported from Japanese Ministry of Education (Monbushou) for the development of this research. We also would like to thank Central Transportation Planning Staff (CTPS) of Boston for all support and travel data used in this research.

REFERENCES

- [1] R.Anand, K.G.Mehrotra, C.K.Mohan, and S.Ranka, "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets", *IEEE Trans. on Neural Networks*, Vol.4, N.6, pp.962-969, Nov.1993.
- [2] T. E. Avery and G. L. Berlin, *Fundamental of Remote Sensing and Airphoto Intrepretation*, Maxwell Macmillan International, New York, USA, 1990.
- [3] B. Harris, *Land use Models in Transportation Planning: a review of past developments and current best practice*, Delawae Valley Regional Planning Commision, USA, 1996.
- [4] C. J. Khisty, *Transportation Engineering: an introduction*, Prentice-Hall, 1990.
- [5] Y. Linde, A. Buzo, A, and R. M. Gray, "An Algorithm for vector Quantizers design", *IEEE Trans. on Communications*, V.COM-28, pp.84-95, Jan. 1980.
- [6] S.Openshaw and S.Alavanides, "Applying geocomputation to the analysis of spatial distributions", *Geographical Information Systems: Principles and Technical issues*. Eds Longely, P.A., Goodchild, M.F, Maguirre, D.J., Rhind, D.W., V.1, Second Edition, USA, 1999.
- [7] J. P. Rodrigue, "Parallel modelling and Neural Networks: an overview for transportation / land use systems", *Transportation Research C*, Vol. 5, N. 5, pp.259-271, 1997.
- [8] P. W. G. Taco, Y. Yamashita; A. S. Dantas, and N. M. Souza, "Trip Model with the Application of Remote Sensing and the Geografic Information Systems", CD ROM 004 - *68th ITE Meeting*, Canada, Aug. 1998.
- [9] P. D. Wasserman, *Neural Computing: theory and practice*, Van Nostrand Reinhold, New York, USA, 1989.