Computer-Aided Civil and Infrastructure Engineering **21** (2006) 104–119

Comparison of Neural Networks and Gravity Models in Trip Distribution

Frans Tillema*, Kasper M. van Zuilekom & Martin F. A. M. van Maarseveen

University of Twente, Department of Civil Engineering, Centre for Transport Studies, P.O. Box 217, 7500 AE, Enschede, The Netherlands

Abstract: *Transportation engineers are commonly faced with the question of how to extract information from expensive and scarce field data. Modeling the distribution of trips between zones is complex and dependent on the quality and availability of field data. This research explores the performance of neural networks in trip distribution modeling and compares the results with commonly used doubly constrained gravity models. The approach differs from other research in several respects; the study is based on both synthetic data, varying in complexity, as well as real-world data. Furthermore, neural networks and gravity models are calibrated using different percentages of hold out data. Extensive statistical analyses are conducted to obtain necessary sample sizes for significant results. The results show that neural networks outperform gravity models when data are scarce in both synthesized as well as real-world cases. Sample size for statistically significant results is forty times lower for neural networks.*

1 INTRODUCTION

Transportation engineers are commonly faced with the question of how to extract information from expensive and scarce field or survey data. A common approach is to create a statistical or gravity model that describes the behavior of the phenomenon observed in which the data are used for calibration/validation. Ideally, such an approach leads to a model with the desirable high accuracy. Unfortunately, there is often a discrepancy between this desire and the obtained accuracy; estimating a model, based on scarce data is not an easy job and can lead to results with high deviations. Furthermore, it is not always easy to construct a statistical model from the data, due to the fact that many phenomena are nonlinear, and/or collinear (Huisken and Coffa, 2000).

Classical transport planning, described in the classical 4-step model (Ortuzar and Willumsen, 2001), is characterized by dependency on data. Spatial interaction patterns, for example, person-trips between zones, and the trip distribution, are highly complex and difficult to model without adequate amounts of data. Errors that are generated during the trip distribution estimation process propagate through till the assignment phase. This causes difficulties for good transport planning. Currently used techniques, like the classical gravity models (GM), try to use limited amounts of data. The question arises, whether these techniques are able to give good trip distribution estimations.

Since the beginning of the nineties, neural network models were introduced as alternatives for traditional (statistical) modeling approaches. Recent literature gives an insight into the opportunities of using neural networks to model spatial interactions. Openshaw and Openshaw (1992) give their opinion on the advantages of using neural networks in geographical/transportation analysis. An eye-catching conclusion is the better performance of these models compared to more traditional models.

Research conducted by Miller et al. (1995), Dougherty (1995), Collins et al. (2001), Pijanowski et al. (in press), Raju et al. (1998), Huisken and Coffa (2000), Currit (2002), and Faghri and Sandeep (1998) supports this notion. These studies carefully reveal the opportunities of applying neural networks in a land-use and traffic and transport context. In addition, Rodrique (1997), Tillema et al. (2002), and Tillema (2004) give an insight into the possibilities of both parallel computing and using neural networks for integrated land-use/transport systems.

To whom correspondence should be addressed. E-mail: *f.tillema@ctw.utwente.nl*.

Based on the assumption that most spatial systems are parallel, a conceptual model using neural networks is developed. So far, these ideas have not surpassed the level of conceptual models.

Several studies have explored the usefulness of neural networks in the context of trip distribution modeling. However, the empirical results leave questions open about whether neural networks give better results than traditional trip distribution methods. Black (1995) asks the question if the basic purpose of neural networks, identifying patterns in data, and to replicate those patterns for new data, can be utilized in a spatial context. He makes a comparison between a gravity model and neural networks. Black uses two case studies: (i) a three-region flow problem; and (ii) a commodity flow problem. The first problem is a very simple three-region flow problem. Both doubly constrained GM as well as a neural network model give excellent results. Black emphasizes that one should not lose sight of the fact that the matrices have only nine flow values. The commodity flow problem gives similar results, ranking the scores of artificial neural networks (ANN) above the scores of GM. Finally, Black concludes that neural networks are capable of high levels of accuracy based on their use in other fields and are suitable for future flow forecasting.

Fischer and Gopal (1994), Gopal and Fischer (1996), and Fischer (1998) compare the forecasting results of neural networks to those of a traditional gravity model. Research is done into the distribution of interregional telecommunication flows. Although the test case is not a traffic- and transport-related problem, the problem is to a large extent comparable with a trip distribution problem. The basic conclusion is that the neural network models outperform the conventional gravity model.

Mozolin et al. (2000) compare the performances of neural networks and maximum likelihood doubly constrained models for commuter trip distribution. The authors state that their approach differs drastically from others in several respects: (i) the models are used in a predictive mode and calibration is done on observed data, while testing is conducted on data for a projection year; (ii) the baseline problem is a doubly constrained model estimated by maximum likelihood; (iii) the models are evaluated on origin-destination matrices of different sizes to be able to test the sensitivity of the conclusions to the size of the interaction system; and (iv) the model applies an adjustment factor to flows predicted by the neural network output to satisfy constraints. It is concluded that neural networks exhibit good to very good ability to predict future commuter flows. Yet, none of the tested neural networks outperforms the corresponding doubly constrained model. The authors find this fact puzzling and unexpected. After further data analysis the following results are formulated: (i) due to over-fitting the ability to generalize is rather poor and the prediction accuracy is low particularly where training data are scarce; (ii) networks fail to extrapolate around and beyond the limits of the training sample; (iii) networks with less hidden nodes are less prone to overfitting; (iv) the ability to approximate data structures with great accuracy is also their weakness.

So, ANN are increasingly used as data analyzing techniques in a spatial interaction, trip distribution context. Yet, the conclusions whether neural networks outperform more traditional models are still under discussion. The aim of this study is to explore the performance of neural networks in trip distribution modeling and to compare the results to more commonly used doubly constrained GM. In addition, the aim is to explore the usefulness of neural networks when field data are scarce.

The approach differs from other research in several respects. Firstly, the evaluation is done based on large synthetic data sets, as well as a real-world data set. Secondly, synthetic data (OD matrices) are used to explore neural network performances under circumstances of increasing complexity. The well-defined differences between OD matrices increase the controllability of the test; differences in results can easily be attributed to the buildup of the data. Thirdly, statistical analysis is conducted to find minimum necessary sample sizes for both models. Fourthly, like Thill and Mozolin (2000) mentioned, the neural network output is enforced on the production and attraction constraints, in this case by using the Furness method (Orthuzar and Willumsun, 2001). Finally, the neural networks and GM are calibrated using different percentages of hold out data. In this way one of the biggest advantages of neural networks, extrapolating/forecasting of missing data (patterns), can be examined.

A complete OD matrix is generated using a gravity model. This results in a completely known OD matrix, without noise and measurement errors. The basic test is a synthetic spatial network of 15 regions, combined with synthetic impedances and attraction/production values. The second test is a comparison of different estimation methods on observed trip patterns in a real-world network, Rotterdam Rijnmond (National Regional model, NRM). The known data from the generated OD matrix are split up into calibration and test data. The calibration percentage is varied between 10 and 90. A complete OD matrix will be estimated, using limited (observed) data and trip attraction and production totals.

The article is organized as follows. The first section gives an introduction into spatial interaction modeling and trip distribution modeling. The following section of this article presents the organization of the test. Finally, the third and most important section discusses the performance of a trip distribution model using neural

networks. Preprocessing is conducted to find a good neural network topology and configuration. Comparisons are made between the performances of both traditional doubly constrained GM and neural networks. In addition, answers are given to the question of whether performances increase or decrease when real-world data are used instead of synthetic data and if the problem of separable matrices deteriorates the results.

2 SPATIAL INTERACTION MODELING

Spatial interaction may be defined in general terms as any flow of commodity, people, capital, or information over space resulting from some explicit or implicit decision process (Fotheringham and O'Kelly, 1989). Productions and attractions provide an idea of the level of trip making in a study area, but this is often not enough for modeling and making decisions (Ortuzar and Willumnsen, 2001). What is needed is a better idea of the pattern of trip making, from where to where do trips take place (trip distribution), the modes of transport chosen (model split), and the routes taken (assignment). A high-quality trip distribution model is a necessary prerequisite for an accurate and usable travel demand model. But how is trip distribution modeled?

An interesting problem is generated when information is available on the number of trips originating and ending in each zone. The sum of all trips in a row, the trip production, should equal the total interaction flows exiting a particular zone.

$$\sum_j T_{ij} = O_i, \forall i \quad \text{(trip attraction constrained)} \qquad (1)$$

The total number of all trips in a column, the trip attraction, should equal the total interaction flows entering a particular zone.

$$\sum_i T_{ij} = D_j, \forall j \quad \text{(trip production constrained)} \qquad (2)$$

When both Equations 1 and 2 hold, the model is called a doubly constrained gravity model:

$$T_{ij} = A_i O_i B_j D_j f(c_{ij}) \qquad (3)$$

where A_i and B_j are balancing factors, c_{ij} is the travel impedance, $f(c_{ij})$ is the distribution function.

Trip distribution can be modeled with any number of constraints. It has been shown that estimating spatial interaction with a doubly constrained gravity model yields the most accurate results (Ortuzar and Willumnsen, 2001). So, the doubly constrained model is used as the benchmark in this article.

Problems arise in trip distribution modeling when data are scarce. Unfortunately, field data are often scarce due to the fact that it is difficult and furthermore expensive to obtain. This influences the quality of trip distribution modeling strongly.

3 ORGANIZATION OF THE TEST

The focus of this research is on comparing the performances of neural networks and GM in well-defined basic and real-world cases (Figure 1). Firstly, synthetic input data are generated: (i) synthetic network; (ii) synthetic skim matrix (impedance); and (iii) synthetic input data of different complexities (OD matrices). Synthetic data give the opportunity to play with complexity. This approach gives an insight into the impact of complexity, without modeling noise or unclear relations between variables. Neural networks and GM are calibrated on different percentages of the input data. GM are calibrated using the Hyman (1969) algorithm. Finally, conclusions are drawn on overall performances and the impact of different percentages of hold out data with respect to the model performances. Neural networks are trained and tested with Matlab 5.3 (The Math Works, Inc. 1996, 1998).

3.1 Synthetic network

A synthetic network combined with synthesized impedances (schematized in a skim matrix) and synthesized trip attraction and production values define trip distribution modeling inputs. The choice for 15 regions results in a 225 cell origin-destination (OD) data set. The use of a simple synthetic 15-region network gives us the opportunity to carefully explore neural networks usefulness in trip distribution modeling. The regions are located on a straight line and distances in between regions are equally distributed. Spatial impedance between regions is mostly measured by the Euclidean distance. Normally the use of the Euclidean distance, instead of the average travel time, enhances the results of neural networks and vice versa for a gravity model (Mozolin et al., 2000). The impedance of the synthetic network is defined as nondimensional.

The logistics of spatial interaction modeling requires clearly defined regions with no, or small, flows across the borders. In the case of the synthetic network, this assumption is not violated. Setting intrazone distance to zero is known to generate systematic measurement errors. Therefore spatial separation within the regions, an interzone distance greater than zero, is introduced within the network.

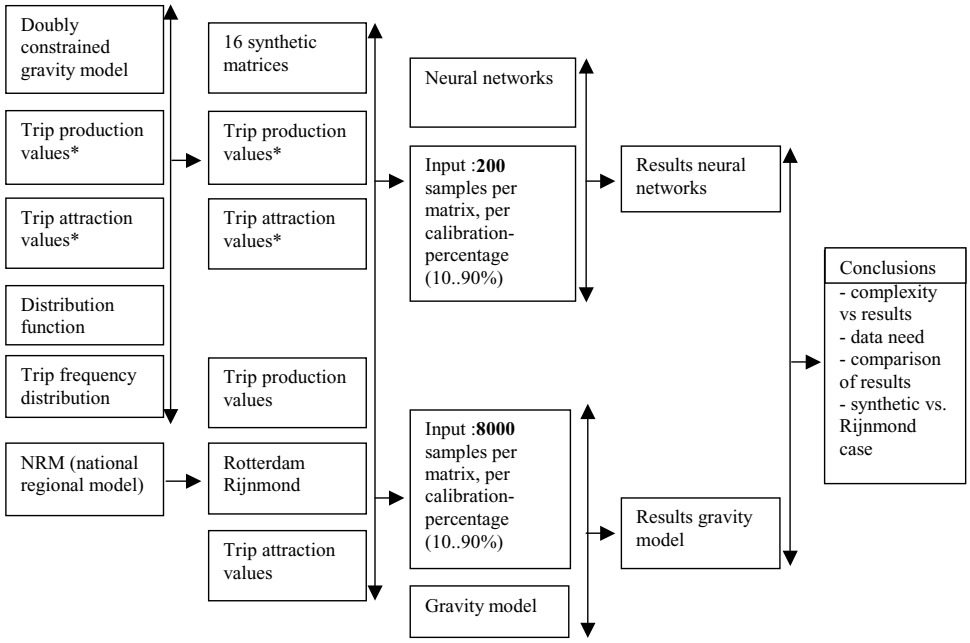

Fig. 1. Organization of the test.

3.2 Synthetic OD matrices

Trip distribution estimation requires input values for the distances between regions as well as trip generation and attraction values. Trip generation and attraction have been synthesized. A total of 15,000 trips is distributed among the zones as schematized in Figure 2.

Each cell in Figure 2 is a synthetic OD matrix. The icons on the edge of the figure (rectangular, triangular, or a combination of these two) show the distribution of the 15,000 trips divided over origins, respectively, the "destinations." From 1 to 16, the matrices' complexity increases: for example, matrix 1, cell 1, is built with evenly distributed origins/destinations; matrix 3, cell 3, is built with evenly distributed origins and a descending pattern for destinations. The well-defined differences between OD matrices increase the controllability of the test. Differences in results can easily be attributed to the buildup of the matrices. Cells in the figure indicated by lines are duplicates; configurations of these matrices are already tested in one of the matrices, matrix 1–16.

The complexity of the matrices is shown in Figure 2 between brackets. The complexity is based on the patterns for destination and origins and the interactions within the matrix.

3.3 Input data format

Consider the flows between the regions as given in matrix 1–16. The distances are given in the skim matrix. Input data are set up as productions, attractions, and distances. All network inputs are scaled by dividing the value observed for each example by the input's maximum value in the set. Scaling of the output is required and is done by dividing by the maximum output value.

3.4 Comparison measure

To compare the performances the error definition that was used is the root mean square error (RMSE). The RMSE is mathematically described by:

$$RMSE = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}\left(x_i^{observed} - x_i^{predicted}\right)^2\right)} \qquad (4)$$

with:

N = number of samples per matrix ($15 \times 15 = 225$).

A model is said to outperform the other if its goodness-of-fit is superior, as measured by the RMSE and the standard deviation. A good fit on the trip production and attraction totals and a low RMSE are no guarantee for good estimates. So, extra analyses have to reveal new information on the fit on the OD-cell level. Therefore comparisons are made between both methods on the average trip length and length distribution. Trip length frequencies give insights into the results of both methods on all trip length categories.

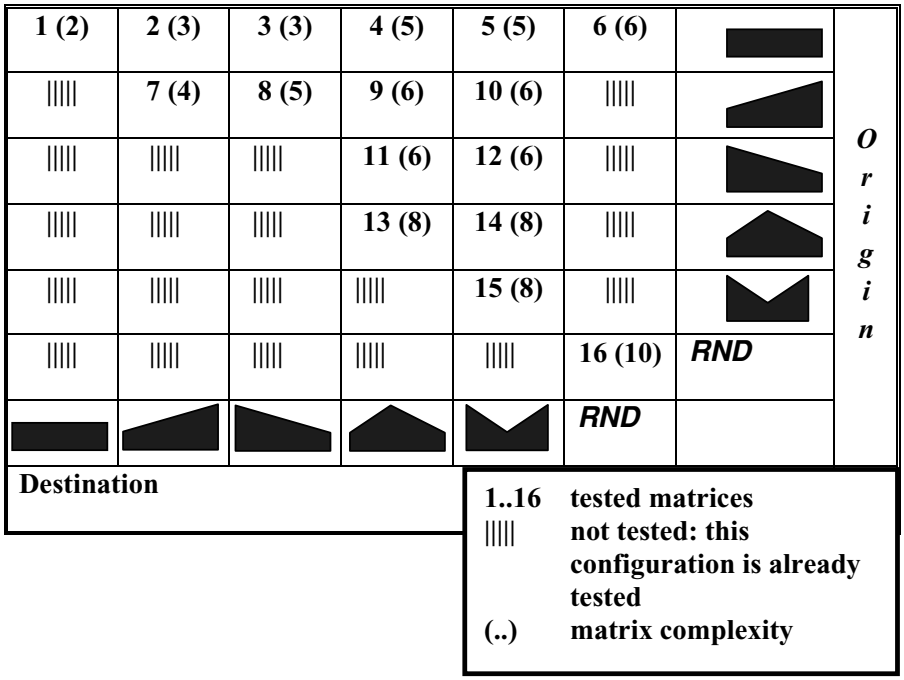

Fig. 2. Synthetic OD matrices.

3.5 Neural network: Input/output

The first step in using a neural network model is the choice for a network topology. The right topology is dependent on the number of relevant inputs and outputs. In the trip distribution problem, topology seems quite clear: (i) three inputs (trip attraction/production and impedance); and (ii) one output (trip distribution). Although all matrices are built around trip attraction and production values and a skim matrix, not all matrices have the same network topology. Input variables that are constant factors cannot be used for the fact that they contain no discriminating information. Dependent on the fact of whether the trip production or attraction can be discriminated upon, the matrices have one (only impedance), two, or three inputs. For example, matrix one is built around evenly distributed destinations and origins, with a value of 1,000. In that case, the only discriminating data on which trip distribution can be estimated are the impedance values. The neural network topology of matrix 1 will be: (i) one input (impedance) and (ii) one output (trip distribution). The topologies are:

- 1 input, 1 output → matrix 1;
- 2 inputs, 1 output → matrix 2, 3, 4, 5, and 6;
- 3 inputs, 1 output → 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, and Rotterdam Rijnmond.

3.6 Real-world matrix: Rotterdam Rijnmond region

Rotterdam, famous for its harbor, is the second city of The Netherlands with a population of 0.6 million inhabitants. Rotterdam Rijnmond (Figure 3) is the whole area of Rotterdam including the harbor and suburbs. Using the NRM zoning method, Rijnmond is divided into 15 zones. A total number of nearly 1.9 million car trips per 24 hours is made, calculated over all motives. Spatial impedance between counties is simply measured as time between the zone centroids.

4 PREPROCESSING

4.1 Reasons for preprocessing

The next step is selecting a NN configuration and parameters. This is difficult and often based upon only a limited number of criteria. It is common practice to proceed by trial and error to select the number of hidden nodes, and to test networks with layers of varying size.

Preprocessing was conducted for several reasons. Firstly, both neural network performance and computer time are strongly related to the network configuration. Preprocessing gives an indication of what the best number of hidden neurons is. So computer and analysis time can be decreased. Secondly, computer time is dependent on the number of training epochs. Preprocessing gives an indication of the minimum necessary number of epochs.

Fig. 3. Rotterdam Rijnmond area (dots).

Thirdly, we search for statistically significant conclusions. Because of the random aspect of the calibration process of both neural networks and GM, we need an indication for the minimum sample size for statistically significant conclusions. Finally, preprocessing gives preliminary results on performances of neural networks and GM and helps to draw conclusions in the end.

4.2 Preprocessing setup

Preprocessing was conducted using 6 of the available 16 synthetic OD matrices, varying in complexity and representative of other matrices. The neural networks were calibrated using both 30% and 70% of the data set; 68 and 158 data vectors, respectively. Epochs are varied between 20 and 220, with a step size of 40. A total number of 200 random draws of 68 and 158 data vectors are performed. This results in 200 random input data sets per configuration. These are used to determine the average performance of neural networks. In addition, tests are done to determine the average necessary number of calculations/draws for statistically significant conclusions. To streamline the preprocess and to get a first impression of a suitable network configuration, a rule of thumb for the number of hidden nodes was used:

$$Number\_of\_hidden\_nodes = \sqrt{(m * n)} \qquad (5)$$

With:

1. $m =$ the number of output neurons (always 1 in this case); and
2. $n =$ the number of input neurons (1–3).

According to this rule, the number of hidden neurons would vary between 1 and 3. However, preliminary experimentation leads to the conclusion that this rule of thumb is not generally applicable: the optimum number of hidden nodes varies per matrix and test percentage. Therefore the total number of hidden nodes was varied between 1 and 20. Matrices 1, 5, 7, 9, 11, and 14 were used in preprocessing. The setup was organized as follows:

(i) Data percentage → 30/70%;
(ii) Epochs → 20–220;
(iii) Hidden neurons → 1–20;
(iv) Number of samples per configuration → 200.

4.3 Results preprocessing

Epochs. The conclusions on the number of epochs are based on a Kruskal-Wallis analysis of variance (ANOVA) and median test. Networks trained on 100–220 epochs do not show significantly different results. Besides that, networks trained on less than 100 epochs show significant differences in results with networks trained on more than 100 epochs. The performance (RMSE) of networks trained on less than 100 epochs showed to be significantly worse than networks trained with more than 100 epochs. It is therefore concluded that in further and future research, networks can be safely trained with only one epoch configuration: 100 epochs.

Hidden neurons. Training and testing the different matrices revealed interesting results. Figure 4 gives information on mean RMSE per network configuration (matrices 1, 5, 7, 9, 11, 14).

Although at first sight all figures give different results, three general preprocessing conclusions can be drawn. Firstly, neural networks trained on 70% of the train set give better results on the RMSE, as expected. Secondly, neural networks trained on 30% give good results when the number of hidden nodes is between 3 and 5. Finally, contrary to the former conclusion, neural networks trained on 70% perform, in general, better when the number of hidden nodes increases. However, results do not significantly improve when more than 10 hidden nodes are used. In general, no conclusions can be drawn on one "generic" best network configuration. Taking into account the computer time that increases with increasing number of hidden nodes, the following network configurations will be used in the comparison with the gravity model:

(i) 10–30% data → 5 hidden nodes;
(ii) 40–70% data → 8 hidden nodes;
(iii) 80–90% data → 10 hidden nodes.

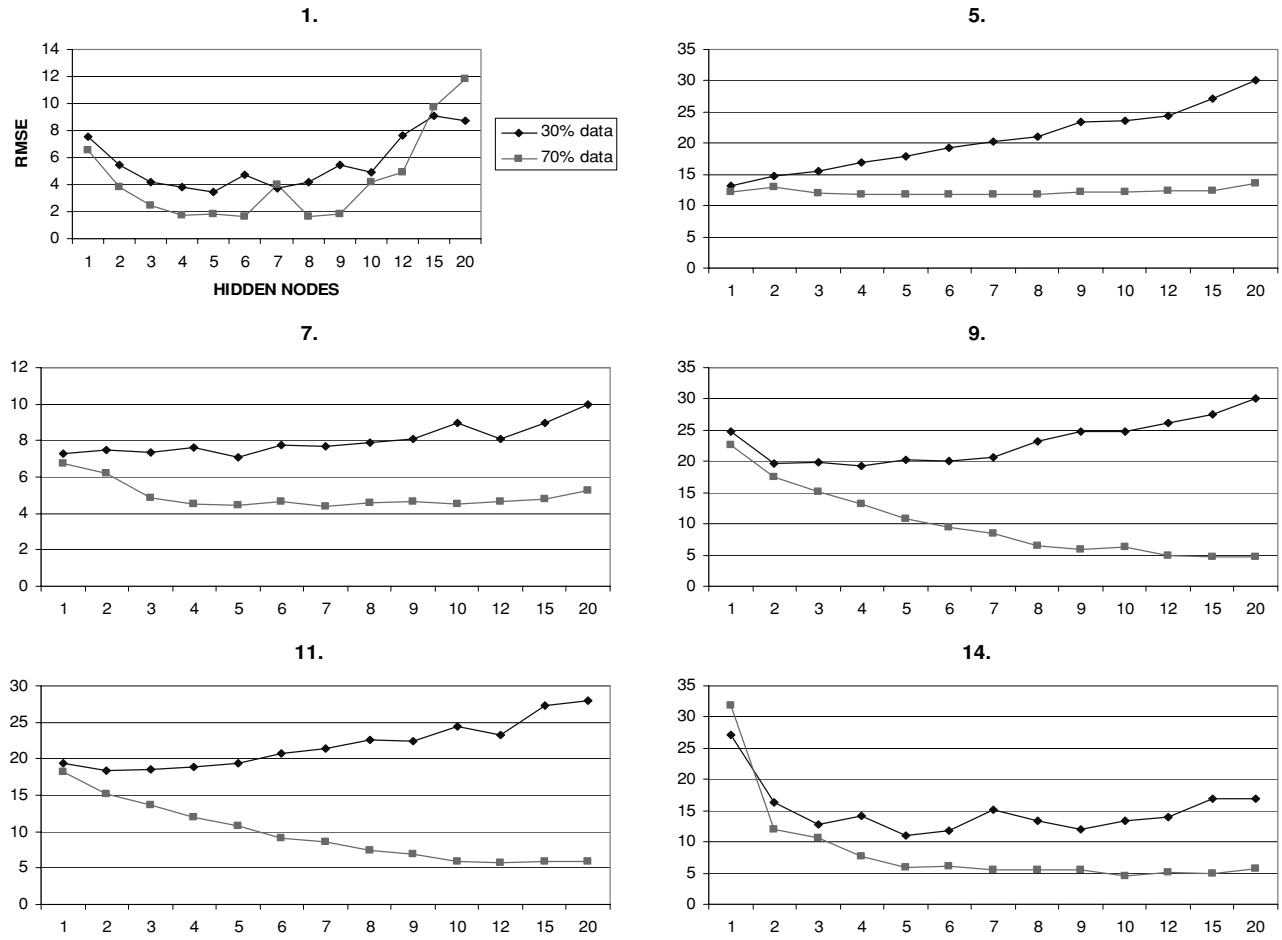

Fig. 4. RMSE when hidden neurons are varied between 1 and 20 (matrices 1, 5, 7, 9, 11, 14).

4.4 Number of calculations per OD matrix

Research is done into the average necessary number of calculations for statistically significant conclusions. Due to the random initialization process of neural networks and the randomly selected draw from the full data set for both neural networks and GM, only an average result out of a number of calibration runs can give comparable results. To determine the necessary number of calculations it was assumed, although not statistically proven, that both neural networks and GM outcome are distributed normally. This results in the following definition of the necessary number of calculations (or sample size):

$$n > \frac{Z^2}{d^2}\sigma^2 \tag{6}$$

With:

1. $n =$ minimum number of samples;
2. $Z =$ distribution value, dependent on statistical confidence and distribution;
3. $d =$ desired accuracy;
4. $\sigma =$ standard deviation.

For consistency reasons, d is related to an absolute value because of the different values, from 2 to 40, of the calculated RMSEs. As a fixed value we calculated the RMSE when the total number of trips (15,000) is equally distributed amongst the 225 OD pairs; the so-called best guess. Conclusions are drawn on the number of calculations for statistically significant results for neural networks and doubly constrained GM:

(i) 10–30% data → 200 neural network, 8,000 gravity model;
(ii) 40–70% data → 100 neural network, 2,500 gravity model;
(iii) 80–90% data → 25 neural network, 1,500 gravity model.

4.5 Data management after training

Contrary to the doubly constrained gravity model, neural network models have no internal constraints. This

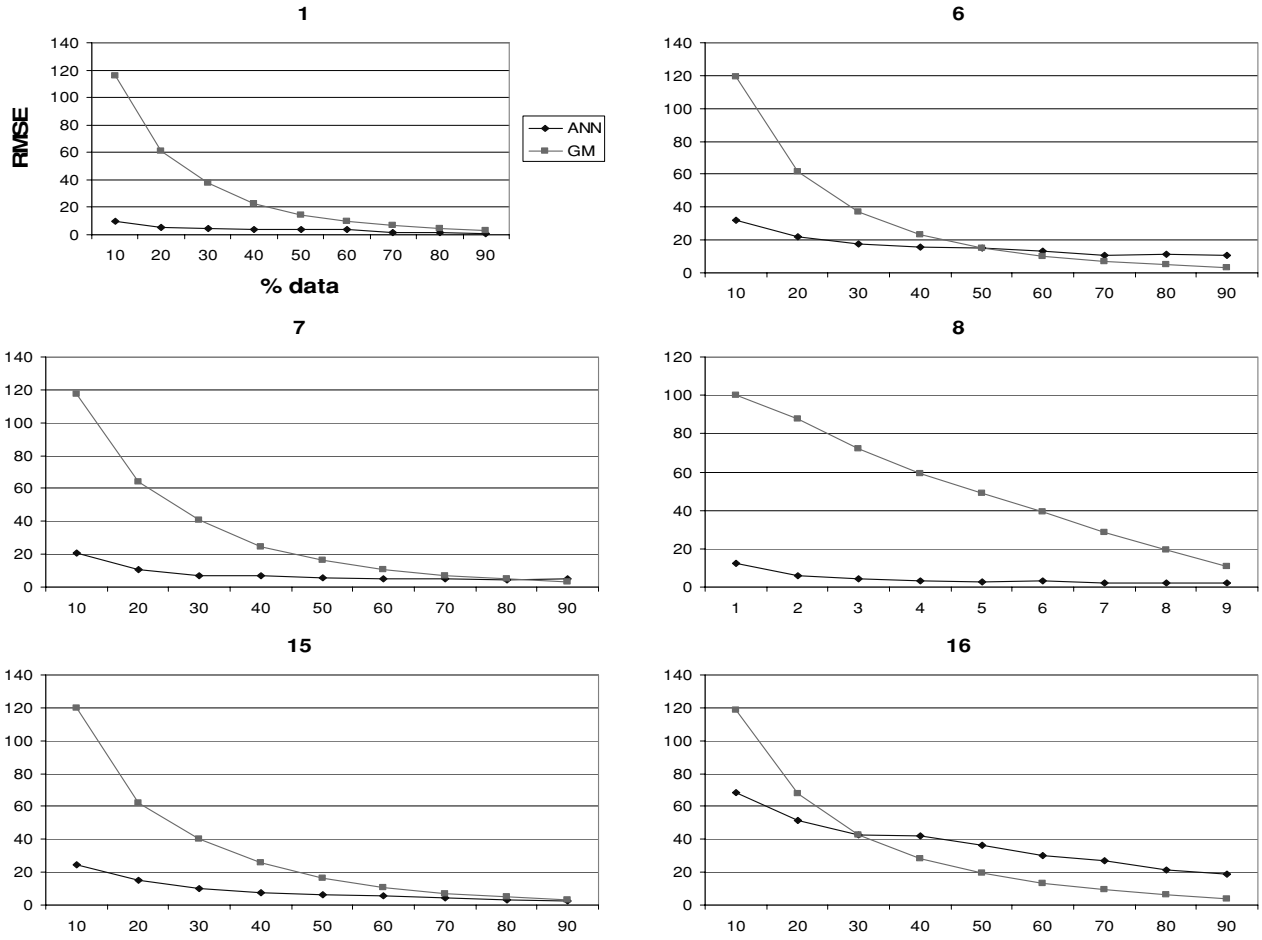

Fig. 5. RMSE of artificial neural networks (ANN) and gravity models (GM) (matrices 1, 6, 7, 8, 15, 16).

means that trained networks are not capable of enforcing the origin and destination constraints. As a consequence, totals of origins and destinations as well as total number of trips in an OD matrix usually differ from the actual input values. So, to enforce the constraints, the first step is to multiply the estimated matrix total (neural network) by the division of input matrix total and estimated matrix total. Secondly, the trip attraction and production totals are enforced using the Furness method.

4.6 Importance of other neural network characteristics

Besides the number of epochs and the number of hidden neurons, other factors determine the quality of the neural networks and consequently the estimations with neural networks. These factors are a.o. neural network charateristics, the activation function, learning method, momentum, and stop criteria. No extensive research is conducted into the optimal values of these factors. Does this limit the conclusions in this research? The answer is twofold. On one hand, it does. We may find a neural network configuration that is not optimal.

On the other hand, not having the optimal neural network structure does not necessarily weaken the conclusions. In a situation where a neural network performs better than a traditional method in (almost) all cases, one can conclude that a neural network performs better. The results might be even better when the network is really optimized. It means that the cases in which the neural networks are compared to traditional methods should be as diverse as possible. In this research, the use of synthetic data helps in providing diverse cases. So, when the results of the study are in favor of neural networks we can conclude that the results might even be better than the shown results.

5 COMPARISON OF MODEL PERFORMANCES

The performance of both ANN and GM are presented in Figure 5. The neural network models outperform

calibrated gravity models especially at low percentages of data. Gravity models outperform neural network models only when sufficient data are at hand to perform a good calibration. Figure 5 shows that most neural network models on average outperform the gravity models when the total percentage of data is under 50%.

In general, the RMSE results are strongly influenced by the percentage of calibration data. At low percentages the results of both methods are less accurate than at high percentages, as expected. Furthermore, the gravity model results are far more influenced by data percentage as are the neural networks.

So, neural networks do not outperform gravity models on the whole scale. When the calibration data set percentage is higher than 80–90%, on average gravity models give better results. This is not surprising, because of the fact that gravity models estimate to a high extent their own creations; at 100%, gravity models always replicate the complete matrix that was created before the result comparison.

Matrix 8 shows strange results. The buildup of the matrix strongly determines the bad results of the gravity model. The trip attraction and production values differ strongly from the theoretical standard distribution function. This standard function gives high trip rates at low distances and low trip rates at high distances. Matrix 8 shows quite the opposite. At low distances people make far less trips than at high distances. Here the advantages of neural networks are shown. Without presuming a certain distance decay function (like the gravity models), the neural network is better capable of estimating the trip distribution. This is a big advantage!

Looking at the Rotterdam Rijnmond matrix, the same observations can be made (Figure 6). Two facts have to be stressed. Firstly, the ANN results are fractionally worse than in the case of the synthetic matrices. Yet, the same pattern is still visible. Secondly, the RMSE values are higher due to the fact that the total number of trips is approximately 130 times higher. However, when the number of trips is related to the RMSE values, the RMSE values are still two times higher.

A good fit on the trip production and attraction totals and a low RMSE are no guarantee for good estimates. So, extra analyses have to reveal new information on the fit on the OD-cell level. Therefore the results of neural network and gravity models trip length frequencies can be studied. Figure 7 gives insights into the trip length frequencies of both methods using 10–50% of the available data. Only results of matrices 1 and 6 are shown. In addition, the data of the Rotterdam Rijnmond case are shown. The remaining matrices show similar results.

Figure 7 gives an insight into three facts. Firstly, for both gravity models and ANN, the performance on the

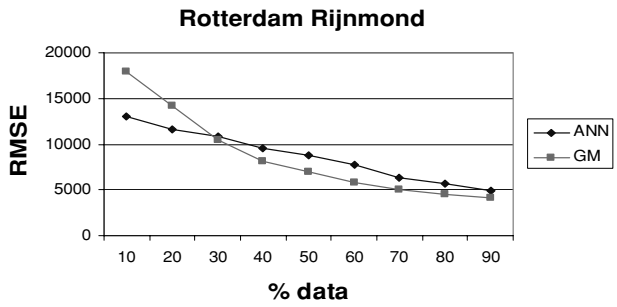

Fig. 6. RMSE of artificial neural networks (ANN) and gravity models (GM) (Rotterdam Rijnmond).

trip length frequency goes up when data percentage go up. Secondly, gravity models seem to have difficulties estimating both the high and the low number of trips. Neural networks only seem to have problems with long distance trips. Finally, the performance at low percentages is much better when neural networks are used, as shown by the smaller range of results. The results perfectly illustrate the conclusions based on Figure 5; when data are scarce (low percentages), the ANN outperforms the GM. When the input data increase, differences in performance decrease.

The results of the Rotterdam Rijnmond case show roughly the same pattern; at low percentages, the ANN outperforms the GM. When the input data increase, differences in performance decrease. Neural networks have more difficulties in estimating extreme values.

6 MORE INSIGHTS IN THE RESULTS

6.1 Matrix complexity

Is there an explanation for these results? Firstly, all matrices show roughly the same pattern; at low percentages neural networks outperform gravity models, at high percentages gravity models outperform neural networks. Neural networks show their ability of extrapolation of data; they can very well cope with small data sets. The performance of gravity models, when calibration data percentage is high, is related to the buildup of the matrices; the matrices were built using gravity models.

So, when the calibration data set nears 100%, the only matrix the gravity model estimates is its own creation. Therefore, the conclusion that gravity models outperform neural networks, when high percentages of data are used, is not very strong. This could favor the use of neural networks, even when data sets are large. Due to the data management after training, neural networks were able to estimate both high and low trip values, also beyond the limits of the training sample.

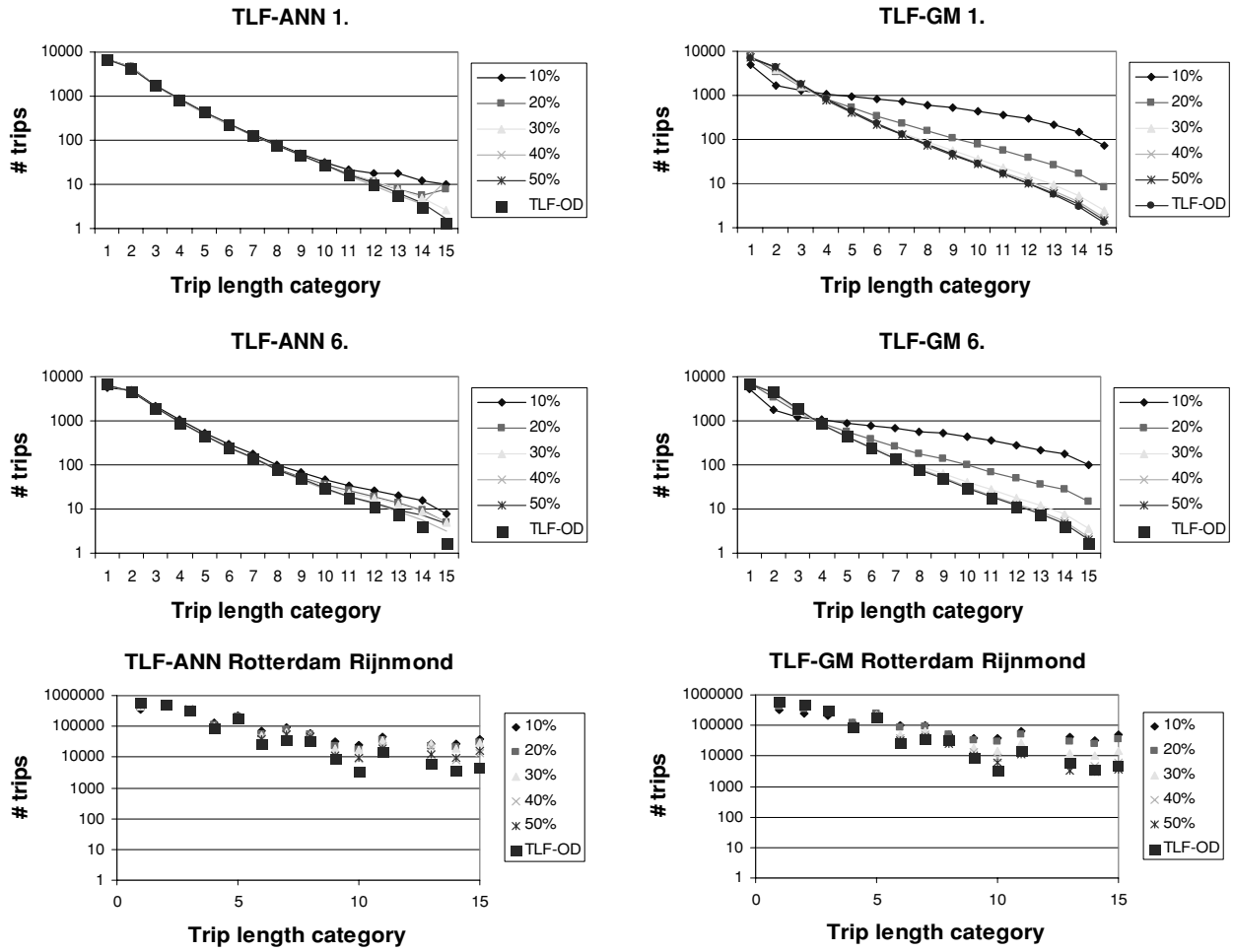

Fig. 7. Trip length frequencies for ANN and GM compared to trip length frequency of the full data set (TLF-OD).

Secondly, there seems to be a difference in performances when matrices vary in complexity. Figure 8 shows the average RMSE per OD matrix versus the complexity of the matrix. The average RMSE value is calculated as the average of all results at percentages 10–90%. So, Figure 8 shows aggregated values, and therefore shows less detail than Figure 5.

In the first instance it seems that no general conclusion can be drawn upon the relationship between complexity and results; no clear relationship is shown for either of the models. The differences in RMSE results between the simplest matrix and the most complex matrix are smaller for gravity models. The gravity models results lie within a smaller range than the neural networks results and therefore appear to be less sensitive to complexity and more stable. The following three conclusions can be drawn: (i) when complexity is minimal, 2, data are most structured and neural network performance is best. This stresses one of the qualities of neural networks: pattern recognition; (ii) contrary to this point, when complexity is at its maximum, 10, average neural network performance over all data set percentages is less good. The complex matrix reveals the least order and therefore the fewest patterns. This results in an RMSE increase; and (iii) the differences in results between neural network models and gravity models are not stable, especially when the difference between both models in matrix 16 is small. This stresses the fact that neural networks perform best in situations where data are most structured; matrix 1, complexity 2, shows the highest difference in (average) RMSE values.

The results shown have to be looked at closely. First of all, the results do not give absolute answers on the estimating abilities of both methods. Figure 8 only gives an answer when the results of all data set percentages are averaged. So, no hard conclusions can be drawn on specific percentages. Furthermore, every calculation for both neural networks and gravity models is based on different random pulls (and calculations) of an x

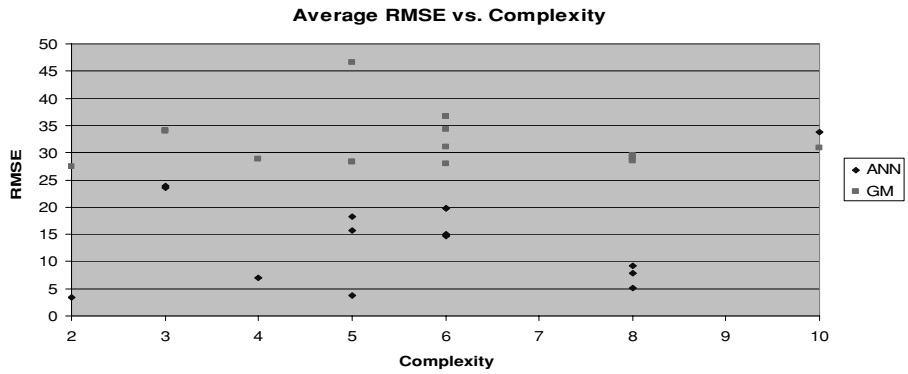

Fig. 8. Average RMSE versus complexity of the matrices.

percentage out of the total data set, as shown before. These results were averaged, and therefore should be seen as relative results. To judge these relative results, we look at the average standard deviation and the distribution of the RMSE. Figure 9 shows that neural network estimations show less variance in performance overall. This favors the use of neural networks.

6.2 Separable matrices

Both neural networks and gravity models use limited data to calibrate. This study shows that calibrating these models, using only limited or incomplete data, is actually possible. However, results show also that there are a number of peaks where results are worse than in other cases. It is interesting to explore whether these results are brought about by a problem called separability of calibration data. Separability of matrices deals with the fact that limited data are used to estimate missing data. Kirby (1979) has shown that there are two basic conditions required for the estimation of missing data. The focus will be on the second condition.

1. The gravity model must fit both the available data and the data that are not available, that is, the model must be a good model for the two regions of the matrix: the observed and unobserved.
2. The two regions of the matrix should not be separable, that is, it should not be possible to split the matrix into two or more independent matrices.

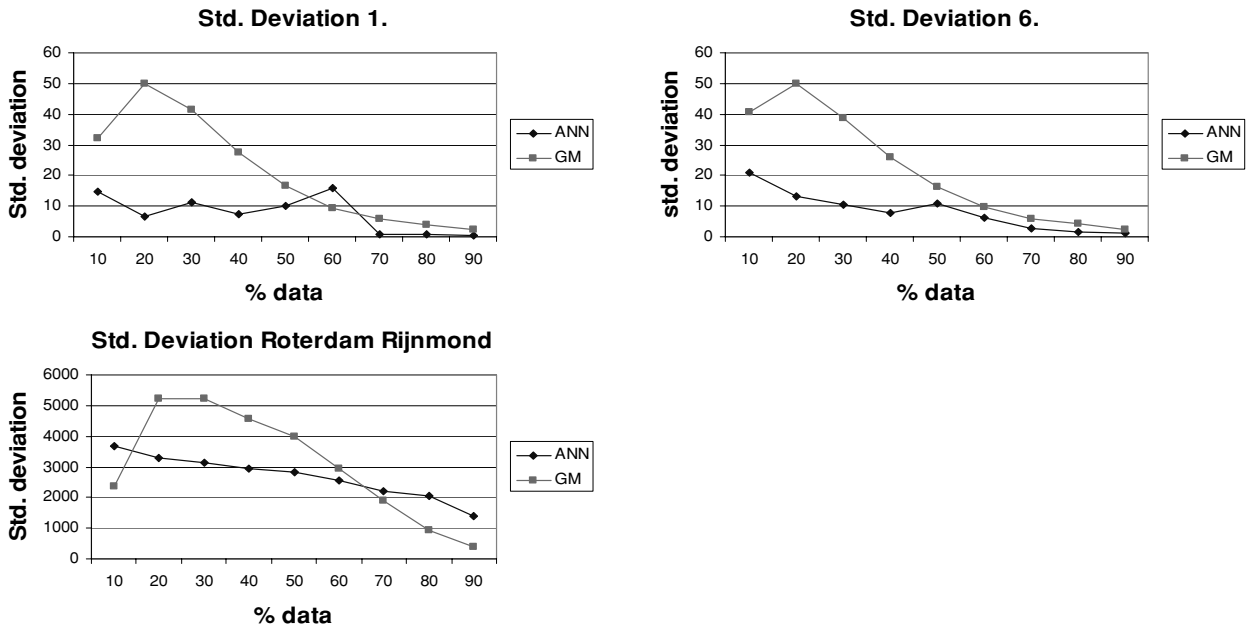

Fig. 9. Standard deviation.

	Internal	*External*
Internal	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	********************
External	********************************	xxxxxxxxxxxxxxxxxxxx
	********************************	xxxxxxxxxxxxxxxxxxxx
	********************************	xxxxxxxxxxxxxxxxxxxx
	********************************	xxxxxxxxxxxxxxxxxxxx

Fig. 10. Matrix separability (Ortuzar and Willumsen, 2001).

Figure 10 schematizes the problem of matrix separability. In the case of doubly constrained gravity models, each separate area has two degrees of indetermination and therefore the balancing factors cannot produce unique estimates, nonidentifiability of unique products for unobserved cell entries. In other words, a matrix can be divided into separate independent parts with few or no interactions to other parts.

Research is conducted into this phenomenon to explain the peak RMSE values in the results. The random selection of calibration data was adjusted. Contrary to the first research part, random matrices instead of random cells are used as inputs. The procedure is simple. Matrices of 25 (5×5, 10%), 49 (7×7, 20%), and 64 (8×8, 30%) are randomly selected as input for calibration. It is only thought useful to research calibration percentages up to 30%. Figure 11 shows the ANN results of the RMSE of matrices 1, 6, and the Rotterdam Rijnmond matrix.

Figure 11 also shows that the effect of separability is large for matrix 1. Matrix 6 and the Rijnmond matrix show different results. The trend in the standard deviations is clear. As expected, the more random the calibration data, the less variance in the results; the chances of picking less suitable data are higher for the separable data cases. Yet, the results in average RMSE show different results. Where matrix 1 shows big disadvantages for separable input, matrix 6 and the Rotterdam case show somewhat distinct results. Regularities in matrix 1 are not that large in small separate parts of the matrix. Therefore, the separable matrices consist of only a small part of the data in the input range, resulting in worse results. Matrix 6 and the Rotterdam matrix show results that are in some cases better than the earlier results.

The statistical results of matrix 1 show that the results of 10 and 20% are significantly different and that the nonseparable input data end in better results. At 30% no differences are found. Matrix 6 shows statistical differences at percentages 20 and 30. So, the results of the separable matrix input even seem better. At 10% no differences were found. The Rijnmond matrix shows the same results as matrix 6. So, no hard conclusions can be drawn on the fact of whether separable input can cast clouds on the results by causing peaks. Separability does not seem to influence the average RMSE of an ANN model in this research.

How about the results for the gravity model? The first striking and most important difference is that gravity models do not give answers with all input data. Certain combinations of data, separable matrices, give no or meaningless results. The used calibration process probably causes this. This was also observed in earlier results. Calibration of the gravity model is done using the average trip length of the calibration set, as a calibration target. To calculate the average trip length, firstly all trips are classified in matching trip length classes. However, when the input data is built up with trips in only a limited number of trip length classes, calculation of the average trip length is very difficult. As a consequence, estimating the whole OD matrix will be difficult. The results are shown in Figure 12.

Examination of the results shows that in this case gravity models show better results with separable inputs. Regularities in matrix 1 are not that large in small separate parts of the matrix. Contrary to neural networks, gravity models seem to give better results when matrices are more irregular. This explains the better results. In general, it can be said that the use of separable matrices does not disturb the results negatively. So, peaks cannot be explained by separable inputs in this research.

7 DISCUSSION AND CONCLUSIONS

This research shows that neural networks outperform gravity models in both synthetic and real situations when

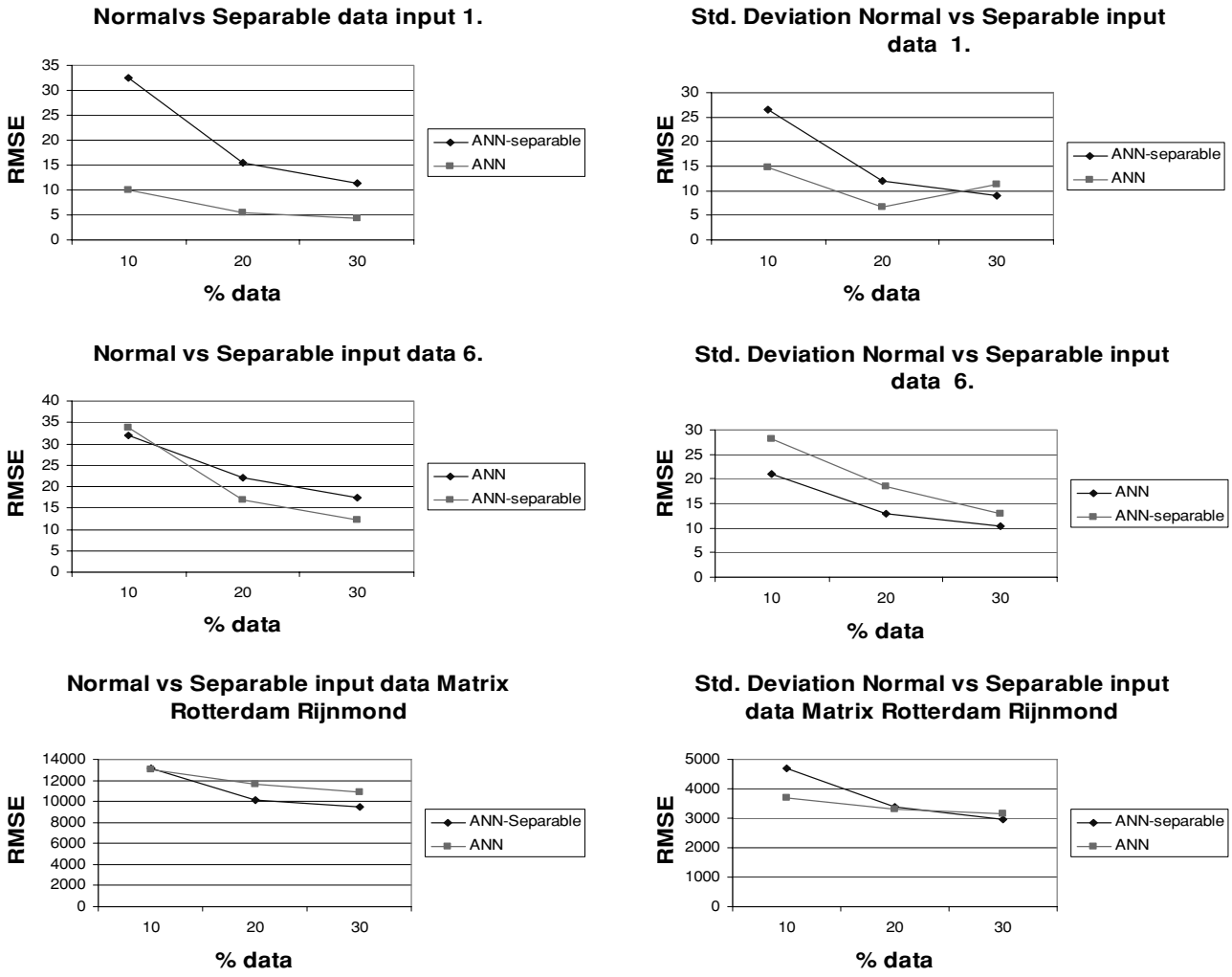

Fig. 11. RMSE results of ANN separable and nonseparable inputs.

data are scarce. These results are promising for future trip distribution modeling, which is an important step for good transport planning. The results were obtained using both synthetic and real-world data sets. This gives the opportunity to control the test. This is a major difference compared to other studies, for example, the study by Mozolin et al. (2000). This article gives more insights into the performance of neural networks in different complexity cases. It is shown that in cases that are quite unusual, as shown by matrix 8 in Figure 5, neural networks perform much better. This can only be concluded if one uses synthetic test cases. The use of a real-world case strengthens these conclusions. Mozolin et al. (2000) for example do not use both real-world data and synthetic data in varying complexity. In addition to other research, this research shows the results of a statistical analysis, which reveals that gravity models need more training samples than neural networks.

The results have to be looked at closely. First of all, they do not give absolute answers on the estimating abilities of both methods. As shown before, each of the calibrated samples for both models is based on different random pulls (calculations) of a predefined percentage out of the total data set. These results are averaged, and therefore can only be used for comparisons.

As seen in the different figures, neural network performance, compared to gravity models, is best when data are scarce. As stated before, the synthetic matrix data were generated using a gravity model. This creates a situation in which gravity models should give good results. However, the gravity model only gives the best results when calibration percentage is high; gravity models only reproduce their own results. In situations close to reality, with only limited amounts of data, neural networks show their abilities. This strengthens the results of neural networks.

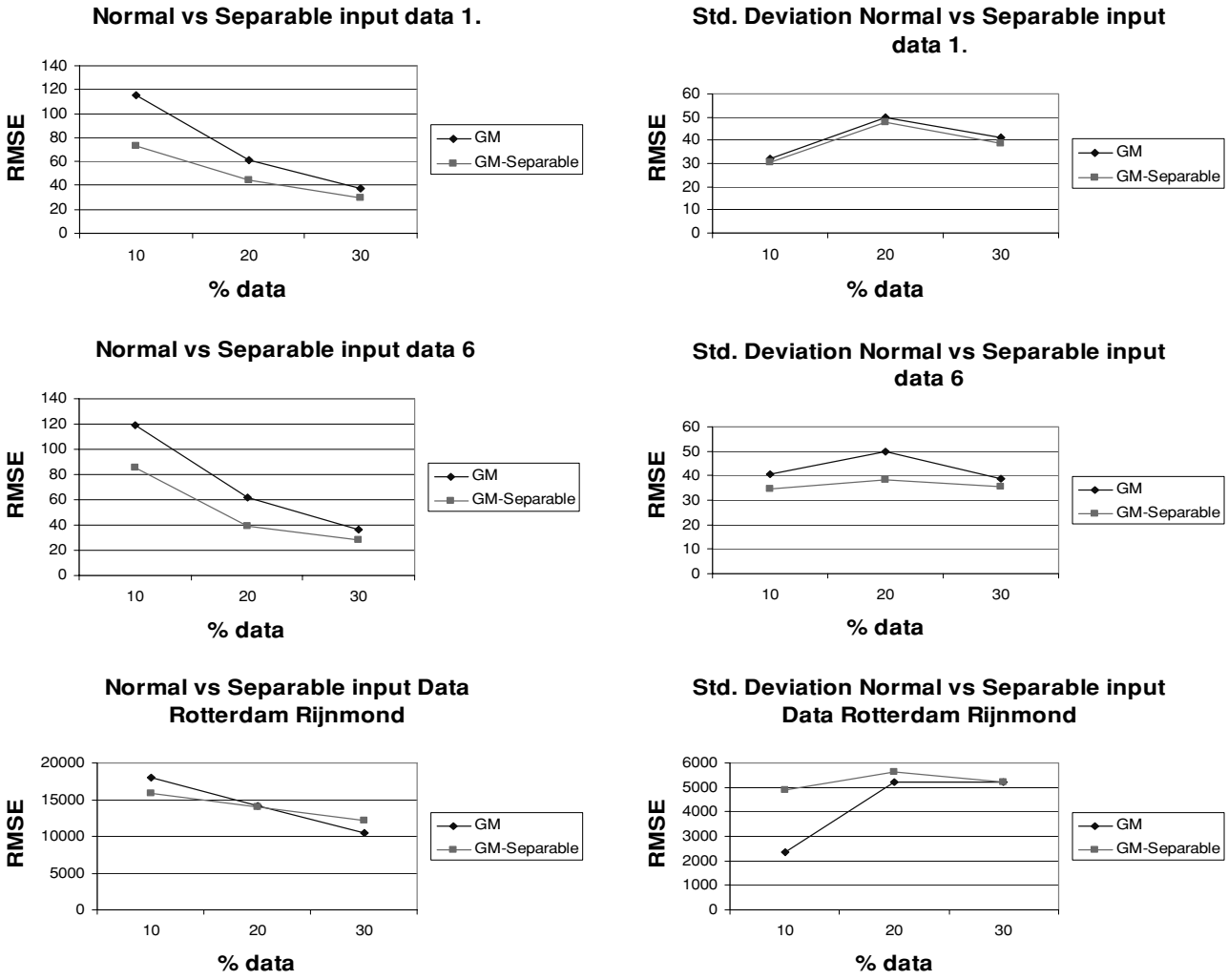

Fig. 12. RMSE results of GM separable and nonseparable inputs.

The investigation into the trip length frequencies and the standard deviations gives an insight into the absolute performances of both methods. Neural networks show better performance on standard deviations and trip length frequencies when data are scarce. Due to the data management after training, neural networks were able to estimate both high and low trip values, also beyond the limits of the training sample.

The behavior of both methods changes when complexity increases. The data sets are complex enough, especially the random matrices, to come close to reality. Results show that neural networks perform better under conditions in which data are structured. But, results show also that even the performance of the random matrix and real-world matrix are good. Due to the large number of trips in the Rijnmond case, the RMSE values were higher than in the synthetic cases.

There is not one best neural network topology and configuration for all proportions of available data. In addition, neural network performance is dependent on factors like the activation function, learning method, and corresponding momentum stop criteria for learning, and software to run neural networks. During this research no extensive extra work has been conducted to investigate the influence of these parameters on neural network performance. Preprocessing showed that no general conclusion was drawn on one best network configuration for all types of data. However, results show that smaller hidden layers, less than 5 neurons, are not preferred. Networks with more than 10 hidden neurons should not be used because of over-fitting problems (Bishop, 1995); the ability to generalize decreases when training data are scarce and too large a number of hidden neurons is used.

The performance of neural networks is promising. The research shows that the trip distribution problem is a complicated one, but also an important step in transport planning. Many errors generated during this distribution

phase are passed on to the next steps. Often, real-world problems have only limited data. And contrary to this research, real-world problems have only one sample of that data. Scarce data can give difficulties during calibration of models and results have high standard deviations. The extent to which the available data suits a calibration process determines the performance. However, if only 20% or less data are available, the calibration process can lead to a large number of different matrices. In this study, the total data set are available as a reference for determining the quality of the estimated data: the RMSE.

Contrary to this situation, real-world data sets are never complete and so no RMSE can be calculated. This is why the quality of the estimated matrix cannot be determined. Ortuzar and Willumsen (2001) conclude that observed trip matrices are almost always scarce; they have a large number of empty cells. If the sampling rate is 20% (1 in 5), then the chances of making no observations on a particular OD pair are very high. Matrix expansion methods can be used to seed empty cells. In addition, it is important to realize that observed trip matrices normally contain a large number of errors and that these are amplified by the expansion process. This pleads for more research into the conditions under which gravity models and neural networks work well.

It is difficult to obtain a good estimation for the total number of samples necessary to be sure about the results. In addition, it can be concluded that a large number of calculations are necessary due to, among others, the random initialization process of neural networks. Furthermore, the calibration process used for gravity models needs 40 times more calculations than the neural network. It shows that the use of gravity models can be dangerous. The chances that small sets of data "fit" a gravity model is smaller than chances for a neural network.

It was expected that the method of neural networks can even be improved. In this research only the trip production and attraction constraints are used. One of the calibration factors for gravity models is the average trip length of the calibration sample. Up to the present time this piece of extra information has not been used for neural networks. To investigate whether the average trip length could be useful for neural networks, an extra constraint can be added. The total number of constraints will be three (tri-proportional): constraints on production and attraction totals and the average trip length. Preliminary research shows that adding a third constraint might not lead to better results.

So, what are the implications of these findings for travel planning agencies? First of all, the research shows that performing a good trip distribution is very difficult, even with traditional methods. One should be aware of the influence that the amount and quality of data have on the quality of the actual trip distribution. In cases where data are scarce, these neural networks are very suitable for planning agencies to be used. In addition, neural networks are less sensitive to small errors in the data, which is a general characteristic of a neural network. There is, however, one major drawback. Neural networks are black boxes. This makes it difficult to understand what is happening inside the networks. Especially in cases where policy is made and clear reasoning is asked for, the results of neural networks are difficult to sell. In these cases traditional methods are more understandable, yet not more accurate.

Finally, this study compared the performances of neural networks and doubly constrained gravity models in a trip distribution context. The study shows that neural networks outperform gravity models when data are scarce. The conclusion that gravity models outperform neural networks when a lot of data are available seems less certain, due to the research method and the generation of the synthetic data. This article adds new inputs to the discussion of trip distribution modeling with neural networks. Neural networks can improve trip distribution; however, the black box character of the models makes understanding the models more difficult.

REFERENCES

Bishop, C. M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK.

Black, W. R. (1995), Spatial interaction modelling using artificial neural networks, *Journal of Transport Geography*, **3**(3), 159–66.

Collins, M. G., Steiner, F. R. & Rushman, M. J. (2001), Land-use suitability analysis in the United States: Historical development and promising technological achievements, *Environmental Management*, **28**(5), 611–21.

Currit, N. (2002), Inductive regression: Overcoming OLS limitations with the general regression neural network, *Computers, Environment and Urban Systems*, **26**, 335–53.

Dougherty, M. (1995), A review of neural networks applied to transport, *Transportation Research, Part C*, **3**(4), 247–60.

Faghri, A. & Sandeep, A. (1998), Analysis of performance of backpropagation, ANN with different training parameters, in V. Himanen, P. Nijkamp, and A. Reggiani (eds.), *Neural Networks in Transport Applications*, Ashgate, Aldeshot, UK.

Fischer, M. M. (1998), Computational neural networks: An attractive class of mathematical models for transportation research, in Himanen, Nijkamp, and Reggiani (eds.), *Neural Networks in Transport Applications*, Ashgate, Aldeshot, England.

Fischer, M. M. & Gopal, S. (1994), Artificial neural networks. A new approach to modelling interregional telecommunications flows, *Journal of Regional Science*, **34**(4), 503–527.

Fotheringham, A. S. & O'Kelly, M. E. (1989), *Spatial Interaction Models: Formulations and Applications*, Kluwer Academic Publishers, London.

Gopal, S. & Fischer, M. M. (1996), Learning in single hidden-layer feed forward network models: Backpropagation in a spatial interaction modelling context, *Geographical Analysis*, **28**(1), 38–55.

Huisken, G. & Coffa, A. (2000), Neural networks and fuzzy logic to improve trip generation modelling, in *Proceedings of the 9th International Association for Travel Behaviour Research Conference*, Institute of Transport Studies, IATBR, Gold Coast, Queensland, Australia, 2–7 July.

Hyman, G. M. (1969), The calibration of trip distribution models, *Environment and Planning*, **1**(3), 105–12.

Kirby, H. R. (1979), Partial matrix techniques, *Traffic Engineering and Control*, **20**(8-9), 422–28.

Miller, D. M., Kaminsky, E. J. & Rana, S. (1995), Neural network classification of remote-sensing data, *Computers and Geosciences*, **21**(3), 377–86.

Mozolin, M., Thill, J. C. & Lynn Usery, E. L. (2000), Trip distribution forecasting with multiplayer perceptron neural networks: A critical evaluation, *Transportation Research, Part C*, **34**(1), 53.

Openshaw, S. & Openshaw, C. (1992), *Artificial Intelligence in Geography*, John Wiley and Sons, Chichester.

de Ortuzar, J. D. & Willumsen, L. G. (2001), *Modelling Transport*, John Wiley & Sons, Ltd, Chichester, UK.

Pijanowski, B. C., Brown, D. G., Manik, G. A. & Shellito, B. A. (2002), Using neural networks and GIS to forecast land use changes: A land transformation model, *Computers, Environment and Urban Systems*, **26**(6), 553–575.

Raju, K. A., Sikdar, P. K. & Dhingra, S. L. (1998), Microsimulation of residential location choice and its variation, *Computers, Environment and Urban Systems*, **22**(3), 203–18.

Rodrigue, J.-P. (1997), Parallel modelling and neural networks: An overview for transportation/land use systems, *Transportation Research, Part C*, **5**(5), 259–71.

The Math Works, Inc. (1996), *Matlab: Using Matlab*, version 5.

The Math Works, Inc. (1998), *Neural Network Toolbox: For Use with Matlab*, H. Demuth, and M. Beale (eds.), version 3.

Thill, J. C. & Mozolin, M. (2000), Feel-forward neural networks for spatial interaction: Are they trustwathy forecasting tools?, in A. Reggiani (ed.), *Spatial Economic Science: New Frontiers in Theory and Methodology*, Springer, Heidelberg, pp. 355–381.

Tillema, F. (2004), *Development of a data driven land use transport interaction model*. Ph.D. thesis, University of Twente, Centre for Transport Studies.

Tillema, F., Huisken, G. & Maarseveen, M. F. A. M. (2002), Neurale netwerk technieken ten behoeve van landgebruik/transport modellen, *Presented at the Dutch Colloquium Vervoersplanologisch Speurwerk*, Amsterdam.