Infering Air Quality from Traffic Data using Transferable Neural Network Models

Miguel A. Molina-Cabello[1], Benjamin N. Passow[2], Enrique Dominguez[1], David Elizondo[2], and Jolanta Obszynska[3]

[1] Department of Computer Science, ETSI Informatica, University of Malaga, Spain.
{miguelangel,enriqued}@lcc.uma.es
[2] De Montfort University's Interdisciplinary Group in Intelligent Transport Systems (DIGITS)
De Montfort University, Leicester, LE1 9BH, United Kingdom.
benpassow@ieee.org, elizondo@dmu.ac.uk
[3] Leicester City Council's Pollution Team, Leicester, LE1 6ZG, United Kingdom.
jolanta.obszynska@leicester.gov.uk

Abstract. This work presents a neural network based model for inferring air quality from traffic measurements. It is important to obtain information on air quality in urban environments in order to meet legislative and policy requirements. Measurement equipment tends to be expensive to purchase and maintain. Therefore, a model based approach capable of accurate determination of pollution levels is highly beneficial. The objective of this study was to develop a neural network model to accurately infer pollution levels from existing data sources in Leicester, UK. Neural Networks are models made of several highly interconnected processing elements. These elements process information by their dynamic state response to inputs. Problems which were not solvable by traditional algorithmic approaches frequently can be solved using neural networks. This paper shows that using a simple neural network with traffic and meteorological data as inputs, the air quality can be estimated with a good level of generalisation and in near real-time. By applying these models to links rather than nodes, this methodology can directly be used to inform traffic engineers and direct traffic management decisions towards enhancing local air quality and traffic management simultaneously.

Keywords: Neural network, inferring pollution concentration, air quality, traffic management

1 Introduction

Detailed air quality simulation models have been developed for many cities. These mechanistic and deterministic models are used as tools to help understand the basic air quality and traffic management. They support legislative and policy requirements which depend on air quality in urban environments. Hardware equipment to measure air quality is normally expensive to purchase and

maintain. Therefore, a model based approach capable of accurate prediction of pollution levels is highly beneficial.

This paper is organized as follows. First of all, the background is presented in Section 2. After that, the neural network model is reported in Section 3, while Section 4 exhibits the data sources. Then, the results are depicted in Section 5. Finally some conclusions are presented in the last section.

2 Background

City councils and local authorities across the world are managing and controlling urban traffic. This involves monitoring traffic conditions and designing, maintaining and controlling traffic signals. It also involves dealing with faults, roadworks, air quality, noise and traffic. Across Europe, a large number of research projects are related to air pollution and its impact on the quality of people's health.

Emissions from motor vehicles are a major source of air pollution. Despite the fact that air quality has improved significantly in the last decades, there is string scientific evidence suggesting that current levels of air pollution are a high risk for the environment and to human health ([1, 2]).

Therefore, there is a need for air quality information in urban environments to meet legislative and policy requirements. Recent research in this topic using Computational Intelligence (CI) techniques can be categorized into two areas: Modelling and Forecasting.

2.1 Air Quality Forecasting

The ability to forecast air quality is becoming increasingly important as links are made between poor air quality and adverse health effects. With accurate forewarning of low air quality levels, it becomes possible to take preventive action thereby reducing the severity of the pollution incident.

There is a myriad of air quality forecasting methods. In particular, Artificial Neural Networks (ANNs) have been employed for this purpose. For example, an ANN with weather parameters and traffic conditions developed by Viotti et al. [17] was used to forecast pollutant concentrations (benzene, carbon monoxide, NO_2 , NO_x, Ozone) at both 1 hour and 24 hours intervals into the future. They point out that different models are required for each pollutant type and forecasting interval.

A good illustration of this can be found in a comparison study, where five different ANN models, one linear statistical model and one deterministic modelling system were assessed in terms of their capability to forecast urban NO_2 and PM_{10} concentrations in Helsinki [7]. The ANN models were generally found to perform better than the linear and deterministic models, especially for forecasting PM_{10} which is considered more difficult to predict than NO_2. However, it is pointed out that ANNs are limited to the forecast window and location for which they were trained, and that they require retraining if either of these parameters change.

Another example is presented in Ibarra-Berastegi et al. [6], where the use of neural networks is investigated for providing air quality forecasts up to 8 hours in advance for five different pollutants (SO_2, CO, NO_2, NO and O_3) at six different locations in Bilbao, Spain. They report a best case for NO_2 with $R^2 = 0.88$.

Other option could be the use of a Genetic Algorithm to optimise the design of an ANN. Niska et al. [11] apply a model of this kind for predicting NO_2 levels in Helsinki. It was found that any improvements offered by the Genetic Algorithm were negligible when compared with a reference model. Moreover, the use of neural networks optimised by genetic algorithms for predicting PM_{10} concentrations at 4 locations in the Greater Athens area is presented in [4]. When compared with linear regression methods, the neural networks were found to be superior, with R^2 in the range $0.50 - 0.67$ compared to $0.29 - 0.35$ for linear regression.

Different neural network topologies are also considered. The performance of three different neural network topologies (MLP, RBF and Modular-ANN) for predicting roadside concentrations of CO and NO_2 in Melton Mowbray, UK, are compared in [19]. The transferability of these models is then considered by applying them to data from another site in Leicester, UK, concluding that whilst performance is degraded, it is still able to provide a useful prediction capability.

2.2 Air Quality Modelling

The behaviour of pollution dispersion, especially in urban areas is complex and difficult to model. Dispersion is influenced by many factors such as temperature, wind speed, wind direction, wind strength and terrain, all of which exhibit significant levels of uncertainty on any mathematical models used. The application of CI methods to such problems, especially those that do not require an underlying model (such as neural networks and fuzzy sets) is therefore becoming the focus of much ongoing research. Some examples of recent work are summarised here.

Pérez-Roa et al. [14] use a neural network to predict the eddy diffusivity (K_v), an important parameter used in the study of pollution dispersion, for which values supplied in literature are often inaccurate. Based on the neural network predicted values of K_v, a dispersion model was able to forecast peak carbon monoxide concentrations with improved accuracy.

Neural networks as a filter for the concentration levels produced by an air pollution model were applied by Pelliccioni and Tirabassi [13] to account for disagreement between the measured and predicted values. Their results demonstrated the efficiency of Neural Networks.

The modelling of the nitrogen dioxide (NO_2) dispersion phenomena using the ANN technique was also assessed by Shiva Nagendra and Khare [10]. A satisfactory performance of the ANN-based NO_2 models on the evaluation data set is shown.

Another example of the successful application of ANN to calculate the average spatial distribution of air pollutants based on diffusive sampling measurements in Cyprus was demonstrated in Pfeiffera et al. [15]. Their results illustrate

the application of ANN resulting in realistic maps of the annual average distribution of NO_2 in Cyprus.

In addition to ANN, fuzzy sets provide another tool in dealing with uncertainty in dispersion modelling. Fisher [3] reviewed the various uncertainties existing in dispersion modelling, and highlighted the feasibility of fuzzy approaches to environmental decisions. Another fuzzy based system for predicting modelling air quality was proposed by [16], where the use of a trapezoidal membership function was proposed. The model is based on data collected over a year using 5 locations in Tehran.

Furthermore, the feasibility of using ANNs in predicting air quality based on the SCOOT data (stops, flows, congestions and delays) per link-intersection and the weather conditions (temperature, wind speed and direction) was performed in a preliminary study at the University of New Castle by Bell et all [18]. The results indicate that this approach might be a promising option.

In this research we propose to expand this work by providing a global against a local approach for modelling air quality which can be easily adapted to new locations. Our model uses a combination of all the SCOOT data (combining all the links) and the Climate data as inputs. Another novelty of this study consists on the use of a Self-Organizing Map (SOM) neural network to filter the dataset.

3 Neural Network Model

Of the neural network algorithms available, two of the most common ones are the multilayer perceptron (MLP) for supervised learning and the Kohonen self organising maps for unsupervised learning. In supervised learning, the inputs are presented to the neural network and the output produced y the neural network is compared with the desired output. After this, the weights of the neural network are adjusted taking into account any error in the output pattern. In unsupervised learning, the weight adjustments are not made based on comparison with some target output. Instead, the neural network self organises its weights.

The proposed neural model to infer the air quality is based on a MLP of feedforwad network, since this kind of neural networks is widely used for the function fitting problem.

The proposed MLP uses the back propagation algorithm and it consists on one single hidden layer of simply interconnected neurons. Both the input and output layers are determined by the problem. In this case, the output layer is composed by an only neuron, which is used for the air quality. The input layer is composed by nine neurons: five for the climate data (wind speed, wind direction, rain, radiation and air pressure) and four for SCOOT data (stops, delays, flows and congestions). The output of a neuron is given by the following expression:

$$y_i = f(\sum_{j=0}^{N_i} w_{ji}x_j + b_i) \tag{1}$$

where N_i is the number of inputs of the neuron i, w_{ji} is the synaptic weight between the j-th input and the neuron i, b_i is the bias and f is the transfer function. In the hidden layer, a tangential sigmoid function:

$$f(x) = \frac{sinhx}{coshx} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{2}$$

is used as transfer function, although a pure linear function is used in the output layer.

Generally, only one hidden layer is required to approximate any measurable function [5, 8]. The optimum number of neurons in the hidden layer depends on the desired accuracy and a trial and error approach is normally used to determine it. In our experiments, a small hidden layer composed by only ten neurons was sufficient to achieve good results.

The learning of the proposed MLP is supervised, hence a training data set is required. This training data set consists of a set of pairs, that is, an input vector (climate and scoot data) and an associated target (the air quality). During the training, the synaptic weights in the neural network are adjusted until the output produced by the network matches the target within a certain error, which it is used for the weights adaptation according to the learning algorithm. In this case, the backpropagation algorithm is proposed to minimise the mean squared error.

Overtraining of the neural network is an undesirable characteristic, which may occur when the neural model learns too much the details of the training data. Therefore, the results produced by network are poor generalization capability when new data are presented. A bad choice of the training data set or an inappropriate learning algorithm typically impair the network's generalization capability and lead to model overfitting. In this sense, a validation data set was used during the training to avoid the overtraining of the neural network and to check the generalization performance. Training is stopped when the performance on the validation data starts to decline.

4 Data Sources

In a long term collaboration with Leicester City Council's Traffic and Air Quality Teams a large amount of data from near real time feeds has been collected over the past years. This data includes:

- Traffic detection via inductive road loops
- Air quality measurements from road-side monitors
- Meteorological measurements from a mast / measurement station

The data collected from street loops is commonly noisy and often corrupted. The level of performance achieved with an ANN model is directly linked with the quality of the data sets. This data is often surrounded by errors which can be caused by noise and equipment malfunction. These errors are referred to as outliers, which need to be identified and eliminated from the data sets in order to improve the models performance. In this work a Self-Organizing Map (SOM) is proposed for filtering and handling outliers[12].

4.1 Road traffic condition data

Road traffic related data has been collected on an ongoing basis from Leicester City Council's traffic management systems. Specifically, a large number of SCOOT M02 messages have been collected via FTP. SCOOT uses raw induction loop data to measure and estimate four traffic flow and congestion related parameters[9]. These parameters are reported every 15 minutes via M02 system messages.

The four parameters collected by every link are flow, delay, stops and congestion. *Flow* is estimated as vehicles per hour. *Delay* is estimated as total delay in vehicles per hour. *Stops* is estimated as the number of vehicles that have stopped at least once on the link in the given period. *Congestion* is measured as percentage of 4 second intervals when a detector is occupied by traffic. This data has been accumulated and used as hourly means in order to match the time-base of the air quality and meteorological measurements. Two possibilities have been addressed in this work:

– Aggregated links. Road traffic data from all links is previously aggregated and then accumulated for every site. Therefore, a fixed number of four traffic inputs is established. Note that this option is needed to build transferable models.
– Separated links. Road traffic data is separately accumulated for every link and site. Consequently, the number of traffic inputs is four times the number of links composing every site.

4.2 Air quality data

A number of pollutants are continuously being monitored and recorded by Leicester City Council's Air Quality team. The ratified data for NO_2 and PM10 has been used for this study from the following sites:

– Abbey Lane
– Uppingham Road
– Glenhills Lane
– AURN, New Walk Centre
– Imperial Av / Narborough Rd
– London Road

Figure 1 shows a map with the locations of these monitors, which are commonly placed in critical sites (e.g. crossing roads) composed by several link roads.

4.3 Meteorological data

A meteorological mast in Leicester (location, see 1) automatically provided weather data as an hourly average. A NRT stream of meteorological data (met data) was used as part of the model to determine and forecast the emission level at a given location. There were 6 indicators for weather condition in the model (see Table 1).

Table 1: Meteorological measurements available and used in this work.

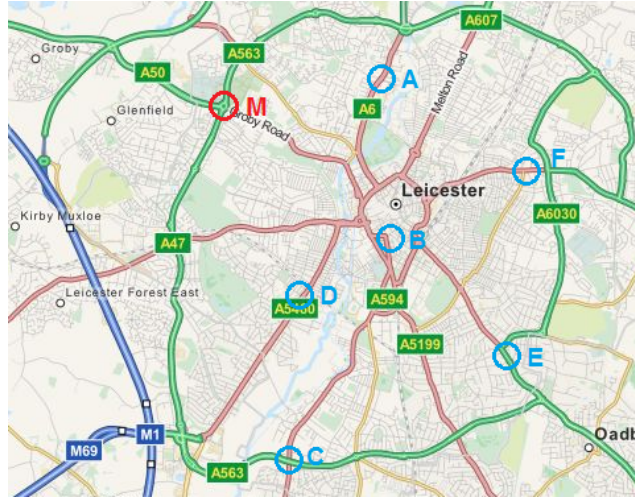| Parameter	Unit of measurement
Cloud coverage	Amount of sunshine reaching the surface $[0 \rightarrow 1]$
Air pressure	millibar
Temperature	degrees Celsius
Rain (amount of)	millimetres per hour
Wind direction	degrees from North
Wind speed	metres per second

Fig. 1: Map of Leicester City, UK, with locations of air quality measurement stations (A to F) and meteorological mast (M). The labels represent the following measurement stations: A = Abbey Lane, B = AURN, New Walk, C = Glenhills Way, D = Imperial Avenue, E = London Road, F = Uppingham Road.

5 Results

Several experiments have been carried out. Three different proposals have been considered. The first one presents a single ANN model with two outputs (NO_2 and PM10), and with SOM based filtering. The second proposal is composed by two separate ANN models, each one with one output (NO_2 or PM10), and without SOM based filtering. The third approach consists of two separate ANN models, each one with one output (NO_2 or PM10), and with SOM based filtering.

Some well-known measures have been selected in order to compare the selected proposals from a quantitative point of view. The traditional R and R^2 measures are the main ones. If the R performance is positive that means a positive correlation exists, where higher is better. On the other hand, R^2 is a positive value, where higher is better. The root mean square error (RMSE), the mean

square error (MSE) and the mean absolute error (MAE) are also taken into account, where lower is better.

Tables 2, 3 and 4 exhibit the performance of the three different considered models, respectively. As it can be observed, the third considered proposal (two separate ANN models with one output and without SOM based filtering) offers the best performance of the three approaches. It must be highlighted that each proposal works with a similar performance in each studied site.

Table 2: Results for a single ANN model with 2 outputs, NO_2 and PM10, and with SOM based filtering.

| | NO_2 | | | | | | PM10 | | | | | |
| | Separated Links | | | Aggregated Links | | | Separated Links | | | Aggregated Links | | |
Site	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE
Abbey Lane	0.65	5.05	3.42	0.60	5.27	3.64	0.61	7.04	4.84	0.39	9.13	5.59
AURN, New Walk	0.36	7.81	5.49	0.42	7.36	5.38	0.47	5.93	4.41	0.45	6.53	4.71
Glenhills Way	0.81	6.17	4.69	0.74	6.82	4.75	0.65	8.12	5.62	0.53	11.05	6.74
Imperial Avenue	0.77	2.84	2.12	0.79	2.70	1.97	0.60	6.39	4.34	0.50	8.20	4.67
London Road	0.78	4.12	2.97	0.78	4.22	3.18	0.70	6.22	4.34	0.62	7.53	4.61
Uppingham Road	0.68	4.55	3.33	0.73	4.22	3.19						
Average	0.68	5.09	3.67	0.68	5.10	3.69	0.61	6.74	4.71	0.50	8.49	5.26

Table 3: Results for two separate ANN models each with 1 output, NO_2 or PM10, and without SOM based filtering.

| | NO_2 | | | | | | PM10 | | | | | |
| | Separated Links | | | Aggregated Links | | | Separated Links | | | Aggregated Links | | |
Site	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE
Abbey Lane	0.65	8.33	6.28	0.64	8.38	6.35	0.54	11.55	8.28	0.57	11.23	7.82
AURN, New Walk	0.33	13.77	10.65	0.34	13.71	10.59	0.51	10.97	7.38	0.53	10.79	7.14
Glenhills Way	0.73	9.88	7.51	0.66	10.79	8.25	0.57	12.97	9.43	0.52	13.42	9.83
Imperial Avenue	0.67	5.20	3.99	0.66	5.21	3.99	0.55	10.59	7.37	0.55	10.61	7.42
London Road	0.68	7.47	5.45	0.68	7.51	5.44	0.58	10.62	7.39	0.58	10.63	7.42
Uppingham Road	0.68	7.13	5.46	0.63	7.25	5.55						
Average	0.62	8.63	6.56	0.60	8.81	6.69	0.55	11.34	7.97	0.55	11.34	7.92

Nevertheless, the site *AURN, New Walk Centre* is the most problematic. The NO_2 MSE of the different studied places are reported in Figure 2. It is shown that *AURN, New Walk Centre* presents the highest MSE. In addition, Figure 3 depicts the NO_2 regressions. It can be noticed that the R^2 performance is extremely low.

Table 4: Results for two separate ANN models each with 1 output, NO_2 or PM10, and with SOM based filtering.

| Site | NO_2 | | | | | | PM10 | | | | | |
| | Separated Links | | | Aggregated Links | | | Separated Links | | | Aggregated Links | | |
	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE	R	RMSE	MAE
Abbey Lane	0.72	4.47	3.08	0.68	4.67	3.29	0.62	6.32	4.68	0.77	5.30	3.75
AURN, New Walk	0.50	7.21	4.97	0.53	6.46	4.72	0.69	4.53	3.21	0.61	5.29	3.79
Glenhills Way	0.86	5.18	3.85	0.81	5.72	4.26	0.63	8.38	5.67	0.50	10.88	6.31
Imperial Avenue	0.81	2.53	1.87	0.81	2.50	1.82	0.58	7.26	4.71	0.64	6.74	4.02
London Road	0.83	3.65	2.66	0.78	4.06	2.93	0.66	6.98	4.66	0.75	5.56	3.84
Uppingham Road	0.76	3.88	2.94	0.74	4.07	5.46						
Average	0.75	4.49	3.23	0.73	4.58	3.75	0.64	6.69	4.59	0.65	6.75	4.34
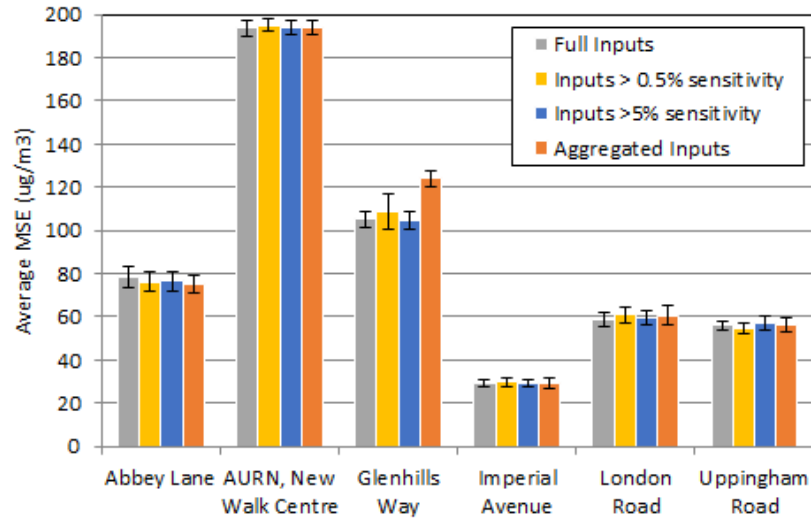
Fig. 2: NO_2 mean squared error.

After the training of the proposed neural models and the evaluation of their performances, the input variables were examined in order to figure out the most important input to estimate the roadside concentration. A simple sensitivity analysis was carried out by increasing a certain percentage (in our case 5%) each of the input variables, and then calculating the change caused in the output. The sensitivity of each input is defined by $S(x) = 100\frac{\Delta V_{output}}{\Delta V_{input}}$ where x is an input variable. Figure 4 shows the sensitivity of each input variable for each monitored location. Note that the wind direction (H) is not important in the estimation, since it is always below 5% for all stations. However, the air pressure (A) is very significant, except for the station placed at Imperial Avenue.
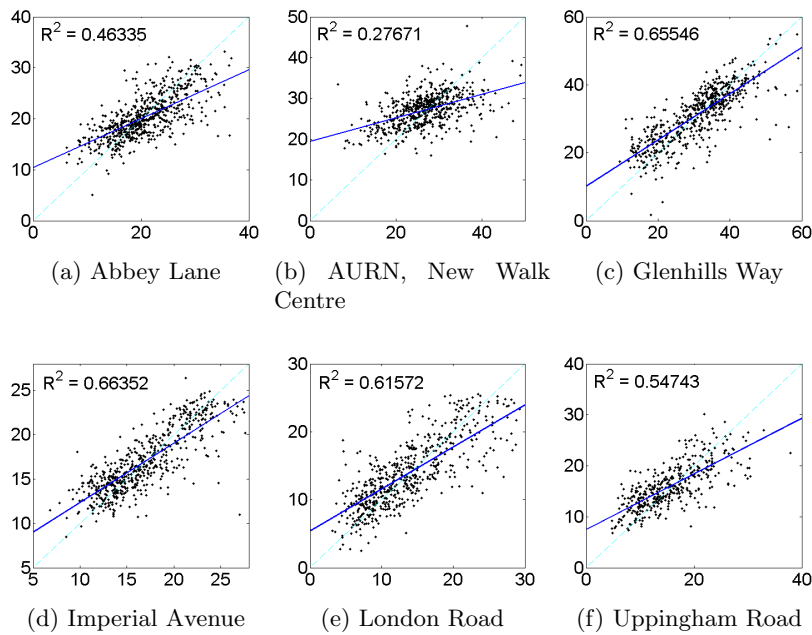
(a) Abbey Lane

(b) AURN, New Walk Centre

(c) Glenhills Way

(d) Imperial Avenue

(e) London Road

(f) Uppingham Road

Fig. 3: NO_2 regressions

6   Conclusion

A neural network model has been presented in order to predict the air quality from traffic data. The input data is formed by climate data inputs (such as wind speed, wind direction, rain, radiation and air pressure), and SCOOT data (such as stops, delays, flows and congestions).

Three different approaches which consist of the proposed neural network model have been considered. Experiments have been carried out according to the data collected in several studied sites. The obtained results indicate that the most suitable proposal is composed by two separate ANN models, each one with one output (NO_2 or PM10) with a SOM based filtering.

Acknowledgment

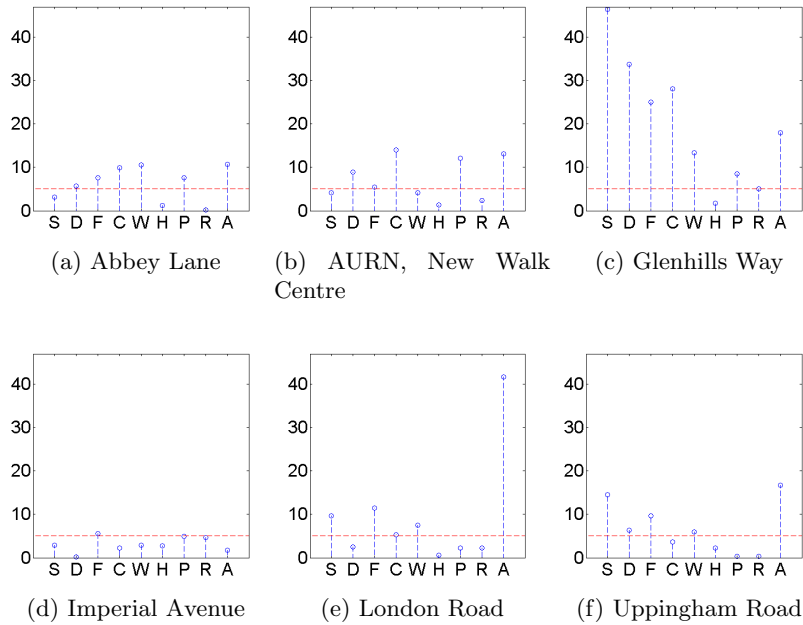| (a) Abbey Lane | (b) AURN, New Walk Centre | (c) Glenhills Way |
| (d) Imperial Avenue | (e) London Road | (f) Uppingham Road |

Fig. 4: Sensitivity analysis (%) of inputs by increasing 5%. The different inputs of the network are shown in horizontal axis: Stops, Delays, Flows, Congestions, Wind speed, Wind direction (H), Precipitations (Rain), Radiation and Air pressure.

References

1. Health effects of particulate matter. policy implications for countries in eastern europe, caucasus and central asia. Tech. rep., World Health Organization (2013)
2. Review of evidence on health aspects of air pollution. Tech. rep., WHO/Europe (2013)
3. Fisher, B.E.: Fuzzy approaches to environmental decisions: application to air quality. Environmental Science & Policy 9(1), 22–31 (2005)
4. Grivas, G., Chaloulakou, A.: Artificial neural network models for prediction of pm10 hourly concentrations, in the greater area of athens, greece. Atmospheric Environment 40(7), 1216–1229 (2006)
5. Hornik, K., Stinchcombe, M., White, H., Tinchcombe, M.: Multilayer feedforward networks are universal approximators. Neural Networks 2(5), 359–366 (1989)
6. Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J.: From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in bilbao. Environmental Modelling & Software 23(5), 622–637 (2008)
7. Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G.: Extensive evaluation of neural network models for the prediction of no2 and pm10 con-

centrations, compared with a deterministic modelling system and measurements in central helsinki. Atmospheric Environment 37(32), 4539–4550 (2003)

8. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural Networks 6(6), 861–867 (Jan 1993)

9. Moore, Ii, J.E., Mattingly, S.P., MacCarley, C.A., McNally, M.G.: Anaheim advanced traffic control system field operations test: A technical evaluation of scoot. Transportation planning and technology 28(6), 465–482 (2005)

10. Nagendra, S.S., Khare, M.: Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecological Modelling 190(1-2), 99–115 (2006)

11. Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M.: Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence 17(2), 159–167 (2004)

12. Passow, B.N., Elizondo, D., Chiclana, F., Witheridge, S., Goodyer, E.: Adapting traffic simulation for traffic management: A neural network approach. In: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013). pp. 1402–1407. No. Itsc, IEEE (oct 2013), http://ieeexplore.ieee.org/document/6728427/

13. Pelliccioni, A., Tirabassi, T.: Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. Environmental Modelling & Software 21(4), 539–546 (2006)

14. Pérez-Roa, R., Castro, J., Jorquera, H., Pérez-Correa, J., Vesovic, V.: Air-pollution modelling in an urban area: Correlating turbulent diffusion coefficients by means of an artificial neural network approach. Atmospheric Environment 40(1), 109–125 (2006)

15. Pfeiffera, H., Baumbacha, G., Sarachaga-Ruiza, L., Kleanthousb, S., Poulidab, O., Beyaz, E.: Neural modelling of the spatial distribution of air pollutants. Atmospheric Environment 43(20), 3289–3297 (2009)

16. Sowlat, M.H., Gharibi, H., Yunesian, M., Mahmoudi, M.T., Lotfi, S.: A novel, fuzzy-based air quality index (faqi) for air quality assessment. Atmospheric Environment 45(12), 2050 – 2059 (2011)

17. Viotti, P., Liuti, G., Di Genova, P.: Atmospheric urban pollution: applications of an artificial neural network (ann) to the city of perugia. Ecological Modelling 148(1), 27–46 (2002)

18. Zito, P., Chen, H., Bell, M.: Predicting real-time roadside co and no_2 concentrations using neural networks. Intelligent Transportation Systems, IEEE Transactions on 9(3), 514–522 (Sept 2008)

19. Zito, P., Chen, H., Bell, M.C.: Predicting real-time roadside co and no2 concentrations using neural networks. IEEE Transactions on Intelligent Transportation Systems 9(3), 514–522 (2008)