Pattern selection strategies for SO_2 forecasting models

M. Božnar and P. Mlakar

Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia

Abstract

SO_2 pollution of the atmosphere is one of the problems that have not yet been solved in Slovenia, especially around large coal fired thermal power plants. In past years we have developed Perceptron neural network based models for short - term forecasting of ambient SO_2 concentrations.

Basically the forecasting model is taught with patterns that represent SO_2 concentrations and the meteorological history of the observed polluted site. Therefore it is very important how these patterns are selected in order to obtain a model that is capable of accurate forecasting for as many different pollution situations as possible.

Three types of pattern selection strategies were developed. The models built according to these pattern selection strategies were tested with actual cases of SO_2 pollution for two different automatic measuring stations around the Šoštanj Thermal Power Plant and compared with reference models. The results showed a significant improvement of the models' forecasting capabilities (up to 20% better probability of successful forecasting of high SO_2 concentrations in relative terms).

1 INTRODUCTION

The problem of atmospheric pollution around large thermal power plants is only partially solved in Slovenia. Very high concentrations of SO_2 appear in their surroundings. These high concentrations are caused by high emissions and are enhanced by the complex terrain in which the thermal power plants are sited. A typical example of such pollution is that around the Šoštanj Thermal power plant (Elisei[4], Božnar[2]).

We have developed a Perceptron neural network based model for short-term prediction of ambient SO_2 concentrations (Božnar[1], Božnar[3], Mlakar[7], Mlakar[8]). The model's inputs (input features) are half hour average values of meteorological and emission measurements. The model's output is a prediction of the SO_2 concentration for the selected automatic measuring station for one measuring interval in advance (Figure 2).

The measurements come from Automated Measuring System of the Šoštanj Thermal Power Plant (Figure 1) (Lesjak[6]).

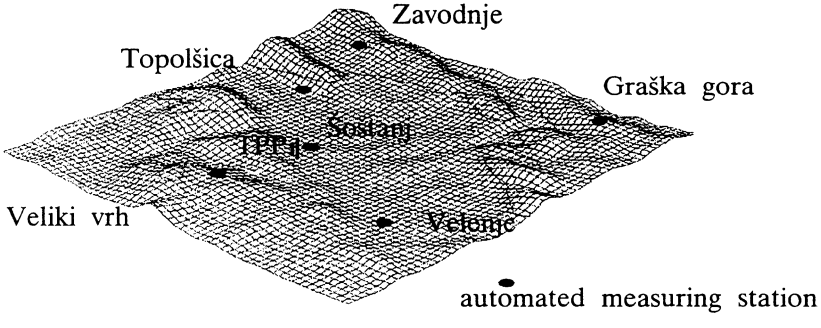

Figure 1 Measuring stations of the Automated Measuring System of the Šoštanj Thermal Power Plant

2 PERCEPTRON NEURAL NETWORK MODEL TRAINING AND PATTERN SELECTION STRATEGIES

The basic principle of building a Perceptron neural network based prediction model for a particular automatic measuring station is the following: First we should determine the proper input features for the models. These are those measurements that include significant information

for SO_2 forecasting at the observed automatic measuring station. Then we should find historical data - patterns (vectors compounded of values for the input features and already measured value for the output feature) that represent information about typical pollution situations at the observed station.

During the training process (backpropagation algorithm) the Perceptron neural network weights (parameters) are adjusted so that later when the training is completed, the model is capable of predicting the output feature's value only by knowing the values of the input features. This is possible if the pollution situation which is represented by the examined pattern is similar to those that were included in the training set. This is the so-called generalising capability of the Perceptron neural network.

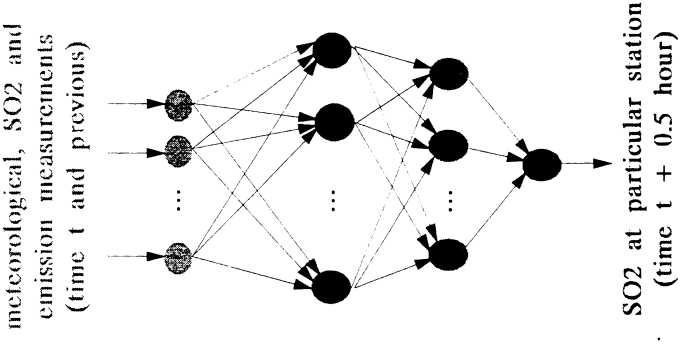

Figure 2 Perceptron neural network based SO_2 prediction model

It is clear that selection of the patterns on which the neural network is trained, determines the capabilities of the model constructed. The patterns used in the training set should represent all expected typical pollution situations.

In the case of the Šoštanj Thermal Power Plant and its Environmental Information System a research data base composed of over 70000 half hour measuring intervals for the period 1990 - 1994 is available. Every half hour interval can form one pattern for training the neural network or testing an already built model.

This amount of data is clearly much too great to be used for model construction at the same time. One of the basic problems of such a data base is that it contains many patterns that represent very frequent

but unimportant meteorological situations and only a few patterns that represent rare but very important pollution situations. It is the task of the pattern selection strategies to find these important patterns and to include them in the training set.

The basic idea of all pattern selection strategies developed is to find clusters of similar patterns and to represent uniformly all clusters in the training set of patterns. This way it is ensured that the model learns rare but important pollution situations as well as frequent ones.

We developed three pattern selection methods.

The first method (meteorological method) is a method based on meteorological knowledge about air pollution mechanisms. Among the available patterns equal numbers of patterns are selected for all known meteorological mechanisms that cause pollution at a particular site.

The second method (Kohonen neural network based method) does not require expert meteorological knowledge. Dividing the patterns into clusters of similar ones is successfully done by a selforganising Kohonen neural network.

The third method (multi - type models) is a modification of the first one. Different models are built for each pollution mechanism determined.

The methods were tested on the Šoštanj and Zavodnje automated measuring stations around the Šoštanj Thermal Power Plant.

3 DETERMINATION OF MODEL PERFORMANCE

The performance of models built according to the above pattern selection strategies were evaluated on independent data sets that were not used in the process of model construction. In order to determine how successful the pattern selection strategies were, the probability of successful prediction of high concentrations was determined for the reference models and for the models trained with selected data. The first type of reference model was the naive (persistent) predictor, and the second type was the Perceptron neural network based model trained with a large but unselected training set of patterns (very large number of successive patterns from a long time interval - this is the same multitude of patterns that was later used for training pattern selection).

From the user point of view, it is most important that the model correctly predicts high peaks of SO_2 concentrations (without time shift) and that it does not make false alarms when the actual concentrations are low.

To fulfil these requirements a new form of criterion termed "p^6" was used defined as **the probability of successful prediction of high concentrations.** It is defined as the number of intervals with successfully predicted high concentrations (actual concentration $\geq$ 0.15 mg/m^3 and the absolute error $<$ 0.1 mg/m^3 or relative error $<$ 0.2) divided by the sum of the number of intervals with high concentrations plus the number of intervals with false alarms (a false alarm occurs when the actual concentration is low and the predicted concentration $\geq$ 0.25 mg/m^3).

4 METEOROLOGICAL PATTERN SELECTION METHOD

The meteorological pattern selection method is now explained for the case of the Zavodnje automated measuring station.

The first task is determination of different air pollution mechanisms that can possibly cause air pollution at the observed station. Zavodnje station lies on the slope of the hill approximately 7 km from the Šoštanj Thermal Power Plant. It is mainly polluted by three typical mechanisms: pollution in the temperature inversion situation, direct wind pollution and fumigation in convective situations (Figure 3). The latter two mechanisms cannot be distinguished reliably, because no SODAR measurements are available.

After the typical mechanisms are determined, patterns representative of the situations in which they occur should be equally distributed in the training set of patterns. For training the Zavodnje model training we took about 2000 patterns. Half of them had high output concentrations and half had low ones. Approximately half of the high concentrations were caused by a nocturnal thermal inversion pollution mechanism and the other half by direct wind or fumigation. Similarly, the patterns with low output concentrations were distributed among different meteorological situations.

The model trained with selected patterns showed up to 20% better probability of successful prediction of high concentrations (p^6) when tested on shorter independent time intervals (these patterns are called the "production set") and approximately 8% when tested on almost 8000 patterns covering one autumn - winter - spring interval in comparison to the reference models.

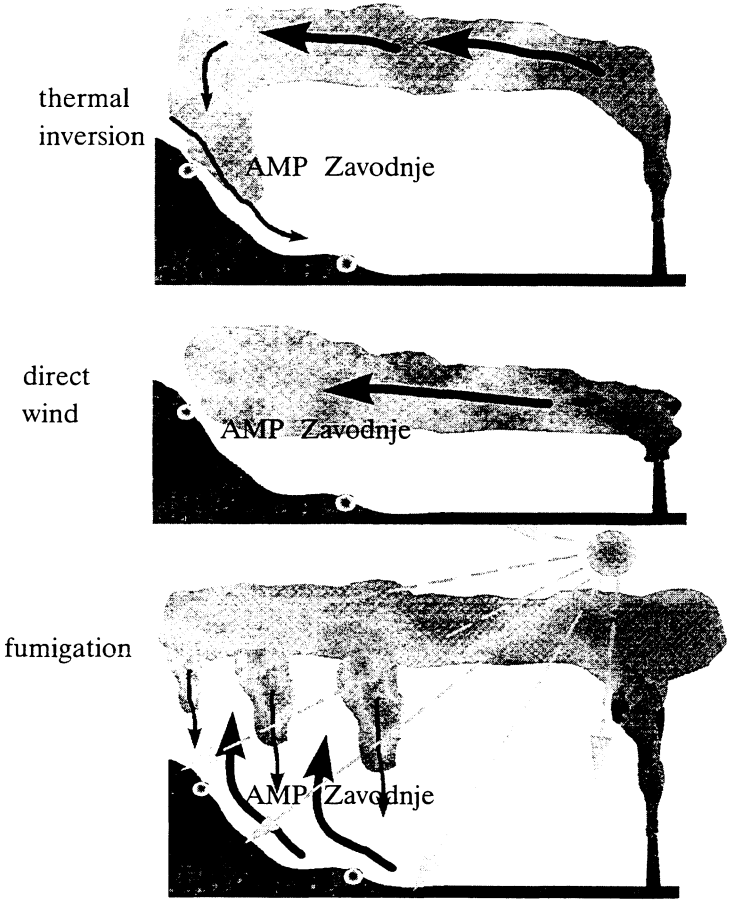

Figure 3 Three typical pollution mechanisms at the Zavodnje station

5 KOHONEN NEURAL NETWORK BASED PATTERN SELECTION METHOD

In the Kohonen neural network based pattern selection strategy the determination of pollution mechanisms that determine clusters of patterns is not needed. Division of patterns into clusters of similar ones is done by a selforganising Kohonen neural network (Kohonen[5]). All the patterns from a sufficiently long time interval (up to 8000 patterns in our case) are processed by a Kohonen neural network that divides them into clusters. The proper number of clusters is not known in advance, so

trials should be made of different numbers and then the meaningful number of clusters should be determined (Mlakar[9]). The method continues with the division of the patterns into a meaningful number of clusters. Clusters usually have a very different number of patterns. From every cluster an equal number of patterns should be selected randomly into the training set, so that the Perceptron neural network can learn all the typical mechanisms equally.

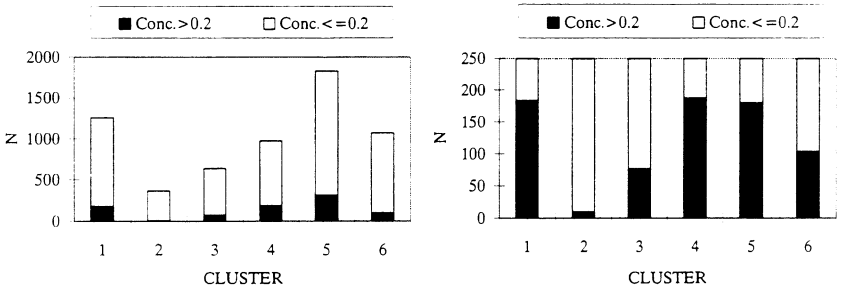

Figure 4 Division of unselected and selected patterns into clusters for the Zavodnje training set

The models built according to this pattern selection method were tested on the same smaller independent set as the models trained by the meteorological pattern selection based training set. This method show a very similar increase (about 20%) of probability of successful prediction of high concentrations (p^6) in comparison to reference models.

6 MULTI - TYPE MODEL PATTERN SELECTION METHOD

This method is similar to the meteorological one explained before. Also in this method we should first divide the patterns into clusters of similar ones representing typical mechanisms that may cause pollution at the observed station. But the second step of the method is the construction of different models - one for each pollution mechanism. Each model is trained only with patterns representing this particular one pollution mechanism (including patterns with low output concentrations, but a similar meteorological mechanism). When using already built models, each independent production pattern is tested with the model built for the

corresponding mechanism. The final result is calculated over all production patterns.

The multi - type model pattern selection method is based on the assumption that the Perceptron neural network based model can learn more successfully from a unique training set representing only one pollution mechanism. Learning of different mechanisms should be more difficult task.

For the Zavodnje station this method was tested by dividing all the patterns into two groups according to the ground level wind direction at the Zavodnje station. The first group represents possible thermal inversion pollution episodes and the second one direct wind or fumigation pollution (the station lies on the slope of a hill so the prevailing wind direction is up the slope, which is also the direction from the Thermal Power Plant). For each group of patterns separate models were built. But the increase of the probability of successful prediction of high concentrations (p^6) was only about 8% for the smaller production set. This is less than for the other two methods, probably because it is not always clear if the pattern belongs to the direct wind or fumigation mechanism. We expect that with more clear data (more features that enable data division) the method can give better results.

7 CONCLUSIONS

Three different training pattern selection methods that enhance neural network based model prediction capabilities were presented. The efficiency of the methods was measured indirectly by evaluating the resulting trained model performances on large independent data sets.

The evaluation of the three methods and two reference models on the same training set is shown in Figure 5.

The best results were obtained by the meteorological method and the Kohonen neural network based method. The pattern selection strategies were developed for the Perceptron neural network based models, but their basic idea can also be used in other types of prediction models.

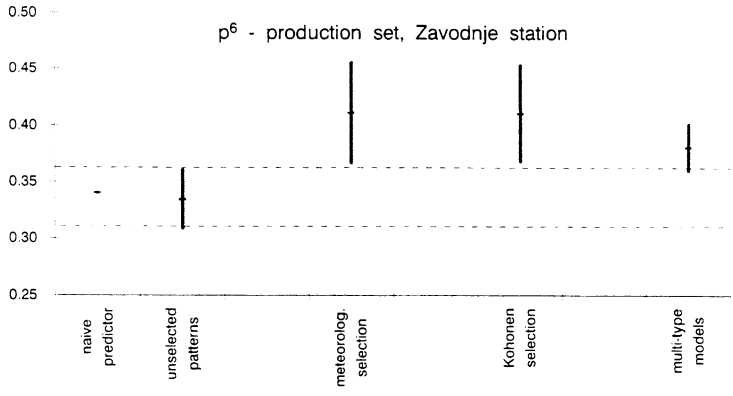

Figure 5 Results of different pattern selection method based models (models tested on independent production sets for the Zavodnje station, 954 patterns, p^6 measure, 95% confidence level)

8 INDEX - KEY WORDS

Perceptron neural network, prediction model, SO_2 pollution, pattern selection strategies, large data sets

9 REFERENCES

1. Božnar M., M. Lesjak, P. Mlakar, A Neural Network-Based Method for Short-Term Predictions of Ambient SO_2 Concentrations in Highly Polluted Industrial Areas of Complex Terrain, Atmospheric Environment, 27B, pp.221-230, (1993).

2. Božnar M., P. Mlakar, Analysis of ambient SO_2 concentrations in the surroundings of the Šoštanj thermal power plant, 4th International Conference on Air pollution, Toulouse, 1996, Air Pollution IV, Monitoring, Simulation & Control / eds. B. Caussadle, H. Power, C. A. Brebbia, Southampton, Boston, Computational Mechanics Publ. pp.727-734, (1996).

3. Božnar M., P. Mlakar, Neural Networks - a New Mathematical Tool for Air Pollution Modelling, 3rd International Conference on Air pollution, Porto Carras, 1995, Air Pollution III, Volume 1, Theory and

Simulation / eds. H. Power et al., Southampton, Boston, Computational Mechanics Publ. pp.259-266, (1995).

4. Elisei et. al., Experimental Campaign for the Environmental Impact Evaluation of Sostanj Thermal Power Plant (1992), Institut Jožef Stefan, ENEL/DSR/CRTN (Milano), CISE (Milano) - Progress Report, (1992)

5. Kohonen T., Self-organizing maps, Springer, Berlin (1995)

6. Lesjak M., B. Diallo, P. Mlakar, Z. Rupnik, J. Snajder and B. Paradiz., Computerised ecological monitoring system for the Šoštanj thermal power plant, in Man and his ecosystem, Proc. 8th World Clean Air Congress, Hague, The Netherlands, (1989), 3-31 to 3-38

7. Mlakar P., M. Božnar, M. Lesjak, Neural Networks Predict Pollution, 20th International Technical Meeting on Air Pollution Modelling and Its Application, Valencia, 1993, Proceedings.Volume 3, pp.531-532, (1993).

8. Mlakar P., M. Božnar, Short-term air pollution prediction on the basis of artificial neural networks, 2nd International Conference on Air pollution, Barcelona, 1994, Air Pollution II.Volume 1, Computer Simulation / eds.J.M. Baldasano et al., Southampton, Boston, Computational Mechanics Publ., pp.545-552, (1994).

9. Mlakar P., M. Božnar, Analysis of winds and SO2 concentrations in complex terrain, 4th International Conference on Air Pollution, 1996, Air Pollution IV.Monitoring, Simulation and Control / eds.B. Caussade, et al., Southampton, Boston, Computational Mechanics Publications (1996), str.455-464.