Neural networks in air quality forecasting

T-61.6060 Seminar presentation

Mikko Korpela, mvkorpel@cc.hut.fi

22 Feb 2006

Outline

- Introduction
- Measurement data
- Analysis methods
- Performance comparisons
- Possible improvements
- Conclusions

Introduction

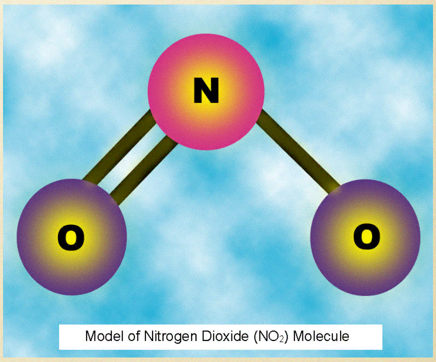
- Presentation based on *Kolehmainen et al., 2001, "Neural networks and periodic components used in air quality forecasting"*
- Air quality is a factor in the quality of living
- Pollution particularly in urban areas
  - Traffic, other human activities
- Forecasting of pollution peaks ➪ evasive action
  - Restrictions on traffic and industry

Measurement data

- Time series measurements, one hour intervals
- Stockholm, Sweden 1994–1998
- NO_2 (average of four measuring stations)
- Meteorological variables
  - Temperature, wind speed, wind direction, solar radiation
- Hour of the day (and the day?), month of the year
- Discontinuous variables transformed (sin, cos)
- Missing values filled in (0,94 %)

Forecasting task

- NO_2 is a toxic gas
- Predict future NO_2 concentration
- Present values of the variables (incl. NO_2) are known
- How far into the future?
    - Not specified in the paper!



Overview of methods

- Preprocessing
    - Extracting periodic components and trend
    - Enhanced performance of neural networks?
- Artificial neural networks
    - Multi-layer perceptron (MLP)
    - Self-organising map (SOM)

Periodic components and trend

- Periodic components are to be expected
  - Traffic, industry, ...
  - Periodic variations in nature
- Regression with zero and first degree curves, sine and cosine functions (year, week, day)
- Estimated from 1994–1997 data

$$y_t = \sum_{k=1}^{m} (a_k \sin k\omega_0 t + b_k \cos k\omega_0 t) + \sum_{l=0}^{n} c_l t^l$$

Regressor coefficients

No trend

Table 1
Regressor coefficients of the curve fitting using training data from the years 1994–1997[a]

l or k	Curves c_l	Year a_k	Year b_k	Week a_k	Week b_k	Day a_k	Day b_k
0	0.04	—	—	—	—	—	—
1	0.00	0.22	− 0.17	− 0.28	− 0.14	− 0.59	− 0.26
2	—	0.09	− 0.16	− 0.13	0.07	− 0.12	0.17
3	—	0.05	0.08	0.01	0.07	0.15	0.03
4	—	− 0.04	− 0.03	− 0.03	− 0.03	− 0.02	− 0.05
5	—	0.02	0.07	− 0.15	− 0.01	− 0.04	0.06
6	—	− 0.10	− 0.03	− 0.01	0.15	0.03	0.01
7	—	− 0.05	0.03	—	—	0.01	0.02
8	—	− 0.02	− 0.04	—	—	− 0.01	0.00
9	—	− 0.10	0.01	—	—	0.01	0.00
10	—	0.04	0.02	—	—	0.01	0.00
11	—	0.03	0.00	—	—	—	—

[a] l = multifold for curve fitting, k = multifold for sine and cosine fitting, a_k = coefficient for sine terms, b_k = coefficient for cosine terms and c_l = coefficient for curves.

Periodic components

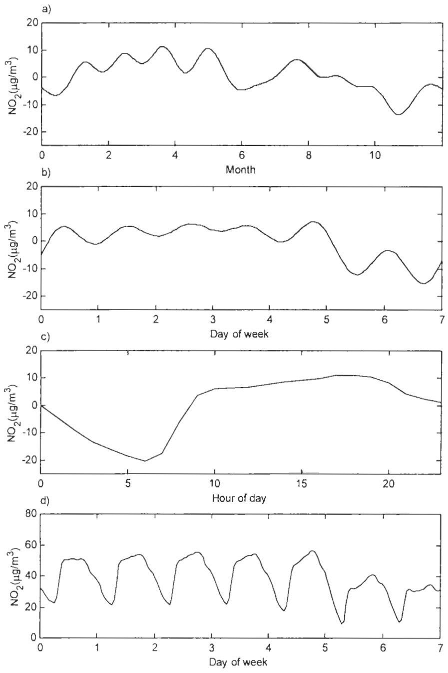
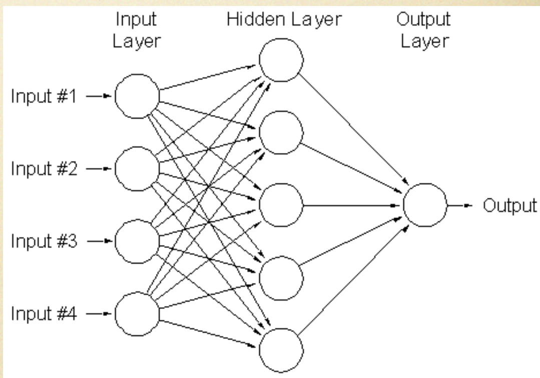


Fig. 1. Periodic components of NO_2 concentration of the data from the years 1994–1997: year-level component (a), week-level component from Monday to Sunday (b), day-level component (c) and a typical week composed of periodic components and an average (d).

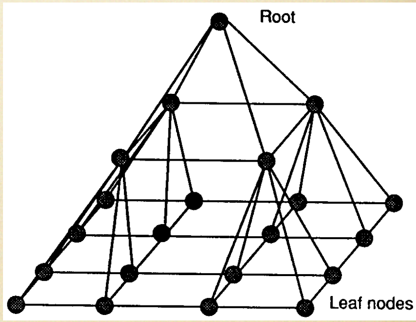
Multi-layer perceptron (MLP)

- Supervised learning
- Error back-propagation algorithm
- Details used in *Kolehmainen et al.*
  - One hidden layer
  - Sigmoid and linear transfer functions



Self-organising map (SOM)

- Unsupervised learning
- Data is represented with a topologically connected lattice
- Ability to visualise the data
- A variation called tree-structured SOM (TS-SOM) used here
    - A number of SOMs organised hierarchically



Application of the methods

- Training with 1994–1997 data
- Residual NO_2 data (periodicities removed)
- Plain NO_2 data
- MLP with 24 or 16 neurons in hidden layer
- SOM
    - Prediction is a function of BMU and time
    - Extra information (NO_2 statistics) stored in addition to SOM weights
    - 4096 or 16384 neurons

Performance indicators

- Performance of the methods was tested
- Data from year 1998 as the test set
- Prediction results were recorded
- Visual inspection of results
- Selected statistical indicators
  - Root mean square error (RMSE)
  - Coefficient of determination (R^2)
  - Index of agreement (d), ...

Predicted vs. observed (1)

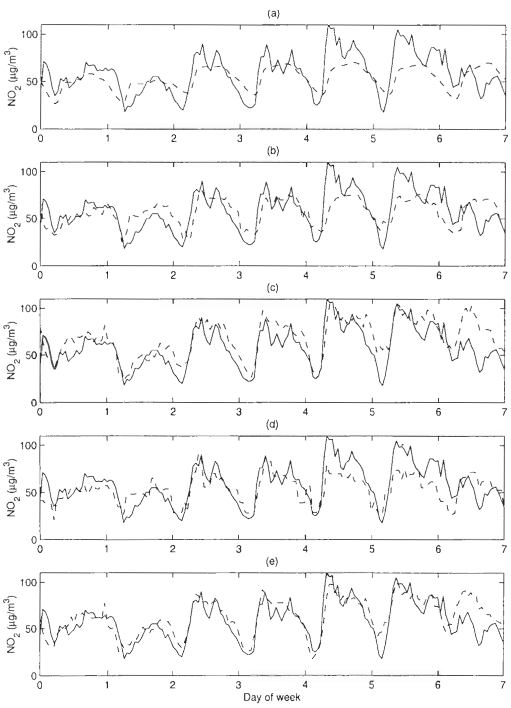


Fig. 2. Predicted (dashed line) versus observed (solid line) signals of NO_2 for the 17th week of 1998. (a) periodic fitting, (b) SOM applied to the residual, (c) SOM, (d) MLP applied to the residual and (e) MLP.

Predicted vs. observed (2)

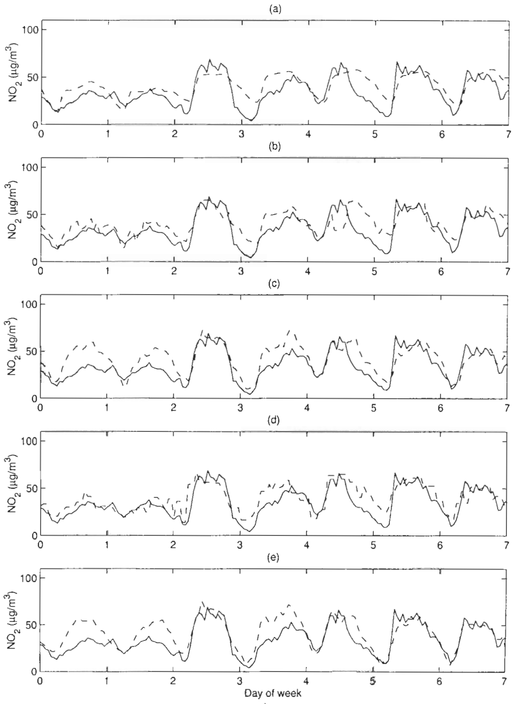
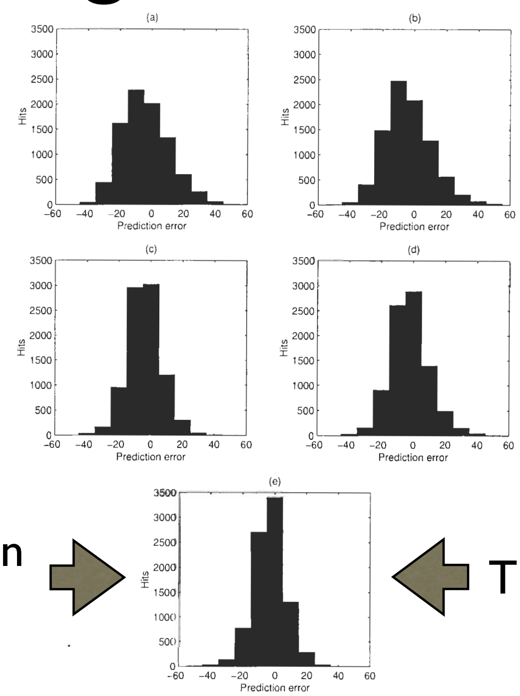


Fig. 3. Predicted (dashed line) versus observed (solid line) signals of NO_2 for the 41st week of 1998. (a) periodic fitting, (b) SOM applied to the residual, (c) SOM, (d) MLP applied to the residual and (e) MLP.

Histograms of error



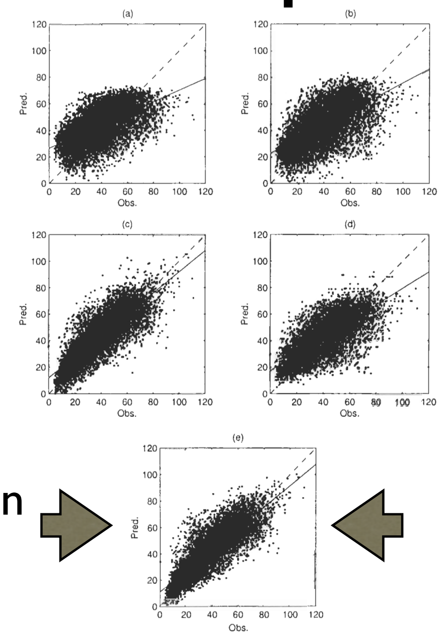
MLP with plain NO2 values

The best

Fig. 4. Histograms of error in NO_2 predictions (signal–estimate) for the year 1998. (a) periodic fitting, (b) SOM applied to the residual, (c) SOM, (d) MLP applied to the residual and (e) MLP.

Predicted vs. observed (scatter plots)

Only periodic



MLP with plain NO_2 values

Smallest bias (11.08)
Best slope (0.80)

Fig. 5. Plots of predicted versus observed signals of NO_2 for the year 1998 using different models. The plots are further enhanced with a least-squares fitting line (solid) and a line showing the perfect fit (dotted). (a) periodic fitting, (b) SOM applied to the residual, (c) SOM, (d) MLP applied to the residual and (e) MLP.

Numerical performance indicators

Estimators for validity of the model for the validation data (1998)

Estimator	Periodic	Periodic + SOM	SOM	Periodic + MLP	MLP
R^2	0.61	0.77	0.72	1.00	0.96
Bias	3.93	3.93	2.17	4.05	3.28
d	0.73	0.77	0.85	0.89	0.90
RMSE	15.38	15.21	12.45	11.41	10.72
$RMSE_s$	10.86	9.34	7.16	5.43	4.72
$RMSE_u$	10.88	12.00	10.19	10.03	9.45
PSE	0.50	0.38	0.33	0.23	0.19

Proportion of systematic error

Best

Worst

- MLP with plain NO_2 values generally the best

Possible improvements

- Improvement of the periodic fitting
- More meteorological variables
- Data from previous days
  - Number of variables increases quickly
- Chaotic nature of the atmosphere
  - Take this into account in the model

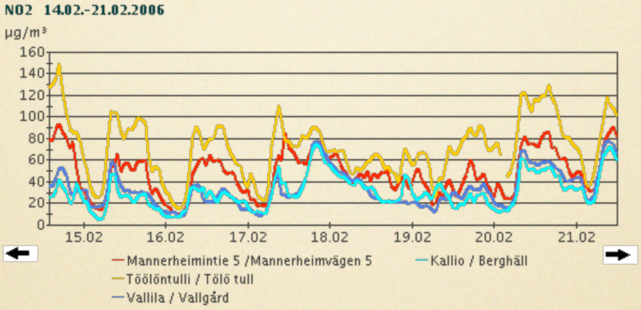
Conclusions

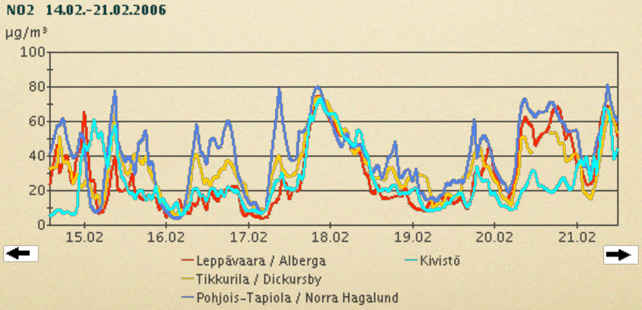
- Different methods for NO_2 forecasting
- Removal of periodic components doesn't improve neural network methods
- MLP better than SOM **in this task**
  - MLP also a more natural choice: input-output mapping, no extra data structures
- Fairly good estimates
- Extreme values hard to predict

Local measurements

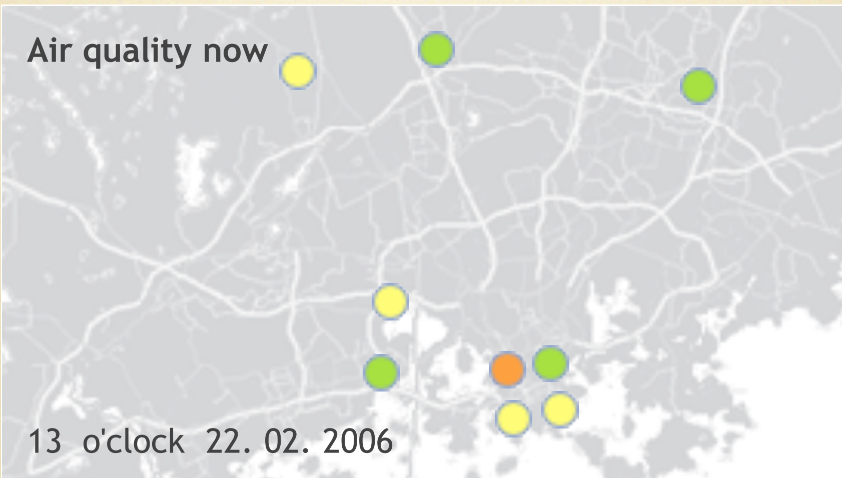
NO_2 concentrations in Helsinki region

© YTV



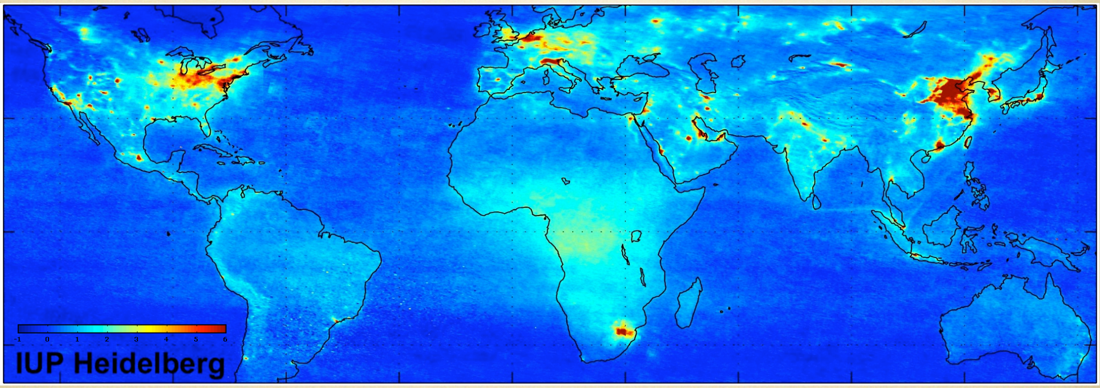


Air quality now



© YTV

The end



Global NO_2 pollution map - Jan 2003 to June 2004
© University of Heidelberg