SMHI

FAIRMODE Technical Meeting, Aveiro, Portugal, 24-25 June 2015

Improvements of SIMAIR with support vector regression, SVR

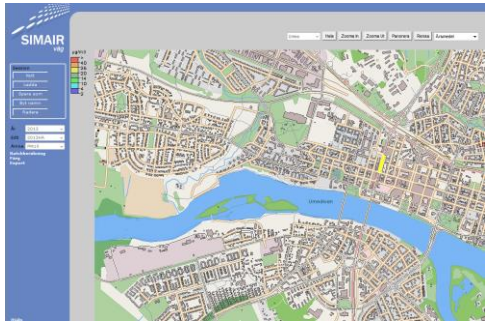
Stefan Andersson and Heléne Alpfjord

Swedish Meteorological and Hydrological Institute

SIMAIR – Swedish national Air Quality model system

Web-based Air Quality model tool

- Can be used by all municipalities and cities in Sweden.
- Simple user interface.
- Fast calculations.
- Applications for road traffic and small-scale residential wood combustion.

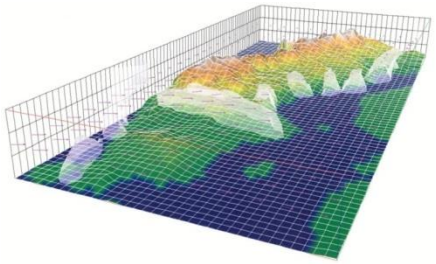


http://www.smhi.se/tema/SIMAIR

SIMAIR - coupled model system

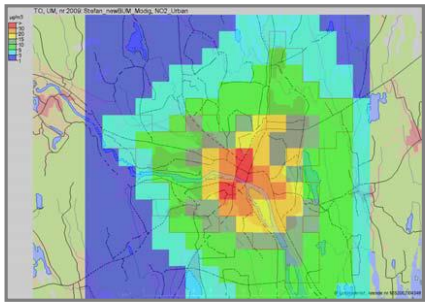
Regional scale

- MATCH: Multi-scale Atmospheric Transport and Chemistry Model
- 44 km x 44 km (Europe)



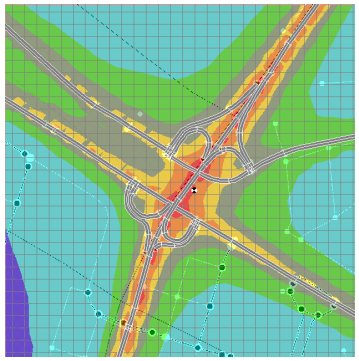
Urban scale

- BUM – back-trajectory model + Gaussian model
- 1 km x 1 km



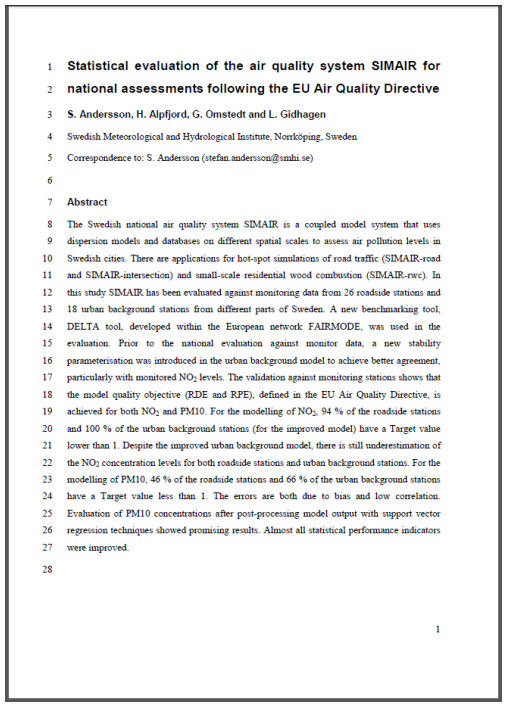
Local scale

- OSPM (street canyon)
- Open road
- Dispersion
- 25 m x 25 m



Statistical evaluation of SIMAIR

- Paper submitted 2015 to ACP.
- Validation including Delta-tool.
- Evaluation of statistical post-processing technique; support vector regression, SVR.



Statistical evaluation of the air quality system SIMAIR for national assessments following the EU Air Quality Directive

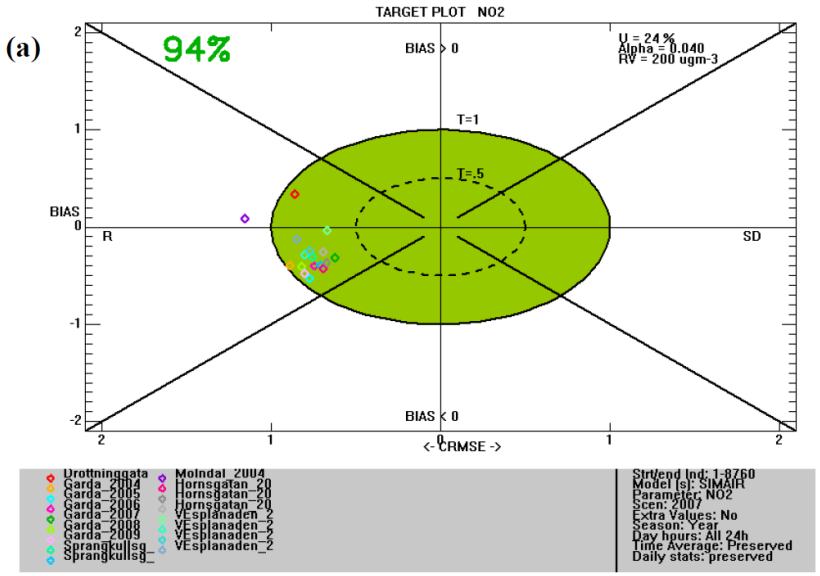
S. Andersson, H. Alpfjord, G. Omstedt and L. Gidhagen

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Correspondence to: S. Andersson (stefan.andersson@smhi.se)

Abstract

The Swedish national air quality system SIMAIR is a coupled model system that uses dispersion models and databases on different spatial scales to assess air pollution levels in Swedish cities. There are applications for hot-spot simulations of road traffic (SIMAIR-road and SIMAIR-intersection) and small-scale residential wood combustion (SIMAIR-rwc). In this study SIMAIR has been evaluated against monitoring data from 26 roadside stations and 18 urban background stations from different parts of Sweden. A new benchmarking tool, DELTA tool, developed within the European network FAIRMODE, was used in the evaluation. Prior to the national evaluation against monitor data, a new stability parameterisation was introduced in the urban background model to achieve better agreement, particularly with monitored NO_2 levels. The validation against monitoring stations shows that the model quality objective (RDE and RPE), defined in the EU Air Quality Directive, is achieved for both NO_2 and PM10. For the modelling of NO_2, 94 % of the roadside stations and 100 % of the urban background stations (for the improved model) have a Target value lower than 1. Despite the improved urban background model, there is still underestimation of the NO_2 concentration levels for both roadside stations and urban background stations. For the modelling of PM10, 46 % of the roadside stations and 66 % of the urban background stations have a Target value less than 1. The errors are both due to bias and low correlation. Evaluation of PM10 concentrations after post-processing model output with support vector regression techniques showed promising results. Almost all statistical performance indicators were improved.



Data fusion using Support Vector Regression (SVR)

- Aim: To implement a method for statistical post-processing of dispersion model output from SIMAIR.
- The compound examined is PM10.
- The statistical method is developed based on data from Hornsgatan in Stockholm 2007-2009. Validation using data from Umeå and Gothenburg.
- Two years used as evaluation set, one year as validation set

Hornsgatan in Stockholm

- Yearly daily average of 28 000 vehicles
- One of the streets in Sweden with the highest concentrations of PM10

Gårda in Gothenburg

- Yearly daily average of 90 000 vehicles

Västra Esplanaden in Umeå

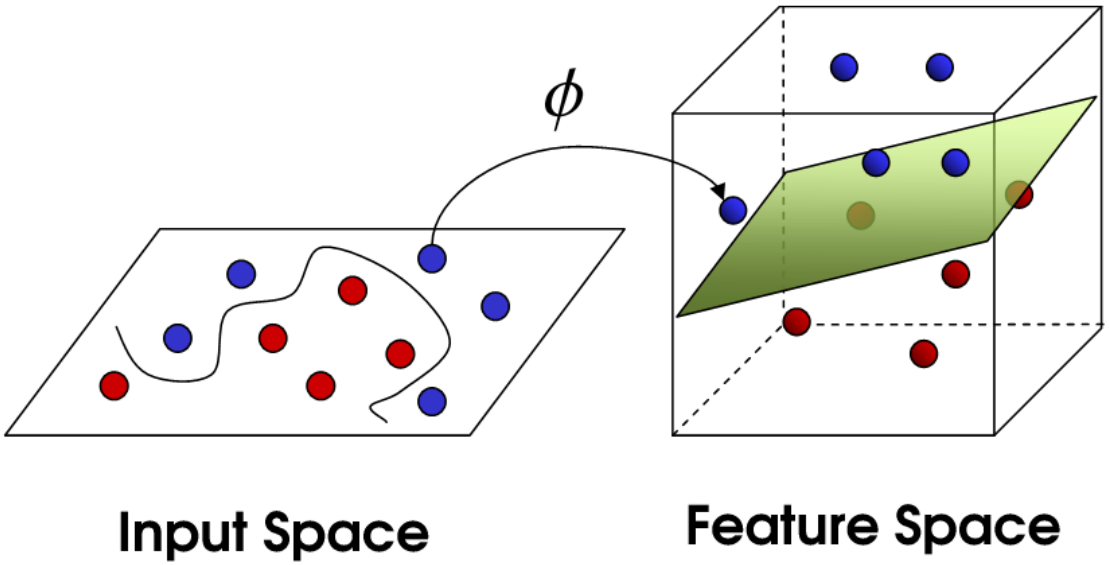
- Yearly daily average of 24 000 vehicles
- Problems with inversions

Support Vector Regression

- Data fusion methods are completely statistical, do not take physical or chemical laws into account.
- Support Vector Regression can be used for both linear and non-linear regression.
- Does not assume normally distributed residuals nor constant variance.

Support Vector Regression

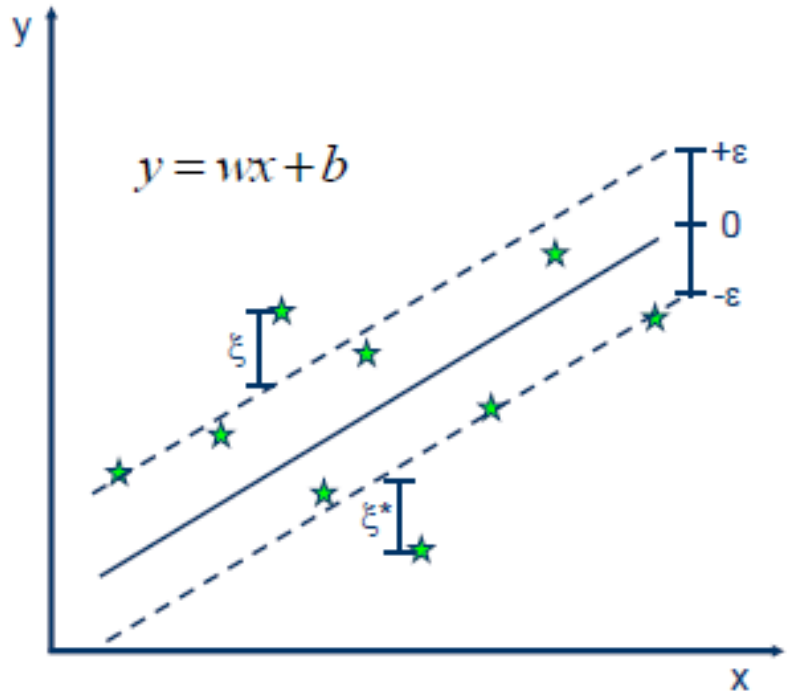
- SVR is a form of supervised machine learning.
  - Training using examples
  - Prediction – input data are given values based on training
- Input data is transformed to a higher dimensional space where a linear regression is performed
- The transformation uses kernels (here radial basis functions)



Source: http://stackoverflow.com/questions/9480605/what-is-the-relation-between-the-number-of-support-vectors-and-training-data-and

Support Vector Regression

- Minimisation is based on both complexity and residual size. Overfitting is avoided.
- The optimisation problem:



• Minimize:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^*\right)$$

• Constraints:

$$y_i - wx_i - b \le \varepsilon + \xi_i$$

$$wx_i + b - y_i \le \varepsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \ge 0$$

Source: http://www.saedsayad.com/support_vector_machine_reg.htm

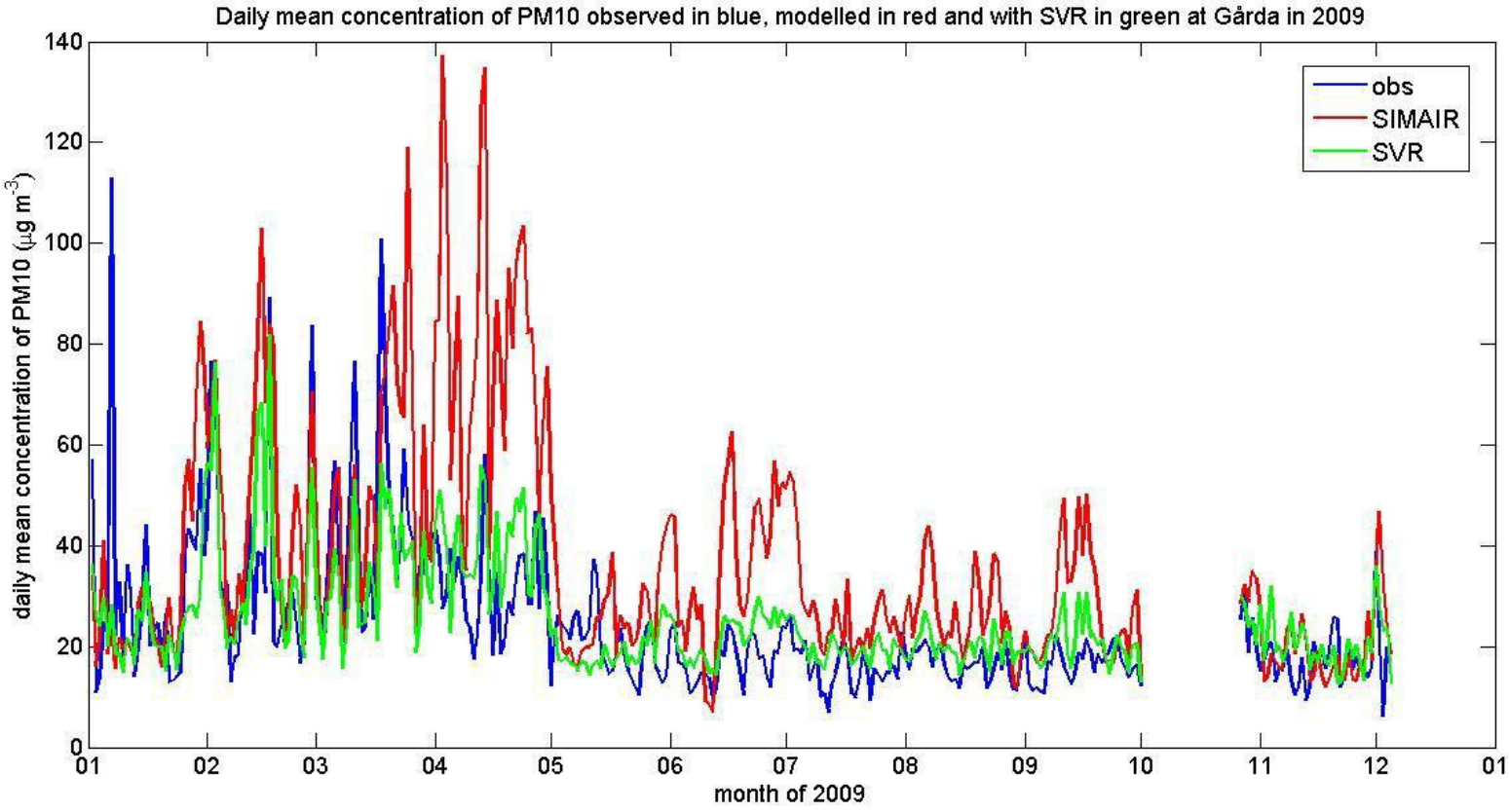
Explanatory variables

- Modelled concentration of PM10 (local and urban background), number of vehicles
- Meteorological: precipitation (exponentially filtered), humidity, wind direction, temperature difference during the day, accumulated global irradiance
- Indicators of day/night and season

Results from Gårda in 2009

	Observations	SIMAIR model	SVR
Yearly mean ($\mu g/m^3$)	23.7	37.0	24.8
90-percentile daily mean ($\mu g/m^3$)	39.3	70.3	39.3
Days > 50 $\mu g/m^3$	15	68	15
RPE %		56	5.6
RDE %		33	3.0
r daily mean		0.60	0.76
r hourly		0.43	0.57
RMSE daily mean ($\mu g/m^3$)		15.0	6.49
RSME hourly ($\mu g/m^3$)		29.5	18.3

Results from Gårda in 2009



SMHI

Conclusions

- The SVR method shows promising results for correcting SIMAIR calculations
- It is important that training data are representative for the period to be predicted
- Independent observation data for validation
- Useful tool for model evaluation
- Future extensions – use the method for sites without measurements by training on adjacent locations?