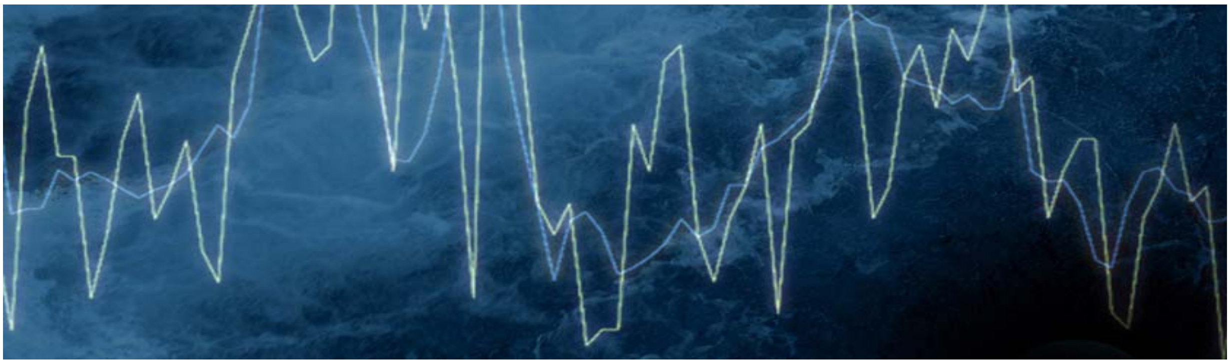
Calculation of pseudo $PM_{2.5}$ annual mean concentrations in Europe based on annual mean PM_{10} concentrations and other supplementary data



ETC/ACC Technical Paper 2010/9
January 2011

Bruce Denby, Giulio Gola, Frank de Leeuw,
Peter de Smet and Jan Horálek

European Topic Centre on Air and Climate Change

The European Topic Centre on Air and Climate Change (ETC/ACC)
is a consortium of European institutes under contract of the European Environment Agency
PBL UBA-D UBA-V NILU AEAT AUTh CHMI MET.NO ÖKO TNO REC

Front page picture:
Cover picture taken from the IntelliSense final report (Kommersialiseringsstipend fra Norges forskningsråd, prosjekt nr 192469)

Author affiliation:
Bruce Denby: Norwegian Institute of Air Research (NILU), Kjeller, Norway
Giulio Gola: Institute for Energy Technology (IFE), Halden, Norway
Frank de Leeuw, Peter de Smet: Netherlands Environmental Assessment Agency (PBL), Bilthoven, The Netherlands
Jan Horálek: Czech Hydrometeorological Institute (CHMI), Prague, Czech Republic

DISCLAIMER

ETC/ACC Technical paper 2010/9
European Topic Centre on Air and Climate Change
PO Box 303
3720 AH Bilthoven
The Netherlands
Phone +31 30 2748562
Fax +31 30 2744433
Email etcacc@pbl.nl
Website http://air-climate.eionet.europa.eu/

Summary

One of the tasks of ETC/ACC is to develop methods that will provide spatial assessment of a number of air pollutants for all of Europe. Currently maps are produced for PM_{10} but not $PM_{2.5}$ due to a lack of monitoring data for $PM_{2.5}$, reported in AirBase, compared to PM_{10}. The lack of monitoring data makes spatial assessment very uncertain and for this reason $PM_{2.5}$ maps are currently not operationally produced. To try to improve this situation we investigate two approaches for producing 'pseudo' $PM_{2.5}$ measurements that can be used for producing European wide maps of $PM_{2.5}$. These pseudo measurements are based first and foremost on the measured PM_{10} concentrations but with the addition of supplementary geographical, population and meteorological data.

The two approaches investigated are 'Empirical Ensemble-based Virtual Sensing' (EEVS), which is based on Artificial Neural Networks, and multiple linear regression (MLR), a standard variational technique. The two approaches are compared and the results are discussed. A number of different supplementary data sources are assessed for their usefulness and a selection is made of these.

The two approaches give similar results, with the EEVS approach showing slightly lower RMSE than MLR (normalised RMSE's of 17% and 19% respectively). For all other metrics assessed there was no significant difference. Neither approach was found to fulfil the monitoring quality objectives as laid out in the European air quality directive (2008/50/EC), but both would fulfil the requirements as laid out for modelling and indicative measurements in the AQ Directive.

From a practical application point of view the EEVS approach is completely opaque, as there is no information available on either the Neural Network models used in the application or the parameters they apply after training. In addition, a patent application is currently pending for the application of EEVS for pseudo $PM_{2.5}$ calculations. This makes its use by third parties unclear. Unless this situation changes in the future, we recommend MLR as a suitable approach for future $PM_{2.5}$ mapping for Europe

Contents

Summary 3

1. Introduction 5

2. Datasets 5

2.1 AirBase air quality data 5

2.2 Population 6

2.3 Meteorological data 6

3. Methodology 6

3.1 Empirical Ensemble-based Virtual Sensing (EEVS) 6

3.2 Multiple Linear Regression (MLR) 7

3.3 Assessment of supplementary sources 7

3.4 Comparison and validation of the two approaches 8

4. Exploration and selection of supplementary data 8

4.1 MLR supplementary data assessment 9

4.2 EEVS supplementary data assessment 13

4.3 Discussion concerning the two supplementary data choices 15

5. Comparison of the MLR and EEVS approaches using the selected supplementary data source...16

5.1 Validation on 2007 data, trained on 2004 – 2006 data 16

5.2 Comparison of predicted $PM_{2.5}$ at PM_{10} stations for 2007 17

6. Discussion and conclusion 18

References 19

1. Introduction

One of the tasks of ETC/ACC is to develop methods that will provide spatial assessment of air pollutants for all of Europe. This task, and its development, is described in a number of ETC/ACC publications (Horálek et al., 2007; 2010; Fiala et al., 2009, de Smet et al., 2010). To create these maps, monitoring data from AirBase is combined with other spatially resolved supplementary data (e.g. EMEP model outputs, altitude, population, meteorology) by first applying multiple linear regression and there after residual kriging. Currently, maps for PM_{10} indicators and ozone indicators are operationally produced.

There is also a need to produce European wide maps of $PM_{2.5}$, as this is considered to be a major contributor to health impact. However, currently there is a lack of monitoring data for $PM_{2.5}$ in AirBase compared to PM_{10}. In 2007, $PM_{2.5}$ annual mean concentrations were reported for 229 background stations (98 urban, 60 suburban and 71 rural). For the same year and station categories 1160 PM_{10} annual mean concentrations were reported. The lack of monitoring data makes spatial assessment very uncertain and for this reason $PM_{2.5}$ maps are currently not operationally produced.

If $PM_{2.5}$ maps are to be produced then some method must be applied to fill in the missing $PM_{2.5}$ data. One suggested method is to use the PM_{10} data and produce 'pseudo' $PM_{2.5}$ observations based on these data. Since the measurement of PM_{10} also contains the measurement of $PM_{2.5}$, it may be possible, to some extent, to derive the fraction of $PM_{2.5}$ that is contained in PM_{10}.

In a preliminary assessment by ETC/ACC (de Leeuw and Horálek, 2009), the ratio of $PM_{2.5}$:PM_{10} was assessed based on station type and based on a basic North-South-East-West geographical distribution. These results indicated that $PM_{2.5}$ could be derived from the PM_{10} measurements but with an uncertainty that was considered unacceptable for use in mapping applications.

In 2009, NILU took advantage of a co-operative project with IFE (Norwegian Institute for Energy Technology) to explore the possibility of applying Neural Network approaches to derive $PM_{2.5}$ concentrations from PM_{10} concentrations using the AirBase database and a number of supplementary input data. This approach, developed at IFE, is called 'Empirical Ensemble-based Virtual Sensing' (EEVS) and bases the estimation of $PM_{2.5}$ on the use of an ensemble of Neural Networks. EEVS has proved successful in a number of practical engineering applications where direct monitoring of critical quantities is difficult or expensive to undertake.

In this report, we provide the results of the application of the EEVS methodology to estimate annual means of$PM_{2.5}$, along with an alternative simpler methodology that makes use of multiple linear regression. The two approaches are compared and the results are discussed. A number of different supplementary data sources are assessed for their usefulness and a selection is made.

2. Datasets

2.1 AirBase air quality data

The major datasets used in the study are the PM_{10} and $PM_{2.5}$ annual mean concentrations for Europe taken from AirBase (AirBase, 2010). The years 2004 – 2007 have been used. In addition to the PM data, both NO_2 and ozone (annual means) were extracted from AirBase for use as supplementary air quality data. The geographical positions and altitude of the stations were also used as supplementary data.

2.2 Population

Population density was used as supplementary data. This was provided on a 10 x 10 km^2 grid resolution. (Source EEA, pop01c00v3int, official version Aug. 2006; Owner: JRC).

2.3 Meteorological data

Both climatological data (years 1961 – 1990; New et al., 2002) (resolution 10 x 10 minutes) and current meteorological fields for the particular years (2004 – 2007) taken form ECMWF re-analysis (15 x 15 minutes) were used as supplementary data sources (European Centre for Medium-range Weather Forecasts; http://www.ecmwf.int/). The meteorological data used includes: wind speed, temperature, relative humidity, sunshine duration and precipitation

3. Methodology

3.1 Empirical Ensemble-based Virtual Sensing (EEVS)

To estimate the concentration of $PM_{2.5}$, the empirical model-based EEVS technique developed at the Norwegian Institute for Energy Technology (IFE), on which a patent is currently pending, has been applied. The approach is based on the use of feed-forward Artificial Neural Networks (ANNs) (Rumelhart and McClelland, 1986). In general, given a number of input parameters correlated to a quantity of interest (in this case the $PM_{2.5}$ concentration), the method aims at providing a reliable estimate of that quantity.

In the EEVS system employed by IFE, an ensemble of Neural Networks is applied. The performance of the ensemble of models is generally better than that of any of the individual models, i.e. it tends to be more robust with respect to the risk of over-fitting the data and it reduces the estimation uncertainty and avoids the instability problems connected to the ANN training procedure.

In the ensemble, each model is trained separately and the predictions of the individual models are then aggregated to produce the output of the ensemble. However, combining the output of several models is useful only if there is some form of “disagreement” between their predictions. Obviously, the combination of identical models produces no gain. In the EEVS technology, diverse models are generated simply by randomly sampling the values of the initial parameters of the ANNs and the ensemble aggregated output is obtained by retaining the median of the predictions of the individual models.

In the training phase of the empirical model, the average annual PM_{10} concentrations and the corresponding supplementary parameters are fed as inputs to the ANN to generate the estimate of $PM_{2.5}$. This is then compared to the corresponding actual measured $PM_{2.5}$ concentration (i.e. the target of the model’s learning) and the mismatch between these values is used to tune and adjust the parameters of the ANN in order to obtain a more accurate $PM_{2.5}$ estimate as the training proceeds (Rumelhart, and McClelland, 1986).A schematic view of EEVS is shown in Figure 1. More information concerning the technique can be found in Roverso (2009).

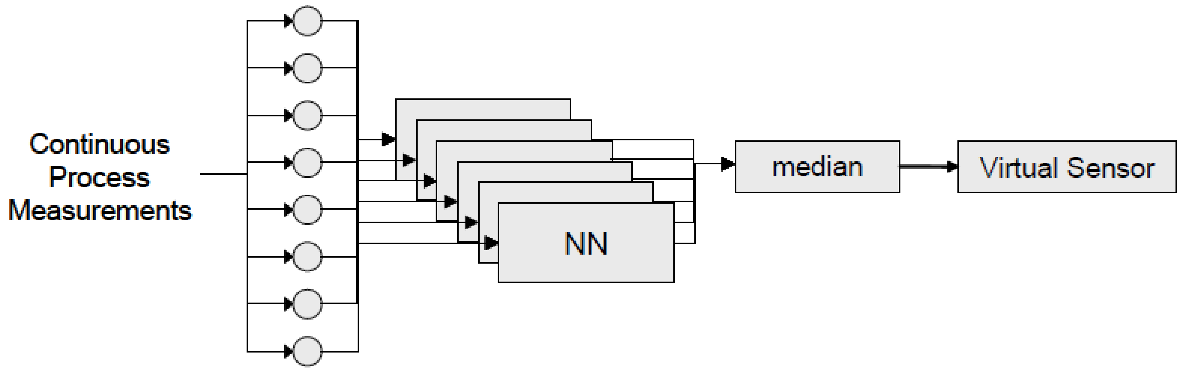


Figure 1.The EEVS concept, taken from Roverso (2009).

3.2 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is an optimization technique that fits a linear model to the data, minimizing the mean square error. It is straightforward and can be found in virtually all statistical software packages. The application of MLR is usually a stepwise procedure where new parameters are introduced to the regression and their impact on reducing the mean square error or on improving the correlation is assessed. It can be written in its simplest form as:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i$$

Where n is the number of regression parameters (x_i) used to fit y by adjusting the regression coefficients a_i.

As in the EEVS, PM_{10} and other supplementary data are provided as regression parameters and regression coefficients are determined to give the best fit. It is important to note that since the technique minimizes the mean square error, then the root mean square error (RMSE), which is one of the indicators used to assess the performance of MLR, will always remain the same or decrease with any additional parameter. For this reason it is also important to assess the performance of MLR using other metrics.

3.3 Assessment of supplementary sources

For both MLR and EEVS a number of supplementary data sources are assessed and a selection of these sources is made that are most useful for improving the prediction of $PM_{2.5}$ for each approach individually (Section 4). In the case of MLR, this selection was carried out by adding additional sources and assessing their impact on the RMSE and the correlation of the regression model, using all years of data. In the case of EEVS, this was carried out using a number of different indicators and the method applied was a cross validation where 75% of the data was randomly sampled and compared to the remaining 25%. This re-sampling was carried out 50 times.

In this regard, the two approaches and their selection of supplementary data are generally not consistent as different indicators and methods have been used (it should be noted that 'leave one out' as a validation method for MLR does not usually change the results to any significant degree when the data set is sufficiently large).

In addition to the different metrics, the supplementary sources for EEVS were assessed using both background and traffic stations, whilst the supplementary data selection procedure for MLR was based on background stations only. For comparative purposes, MLR was also applied to traffic sites but it was not optimised on these (see Section 4). The inclusion of traffic stations in the EEVS assessment was intended to test the validity of the method for all station types and not just the stations used for European mapping applications.

3.4 Comparison and validation of the two approaches

To assess the ability of both MLR and EEVS to reproduce the $PM_{2.5}$ observations, the model training (regression coefficient determination, in the case of MLR) is carried out on the 2004-2006 data and compared to the 2007 data set, using the supplementary data sources selected in the initial analysis (Section 4). This will provide an independent validation of the two approaches in a comparable way. Assessment is carried out using several indicators, these are:

1. Root mean square error (RMSE) and also normalised RMSE (NRMSE). Normalisation is carried out using the mean of the observed values.
2. Correlation (r^2).
3. Fractional bias (FB)
4. Fraction of station predictions within $\pm$ 25% of the observed $PM_{2.5}$ (FAC25%).
5. Number of stations predictions that fulfil the European AQ Directive (2008/50/EC; EC, 2008) monitoring quality objective criteria (ECQO). This is similar to FAC25% above except that for concentrations smaller than the lower assessment threshold the allowable error is $\pm$ 100%.

In addition to the training and validation against 2007 $PM_{2.5}$ data, all the available data was also used to produce $PM_{2.5}$ estimates at all the PM_{10} stations for the year 2007. Though the results cannot be compared to measured$PM_{2.5}$ at these stations, the difference between the two approaches is assessed at PM_{10} stations where no $PM_{2.5}$ monitoring exists.

4. Exploration and selection of supplementary data

In this section, MLR and EEVS are applied to determine the most useful supplementary data sources in terms of the statistical parameters provided above. Because the EEVS and MLR applications have been carried out separately and at different times, the approaches and metrics used to determine the most useful supplementary data differ. In particular, the EEVS supplementary data assessment was carried using both traffic and background stations. For MLR only background stations were used to assess the usefulness of the supplementary data (though the selected supplementary data set is also used to calculate pseudo $PM_{2.5}$ for both traffic and background stations). The inclusion of traffic data may have an impact on the selection of supplementary data sources. In any case, the two approaches for calculating pseudo $PM_{2.5}$ cannot be compared at this point. In Section 5, a comparison will be carried out, based on the selected supplementary data sources, in a more controlled way.

4.1 MLR supplementary data assessment

4.1.1 Exploratory assessment of individual parameters using three years of data

In Tables 1 and 2, the results of the supplementary data exploration are shown. In Table 1, three years of data (2004 - 2006) are used to assess a number of meteorological, air quality and spatial parameters. Only three years are used in this case because annual meteorological data from ECMWF was not available at the time for the 2007 dataset. The values given in the table indicate the mean of the various parameters for the three years, with the exception of N (the number of data points) which is the sum of all data points used. The results of the assessment can be summarized as follows:

- Regression with only PM_{10} data already provides quite reasonable estimates of $PM_{2.5}$ (NRMSE = 22%).
- The inclusion of latitude and longitude introduces the largest improvement in the assessment parameters. Because of correlations between latitude and longitude and the other spatial parameters these are kept in all the subsequent tests.
- The best improvement in the regression is obtained using the climatological sunshine duration parameter (NRMSE = 19%).
- Annual meteorological data (meteo), rather than climatological data (clim), did not show any improvement on the climatological results.
- The inclusion of NO_2 or annual mean ozone appears to improve the results as well, however because these data are not available at every station the subset of stations used is different. As a result these cannot be directly compared to the other data.

Table 1. Assessment of supplementary data parameters with multiple linear regression using the years 2004-2006 (background stations only). Shown are the mean results of the individual regressions for each year. Green indicates the set of 'best' results for the respective parameters. Note that FB for MLR, applied to the same data set on which it is trained, will always be 0.

Parameters	N	MEAN	FB	RMSE	r^2	FAC25%	ECQO
PM10	320	17.2	0.00	3.80	0.805	0.748	0.857
PM10 + lat + lon	320	17.2	0.00	3.40	0.842	0.778	0.895
PM10 + lat + lon + pop	320	17.2	0.00	3.39	0.843	0.788	0.900
PM10 + lat + lon + sun(clim)	320	17.2	0.00	3.32	0.850	0.816	0.908
PM10 + lat + lon + sun(meteo)	320	17.2	0.00	3.33	0.849	0.808	0.901
PM10 + lat + lon + alt	320	17.2	0.00	3.38	0.844	0.771	0.890
PM10 + lat + lon + wind(clim)	320	17.2	0.00	3.39	0.843	0.781	0.902
PM10 + lat + lon + wind(meteo)	320	17.2	0.00	3.36	0.845	0.780	0.894
PM10 + lat + lon + temp(clim)	320	17.2	0.00	3.38	0.844	0.776	0.892
PM10 + lat + lon + temp(meteo)	320	17.2	0.00	3.38	0.845	0.785	0.895
PM10 + lat + lon + NO2	282	16.7	0.00	3.30	0.847	0.785	0.900
PM10 + lat + lon + O3	247	15.7	0.00	2.94	0.817	0.809	0.921
PM10 + lat + lon + Precip(clim)	320	17.2	0.00	3.38	0.844	0.781	0.892

4.1.2 Exploratory assessment of selected combinations of parameters using four years of data

In Table 2a, a number of the promising parameters determined above are included in a backwards assessment, this time covering all four years of data since the annual meteorological data has been excluded from this further assessment. As before, the given values represent the average of the four years of assessment. In Table 2b, the standard deviation of the four years is shown to indicate the variability in the statistical metrics from year to year. The results can be summarized as follows:

- The minimum RMSE is found for the MLR using the largest number of parameters. This is expected since it is the mean square that is minimized in the MLR.
- The maximum FAC25% and ECQO metrics are found for the case using latitude, longitude, sunshine duration and population density.
- The inclusion of altitude does not improve the results for FAC25% and ECQO.
- The variability in the RMSE reflects the variability in the mean concentrations. The variability in the Normalised RMSE is around 3% compared to the mean of 19%.
- The variability in the FAC25% and ECQO parameters from year to year is around 5%, which is roughly equivalent to 6 stations being inside, or outside, of the threshold borders.

Table 2a. Selected assessment of supplementary data parameters with MLR using the years 2004-2007. Shown are the mean results of the individual regressions for each year. Green indicates the set of 'best' results for the respective parameters. Background stations only are used.

Parameters	N	MEAN	FB	RMSE	R^2	FAC25%	ECQO
PM10 + lat + lon + alt + sun (clim) + wind (clim) + pop	452	16.8	0.00	3.12	0.856	0.801	0.907
PM10 + lat + lon + sun (clim) + wind(clim) + pop	452	16.7	0.00	3.14	0.854	0.804	0.913
PM10 + lat + lon + sun (clim) + pop	452	16.7	0.00	3.17	0.850	0.816	0.915
PM10 + lat + lon + sun (clim)	452	16.7	0.00	3.19	0.849	0.816	0.914
PM10 + lat + lon + pop + alt	452	16.8	0.00	3.25	0.843	0.797	0.909
PM10 + lat + lon + pop	452	16.8	0.00	3.26	0.841	0.807	0.914
PM10 + lat + lon	452	16.8	0.00	3.28	0.839	0.797	0.908
PM10	452	16.8	0.00	3.67	0.800	0.765	0.868

Table 2b. As in Table 2a but here the standard deviation of the four years is presented instead of the mean to indicate the variability from year to year.

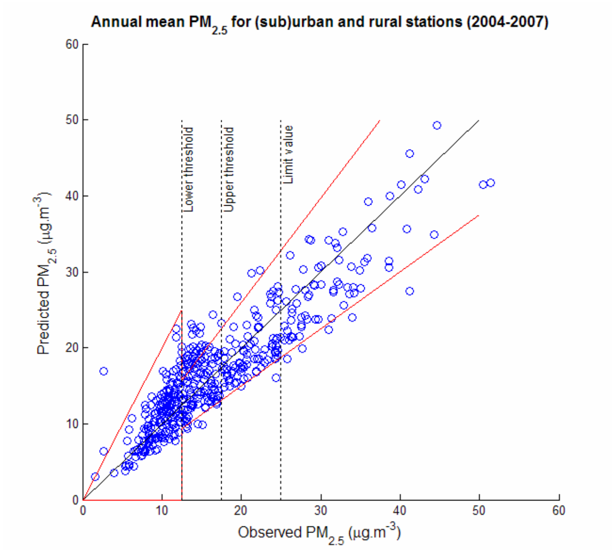
Parameters (standard deviation)	N	MEAN	FB	RMSE	R^2	FAC25%	ECQO
PM10 + lat + lon + alt + sun (clim) + wind (clim) + pop	26	1.6	0.00	0.36	0.027	0.043	0.041
PM10 + lat + lon + sun (clim) + wind(clim) + pop	26	1.6	0.00	0.37	0.027	0.054	0.046
PM10 + lat + lon + sun (clim) + pop	26	1.6	0.00	0.34	0.027	0.055	0.046
PM10 + lat + lon + sun (clim)	26	1.6	0.00	0.33	0.027	0.044	0.041
PM10 + lat + lon + pop + alt	26	1.6	0.00	0.31	0.028	0.067	0.058
PM10 + lat + lon + pop	26	1.6	0.00	0.32	0.029	0.054	0.052

PM10 + lat + lon	26	1.6	0.00	0.31	0.030	0.051	0.048
PM10	26	1.7	0.00	0.33	0.039	0.053	0.062

4.1.3 Assessment of the selected parameter set using all background stations

Using the above selected parameter set (PM10+lat+lon+sun(clim)+pop) the regression is recalculated using all the data simultaneously, i.e. all the data is combined in a single regression rather than by year. The result for background stations is shown in Figure 2 and in Table 3. Table 4 provides the resulting regression parameters when applied to both 3 and 4 years of data. Based on this analysis, 91.4% of all data falls within the region defined by the AQ Directive quality objectives, which would not be sufficient to fulfil the 95% requirement set out in the AQ Directive for monitoring.

For the background stations there is only limited improvement when including the most important supplementary data, compared to using just PM_{10} (Table 3). There is only a 3 – 5 % improvement in the FAC25% and the ECQO indicators and only a 10% reduction in RMSE.



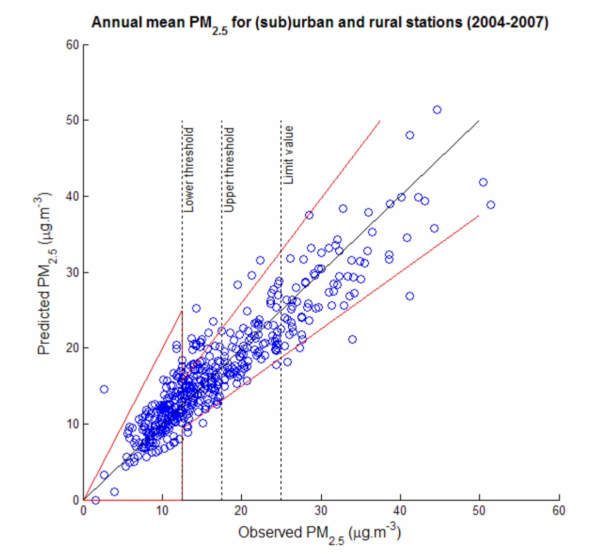


Figure 2. Results of the MLR using PM_{10} only (left) and using latitude, longitude, population density and sunshine duration (right) when applied to all years of data (background stations only). Red lines represent the boundaries of the European AQ Directive monitoring quality objective (ECQO).

Table 3. As in Tables 1 and 2 except the results are for all four years of data with a single regression applied (background stations only).

Parameters: BACKGROUND stations	**N**	**MEAN**	**FB**	**RMSE**	**R^2**	**FAC25%**	**ECQO**
PM10	452	16.9	0.00	3.66	0.811	0.765	0.881
PM10 + lat + lon + sun (clim) + pop	452	16.9	0.00	3.29	0.847	0.812	0.914

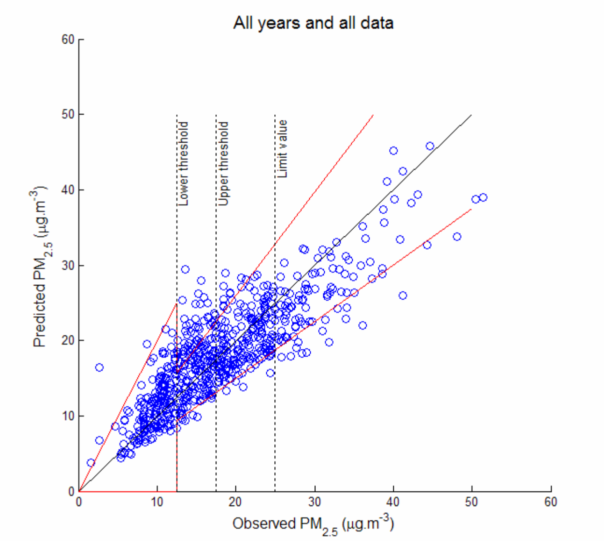
Table 4. Regression coefficients for the MLR regression for 2004 -2007 and 2004-2006 training sets (background stations only).

Regression coefficients	**a_0**	**a_1**	**a_2**	**a_3**	**a_4**	**a_5**
Units	**(μg/m^3)**	**PM_{10} (μg/m^3)**	**Latitude (°)**	**Longitude (°)**	**Sunshine duration (%)**	**Population (10x10km^2)**
Trained on 2004 – 2007	27.79	0.589	-0.451	0.219	-0.612	-1.79x10^{-6}
Trained on 2004 – 2006	29.85	0.585	-0.493	0.253	-0.159	-1.26x10^{-6}

4.1.4 Assessment of the selected parameter set using traffic and background stations

Even though the eventual application of the pseudo $PM_{2.5}$ calculations is for mapping purposes, which use only background stations, we also show here the results when including traffic stations. The reason for this is that the EEVS supplementary data source assessment was carried out on both traffic and background stations and a comparison can be made between the two approaches when both traffic and background stations are included.

In Figure 3 and Table 5 (as in Figure 2 and Table 3), the results of the MLR for all four years of data are shown including both traffic and background stations. More significant improvements on the basic PM_{10} regression are obtained here compared to when MLR is applied to just the background stations. There is a 6 – 9 % improvement in the FAC25% and the ECQO indicators and a 16% reduction in RMSE. In general, very similar results for the statistical metrics are obtained when traffic stations are included.



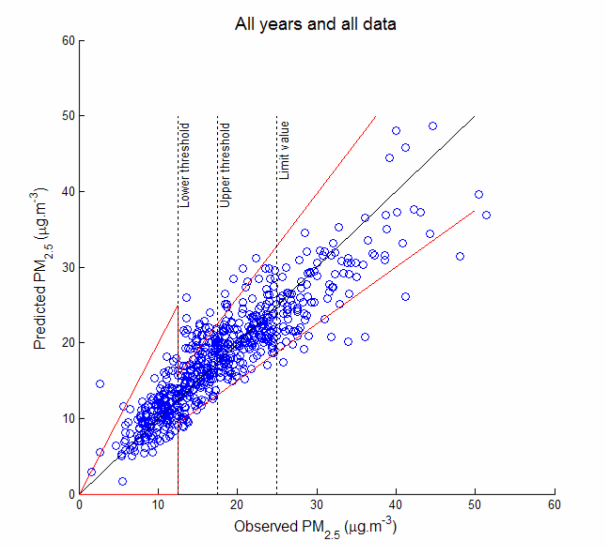


Figure 3. Results of the MLR using PM_{10} only (left) and using latitude, longitude, population density and sunshine duration (right) when applied to all years of data. Both traffic and background stations are included. Red lines represent the boundaries of the European AQ Directive monitoring quality objective (ECQO).

Table 5. As in Table 3 except the results are for all both background and traffic stations.

Parameters: TRAFFIC and BACKGROUND stations	**N**	**MEAN**	**FB**	**RMSE**	**R^2**	**FAC25%**	**ECQO**
PM10	642	18.0	0.00	4.23	0.722	0.729	0.827
PM10 + lat + lon + sun (clim) + pop	642	18.0	0.00	3.53	0.805	0.821	0.893

4.2 EEVS supplementary data assessment

For the EEVS assessment Monte Carlo re-sampling methods were used with all four years of data (2004-2007). This means that the EEVS was trained on 75% of the station data and the results were compared to the remaining 25% for all four years of data. The random selection of the training and validation data sets was carried out 50 times. The predicted $PM_{2.5}$ was compared at the 'left out' stations for the four years and statistics based on FAC25% and ECQO were built up.

The following columns are indicated in the assessment in Table 6:

1. **Average fraction of wrong estimates (%):** this is the percentage of cross-validated stations giving an estimate outside the specified tolerance level.
2. **Number of stations with at least one wrong estimate** (%): since for each station we can have from 0 to 4 years of data, the results in this column tell how many stations had at least one of those years wrongly estimated together with the percentage in brackets computed over the total number of stations effectively considered for that parameters set.
3. **Number of stations with wrong estimates greater than or equal to half the total test measurements (%):** similar to above but showing the number and percentage of stations where the number of wrong estimates is greater than or equal to the number of correct estimates.

Table 6. Showing the results of the EEVS assessment of supplementary parameters. Green indicates the set of 'best' results for the respective parameters.

Model number and parameters set - 4-year data (2004-2007) - All parameters are annual means - 352 total stations (French Guiney and Iceland stations removed)		**Results** (50 cross-validations, 75% stations for training)						
		Number of stations with at least one complete measurement	Average fraction of wrong estimates (%)		Number of stations with at least one wrong estimate (%)		Number of stations with wrong estimates greater than or equal to half total test measurements (%)	
			100 x (1-ECQO)	100 x (1-FAC25%)	100 x (1-ECQO)	100 x (1-FAC25%)	100 x (1-ECQO)	100 x (1-FAC25%)
A	**1 parameter:** PM10	246	11.35	27.52	**62** (25.2%)	**123** (50%)	**29** (11.8%)	**74** (30.1%)
B	**2 parameters:** PM10, NO2	224	9.70	27.21	**55** (24.5%)	**118** (52.6%)	**25** (11.1%)	**64** (28.5%)
C	**2 parameters:** PM10, O3	157	8.81	25.14	**34** (21.6%)	**84** (53.5%)	**14** (8.9%)	**40** (25.4%)
D	**3 parameters:** PM10, NO2, O3	155	8.72	25.89	**44** (28.4%)	**90** (58.1%)	**13** (8.4%)	**38** (24.5%)
E	**5 parameters:** PM10, Lat, Lon, Alt, Pop	246	6.76	16.18	**53** (21.5%)	**122** (49.5%)	**22** (8.9%)	**48** (19.5%)
F	**6 parameters:** PM10, NO2, Lat, Lon, Alt, Pop	224	6.79	17.61	**57** (25.4%)	**115** (51.3%)	**18** (8.1%)	**43** (19.2%)
G	**6 parameters:** PM10, O3, Lat, Lon, Alt, Pop	157	6.53	17.56	**39** (24.8%)	**85** (54.1%)	**9** (5.7%)	**28** (17.8%)
H	**7 parameters:** PM10, NO2, O3, Lat, Lon, Alt, Pop	155	7.24	18.54	**42** (27.1%)	**84** (54.1%)	**10**(6.5%)	**31** (20%)
I	**8 parameters:** PM10, Lat, Lon, Alt, Pop, sun, Temp, RH (meteo)	244	8.79	19.91	**60** (24.6%)	**124** (50.8%)	**21** (8.6%)	**46** (18.8%)
J	**8 parameters:** PM10, Lat, Lon, Alt, Pop, Sun, Temp, RH (clim)	246	7.64	17.30	**66** (26.8%)	**132** (53.6%)	**21** (8.5%)	**45** (18.3%)

K	**10 parameters:** PM10, NO2, O3, Lat, Lon, Alt, Pop, Sun, Temp, RH (meteo)	154	7.17	18.54	**37** (24.1%)	**81** (52.6%)	**14** (9.1%)	**31** (20.1%)
L	**10 parameters:** PM10, NO2, O3, Lat, Lon, Alt, Pop, Sun, Temp, RH (clim)	155	6.65	18.31	**40** (25.8%)	**89** (57.4%)	**10** (6.5%)	**31** (20%)
M	**7 parameters:** PM10, NO2, Lat, Lon, Pop, Wind, Sun(clim)	224	7.20	17.49	**58** (25.8%)	**119** (53.1%)	**19** (8.5%)	**39** (17.4%)

The following comments and conclusions can be made:

- Adding other air quality compounds to PM_{10} improves the estimation (A → B, C, D and E → G, H with the exception of E → F).
- The effect of inclusion of NO_2 is generally detrimental (B vs. C, F vs. G, E → F and G → H).
- Adding O_3 generally improves the estimation (A → C, E → G). However, O_3 is measured in less than half stations (157/352) and is thus not so useful for the application.
- Adding the geographical parameters generally improves the estimation (A → E, B → F, C → G and D → H) but is detrimental if coupled with NO_2 (E → F and G → H).
- Adding the meteorological parameters is generally detrimental (E → I, J and H → K, L).
- The climatological parameters have a better effect than the annual meteorological ones (I vs. J and K vs. L).
- In general, G and E seem the best parameter choices for developing a general model for Europe, i.e. one or two compounds and only the geographical parameters and populaion.
- Based on the simplicity of E (**5 parameters:** PM10, Lat, Lon, Alt, Pop) this was chosen for further application. This choice shows a 5 – 10% improvement on the FAC25% and the ECQO indicators, better than that found in the MLR.

4.3 Discussion concerning the two supplementary data choices

The eventual choice of supplementary data sources independently arrived at using the two approaches are quite similar. Both include PM_{10}, latitude, longitude and population density. The difference is the climatological sunshine duration was chosen in the MLR applications whilst altitude was chosen in the EEVS application. This may reflect more on the methodology used for selection. In MLR, individual geographic and meteorological parameters were tested, whereas in the EEVS selection method these were tested in blocks, e.g. all climatological data and all geographical data. The improvement from altitude was not tested. Nonetheless, statistically there is only a slight difference between the inclusion, or not, of many of the supplementary data sources.

The only element that is in some way comparable between the two approaches are the FAC25% and ECQO values provided by the MLR (when including traffic, Table 5) and the ‘Average fraction of wrong estimates %’ provided by the EEVS, for the selected supplementary data sources. In that case, we find EEVS to provide superior results compared to MLR. For FAC25% this is 0.838 for EEVS compared to 0.821 for MLR and for ECQO this is 0.932 compared to 0.893. Obviously, the validation methods (one using re-sampling cross-validation and the other all the data) are not strictly comparable, but this does provide an indication.

In addition, it is worth noting that the improvement in the metrics for FAC25% and ECQO when including supplementary data, when compared to just using PM_{10}, is similar for both the EEVS (5 – 10 %) and the MLR (6 – 9 %) when including traffic data. In the following section, a stricter analysis will be carried out based on the supplementary data chosen in this section, that will allow a more direct and honest comparison of the results.

5. Comparison of the MLR and EEVS approaches using the selected supplementary data source

In this section, MLR and EEVS are compared directly using the choice of supplementary data determined in the previous section. To do this two comparisons are made:

1. The two separate approaches are trained on 2004 – 2006 data and validated against 2007 $PM_{2.5}$data.
2. The two separate approaches are trained on 2004 – 2007 data and the predicted $PM_{2.5}$ from each approach is compared at the unused 2007 PM_{10} station data.

The first provides an independent validation set, whilst the second provides a comparison of the predicted results. This is intended to assess the difference in the results when all the data is used. In this case only urban/suburban and rural background stations are used in the comparison.

5.1 Validation on 2007 data, trained on 2004 – 2006 data

Figure 4 and Table 7 show the results of the validation of the approaches against observed $PM_{2.5}$ concentrations for 2007, trained on 2004 – 2006 data. The result can be summarized as follows:

- EEVS gives an improved relative RMSE (17%) compared to MLR (19%).
- There is no difference in correlation.
- FAC25% is slightly higher for MLR.
- ECQO is slightly higher for EEVS (note that with 132 stations, the fraction represented by one station, in or out of the boundaries, is equivalent to 0.008).

The most significant difference is the RMSE, which is the result of improved predictions by EEVS for the higher $PM_{2.5}$ concentrations.

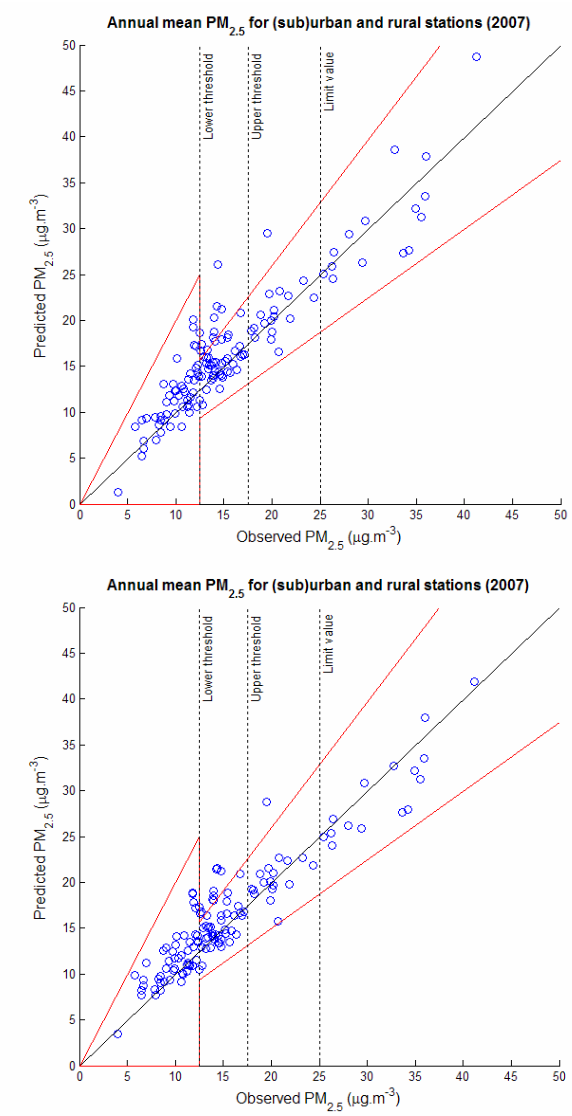


Figure 4. Results of the MLR (left) and EEVS (right) predicted annual mean background station $PM_{2.5}$ concentrations for 2007, trained on 2004 – 2006 data (background stations only). Red lines represent the boundaries of the European AQ Directive monitoring quality objective (ECQO).

Table 7. Validation of the MLR and EEVS approaches on 2007 background station data.

Trained on 2004 - 2006	N	MEAN	FB	RMSE	r^2	FAC25%	ECQO
MLR (PM10+lat+lon+sun (clim)+pop)	132	16.68	0.075	3.10	0.859	0.811	0.909
EEVS (PM10+lat+lon+alt+pop)	132	16.40	0.059	2.72	0.859	0.795	0.917

5.2 Comparison of predicted $PM_{2.5}$ at PM_{10} stations for 2007

To compare the two approaches on a larger dataset both MLR and EEVS, trained on 2004-2007 data, are used to predict $PM_{2.5}$ concentrations at all the available background PM_{10} stations for 2007 (1161 stations). The results are shown in Figure 5. There is a tendency for EEVS to predict slightly lower $PM_{2.5}$ concentrations at higher values and for MLR to predict slightly lower $PM_{2.5}$concentrations at lower values. The differences are never greater than around 5 $\mu g/m^3$ at any of the stations with a

RMSE = 0.92 μg/m^3 or NRMSE = 5.5%. Interestingly the extreme differences occur at either the high or the low concentrations levels and the differences are smaller at the limit value of 25 μg/m^3.

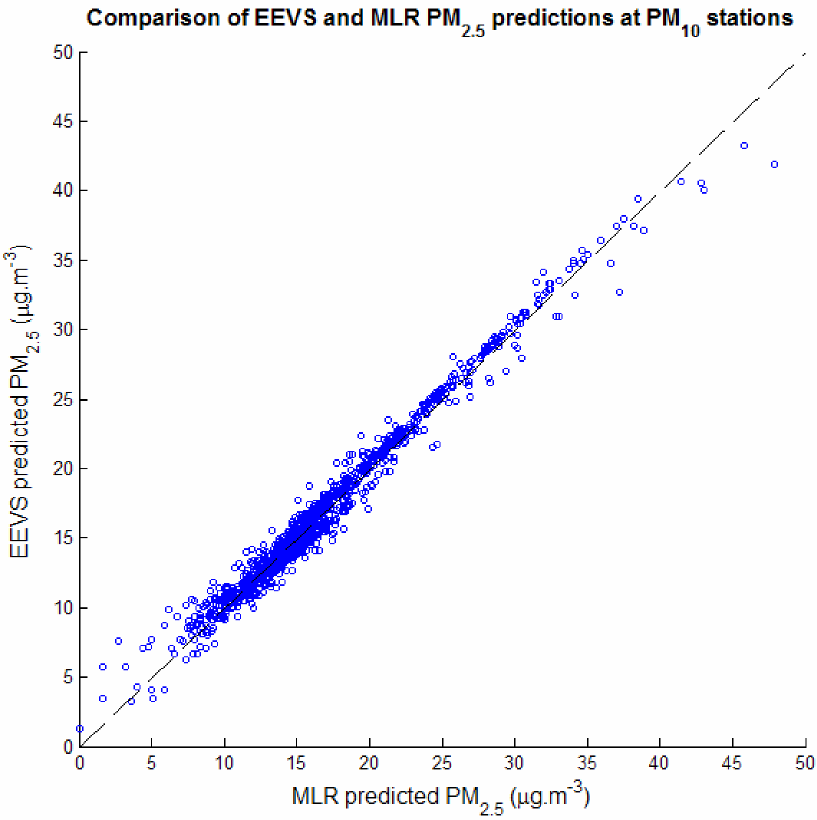


*Figure 5. Comparison of the predicted annual mean background station $PM_{2.5}$ concentrations for 2007 using MLR and EEVS approaches at all PM_{10} stations where $PM_{2.5}$ is **not** measured. Both approaches are trained on 2004 – 2007 data.*

6. Discussion and conclusion

In this report, two approaches for creating 'pseudo' $PM_{2.5}$ concentrations are described and evaluated. These are Empirical Ensemble based Virtual Sensing (EEVS) (based on an ensemble of Neural Networks) and Multiple Linear Regression (MLR). The aim of the pseudo $PM_{2.5}$ data is to improve the spatial coverage of monitoring data for the application of $PM_{2.5}$ mapping in Europe.

A number of supplementary data sources, in addition to PM_{10}, were assessed for their usefulness in improving the estimate of pseudo $PM_{2.5}$ concentrations. It was found for the EEVS approach that the most useful supplementary data sources were latitude, longitude, population density and altitude. For MLR, the selected supplementary data sources were latitude, longitude, population density and sunshine duration. Supplementary air quality data was also assessed but since these are not available at a large number of PM_{10} stations these were not used in further analysis.

The two approaches were validated against 2007 $PM_{2.5}$ concentrations having been trained on 2004 – 2006 data. It was found that EEVS provided a slightly reduced RMSE but the two methods were similar for the other metrics assessed. It is worth noting that though the 2007 data was independent of the training set most of the stations using in the training and validation are the same stations. In this regard, they are likely to show the same characteristics from year to year and as such are not completely independent. The alternative to this method is the random sampling carried out for the EEVS in Section 4.2. The choice of the validation method employed here was to ensure comparability of the two separately conducted assessments.

Neither approach was found to fulfil the monitoring quality objectives as laid out in the European air quality directive (2008/50/EC; EC, 2008), but both would fulfil the requirements as laid out for modelling and indicative measurements in the AQ Directive. The assessment does not take into account any uncertainty in the PM_{10} or $PM_{2.5}$ monitoring data used for the assessment.

The two approaches were compared by predicting the $PM_{2.5}$ concentrations at all available suburban/urban and rural background stations in AirBase for 2007. Slight differences were found in the extreme values (low and high concentrations) of up to 5 $\mu g/m^3$.

The aim of this study was to determine if pseudo $PM_{2.5}$ data was of sufficient quality for it to be employed in the mapping activities of ETC/ACC for the EEA. Though the resulting pseudo measurements do not fulfil the official AQ Directive requirements for monitoring data, the results are considered useful for mapping purposes. The reason for this is that the relative RMSE of the pseudo $PM_{2.5}$ concentrations were found to be from 16% - 19%, dependent on the approach used. This can be compared to the estimated uncertainty in the maps currently produced for PM_{10} which is around 25% (Horálek et al., 2010) and we expect the uncertainty in any $PM_{2.5}$ maps to be similar. Any inclusion of monitoring data with lower uncertainty should lead to improved mapping. However, in making these maps it will be necessary to take into account the uncertainty of the pseudo $PM_{2.5}$ measurements as these will be larger than the direct monitoring data. Currently uncertainty in the monitoring data is not explicitly included in the maps, but is indicated by the residual kriging nugget value, that represents both monitoring and spatial representativeness uncertainty.

There remains the choice of the approach that can be applied for future mapping applications. From a statistical assessment perspective the EEVS method, using an ensemble of Neural Networks, gives a slightly reduced RMSE compared to MLR. For all other metrics assessed there was no significant difference. For this reason EEVS should be recommended over MLR. However, from a practical application point of view the EEVS methodology is completely opaque, as there is no information available on either the Neural Network models used in the application or the parameters they apply after training. This means that the approach is not reproducible. In addition, a patent application is currently pending for the application of EEVS for pseudo $PM_{2.5}$ calculations. This makes its use by third parties unclear. Unless this situation changes in the future, it seems adequate to apply MLR for future $PM_{2.5}$ mapping for Europe.

References

AirBase (2010), European air quality database, http://airbase.eionet.europa.eu/

de Leeuw, F. and J. Horálek (2009) Assessment of the health impacts of exposure to PM2.5 at a European level ETC/ACC Technical Paper 2009/1. http://air-climate.eionet.europa.eu/reports/ETCACC_TP_2009_1_European_PM2.5_HIA de Smet, P., Horálek, J., Conková, M., Kurfürst, de Leeuw, F. and Denby, B. (2010). European air quality maps of ozone and PM10 for 2007 and their uncertainty analysis. ETC/ACC Technical paper 2009/9. http://air-climate.eionet.europa.eu/reports/ETCACC_TP_2009_9_spatAQmaps_2007

EC (2008) Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. Official Journal L 152, 11.6.2008, p. 1–44

Fiala, J., B. Denby, J. Horálek, P. Kurfürst, P. de Smet, F. de Leeuw and A.Lükewille (2009). Spatial assessment of PM10 and ozone concentrations in Europe (2005). Technical report No 1/2009. http://www.eea.europa.eu/publications/spatial-assessment-of-PM10-and-ozone-concentrations-in-europe-2005-1

Horálek, J.; Denby, B.; de Smet, P. A. M.; de Leeuw, F. A. A. M.; Kurfürst, P.; Swart, R.; van Noije, T., (2007). Spatial mapping of air quality for European scale assessment. ETC/ACC Technical paper No 2006/6. http://air-climate.eionet.europa.eu/docs/ETCACC_TechnPaper_2006_6_Spat_AQ.pdf

Horálek, J., de Smet, P. A. M., de Leeuw, F. A. A. M., Coňková M., Denby, B., Kurfürst, P. (2010) Methodological improvements on interpolating European air quality maps.ETC/ACC Technical Paper 2009/16.http://air-climate.eionet.europa.eu/reports/ETCACC_TP_2009_16_Improv_SpatAQmapping

New, M., Lister, D., Hulme, M., Makin, I. (2002). A high-resolution data set of surface climate over global land areas. Climate Research, Vol. 21, pp. 1-25. http://www.cru.uea.ac.uk/cru/data/hrg.htm/; www.cru.uea.ac.uk/cru/data/TH

Roverso, D. (2009) Empirical Ensemble-Based Virtual Sensing – A Novel Approach to Oil-in-Water Monitoring. Oil-in-Water Monitoring Workshop 24th September 2009. http://www.ife.no/publications/2009/mto/publication.2009-09-30.4024994703/view?set_language=no&cl=no

Rumelhart, D.E. and McClelland, J.L., 1986, Parallel Distributed Processing, vol. 1. MIT Press, Cambridge, MA.