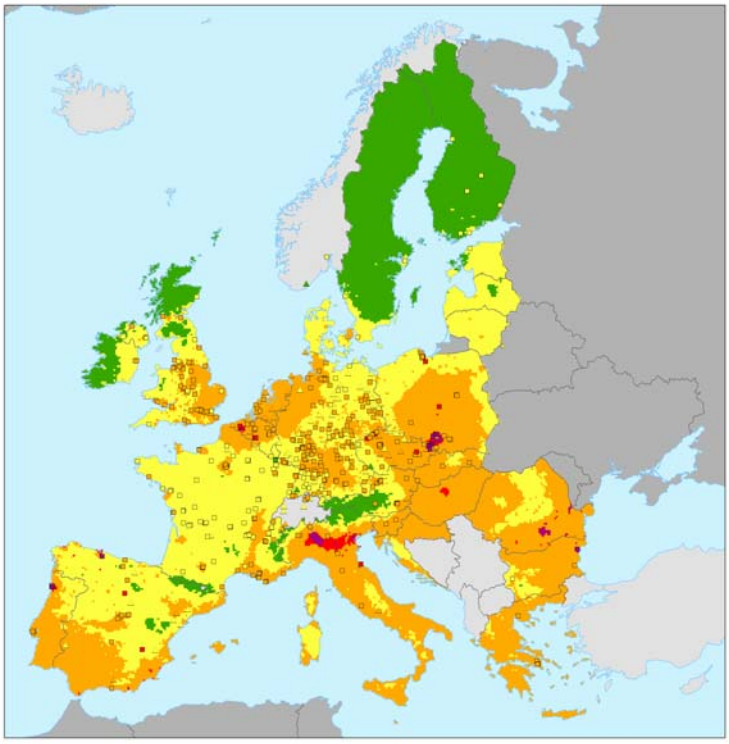
Interpolation and assimilation methods for European scale air quality assessment and mapping

Part I: Review and recommendations



ETC/ACC Technical Paper 2005/7
December 2005

Final draft

Bruce Denby, Jan Horálek, Sam Erik Walker, Kryštof Eben and Jaroslav Fiala

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 Environmental Agency
RIVM UBA-B UBA-V IIASA NILU AEAT AUTh CHMI DNMI NTUA ÖKO IEP TNO UEA

Front page picture

Combined rural and urban concentration map of the annual average PM_{10} for the year 2001 based on spatial interpolated concentration fields and measured values at the indicated measuring points (units in $\mu g.m^{-3}$.days).
(The figure is taken from ETC/ACC Technical Paper 2005/8, Annex Figure 6A. It represents the combined map as created by merging the interpolated rural map (combination of measured values with EMEP model, altitude and sunshine duration, using linear regression and ordinary kriging of residuals) and the interpolated urban map (using interpolation of urban increment Delta by ordinary kriging). Countries with interpolation based on additional data only: BG, GR, HR, HU, RO. Countries with missing population density information and therefore excluded from the mapping calculations: AD, AL, BA, CH, CS, CY, IS, LI, MK, NO, TR.)

DISCLAIMER

Interpolation and assimilation methods for European scale air quality assessment and mapping

Part I: Review and recommendations

Contributing authors
Bruce Denby and Sam Erik Walker, Norwegian Institute of Air Research (NILU), Oslo
Jan Horálek, Czech Hydrometeorological Institute, Praha
Kryštof Eben, Academy of Sciences of the Czech Republic, Praha

(EEA project manager: Jaroslav Fiala)

Final Draft

December 2005

TABLE OF CONTENTS

Preamble

The ETC/ACC manages the air quality database AirBase for the EEA, which contains air quality data reported through the Exchange of Information Decision. AirBase covers thousands of monitoring stations across Europe, but the density of the network varies across regions. For both public information (e.g., the EEA 'In Your Neighbourhood' project) and for air quality impact assessments, the situation between stations must also be known. This paper is intended to examine and recommend methods that can be used to interpolate between stations, to arrive at a spatial representation of air quality information.

1 Introduction

The important task of an air quality information system is to offer the most complete information about the air quality in a given region. Information concerning the regional distribution of air pollutants can be obtained either through modelling, which gives good spatial coverage but poor quantitative certainty, or by measurements, which have limited spatial coverage but are assumed to be more certain. Traditionally assessment is based on monitoring data but when information is required between stations, some form of interpolation is required.

It is clear that the number of measuring stations available throughout Europe is limited. E.g. the EMEP network consists of just over 100 stations, dependent on pollutant, whilst the AIRBASE network has just over 440 rural ozone and 205 rural PM_{10} stations registered. These stations are assumed to be representative of a region with a radius of approximately 50 km, but the average distance between stations is much larger than this. Any spatially distributed analysis reliant on air quality data, e.g. ecosystem or health studies, will naturally require information between the stations. This leads directly to the need for efficient and accurate interpolation methods.

> *The overall aim of this report is to recommend suitable methodologies for the interpolation of regional scale monitoring data for the purposes of health and ecosystem exposure analysis as well as for public information on a European wide scale.*

There are a large variety of interpolation methods, developed for many varied applications, that use individual point datasets and create spatially distributed fields from these. Methods such as kriging, bilinear interpolation and inverse distance weighting can, and have, been used for interpolating air quality data. Methods that only use observational data fall into two categories, geostatistical and deterministic. The difference between these two is that geostatistical methods explicitly utilize the spatial statistical structure of the data.

In addition to observed air quality it is also possible to introduce supplementary data, with better spatial coverage, to improve the interpolation. Typically such supplementary data should be either representative of the data to be interpolated, e.g. the use of PM10 to represent the distribution of $PM_{2.5}$, or should reflect correlation between the physical processes that lead to the spatial distribution of the data to be interpolated, e.g. elevation and precipitation. Whatever the supplementary data represents it must have a significant correlation with the data to be interpolated for it to be applicable. There is currently no standard method for interpolating monitoring data for air quality assessment on the regional, or any, scale.

Models can also provide complex spatial information of a region as they reflect the physical and chemical processes involved. However, a large number of uncertainties in models, such as emissions and model parameterisations, mean that they require validation against monitoring. Different models used for regional scale assessment do not give the same results for the various pollutants (van Loon et al, 2004). In general the results are considered less certain than the measured data, when the models are evaluated at measurement points. The combination of model and monitoring data in an optimised way has the potential to decrease uncertainty in the resulting spatial fields.

The final draft of the EU guidance (EC working group, 2000) distinguishes among the eight different ways of combining measurements with models, figure 1.1. Of these possibilities the most extreme forms, (a) and (h), cannot be used for air quality assessment as isolated measurements without any further interpretation give incomplete information and non validated models cannot be considered reliable. Most modelling activities rely on validation, (f) and (g), to lend credibility to model results.

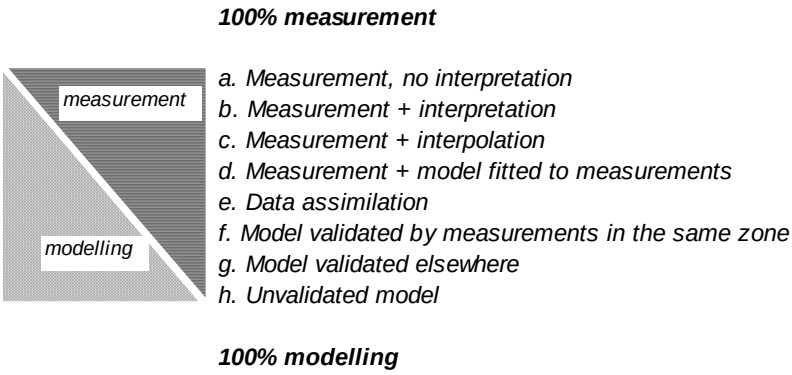


Figure 1.1. Degrees of combining measurements and models, taken from EC working group (2000)

This review will cover the methods (c) Measurement and interpolation; (d) measurement and model fitted to measurement; and (e) data assimilation. Most emphasis will be on the first two areas, as these are the most accessible. In this review we distinguish between 4 different methodologies for combining monitoring and modelling, that lie within the methods (c-e) given above.

These are:

- Chapter 2. Interpolation methods using monitoring data only

  - *E.g. IDW interpolation, kriging, etc.*

- Chapter 3. Interpolation methods using monitoring and other supplementary data

  - *Use of emission data, land use, altitude, population, climate etc. to improve interpolation*

- Chapter 4. Interpolation methods using monitoring and models

  - *Use of chemical transport model output concentration fields to improve interpolation*

- Chapter 5. Data assimilation methods

  - *Inclusion of observations in the model prognosis*

In addition to these methodologies one special area of interest, how to include smaller scale features such as urban conurbations in regional scale interpolation, will be discussed, chapter 6. This is of particular importance for population exposure calculations based on interpolated data. Chapter 7 will specify the best method for assessing the quality of the interpolation methodologies and then finally, in chapter 8, recommendations on the best interpolation method(s) will be made.

The basis for the final recommendations will be built upon both the available literature but also to a large extent on studies carried out in Part II of this report. In Part II the testing, assessment and application of a number of selected interpolation methods is described. The indicators chosen for the studies are indicative of both exposure and eco-system effects, these being the ozone indicators AOT40, SOMO0 and SOMO35 as well as the indicators for PM10, annual mean and 36'th highest daily average concentrations. The studies are carried out within a common framework ensuring comparability between different methods.

2 Interpolation methods using monitoring data only

These methods are based solely on measured data (without any additional information concerning the spatial form of the data) collected at observational stations. Similar problems which arise in the mapping of air pollution concentrations, also appear in meteorology and climatology - especially in the mapping of precipitation and temperature fields. It can be useful to draw inspiration from these areas for use in air quality mapping.

From a historical perspective, in the beginning of the 20th century climatology used the so-called Thiessen method for the estimation of the long-term means in the regions with no measurements. Since the mid-sixties geostatistical methods, especially different types of kriging, have been applied. These methods are based either on the time series of measured quantities (optimum interpolation), or they consider only one realization of the field of measured quantity (kriging). Both optimum interpolation and the different types of kriging are used in air pollution mapping.

Apart from the geostatistical methods, which utilize knowledge of the statistical structure of the concentration field, several other simpler methods can be used, such as the minimum curvature method, different types of splines or the IDW method (interpolation using inverse distance weighting).

2.1 Simple interpolation methods

2.1.1 Thiessen polygon (nearest neighbour method)

One of the simplest methods is the Thiessen polygon or the nearest neighbour method. This method defines individual areas of influence around each of a set of points. These areas correspond to the area around a point that is closest to that point. They are mathematically defined by the perpendicular bisectors of the lines between all the points. Having defined such a polygon the area within is given the same value as the point it contains.

2.1.2 Bilinear interpolation (triangulation)

Given a random field of measurement points it is possible to linearly interpolate between points to produce a concentration field. This can be done by generating Delaunay triangles, triangles that define the closest 3 points to each other. From these points the equation of a plane can be derived and used to determine concentrations within the triangle, or alternatively to directly produce contour maps, as is applied in a number of GIS systems (Burrough, 1986). This form of interpolation introduces discontinuities in the concentration gradients at the edges of the interpolating triangles. Its simplicity though makes it attractive for simple applications.

Some other simple interpolation methods can be found in SURFER (1994).

2.2 Inverse distance weighting (IDW)

The Inverse Distance Weighting (IDW) method improves on the Thiessen polygon method by weighting the linear combination of values measured at all points, where the weight is an inverse function of the distance according to the equation:

$$\hat{Z}(\mathrm{s}_0) = \frac{\sum_{i=1}^{n} \frac{Z(\mathrm{s}_i)}{d_{0i}{}^{\beta}}}{\sum_{i=1}^{n} \frac{1}{d_{0i}{}^{\beta}}} \qquad (2.1)$$

where $\hat{Z}(\boldsymbol{s}_0)$ is the interpolated value of the concentration in the point s_0,

$Z(s_i)$ is the measured value of the concentration in the *i*-th point,

n is the number of surrounding stations from which the interpolation is computed,

d_{0i} is the distance between the interpolated point and the *i*-th station,

β is the weighting power.

In practice the weighting power is mostly equal to two, then the name of the method is "inverse distance squared". This method is used e.g. by Dittmann et al. (1999). According to the EC working group (2000) the method IDW (when $\beta = 4$) has been used for constructing air pollution maps in Belgium.

Nalder and Wein (1998) introduce a developed IDW method, which additionally tries to remove the influence of gradient has been termed **gradient-plus-inverse distance squared method**:

$$Z(s_0) = \frac{\sum_{i=1}^{n} \frac{Z(s_i) + (x_0 - x_i).c_x + (y_0 - y_i).c_y + (v_0 - v_i).c_v}{{d_{0i}}^2}}{\sum_{i=1}^{n} \frac{1}{{d_{0i}}^2}} \qquad (2.2)$$

where x_i and y_i are the relevant geographical coordinates,

v_i is the altitude,

x_0 , y_0 and v_0 are the coordinates of the interpolation point

c_x , c_y and c_v are the regression coefficients for *x*, *y* and *v*.

In the Czech Republic the so-called **modified version of IDW**, Fiala et al. (2000), is routinely used. For each station there a radius of representativeness is set and eventually its 'weight' for air quality assessment. This allows both urban and rural stations to be interpolated together (different types of stations have different representativeness). The interpolation is computed by:

$$\hat{Z}(s_0) = \frac{\sum_{i=1;d_{0i} \le r_i}^{n} \frac{w_i Z(s_i)}{d_{0i}}}{\sum_{i=1}^{n} \frac{w_i}{d_{0i}}}, \qquad (2.3)$$

where $s_1,\ldots,s_n$ is the set of stations nearest to s_0, which satisfy the condition $d_{0i} < r_i$

d_{0i} is the distance between the interpolated point and the *i*-th point of measurement,

r_i is the radius of representativeness of the *i*-th station (individually assessed for each station),

n is the given number of stations, from which the interpolation is computed,

w_i is the weight of *i*-th station (expressing its credibility).

Falke and Husar (1998a) also extended the IDW method to account for groups of monitoring stations associated with urban areas. They suggest a method for **clustering** such urban groups by reducing the area of influence of clusters of stations so that they do not adversely influence regions far from the city, giving extra weight to rural stations that have an area of representativeness that is larger.

They suggest a declustering weight *CW* given by

$$CW_{ij} = \frac{D_{ij} + \sum_{k=1}^{m-1} d_{jk}{}^{p}}{D_{ij} \times m} \qquad (2.4)$$

where i is the interpolated point index,

j is the monitoring site index,

m is the number of sites within D_{ij} of the monitoring site (including the monitoring site)

D_{ij} is the distance between the monitoring site and the interpolation point

k is an index of the $(n-1)$ neighbouring sites around the monitoring site

d_{jk} are the distances between the monitoring site and its neighbouring sites,

p is the weighting power.

This declustering weight is then multiplied with the normal individual IDW to produce a total weight. The resulting interpolation then becomes

$$\hat{Z}(s_j) = \frac{\sum_{i=1}^{n} CW_{ij} \frac{Z(s_i)}{d_{ij}{}^{\beta}}}{\sum_{i=1}^{n} \frac{1}{d_{ij}{}^{\beta}}} \qquad (2.5)$$

They applied this methodology for the interpolation of ozone maps in the USA. They state that by using this methodology the decluster weighting will adjust the total weighting so that a cluster of stations 50 km from the estimation point will have approximately the same weight as a single station 50 km away. By using cross-validation they reported an improvement in the resulting fields, particularly in rural areas.

2.3 Methods of radial basis functions (RBF)

Radial basis functions methods interpolate the measured values while minimizing the total curvature of the surface, so that the resultant surface is continuous and soft. Identically, as in the case of IDW, the interpolation goes exactly through the measured values. The interpolation is described by

$$\hat{Z}(s_0) = \sum_{i=1}^{n} w_i . \Phi(d_{0i}) + w_{n+1} \qquad (2.6)$$

where $\Phi(x)$ is a concrete RBF function,
d_{0i} is a distance of interpolated point from i-th station,
$w_1, \ldots, w_{n+1}$ are the weight parameters,
n is a number of surrounding stations from which the interpolation is computed.

From the formal aspect the calculation of the radial basic functions and the estimation of their parameters is rather complicated, while from the computational aspect it is quite simple and fast. The parameters $w_1, \ldots, w_{n+1}$ are obtained from the system of equations given by

$$\sum_{j=1}^{n} w_j \Phi(d_{ij}) + w_{n+1} = Z(s_i), \qquad i=1,\ldots, n, \qquad (2.7)$$

$$\sum_{j=1}^{n} w_j = - w_{n+1}$$

Some of the radial basis functions are: completely regularized spline, spline with tension, thin-plate spline, multiquadric function or quadric beta-splines. A more detailed description of some radial basic functions are given in Johnston et al. (2001). These methods do not use knowledge of the covariance structure of data, but they include this information implicitly. It is used e.g. by Coyle et al. (2002).

2.4 Beier's intuitive method

Beier and Doppelfeld (1992, 1999) introduced their own methodology of spatial interpolation, which is based on an intuitive approach to the problem. The general relation for estimating the concentration $\hat{Z}(s_0)$ at the location s_0 is given by

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{N} w_{0i} \cdot Z(s_i)}{\sum_{i=1}^{N} w_{0i}} \tag{2.8}$$

The local weighting function w is defined as

$$w_{0i} = 1 - \frac{d_{0i}}{R(s_i)} \quad when \quad \frac{d_{0i}}{R(s_i)} < 1 \tag{2.9}$$

$$= 0 \quad when \quad \frac{d_{0i}}{R(s_i)} \geq 1$$

where d_{0i} is the distance between the point s_0 and the measuring station s_i,
$R(s_i)$ is the radius of influence assigned to the measuring station s_i and the relevant pollutant.

The radius of influence is assessed intuitively for the individual pollutants and types of stations (traffic, industrial, urban background and rural) - e.g. for NO_x it is assessed as 10 km for the urban background stations and 20 km for the rural stations.

Then the resultant interpolation is constructed dependent upon the number of stations around the point X in the following way:

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{N} w_{0i} \cdot Z(s_i)}{\sum_{i=1}^{N} w_{0i}} \qquad \text{if } d_{0i} \leq R(s_i) \text{ for more than one station } s_i ,$$

$$\hat{Z}(s_0) = Z(s_k) \cdot w_{0k} + \overline{Z}(R) \cdot (1 - w_{0k}) \quad \text{if } d_{0i} \leq R(s_i) \text{ for exactly one station } s_k ,$$

$$\hat{Z}(s_0) = \overline{Z}(R) \qquad \text{if } d_{0i} > R(s_i) \text{ for all measuring stations } s_i, \tag{2.10}$$

where $\overline{Z}(R) = \frac{1}{N} \sum_{i=1}^{N} Z(s_i)$ is the arithmetic mean of the rural stations in the whole region.

2.5 Geostatistical methods

These methods are based on knowledge of the statistical structure of a random field $Z(X)$ (where $X = X_1, X_2, \ldots, X_k$. are the points in an Euclidean space R^a). They respect the statistical relationship between the individual points of the field. A number of different interpolation methods can be used for spatial prediction where optimum interpolation and kriging belong to these basic methods. They both assume the

homogeneity of a random field and define an estimate of the unknown value of the field Z at the point x_0 as a linear combination of Z measured at the points $x_i = x_1, x_2,..., x_M$.

This is formulated as

$$\hat{Z}(x_0) = \sum_{i=1}^{M} \lambda_i Z(x_i) \tag{2.11}$$

where the accuracy of the estimation $\hat{Z}(x_0)$ is characterized by the mean squared error

$$Q^2_{\hat{Z}(x_0)} = E[Z(x_0) - \hat{Z}(x_0)]^2 \tag{2.12}$$

where E represents the mean operator. However, when calculating the weights, λ_i, in equation 2.11 optimum interpolation uses mostly the time series of the measurements, whilst kriging in its basic form improves only one value (considered as the realization of a random field). However, despite the different origin of these two methods, Cressie (1993) states the principles of optimum interpolation and kriging are identical.

2.5.1 Kriging and its variants

Kriging in its basic form is used when only spatial, but not temporal, data are taken into consideration. To describe the concentration field variability on the basis of measurements at N stations a semivariogram is constructed. This approach was introduced by a French statistician G. Matheron, Matheron (1963). He named this method of optimal spatial linear prediction after the South African mining engineer D.G. Krige, who applied this method to geostatistical problems in geology (Krige, 1951). A detailed description of this methodology is given by Cressie (1993).

The semivariogram $\gamma(h)$, resp. variogram $2\gamma(h)$, is a measure of the relation between pairs of measuring stations s_1 and s_2, under the condition

$$\mathrm{var}((Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2), \qquad \text{for all } s_1, s_2, \tag{2.13}$$

where *var* is the variance and h is the two-dimensional distance (which can be expressed by distance and direction). If the variogram is the same for all directions, i.e. $\gamma(h) = \gamma(|h|)$, then it is called isotropic.

The empirical semivariogram γ_v is calculated by the equation:

$$\gamma_v(\mathrm{h}) = \frac{1}{2n} \sum_{i=1}^{n} \{Z(\mathrm{s}_i) - Z(\mathrm{s}_i + \mathrm{h})\}^2 , \tag{2.14}$$

where $Z(\boldsymbol{s}_i)$ and $Z(\boldsymbol{s}_i + \boldsymbol{h})$ are the measurements in the points $\boldsymbol{s}_i$ and $\boldsymbol{s}_i + \boldsymbol{h}$,
n is the number of distinct pairs of points, the distance of which is $\mathbf{h} \pm \boldsymbol{\delta}$,
and $\boldsymbol{\delta}$ is the tolerance.

The empirical semivariogram is fitted by an analytical function (model) - e.g. spherical, exponential, Gaussian. (The basic parameters of a semivariogram are called nugget, sill and range. When the fitted curve does not pass through the origin, it denotes the existence of the so-called nugget effect). The estimated semivariogram is consequently used in the interpolation. The parameters used to establish the semivariogram are shown in figure 2.1.

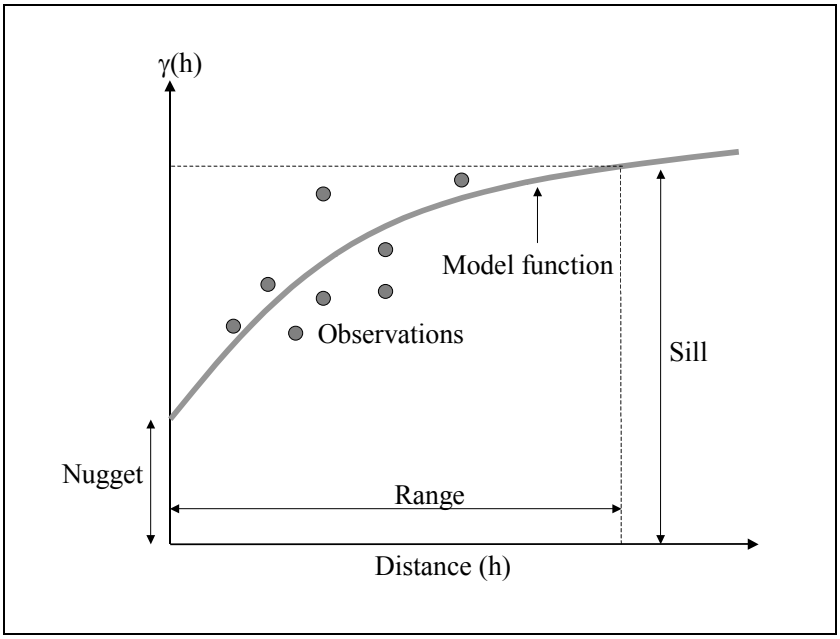


Figure 2.1. Diagram showing the important parameters that describe the semivariogram, γ(h), used in kriging.

Ordinary kriging performs spatial interpolation under two assumptions. These are:

the *model assumption*

$$Z(s) = \mu + \varepsilon(s), \qquad s \in D, \tag{2.15}$$

where μ represents the constant mean structure of the concentration field,
$\varepsilon(s)$ is a smooth variation plus measurement error (both zero-mean)
D is the examining area

and the *interpolator assumption*

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i), \quad \sum_{i=1}^{n} \lambda_i = 1, \tag{2.16}$$

where $\hat{Z}(\boldsymbol{s}_0)$ is the interpolated value of concentration in the point s_0
$Z(\mathrm{s}_i)$ is the measured value of concentration in the i-th point, $i=1,..,n$
n is the number of surrounding stations from which the interpolation is computed
$\lambda_1, \ldots, \lambda_n$ are the weights assumed at the basis of variogram

The weights λ_i are derived from the semivariogram in order to minimize the mean-square-error; n is the number of surrounding measuring sites from which the interpolation is computed. The explicit calculation is achieved by the system of equations

$$-\sum_{j=1}^{n} \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) + \gamma(\boldsymbol{s}_0 - \boldsymbol{s}_i) - m = 0, \qquad i=1,..., n,$$

$$\sum_{i=1}^{n} \lambda_i = 1. \tag{2.17}$$

By this system of equations the parameters $\lambda_1,\ldots, \lambda_n$ can be calculated (as well as m, the co-called Lagrange multiplier, that ensures $\sum_{i=1}^{n} \lambda_i = 1$). By their substitution into equation 2.16 we obtain the estimation of concentration $\hat{Z}(s_0)$.

Kriging is an often used standard method, which has been used e.g. in the materials of EMEP for constructing maps of background concentrations (Hjellbrekke, 2000). The area of Europe is in that case divided into nine regions, which are processed separately (some poorly representative stations are not taken into consideration). For each region the variogram is constructed and the interpolation using kriging is performed. Consequently the results from the regions are united into a complete map.

An application of kriging is presented by van Leeuwen et al. (1994). They also document the fitting of variograms. Kriging is used also by e.g. Dittman et al. (1999), Atkins and Lee (1995), Lefohn et al. (1988) and Casado et al. (1994).

To apply ordinary kriging it is necessary that some essential assumptions be fulfilled, especially stationarity. (i.e. the assumptions given in equations 2.13 and 2.15 are fulfilled). As Federov (1989) shows, this is often not true. (He therefore presents an alternative to kriging called the method of moving least squares).

The situation, where the spatial trend exists, is dealt with by **detrended kriging**, which is mentioned by Nalder and Wein (1998), and especially by **universal kriging**, the description of which is given by Cressie (1993). In universal kriging the following model assumption is considered:

$$Z(s) = \mu(s) + \varepsilon(s), \tag{2.18}$$

where $\mu(s)$ is a *trend*.

Universal kriging has been used e.g. Bilonick (1985) for the interpolation of sulphate deposition data.

Şen (1995) tries to solve the problem of nonstationarity in a different way. He constructs a so-called **cumulative semivariogram**:

$$\tilde{\tau}_{cum}(d_k) = \frac{1}{2}\sum_{l=1}^{k}((Z(s_i) - Z(s_j))^2 \tag{2.19}$$

where d_k are the distances ranked in ascending order between the measuring stations for all combinations of the pairs of the stations, $k = 1, 2,\ldots, n(n-1)/2$,
$Z(s_i)$ and $Z(s_j)$ are the concentrations measured at the stations s_i and s_i, the distance of which is d_l .

The graph of the cumulative semivariogram enables the first analysis of the relations among the stations and enables the determination of the local dependencies. The values of the cumulative semivariogram can become the weighting factors in the interpolating methods.

The estimation of the unknown concentration $\hat{Z}(s_0)$ in the point s_0 is calculated by the equation

$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} \lambda_i Z(s_i)}{\sum_{i=1}^{n} \lambda_i} \tag{2.20}$$

where d_{0i} is the distance between the measuring station s_i and the estimation point s_0,
$\lambda_1, \ldots, \lambda_n$ are the weights computed at the basis of cumulative semivariogram.

Şen (1998) applies this method to estimate the SO_2 and SPM concentration fields in the Istanbul urban area.

The last method to be mentioned in this section is **cokriging,** which adds the property of correlation between the variables. Sometimes it is used for interconnection of different pollutants, otherwise for spatio-temporal analysis of the one pollutant. Its theoretical description is also given by Cressie (1993). Cokriging

is in practice used e.g. by Pardo-Iguzquiza (1998) for the construction of precipitation maps using both precipitation measurements and the altitude of the measuring stations. It's general form can be written as

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) + \sum_{j=1}^{m} \eta_j Y(x_j) \qquad (2.21)$$

where $Z(s_i)$ are the primary data at a measurement point (in this case rainfall),
$Y(s_i)$ are the secondary data (in this case altitude),
λ_i and η_j are the weights, which are based on knowledge of the variograms and the crossvariogram.

The crossvariogram is determined using the covariance of the two quantities Z and Y in a similar manner to the use of the variance, equation 2.13, to determine the semivariogram. Bytnerowicz et al. (2002) investigated a large number of interpolation methods, including both simple and geostatistical methods, for the interpolation of ozone measurements in the Carpathian mountains of Eastern Europe. In total 28 different methods were evaluated, all of which were available in standard ArcGIS software (Johnston et al., 2001). Results from that study indicate a number of suitable interpolation methods but a spherical model of cokriging, with altitude as the secondary variable, was finally recommended for use in that analysis.

2.5.2 Optimum interpolation

The theory of optimum interpolation was introduced by Gandin (1963) and further developed by e.g. Alaka et al. (1972). In the area of air quality it has been used by e.g. Hrdà (1981). The optimum interpolation assumes that the assumption of homogeneity for the field is fulfilled: i.e. the mean of the field must be constant and the correlation function must depend only on the vector d_{ij} between the relevant points, i.e.

$$E[Z(X)] = \mu \qquad (2.24)$$

$$K_Z(\mathrm{X_i}, \mathrm{X_j}) = E[Z(\mathrm{X_i}).Z(\mathrm{X_j})] = B_Z(\vec{h}_{ij}) \qquad (2.25)$$

In most cases fulfilment of the assumption of isotropy is also required, i.e. the correlation function of the concentration field must depend only on the size of a vector d_{ij} (independent of the direction and orientation).

If these assumptions are fulfilled, this method of interpolation is really objective, because it minimizes the size of the mean squared error.

The weights $\lambda_\iota = \lambda_1, \ldots, \lambda_M$, which are necessary for equation 2.11, are calculated from the equation

$$\sum_{i=1}^{M} \lambda_i r_{ij} = r_{0j} \qquad j = 1, 2, \ldots, M \qquad (2.26)$$

where r_{01} and r_{ij} are the values of the correlation function for the consequent distances between the stations.

Beier and Dopperfeld (1992) proceed from optimum interpolation, they construct the so-called **modified structure function (MOST)**:

$$MOST(x_j, x_i) = \frac{1}{M} \sum_{\substack{k=1 \\ c_i(t_k) > h}}^{N} \left(\frac{c_j(t_k)}{c_i(t_k)} - 1 \right)^2 \qquad (2.27)$$

where $c_i(t_k)$ and $c_j(t_k)$ are the time series of the measured concentrations at x_i and x_j , $k=1, 2,\ldots,N$
M is the number of measurements in the time series,
h is the chosen positive constant.

This function is consequently used in the calculation of interpolation.

2.5.3 Other spatio-temporal methods

Spatio-temporal mapping includes many problems, as Rouhani et al. (1992) shows. Among these problems are e.g. non-stationarity of the spatial means, imbalance between the availability of spatial and temporal

data, the presence of periodic and non-periodic temporal trends (e.g. seasonal trends) in data. Rouhani et al. (1992) introduce a **multivariate geostatistic model**. This model is based on a multi-scale temporal approach (and it is in many ways analogous to an optimum interpolation). The values, which are measured at the individual stations, are considered as realizations of separate, but correlated random variables. The collection of one-dimensional random variables can be considered as a set of correlated random functions. The only disadvantage of that approach is a great number of variograms and covariance functions, which need to be estimated.

Let's consider spatio-temporal data set $\{z_i(t_\alpha);\ i = 1,\ldots, N;\ \alpha = 1, \ldots, T\}$, measured at N stations at T time intervals. These variables can be viewed as a realization of a set of one-dimensional random functions $\{Z_i(t);\ i = 1, \ldots, N\}$. It is useful to postulate the hypothesis that the increments, $Z_i(t_\alpha) - Z_i(t_\alpha+\tau)$, are second-order stationary:

$$E[Z_i(t) - Z_i(t+\tau)] = 0 \tag{2.28}$$

$$E[(Z_i(t) - Z_i(t+\tau).((Z_\varphi(t) - Z_\varphi(t+\tau))] = 2\gamma_{i\varphi}(\tau) \tag{2.29}$$

where $2\gamma_{i\varphi}(\tau)$ is defined as the cross variogram.

In the particular case where the variables themselves, Z_i and Z_φ , can be considered as second-order stationary and uncorrelated for large time lags then

$$\gamma_{i\varphi}(\tau) \to \sigma_{i\varphi} \qquad \text{for } \tau \to \infty$$

where $\sigma_{i\varphi}$ is the covariance of Z_i and Z_φ . The value of this variogram can be calculated directly as

$$\gamma_{ij}(\tau_k) = \frac{1}{2T_k}\sum_{\alpha=1}^{T_k}\{z_i(t_\alpha) - z_i(t_\alpha + \tau')\}\{z_j(t_\alpha) - z_j(t_\alpha + \tau')\} \tag{2.30}$$

where τ' is the time lag belonging to a class of lags τ_κ,
T_κ is the number of increment pairs in such a class.

Equation 2.28 is based on the assumption, that the increments are second-order stationary. If the assumption is not fulfilled, there are non-periodic trends in the set.

On the basis of this theory different forms of interpolations can be performed, which are based on cokriging of the measured values.

$$Z_i(t_0) = \sum_{j=1}^{N}\sum_{\alpha=1}^{T}\lambda_\alpha^j z_j(t_\alpha) \tag{2.31}$$

where λ_α^j is the weight of the observed value at the j-th station at the α-th time interval.

Christakos and Vyas (1998) developed a method, which is also based on a spatiotemporal random field. For their method stationarity is not necessary (this is important, because the assumption of stationarity often is not fulfilled, see section 2.5.1). They use this method e.g. for constructing maps of short-term ozone concentrations or the annual wet deposition of sulphur.

Another method introduced by Christakos and Serre (2000) is the approach of Bayesian maximum entropy (when the hard and soft data are distinguished). On the basis of this method they construct the map of short-term PM_{10} concentrations in Northern Carolina.

3 Interpolation methods using monitoring and supplementary data

In section 2 the interpolation between monitoring stations was carried out purely on the basis of the monitoring data themselves. However, there can be well correlated physical relationships between concentrations and other distributed fields, which may be better known. A typical example is given in the previous section in regard to cokriging, where rainfall has a strong relationship to altitude. When discussing supplementary, or additional, data we define two types. Ancillary data is information such as land use, climatology and elevation that are relevant for the processes that determine concentrations and surrogate (or proxy) data is data that can be used to represent the assumed distribution of the field, e.g. using PM_{10} as a surrogate for $PM_{2.5}$ or using emission fields, through a regression relationship, to represent the distribution of a concentration field.

Relationships may exist between measured concentrations and spatially distributed fields of: emissions, land use, population, road network distribution, altitude, rainfall, latitude, climatology, other observed air quality data, etc. Use can be made of these spatial relationships to improve the interpolation between measured concentrations points. In this chapter examples of some of these methods are reviewed. They generally involve the use of regression relationships between known spatial fields and limited monitoring data. The regression relationship, or model as it is often referred to, can then be used to create a spatial map of data, based on the ancillary data field.

When referring to ancillary data, this infers extra data that is useful for the interpolation. When referring to surrogate data this implies data that can effectively replace, given some conversion function (regression model), the monitored data. When the ancillary data has a spatial resolution analogous to the desired interpolation field, e.g. satellite or land use data, then the difference between using the ancillary data as an interpolation field and using the monitored data as a conversion function (surrogate data) for the field data becomes less well defined.

3.1 Regression based on emission fields

The combination of emission and observed air quality data has, for a long time, been applied by Stedman (1998). His approach, used for many pollutants (SO_2, NO_2, NO_x, PM_{10}, CO, benzene, 1.3-butadien, CO, lead), is based on a 1x1 km grid resolution emission inventory. The inventory is used to calculate the contribution from sources within a 25 km^2 region around the relevant measuring point for local air quality (Stedman states, that this area gives the most robust estimation). Thus there is a distinction between the contribution of large distant sources (such as power plants or agglomerations) and the contribution of local emission sources.

At first the annual mean background field of concentrations is calculated on the basis of the rural background stations (using kriging or bilinear interpolation). Subsequently, the difference (*dif)* between the ambient pollutant concentrations measured at urban monitoring sites (not roadside or industrial sites) and the underlying rural concentration field is calculated where monitoring data are available. The relation between this difference (*dif*) and the estimation of emissions (*em)* in the vicinity of the monitoring stations is then examined through regression analysis:

$$dif = k.em \tag{3.1}$$

Then, with the aid of the coefficient *k*, the map of annual average concentrations is constructed by combining the emission database and the background map obtained by interpolating data from rural background stations:

$$Z(s_0) = P(s_0) + k.\ em(s_0) \tag{3.2}$$

where *em* is the sum of emissions in the area 25 km^2 around point s_0 (in kilotons)
k is the regression coefficient
P is the rural background field.

To estimate the annual rural background field of lead and benzene concentrations (which are measured in Britain at much fewer stations than the basic pollutants) empirical relations of these pollutants with nitrogen dioxide are utilised. For benzene Stedman states the relation (for annual averages)

$$BZN\ [ppb] = 0.031 * NO_2\ [ppb]. \quad (3.3)$$

To estimate emission fields of lead and benzene empirical relationships with other pollutants, in the case of traffic emissions, are utilized. For lead emissions Stedman used NO_x emissions. Originally he related nitrogen oxides computations to benzene emissions (Stedman et al., 1997) but in further work (Stedman, 1998) he states that the relationship between benzene and NO_x differs in dependency on the speed of cars – the emissions of nitrogen oxides from road transport are larger at higher speed, while the emissions of benzene are higher at lower speeds (this leads to overestimation of benzene concentration at highways). As a better alternative he states the relationship between benzene and VOC (benzene is estimated as 1.55 % of total VOC) and uses the emission inventory of VOC for the mapping of benzene. This approach to create the benzene emission field is further improved in Stedman et al. (1998). To assess the contribution of emissions from traffic an empirical model is constructed, which utilizes the relationship between VOC emissions from traffic (obtained from an emission database) and the concentrations of benzene measured at traffic stations.

Stedman states that the most complicated field is the construction of PM_{10} maps. The main reason is the large number of pollution sources. He separates PM sources into the primary particles emissions from traffic, primary particles from stationary combustion sources, primary particles from non-combustion sources (e.g. mining, demolition, dust drifted in by wind) and secondary particles. To estimate the contribution of secondary particles into the background field Stedman suggests the use of ground-level ozone or background measurements of SO_4, the second of which he consequently prefers. The method for calculation of annual averages of PM_{10} and NO_2 is further developed in Stedman et al. (2001), see figure 3.1 and 3.2 The sum of emissions (*em)* in equation 3.2 is considered for a larger region (1225 km^2) and in addition a dispersion model is utilized. The coefficient *k* is computed separately for agglomerations and the remaining area. Verification of the PM_{10} maps was carried out. For background concentrations a regression dependency was detected with the value of the coefficient R^2 between 0.50 a 0.65. The statistical dependency at the traffic stations was noticeably weaker.

Lloyd and Atkinson (2004) make a comparison of interpolation methods including IDW, linear regression, ordinary kriging, kriging with an external drift and simple kriging with a local varying mean. Emission fields of nitrous oxides are used as ancillary data in the interpolations in order to calculate NO_2 fields for the United Kingdom. The conclusion from that study is that simple kriging with a locally varying mean gives the best results, the local mean being determined from the ancillary emission data. This has been converted from NO_x emissions to NO_2 by way of a linear regression with NO_2 observations.

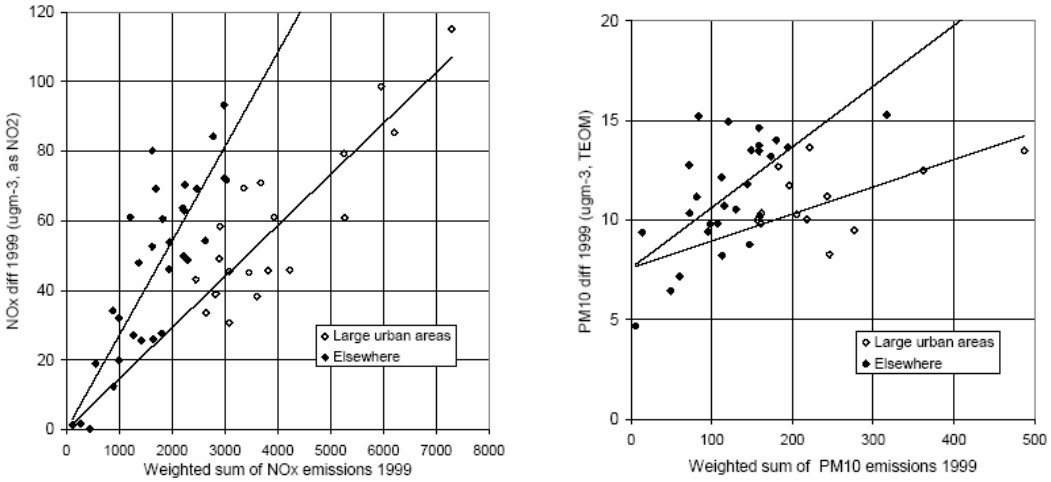


Figure 3.1. The relationship between ambient NO_2 (left) and PM_{10} (right) concentrations against the weighted sum of local emission (Stedman et al., 2001). These data are used to establish the regression coefficients applied to produce the air quality maps shown in figure 3.2. Measurements for urban and elsewhere (rural) have been separated and given different regression coefficients.

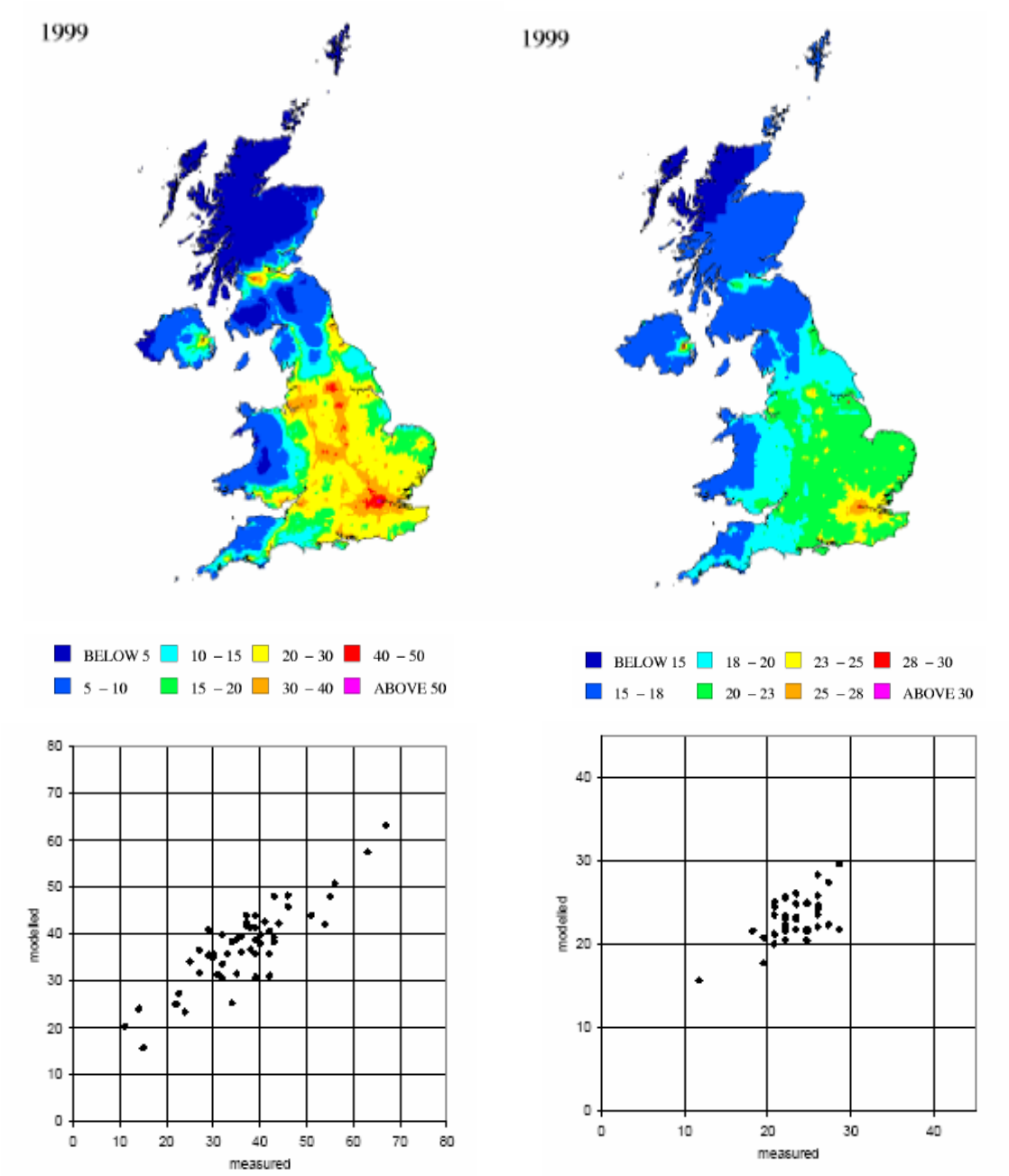


Figure 3.2. Estimated annual mean (1999) of NO2 (left) and PM10 (right) concentration (ugm-3) using the regression model from Stedman et al. (2001). The top maps show the concentration fields calculated from the emission fields in combination with the regression relationship. The bottom graphs show the comparison of measured and modelled annual means at rural sites (ugm-3).

3.2 Regression based on land use

Jerrett et al. (2003) applies a regression model to measured concentrations and land use data. In the study described, 95 passive samplers for NO_2 were placed in Toronto over a 2 week period. Measured concentrations were tested against 85 different land use parameters (e.g. traffic density, distance from roads, population density, physical geography, wind direction, etc.). The final regression model used 8 different parameters and gave a coefficient of determination (r^2) of 0.69 with the most positive associations from traffic counts, road length and distance, numbers of dwellings and land us (industrial), as well as up- or

downwind indicators. A comparison of the land use regression model with ordinary kriging is given in figure 3.3.

Similarly Briggs et al. (2000) and Lebret et al. (2000) have carried out regression studies in 4 European cities as part of the SAVIAH project. Here a large number of passive samplers for NO_2 were distributed throughout the cities and the resulting monitored concentrations were used to build up a regression model on the basis of land use data that included traffic volume within 300 m, land use (industrial, commercial, housing density) and elevation. The study from Briggs et al. (2000), that looked at 4 areas within London, acquired coefficients of determination for the 3 regions that ranged from 0.51 to 0.76 for the mean annual concentrations.

A method similar to this is the approach of Kopeckà et al. (1995). It is based on the linear regression among the measured air pollution and the emission density, the roughness and the percentage woodiness. If this regression dependency is significant, it is used to calculate the concentrations at locations without measurements.

These studies all indicate that the interpolation of monitoring data can be significantly improved by the inclusion of land use data. These studies have all been carried out in urban regions and indicate the importance of the emission field in improving interpolation.

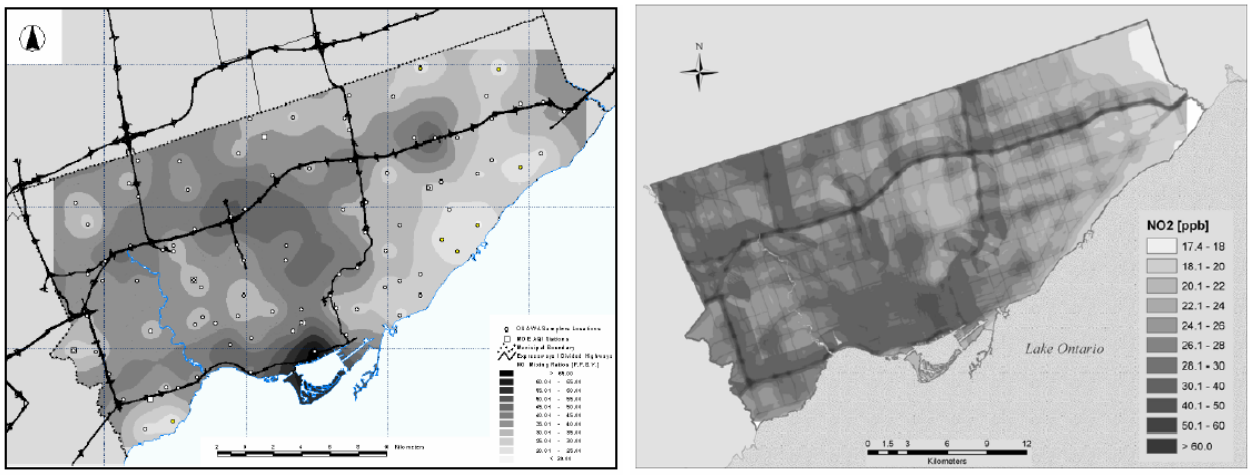


Figure 3.3. Resulting maps of NO2 from the study by Jerrett et al. (2003). Left is the kriged surface generated with a spherical model of NO2 using 95 data points. Right is the final operational map that uses 8 different land use parameters, including wind direction, for the regression model.

3.3 Regression based on altitude

The Czech Hydrometeorological Institute in Prague has developed a method, Fiala et al. (1995) and Kveton et al. (1998), which uses linear regression between the values measured at the stations and altitude. Regression is calculated for all stations individually, while only the stations in the given area ambient to the particular station are considered in each calculation. The resulting regression coefficients are interpolated by the IDW method to the nodal points. In the nodal points relating to the stations the absolute members of regression equations are calculated. They are also interpolated by the IDW method. From the above parameters, a linear regression equation of dependency of monitoring value on the altitude is constructed for each nodal point. The final values for each nodal point are then calculated by substitution of the altitude.

A similar methodology has been applied in hydrology for the interpolation of precipitation data, using altitude as secondary data (Goovaerts, 2000). The linear relation between the model and measurements was estimated at the points of measurements, and then ordinary kriging was applied on the residuals. This is sometimes termed kriging with varying local means. This methodology is the same as one of the methods applied by Lloyd and Atkinson (2004), see section 3.1, but uses altitude data instead of emission data.

A methodology was applied by Loibl et al. (2000) for the interpolation of ozone fields from monitoring stations. This correction was based on a functionality of altitude and time of day. The interpolation method is currently applied in the Austrian Ozone Monitoring Network- visualisation interface.

3.4 Regression based on other data

Abraham and Comrie (2004) improve standard kriging of O_3 fields by using a hybrid regression-interpolation methodology. The mapping of local patterns is enhanced with pre-interpolation regression modelling of local-scale deviation-from-mean variability, preserving variation in the monitor data that is ubiquitous across the modelling domain (i.e., the areal mean). The model is trained on several years of deviation-from-mean hourly O_3 data, and predictor variables are developed using theoretically and empirically derived proxy regression variables. The regression model explains a significant proportion of the variation in the data ($r^2 = 0.54$), with an average error of 7.1 ppb. When augmented with the areal mean, the r^2 of the pre-interpolation model increases to 0.847. Model residuals are then spatially interpolated to the extent of the modelling domain. The authors report significant improvements to the coefficient of determination when applying this method

3.5 Use of stations measuring alternative air quality data

Falke and Husar (1998b) make use of surrogate data to improve interpolation. Apart from using directly measured concentrations of the species to be interpolated, other supplementary data can be introduced – e.g. other pollutants that are measured in a denser monitoring network. They introduce the following formulation, which assumes a direct relationship between the two compounds (Z) and (Y),

$$\hat{Z}(s_0) = \left(\frac{\sum_{j=1}^{n} w_{ij} . \frac{Z(s_i)}{Y(s_i)}}{\sum_{j=1}^{n} w_{ij}} \right) . Y(s_0) \qquad (3.4)$$

where $Z(s_i)$ is the measured value in measuring points s_i, $i = 1, \ldots, n$
$Y(s_i)$ is the supplementary value in the measuring points s_i
$Y(s_0)$ is the supplementary value in the estimated point s_0
w_{ij} is the weight (which depends on the distance of points s_i , s_j).

The new supplementary station values can consequently be used in any of the interpolation methods listed in Chapter 2. To apply this method to air quality problems the surrogate compound must undergo similar physical processes and have a similar spatial emission field as the pollutant being mapped.

An example (Falke and Husar, 1998b) are $PM_{2.5}$ maps for Eastern U.S. using $PM_{2.5}$ (30 stations) as well as PM_{10} (1500 stations) and visibility measurements (280 stations) as supplementary measurements. This methodology does not use regression methods but directly applies measured PM ratios and visibility conversion factors to determine a ratio field, using IDW. The authors report an improvement in the cross-validation regression coefficient (R^2) of 0.59 to 0.69 for the third quarter when applying this method, figure 3.4.

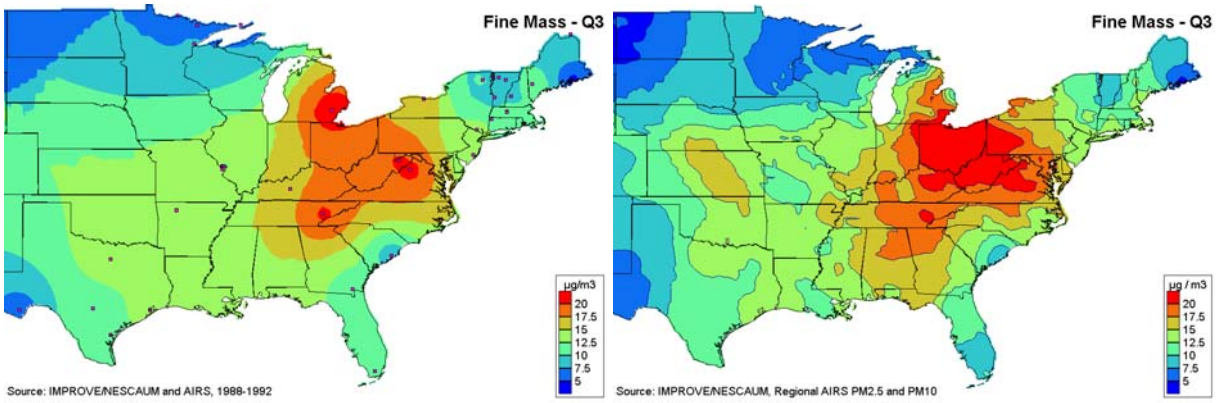


Figure 3.4. Comparison of PM2.5 maps for Eastern United States based on monitored PM2.5 data (left) and after the inclusion of PM10 measurements as surrogate data (right).Taken from Falke and Husar (1998b).

3.6 Satellite data

The use of satellite data, which has a very good spatial coverage but generally poor temporal coverage, has also been suggested and applied in interpolation. In general the actual concentrations of pollutants cannot be directly measured using satellite remote sensing but other indicative parameters can be, for example optical thickness for aerosol measurements.

In an application to Strasbourg (Weber et al., 2001) images from Landsat TM instrument were used to interpolate maps of PM_{10} based on 3 monitoring stations. The process there involved:

1. Identifying pixels in the image, using channels TM1-TM5, which were similar to the pixels corresponding to measurement sites.
2. Creating 'virtual stations' at these similar pixel positions and providing these virtual stations with PM_{10} concentrations based on a regression relation between the thermal infrared Landsat channel TM6 and monitored PM_{10} values
3. These virtual stations are then used to create a spatial map, based on the thin plate interpolation method. An interpolation method similar to the quadratic beta spline method outlined in Section 2.2.

This method is based on previous studies that indicate a correlation between TM6 data and PM measurements. The use and basis for the 'virtual stations' is not clearly defined.

Sarigannis et al. (2003) use ground based measurements of particles and other gases to develop a regression model for PM concentrations and satellite measured values of aerosol optical thickness. The resulting map, based on the spatial AOT measurements can then be used as a surrogate for PM concentrations.

4 Interpolation methods using monitoring and modelling data

The correlation with ancillary data, such as emissions, precipitation, wind speed etc. indicates that there are a number of physical and chemical processes governing atmospheric concentrations. This is, of course, not surprising and the next step to improve interpolation is to use model generated concentration fields that better reflect these processes. This can be done in several ways and some of the interpolation methods described in chapter 2 have been utilized in the application of these methods.

The use of model fields for interpolation relies on a certain amount of confidence in the models ability to correctly reflect the physical processes governing air quality and, in addition, that input data such as emissions are representative of the real world. The model confidence is reflected in an assessment of the uncertainty in model calculations. Uncertainty in model results should be taken into account when applying interpolation methods. This is however only carried out in more advanced data assimilation techniques, see chapter 5.

4.1 Combination with models using interpolation of residual fields

To produce European wide maps of pollutant concentrations as accurately as possible, a methodology has been developed to combine both EMEP monitoring and modelling data through spatial interpolation, to produce geographically distributed concentrations fields (Tarrason et al., 1998). After an initial selection process that excludes inconsistent monitoring data, a comparison is made between modelled and monitored data. The difference between the modelled and measured values at each measurement point is taken and this difference field is interpolated using radial basis functions, similar to the quadratic beta-spline method described in section 2.3, to create a difference field. The model and difference field are then added together using a distance weighting and a defined region of influence, similar to the 'Intuitive method' described in section 2.4. This influence distance is compound dependent as it coincides to the spatial covariance. For NO_2, the example given here in figure 4.1, the influence distance is 300 km but this distance can be as large as 750 km, depending on the typical time scale of the pollutant. The methodology followed is given below.

Let $m(x_t,y_t)$, $t=1,\ldots,n$ be the measurement value at each measurement site (x_t,y_t) and $mod(i,j)$, $i=1,\ldots,i_{max}$, $j=1,\ldots,j_{max}$ be the modelled values. For each measurement point (x_t,y_t), let the difference between the measured and modelled value in the corresponding grid cell (i_t,j_t) be

$$diff(x_t,y_t) = m(x_t,y_t) - mod(i_t,j_t). \tag{4.1}$$

The differences are interpolated using radial basis functions

$$f(x,y) = \sum_{k=1}^{n} c_k \varphi\left(d_k\left(x,y\right)\right) \tag{4.2}$$

where $d_k(x,y)$ is the Euclidean distance to measurement point (x_k,y_k) and $\phi(r) = \sqrt{(r)}$. The parameters c_k are calculated by solving the linear system of equations obtained by letting $diff(x_t,y_t) = f(x_t,y_t)$, $t=1,\ldots,n$. The function f is a two dimensional continuous function describing the difference at any point (x,y) within the modelled grid, and equals the difference between observed and modelled values at all measurement points.

The new optimal maps are produced by adjusting the model calculations with the interpolated difference. Close to measurements, larger weight is given to observed values whereas in regions with no observations modelled results are preferred. The actual region of influence of measured values depends on the type of component, in particular, on its characteristic transport distance, and is determined by the range of spatial covariance used in kriging.

The new concentration fields $N(x,y)$ are derived as follows:

$$\begin{aligned} N(x,y) &= f(x,y) + mod(x,y) && d \le 1 \\ N(x,y) &= f(x,y)\,(D_{comp}-d)/(D_{comp}-1) + mod(x,y) && 1 < d \le D_{comp} \\ N(x,y) &= mod(x,y) && d > D_{comp} \end{aligned} \tag{4.3}$$

where d is the distance to the nearest measurement point, in grid squares, and D_{comp} defines the region of influence for a given component. D_{comp} is set to 2 grid squares (300km) for NO_2, while for nitrate in precipitation D_{comp} is set to 5 grid squares (EMEP grid squares of 150 x 150 km).

No objective assessment of the method, e.g. cross-validation, was carried out using this method but the authors conclude the results are improved.

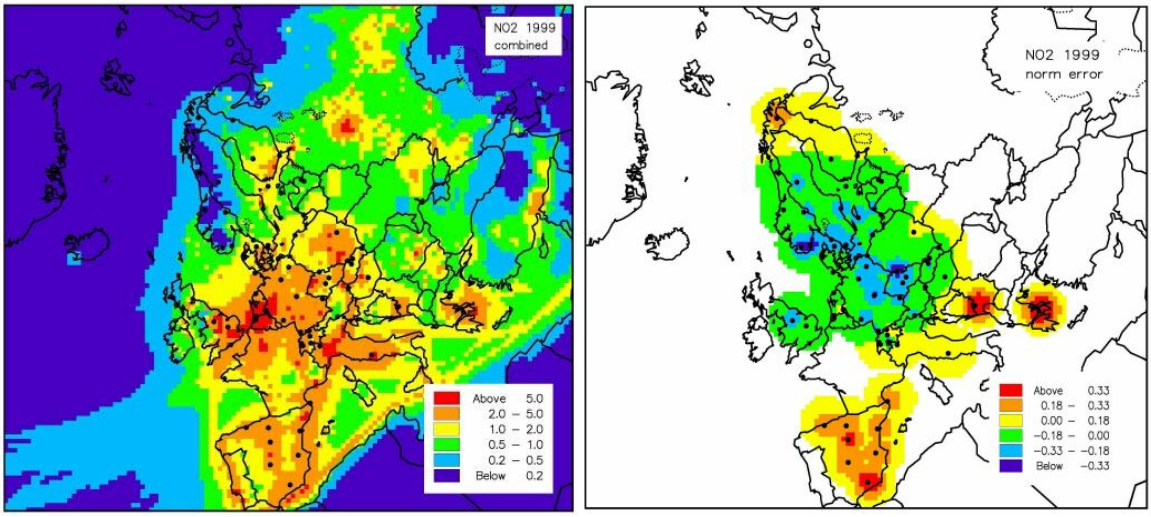


Figure 4.1. Example of interpolation field using the methodology described in the text above (yearly average concentration of NO2, 1999). Left: the final interpolated field where the interpolated difference field has been added to the model field. Right: the interpolated difference field calculated by subtracting the model field from the observations and interpolating using radial basis functions and the intuitive approach.

4.2 Combination with fitted models using interpolation of residual fields

This methodology is similar to the method described in section 4.1, but it first calibrates the model. The model concentration fields are fitted to measurement data, using (mostly) linear regression between measured data and the model, at the points of measurement:

$$M'(s) = c + a.M(s) \tag{4.4}$$

where *M'(s)* are the values of the fitted model at the point of grid *s*,
M(s) are the original values of the model at the point *s*,
c and *a* are the parameters of linear regression.

For every measuring point $s_1, \ldots, s_N$ the differences are calculated

$$diff(s_i) = Z(s_i) - M'(s_i). \tag{4.5}$$

These differences are then interpolated (using different approaches, see chapter 2) into the field of differences *D*. This interpolation field is added to the fitted model *M'* to produce the final interpolated field:

$$\hat{Z}(s_0) = M'(s_0) + D(s_0) \tag{4.6}$$

In the Czech Republic this method is used for the routine mapping of air-quality under the name *method of the weight field,* see Fiala et al. (2000). The mapped area is divided into several regions (e.g. by different climatological regime). The fitted model (4.4) is computed for each of these regions separately to give the linear regression parameters for each region. The interpolation of differences from (4.5) is carried out using the modified version of IDW (see section 2.1.2).

A similar approach is applied by Stedman (2005), based on the methodology described by Abbott and Vincent (1999), for rural SO_2 data. They use ordinary kriging as the interpolation method. Results from Abbott and Vincent (1999) show that the determined regression factor c=3 ugm^{-3} and a=1.1, indicating that

the model was generally capable of predicting SO_2 concentrations but tests against independent data sets did not show a high correlation between these and the resulting interpolated fields.

Horalek et al. (2003) also apply this method using kriging, IDW and radial basis functions as the interpolation methods for the Czech republic. Results from this study are shown in figures 4.2 and 4.3.

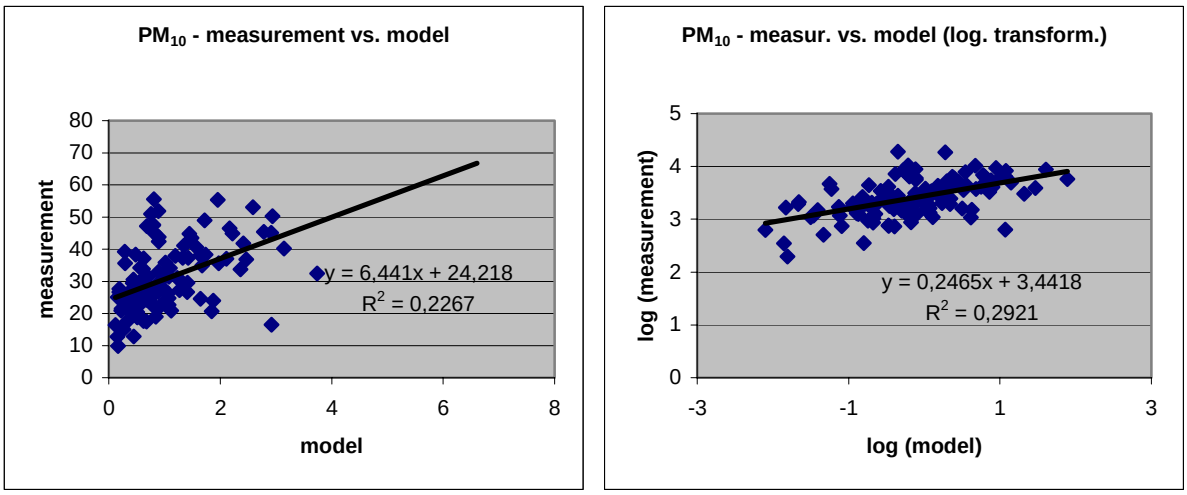


Figure 4.2. Comparison of modelled verses measured annual mean concentrations of PM_{10} for all observational sites within the Czech Republic, 2002. Left: the linear comparison, right: the log comparison. The lognormal regression relation is used in the fitted model (Horalek et al., 2003). Note the model results are significantly lower than the observations.

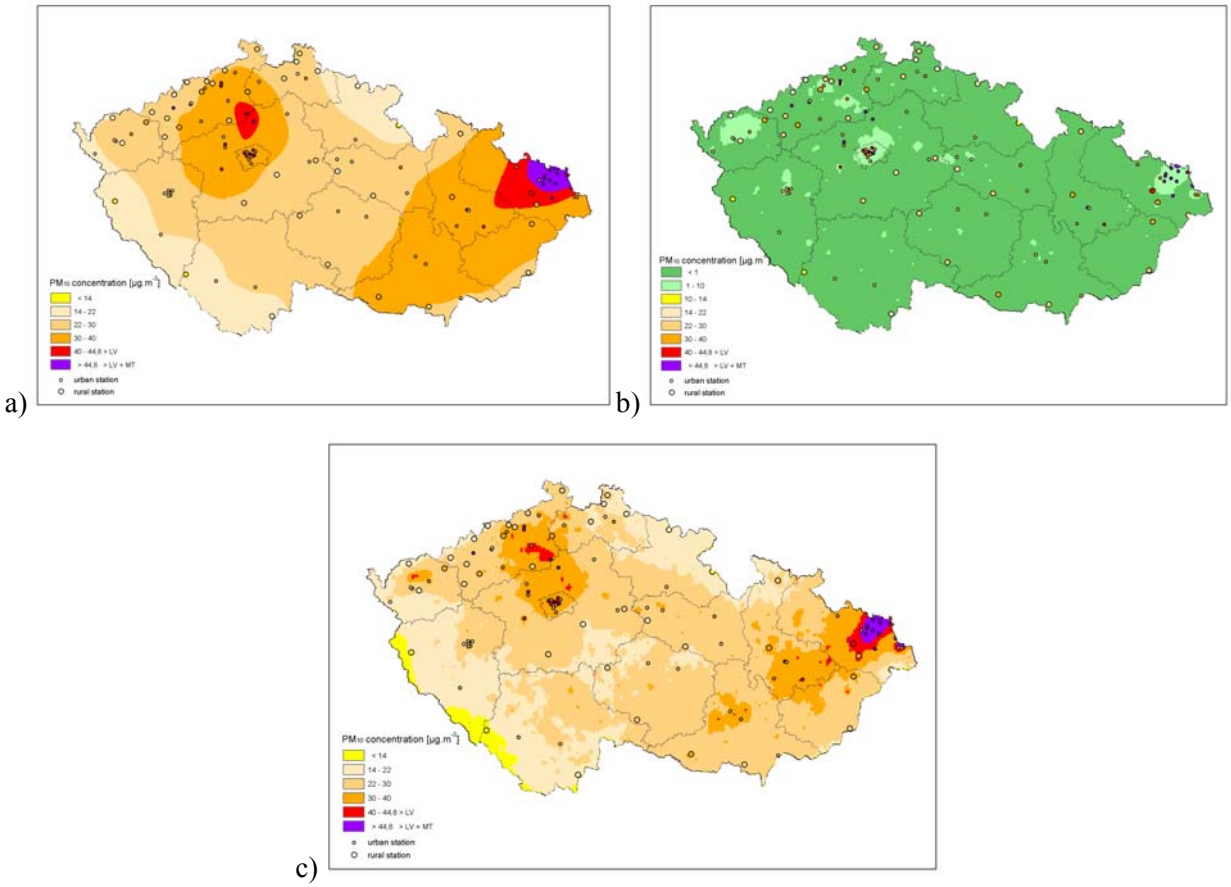


Figure 4.3. Annual mean PM_{10} maps for 2001 in the Czech Republic a) made by interpolation using cokriging with altitude, b) from the pure unfitted model, c) the final combined map using fitted model and cokriging of residuals with altitude. Maps are taken from Horalek et al. (2003).

The major difference between this method and that described in section 4.1 is that the model is first fitted, through linear regression, to the observations and it is this model field that is used to determine the difference, or residual field, that is interpolated. This first step can improve the model field in a general sense, including overall effects of missing emissions, incorrect descriptions of boundary layer parameters or other chemical and physical processes.

4.3 Other methods for combining models and monitoring

4.3.1 FLADIS

The German system "FLADIS" (Wiegand and Diegmann, 2000) computes both short- and long-term concentrations of basic pollutants (including PM_{10}) by a combination of interpolated measured data and the result of both statistical and physical models based on emission, meteorological and topographical data. The combination is carried out using a linear weighing factor. This factor is given by the coefficient of determination between the measured and modelled data at the measuring points such that when the coefficient of determination is high, based on all the measurement sites, then the model field is more heavily weighted. When the coefficient of determination is low then the interpolated observational field is more heavily weighted. The system has been used to generate NO_2 and O_3 maps in a number of German states.

4.3.2 AirQUIS

A now-cast system, developed by NILU and applied in Haifa, Israel, also combines observations and dispersion model calculations to improve concentration fields that are produced and displayed on an hourly basis for the city of Haifa (Denby and Flicstein, 2005). The methodology employed is dependent on the accuracy requirements defined by the user. In this case calculated concentration fields are considered acceptable if the model concentrations, within a 1 km radius of the monitoring sites, are within ±25% of the observed values. Based on these requirements the assimilation of observations is carried out in a 3 step process.

1. For each site the model concentration that is closest to the measured value, within a 1 km radius of the monitoring site, is chosen.
2. This value is used as basis for the model field adjustment. This assumes that the model field can be adjusted by a constant scaling factor that is derived by minimisation of the root mean square error of modelled and measured values.
3. A final local adjustment, based on a local weighting around the monitoring sites, is carried out to insure compliance with the specified acceptance criteria.

The most important aspect of the assimilation is step 2, which adjusts the model fields according to the best fit with observations. The application of this method suggests that certain parameters in the model, or model input data, have an overall effect on the model region, which is 20 x 20 km^2. Such factors as wind speed, mixing height and turbulent diffusion will all have a general influence on the concentration levels. The resulting fields for dispersed pollutants, e.g. traffic related compounds, are improved with assimilation but the method is less successful when applied to plumes, e.g. SO_2, as the monitoring network, even with 15 stations, is not sufficiently dense to detect and adjust plume concentrations.

4.3.3 Bayesian approach

Brabec (2002) introduced an approach for combining measured and modelled data, which distinguished hard (measured) and soft (modelled) data and works with the spatial structure of both of them. This approach is more sophisticated and also more computational demanding than the other methods described here. It is based on the Bayesian statistics, see e.g. Gelman et al. (1997). This methodology is used in order to join the measured data (so-called prior) and the result of the model (apriori information) into the updated version of prior, so-called posterior.

4.3.4 Background estimation, RIVM

In the Netherlands (EC working group, 2000) the unknown background contribution is estimated using modelled and observed fields. The background field of PM_{10} is estimated by taking the difference between the measured and modelled data at every rural background station. These residuals, which indicate a missing background contribution, are spatially interpolated into a 5x5 km grid. This interpolated field is then added into the modelled field, thereby giving an improved PM_{10} field. The urban stations are used for verification of the model.

5 Assimilation methods using monitoring and models

Data assimilation, in order to provide spatial distribution of pollutants, takes the interpolation of observational data with model results to the next logical level. That is that observational data is used to influence the model results, allowing for uncertainties in input data and process representation, to obtain the best physical representation of concentration fields. The previously described assimilation methods, chapters 2 to 4, are by no means consistent with physical principles, e.g. independently interpolated ozone and NO_2 fields may not be physically realisable under meteorological and emission conditions known to exist.

Data assimilation often requires the use of advanced models, especially on the regional level where chemistry is most important, as well as substantial computer resources. Meteorology is one of the scientific fields that has benefited enormously from the use of data assimilation. The subject area of assimilation in atmospheric chemistry modelling is large and ever increasing. In this review a summary of current methods and applications is provided.

5.1 Optimal Interpolation

The optimal interpolation (OI) method (Gandin, 1963) has been quite popular for many years, see section 2.5.2. Applied to a regional scale grid model the method is based on finding an optimal weight-matrix W for which the innovation vector $y^o - H(x^f)$ (y^o is the observational vector and H is the model operator) will be multiplied and added to the forecasted model state x^f in order to form a new assimilated model state x^a. The weights used in the method are optimal in the sense that the prediction variance is minimized using these weights.

The technique relies upon the user to specify, in advance, covariances between the model calculated concentrations (model errors between different grid cells), and to specify the observational errors involved. The model error covariances are usually specified in the method by using one or more spatial correlation or covariance functions, modelling the covariances horizontally and vertically in the grid as a function of distance.

The OI-method is formally equivalent to the 3D-Var and PSAS data assimilation methods (see section 5.3 below), both of which are now considered to be more modern and flexible than the OI approach (Kalnay, 2003). Several numerical weather prediction (NWP) centres have in recent years started to use these other methods instead of OI due to their greater flexibility and ease of operation. However, the OI-method may be of use in order to perform regional scale data assimilation. Application of the method is relatively straightforward since, once the model error covariances have been determined, only a simple system of linear equations needs to be solved.

The assimilation of observed satellite aerosol depth into regional models of PM is one of the areas where optimal interpolation is currently used. Optimal interpolation of satellite data has been applied to the MATCH (Collins et al., 2001) and TM (Verver et al., 2002) models. Both the MATCH and TM models include sulphate, soil dust, black and organic carbon and sea salt aerosols in their particle description. The assimilation procedure introduces extra ‘unknown’ sources and sinks of aerosol mass in the model. These are then used to adjust the model field for an optimal fit to observations. The scheme has 7 tuneable parameters.

5.2 Variational methods

Variational methods are based on the minimization of a cost function for the difference between model concentrations and observations (Lorenc, 1986). One of the disadvantages of these techniques is the requirement of developing a so-called adjoint version of the model, which is a complicated task for chemistry transport models. In contrast to sequential methods the model uncertainties must be specified beforehand, which is not straightforward. In addition, the variational methods are less flexible than ensemble methods, since the model adjoint model needs to be revised every time the model is changed or improved. However, the computational effort is only a few times the normal effort to run the model, and thus much lower than using the ensemble Kalman filtering techniques, see section 5.3.

The application of variational techniques in a chemistry transport model was pioneered by Elbern et al. (1997; 1999; 2000). These authors implemented a 4D-var smoother technique for the EURAD model. The system was used to assimilate ozone concentrations during the first hours of an episode over central Europe. The assimilated state was used to perform forecast simulations. The authors showed that the system was able to reduce the absolute difference between observed and modelled concentration fields in the assimilation window considerably. The influence of the assimilated initial state diminishes after about 24-36 hours. At that time the difference between the first guess of the free running model, using a background ozone concentration field as first guess, and the forecast simulation is very small. Hence, even the uncertainties in the model description of a high-level chemistry transport model like EURAD confine the period in which data assimilation improves forecasts to 1-2 days.

5.3 Kalman filter methods

Kalman filter (KF) methods represent a large class of different techniques for performing data assimilation. The main idea of the KF-methods is an automatic estimation and updating of the model error covariance matrix B from one time step to the next using the model operator (the regional scale model) itself. The assumptions in the methods are essentially the same as in the variational methods (3D-Var, 4D-Var, PSAS), namely that the model state and observation errors (uncertainties) involved must have a Gaussian distribution.

The original Kalman filter (Kalman, 1960) was proposed for a pure linear system of evolution. Later it was extended to non-linear systems, and is then known as the Extended Kalman filter. The Kalman filter in its original form (whether linear or extended) has too large a computational complexity to be implemented in practice, so simplifications of the original filter are necessary.

The first application of Kalman filtering techniques was in oceanography. The successful application of this methodology led to the introduction of similar techniques for air pollution studies at the regional scale. The Dutch research community has been active in the application of Kalman filtering to ozone formation. Several studies have been performed using the regional scale chemistry transport models LOTOS (Van Loon et al., 2000; Segers et al., 1997, 2000; Segers, 2002), and EUROS (Hanea et al., 2004). Further, the number of applications of Kalman filtering techniques is growing and applications to other models start to appear (Eben et al., 2005).

There are a number of variations on Kalman filters of which the ensemble Kalman filter and the Reduced Rank Kalman filter have been the most widely applied.

5.3.1 Ensemble Kalman filters

The ensemble Kalman filter method (EnKF) was originally proposed in (Evensen, 1994). The idea behind the method is to apply an ensemble of N model states $\{x^{(i)}, i = 1,\ldots,N\}$ to represent the model error covariance matrix. The method thus avoids completely the complications involved in updating this matrix. Instead one must run the regional scale model N times in order to propagate the ensemble of model states $\{x^{(i)}\}$ to the next time step using Monte Carlo random draw procedures to simulate model errors. The solution converges towards the exact solution of the Extended Kalman filter when N increases.

Hanea et al. (2004) implemented and compared two Kalman filtering techniques, ensemble (EnKF) and Reduced Rank (RRSQRT-KF) to assimilate ozone concentrations over Europe. For this purpose EMEP and AirBase observations were divided in assimilation and validation monitoring data. Kalman filtering requires applying random noise to key model parameters to build an ensemble of simulations. In this study, noise was applied to the emissions of NO_x and VOC as well as the NO_2 photolysis rate. An extensive description of these data assimilation methods, the chemistry transport model and experimental setup can be found in Hanea et al. (2004) and the references therein.

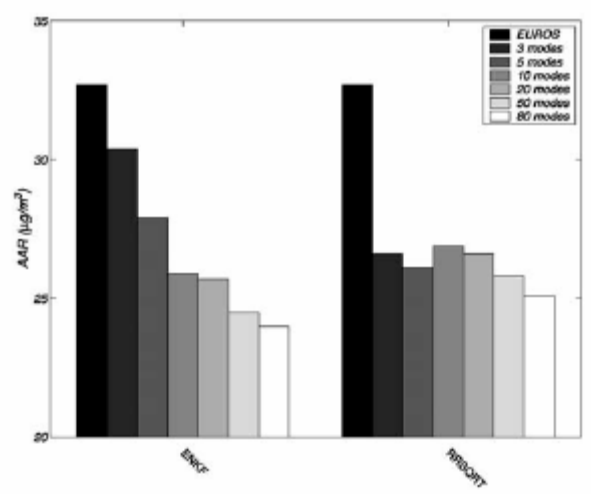


Figure 5.1. Comparison of the performance of the ensemble Kalman filter methods EnKF and RRSQRT-KF using different numbers of modes/ensembles. The average absolute residuals AARs (ug/m3) are averaged over all the validation stations in Europe. Taken from Hanea et al. (2004).

The results of the simulations, figure 5.1, showed that the assimilation system is able to significantly reduce the average absolute residual (AAR), equivalent to MAE (see chapter 7), between the assimilated and observed concentrations at the validation stations. Figure 1 shows the dependence of the improvement in AAR as a function of the number of ensemble members, and therewith the increase in computational time compared to a single simulation. For the Ensemble Kalman filter the residuals decrease with the number of ensemble members or modes. For the reduced rank square root filter (RRSQRT-KF), which constantly reduces the number of modes to those with the most information, the improvement is higher than the EnKF setup with a limited number of modes. However, one still needs a considerable number of modes that increase the computational time considerably. The authors show that adjustment of the emissions strengths gave the best improvement of the modelled ozone concentrations. Moreover, they also showed that the area of influence of an observation station was small, which can be explained by the local characteristics of the photochemical equilibrium.

5.3.2 Reduced Rank Kalman filters

The idea behind reduced rank Kalman filter methods (Verlaan and Heemink, 1997) is to use a low-rank square-root approximation of the model error covariance matrix, which is easier to handle and update than the original matrix. The low-rank approximation is based on representing only the $N < n$ main eigenvalues of the covariance matrix, and assume that the model is linear enough so that the associated eigenvectors are sufficient to represent the probability distribution. Using this method the regional scale model must be run N times instead of ~n. In practical applications this represents a huge saving (as compared with the Extended Kalman filter), since N is usually much smaller than n. The implementation of the technique is, however, less straightforward than the EnKF-method.

Many different variants of reduced rank Kalman filters exists, such as the RRSQRT (Verlaan and Heemink, 1997), POEnKF and COFFEE (Heemink et al., 2001), SEEK and SEIK (Pham et al., 1998) filters. Some of these are also combinations with the EnKF-method. They differ basically in the way the ensemble is being created and maintained. The RRSQRT Kalman filter has been applied in data assimilation of tropospheric ozone in the European regional scale atmospheric transport chemistry model LOTOS (Segers, 2002). Ozone observations were used in order to reduce the uncertainty of emissions and depositions in this model. In this study the RRSQRT Kalman filter was found to work well and was also found to be somewhat faster than the (original) Ensemble Kalman filter method EnKF.

One of the main practical advantages of the EnKF and the reduced rank Kalman filter methods is the separation of the model implementation and filter algorithm. Neither a complicated tangent linear model nor an adjoint operator is required (as in the 4D-Var approach) which makes the coding of the method simpler. For all the ensemble based methods the choice of ensemble size N is very important. It is difficult to say in

advance how large N needs to be, sconsequently trial and error is necessary to a certain extent. The computational effort of the method is roughly proportional to N runs of the regional scale model (including some extra time needed to update the low-rank approximations).

6 Urban scale interpolation within the regional scale

In principle all the interpolation methods described in the previous sections can also be used for interpolation on the urban or agglomeration scale. Some of the examples in these sections are indeed based on urban scale interpolations. The major differences between interpolation on the urban and regional scale are:

1. There are often a very limited number of stations available in particular urban regions, sometimes none at all.
2. The urban region is limited in extent and some form of boundary or weighting scheme must reflect this.
3. Concentration gradients are much higher on the urban scale than the rural scale due to the high spatial variability of emission sources within urban regions.
4. It is not practically possible to use models on the urban scale for all cities, as described in chapters 4 and 5, and so these interpolation methods cannot be used. Simple types of models, e.g. mixing height etc, may possibly be used.
5. It is not practical to collect a large number of ancillary data sets, as discussed in chapter 3, for all cities in Europe and so the methods described in chapter 3 must be limited to ancillary data that can be taken from a centralised database.

As a result some form of interpolation, as described in chapters 2 or 3, is required for the urban scale. To implement this only urban and suburban background stations should be used as these have a much larger region of representativness, compared to traffic stations. The following methods may be applicable for the inclusion of refined urban scale interpolation within a regional scale field. Some of these methods have already been applied and some are conceptual and would require further testing.

6.1 The cluster method

This method assumes that conglomerations have a number of monitoring sites close together. These sites are included in the regional scale interpolation but their distance of influence is limited by their proximity to one another. In this way an urban site will not have much influence further out into the rural region. This is described in Section 2.1.2. One problem with this method is it does not take into account city size and shape nor is it effective when only 1 or 0 stations are available.

6.2 IDW with differential weighting

A variant on the cluster method is that station types are used to define their influence distance. If IDW methods are used then the weighting distance of urban stations may be decreased. The influence distance must be defined by city size, if this is available, or on population, assuming this to be related to physical size. An example of this method is described in section 2.1.2, Fiala et al. (2000).

6.3 Bounded interpolation

If the rural scale concentration field is calculated then this can be used as background and subtracted from the urban stations. One is then left with the urban contribution. Interpolation within the urban region can then be carried out using any of the interpolation methods described in chapters 2 or 3 but with appropriate boundaries. These boundaries can be set either by a population threshold or by other methods, e.g. polygons or circles describing the urban extent.

6.4 No monitoring data available

When considering a European wide interpolated concentration field with refinement on the urban scale, it is necessary to take into account cities that have no monitoring data available. Estimates of concentrations within these cities will require the availability of ancillary data, such as emissions, population, land use, climatology, etc. to estimate urban scale concentrations. It is possible, based on measurements made in

other cities, to determine regression relationships between concentrations and a set of particular parameters, such as population, in order to estimate urban scale concentrations within cities without measurements.

Relationships between total population and urban background concentrations have been established in the Auto-Oil II Programme: Air quality report (2000). In this report regression relationships were established for NO_2 and CO for 1995 on the basis of population density, total settlement population and observed air quality. In total 1065 settlements were taken into account. This report found a normalized relationship between population and concentration given by

$$C = a\, p_d\, u_p^{b+1} \tag{6.1}$$

where C is the concentration

a and b are empirical coefficients

p_d is the population density (inhabitants/km2)

u_p is the percentage of the total settlement population of the total EU15 urban population

Values of the coefficients for the yearly average NO_2 concentration, for example, were given as $a = 1.6291 \times 10^{-4}$ and $b = -2.21239$.

6.5 Superposition of urban and rural interpolated fields

In addition to the use of regression relations where no monitoring data is available (section 6.4), it is also possible to create directly an urban interpolation field that covers a region much larger than individual urban areas, e.g. Europe wide. By taking urban monitoring concentrations, or the difference between urban and rural concentrations (urban delta), an interpolated urban concentration field can be made. The two fields, an interpolated urban or urban delta field $\hat{Z}_u$ and a rural field $\hat{Z}_r$, can then be superimposed by combining the two fields with a weighting function w. This can be written as

$$\hat{Z}(\mathrm{s}_0) = w(\mathrm{s}_0)\hat{Z}_u(\mathrm{s}_0) + (1 - w(\mathrm{s}_0))\hat{Z}_r(\mathrm{s}_0). \tag{6.1}$$

where s_o is the position at which the final concentration $\hat{Z}$ is determined The weighting function will be dependent on population density such that when population density decreases to rural levels the weighting function will approach 0, e.g. <100 inhabitants/km^2. For typical urban population densities, e.g. > 500 inhabitants/km^2, the weighting field will approach unity.

Implicit in this form of interpolation is that there exists some relationship between population density and the urban, or urban delta concentrations. However, if this is the case, then it has the advantage over other methods that it does not require the definition of urban borders or influence distances, as do the methods described in sections 6.2 and 6.3. It also can include effects that are more closely linked to the local region, e.g. climatology, industrial typology, emission factors. One of the disadvantages is that large industrial centers may have a strong influence on nearby non-industrial centers.

7 Objective quantification of interpolation quality

There are several approaches of how to evaluate the quality of interpolation. Some approaches are empirical, others are closely linked with using models - e.g. geostatistical methods allow a quantification of the error of interpolation, simultaneously with the interpolation.

The most often used measure of interpolation quality is cross-validation. This approach is empirical and objective. The cross-validation method computes the spatial interpolation for each measured point using all the available information except from that one point, i.e. it withholds one data point and then makes a prediction at the spatial location of that point). The predicted and measured values are then compared and the procedure is repeated for all points. In this way the behavior of the investigated methods, excluding the influence of the measurement, can be simulated.

The values predicted by cross-validation are compared to measurements in several ways. On the one hand by linear regression (in the ideal case there would be a relation $y = x$), on the other hand by forming residuals and integrating them into the statistical indicators. The most common indicator is RMSE, i.e. root mean squared (cross-validated) error:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Z(s_i) - \hat{Z}(s_i))^2}\ , \tag{7.1}$$

where $Z(s_i)$ is the measured concentration value in the i-th point ,

$\hat{Z}(s_i)$ is the estimated concentration value in the *i*-th point using other points.

A smaller RMSE generally means a better estimation.

Other cross-validation summary statistics are, e.g. mean prediction error (MPE), which is the same as average bias, or absolute error (MAE):

$$MPE = \frac{1}{N}\sum_{i=1}^{N}(Z(s_i) - \hat{Z}(s_i))\,, \tag{7.2}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|Z(s_i) - \hat{Z}(s_i)\right|\ , \tag{7.3}$$

MAE should be the smallest and MPE should be the nearest to zero.

In the case of geostatistical methods it is possible to use other cross-validation parameters (e.g. mean standardized prediction error), which use the prediction standard error (calculated on the basis of these methods). It is also possible to perform validation: a part of the data is removed (so-call “test dataset”) and the rest of the data (“training dataset”) is used to develop the geostatistical model (and its parameters – e.g. nugget, sill and range). Subsequently the model is tested using the removed “test dataset”.

As it is often relevant to compare different methods, including geostatistical and other interpolation methods, it is necessary to use a common parameter or set of parameters that are generally applicable. For this reason the cross-validation methodology should be employed and the parameters of RMSE and MPE should be used.

8 Conclusions and recommendations

The following is a selection of conclusions and recommendations based on the literature review and on tests carried out in Part II of this report. Conclusions from the review will be given and then the selected methodologies for testing will be briefly described. A summary of the results from the tests, which cover the pollutants ozone and PM_{10}, will be given for both rural and urban interpolation tests and finally an overall conclusion and recommendations for future work will be stated.

8.1 Conclusions from the review

A large number of interpolation methods exist and many of these methods are currently available in commercial software. The interpolation methods most often used, in various forms, include inverse distance weighting (IDW), radial basis functions (RBF) and kriging. Clearly when standard software is available the use of such methods will be enhanced. From the studies carried out there is a general tendency towards kriging as an interpolation method above other methods. However, the differences between these methods is not always large in the cases when pure interpolation of monitoring data is carried out. Kriging is usually given as the preferred method.

The studies described in this review, particularly chapter 3 and 4, show that the inclusion of additional data can lead to significant improvement in the interpolated field. It is a logical consequence that if extra data, which is correlated in some way with the primary data to be interpolated, is included in the interpolation then this will lead to improvements in the final interpolated field. This is providing a sensible and statistically sound method is employed. The important point is to identify such additional data and include it correctly in the interpolation.

Methods described in chapters 3 and 4 that include additional data are cokriging, regression models and the interpolation of monitoring-modelling residuals. There are a large number of variants or combinations of these methods, as well as an array of additional data, possible. Such is the variety that their use for particular pollutants and at particular scales should be tested on a case-by-case basis in order to achieve the optimum interpolation method. It is therefore recommended that such tests be carried out to achieve the particular aim of a given study.

Assimilation methods, chapter 5, are the most physically consistent but also the most computationally demanding methods for the combination of measurements and models. The aim of such assimilation methods is to guide a model, which is assumed to reflect to some degree reality, towards observed values, both spatially and temporally. This is perhaps the most promising method for creating the best concentration fields of pollutants but is not easily accessible and is still a developing research area.

There is no clear or tested methodology for combining urban and rural scale observations in the same interpolation. The need to do so is reflected in the typical spatial representativeness of the pollutants. Often rural scale concentrations are transported long distances whilst urban emissions of pollutants, which effect urban measurements, are locally determined. This leads to the need for a more local definition of the urban contribution to pollutant concentrations. Testing of some of the methods described in chapter 6 is required before recommendations can be made.

8.2 Selection of methodologies for testing

As already presented in this review there are a large number of possible interpolation methods, and variations on these, available. A selection has been made of a number of fundamental interpolation methods, listed below, which cover a range of techniques including: pure interpolation, interpolation using supplementary data, interpolation using modelled data and interpolation using regression techniques. They are divided into the following sections with increasing order of complexity:

1. Pure interpolation methods using
    a. IDW
    b. Ordinary kriging
    c. Ordinary cokriging with supplementary data

2. Interpolation using model data. This is tested using the above interpolation methods in two different ways, namely:
    a. Plain subtraction of model and measurements with interpolation of the residual using:
        i. IDW
        ii. Ordinary kriging
        iii. Ordinary cokriging with supplementary data
    b. Combination using fitted model with interpolation of the residual using:
        i. IDW
        ii. Ordinary kriging
        iii. Ordinary cokriging with supplementary data
3. Interpolation using linear regression, including both model and supplementary data, with interpolation of the residual using:
    a. IDW
    b. Ordinary kriging

The methods described in 1 and 2 above can be found in the relevant review chapters 2 and 4. The regression method listed in 3 is a variation on the regression techniques described in Chapter 3 but uses both model and supplementary data to define a regression relation. This is relationship is subsequently used to calculate the residual before interpolation. All these methods are described in detail in Part II, chapter 2.

A number of possible methodologies have not been tested, including:

- Other deterministic interpolation methods than IDW and kriging, such as RBF. The two chosen methods well represent the two types of interpolation, namely deterministic and geostatistical.
- Other regression methods, e.g. with emission data, were not tested (section 3.2) as chemical transport models should provide the best spatial distributions in this regard
- The use of satellite data is not investigated as the focus is on land based measurements
- No model assimilation studies are carried out as this is beyond the scope of the current study
- No temporal-spatial interpolation tests are carried out as this study concentrates on long term, annual, fields

8.3 Summary of rural interpolation studies

This section summarizes the results of tests carried out in Part II of this report to establish the preferred method of interpolating rural monitoring data with regard to ozone and PM_{10} at the European level. The interpolation methods tested, outlined in section 8.2, are applied to the three ozone indicators (AOT40, SOMO0 and SOMO35) and to the two PM_{10} indicators (Yearly mean and 36^{th} highest daily maximum). Four years of data have been selected from the AIRBASE dataset along with EMEP unified model calculations (2000-2003).

The additional supplementary data used in cokriging and in regression calculations includes elevation and 30 year climatological means of sunshine duration, relative humidity, temperature, precipitation and wind speed.

The results of these tests and analysis are described in detail, along with the resultant maps, in Part II of this document. Each of the interpolation methods is analysed in terms of the root mean square error (RMSE) derived by cross-validation analysis. RMSE is used to define the quality of the interpolation.

8.3.1 Supplementary data

Regression analysis is carried out on the supplementary data. The following conclusions are drawn from the analysis.

- Elevation, sunshine and humidity give the most significant correlations for ozone indicators
- Temperature, precipitation and wind speed have very weak correlation with ozone indicators

- Only altitude shows a significant correlation with PM_{10}. None of the other climatic variables show significant correlation
- Annual meteorological data, rather than 30 year climatological data, may be more appropriate as supplementary data for interpolation

8.3.2 Interpolation excluding model data

The following conclusions are drawn from the interpolation methods that do not include model data.

- Of the two pure interpolation methods tested, kriging is always found to be slightly superior to IDW
- For PM_{10}, interpolation using a logarithmic transformation is slightly superior to interpolation without this transformation
- Cokriging with supplementary data, such as elevation and sunshine, significantly improves the interpolation results

8.3.3 Interpolation including model data

The following conclusions can be drawn from interpolations carried out in combination with EMEP model data

- Cokriging with supplementary data significantly improves the results for both plain and fitted model combinations
- The combination of both model and supplementary data using linear regression and ordinary kriging of the residual gives the best result
- The exclusive use of EMEP model data, without any other supplementary data, does not lead to improvement of the interpolated field

8.3.4 Best interpolation methods

The following overall conclusions can be drawn from the interpolation tests

- The best interpolation method, for all the indicators examined, is the linear regression that includes both supplementary and EMEP model data in combination with ordinary kriging of the residual field (Method 3.b)
- When model concentration fields are not available the best interpolation method tested is cokriging with the relevant supplementary data being elevation and, for ozone, sunshine duration (Method 1.c)
- When no supplementary data is available for interpolation ordinary kriging is always found to be slightly superior to IDW (Method 1.a)

8.4 Summary of urban interpolation studies

A number of methodologies are tested for the interpolation of the urban/suburban concentration fields. Monitoring data from AIRBASE classified as urban or suburban background stations are used for the studies along with population density fields and a number of relevant supplementary data sets including geographical position and climatological parameters. In this case only the two ozone indicators SOMO0 and SOMO35 along with the annual mean and 36th highest daily mean PM_{10} concentrations are examined in the studies. The final interpolation methodology is only tested for annual mean PM_{10} concentrations.

Based on the results of the rural interpolation studies the main focus of urban interpolation is to establish regression relationships between measured concentrations and other datasets, e.g. population density, climate etc.. Once defined, estimates of urban concentrations (or deltas) can then be established and used to create interpolated urban concentration fields where the interpolation is not limited to the city borders, as is proposed in section 6.5.

Testing of the methodologies is carried out in a similar fashion to the rural interpolation studies. Regression relationships between measured concentrations and the various supplementary datasets are established and tested using cross-validation. The studies carried out include:

1. Regression relationships between measured urban background concentrations and
    a. population density
    b. other supplementary data
    c. rural background concentrations
2. European wide urban concentration fields based on supplementary data and residual interpolation using:
    a. IDW
    b. Ordinary kriging
3. European wide urban delta concentration fields based on rural background and residual interpolation using:
    a. IDW
    b. Ordinary kriging

Having established the urban (or delta) concentration field the urban and rural concentration fields are then combined using the methodology described in section 6.5 (equation 6.1) that allows a smooth merging of rural and urban fields based on population density. This method is chosen based on an analysis of AIRBASE data, which demonstrates a convergence of rural and urban background station measurements at population densities between 100 and 500 inhabitants/km^2, and for being a practical methodology that allows merging of the two interpolation fields on a European wide scale.

8.4.1 Regression analysis

The following conclusions can be made concerning the regression analysis between urban background monitoring stations and population density and supplementary datasets

- There is very almost no correlation between population density and measured concentrations for both ozone and PM_{10}
- When regions of similar population density are binned a trend for decreasing ozone, increasing PM_{10}, becomes apparent for increasing population density but with large variation
- Similar to the rural stations, there are significant regression relationships between environmental factors such as climate and geographical position
- There is a strong correlation between the measured urban background concentrations and the interpolated rural field (Method 3.2).

8.4.2 Best interpolation methods

Tests are carried out to establish the RMSE, based on cross-validation, of the different regression methodologies for annual average PM_{10} concentrations. The results of this can be summarized as:

- The best interpolation urban fields are derived using delta fields based on a linear regression of the urban background fields with ordinary kriging of the residuals (Method 3.b)
- The best method based purely on regression, i.e. no interpolation of residuals, is based on the linear regression of the urban background fields (Method 1.c)
- The worst method based purely on regression, i.e. no interpolation of residuals, is based on the linear regression of the population density fields (Method 1.a)

8.5 Summary and future recommendations

The mapping methods explored in this paper are intended for use in both the area of public information and policy assessment. There are thus a number of relevant application areas both now and in the future. Of specific relevance to the EEA is the inclusion of air quality in its ‘In Your Neighbourhood’ project. In

addition, the European Union air pollution Directives and the EoI will, in the future, be fully integrated. The INSPIRE Directive will also have implications for spatial air quality reporting in the future.

The current review (Part I) and associated studies (Part II) of interpolation methods have laid the basis for a selection of interpolation methods that are effective and applicable on the European scale, for use in health and ecosystem effect studies. AIRBASE has proven to be an essential source of information for the analysis carried out. The interpolated fields under study have been limited to ozone and PM_{10} indicators but the methodologies can be further extended to other compounds and indicators.

> *The best methodology for interpolation, as a result of this work, is to establish a linear regression model based on relevant supplementary data (model, climatic, elevation, positional) and to use ordinary kriging for interpolation of the residual field.*

It is clear that not all methodologies have been tested and there may be a number of untested supplementary datasets that can improve the interpolation results. A number of questions and possibilities have yet to be explored. The following list indicates areas that have not been completely addressed in this report and will require future attention

- The use of concurrent meteorological data rather than climatological data as a supplementary source of information for regression analysis
- Detailed analysis of uncertainty in the resulting interpolation fields
- Establishment of a robust system for the production of interpolated fields for a variety of compounds and indicators
- Further assessment of the methodology for merging rural and urban maps
- Assessment of interpolation methodologies that can be used on short time scales and that include temporal as well as spatial variability, in support of the EEA and its consultant to apply the mapping methods in the 'In Your Neighbourhood' project.

References

Abbott, J. and K. Vincent (1999). Annual average sulphur dioxide concentration maps derived by dispersion modelling. AEA Technology, National Environmental Technology Centre. Report AEAT - 4629 (http://www.aeat.co.uk/netcen/airqual/reports/kvann1/so2ann.html)

Abraham, J.S. and Comrie, A.C., (2004). Real-time ozone mapping using a regression-interpolation hybrid approach, applied to Tucson, Arizona. Journal of the Air and Waste Management Association 54, 914–925.

Alaka, M. A., Elvander, R. C. (1972). Optimum interpolation from observations of mixed quality. Monthly Weather Review 100, No. 8, 612-624.

Atkins, D.H. F., Lee D. S. (1995). Spatial and temporal variation of rural nitrogen dioxide concentrations accross the United Kingdom. Atmospheric Enviroment 29, 223-239.

Auto-Oil II Programme: Air quality report (2000). Final version 7.1. Report of the Working group1 on Environmental Objectives, Directorate General of the Environment. http://autooil.jrc.cec.eu.int/finalaq.htm

Beier, R., Doppelfeld, A. (1992). Räumliche Übertragbarkeit und Interpolation von Luftqualitatsdaten im Messnetz TEMES. LIS-Berichte 101, p.103.

Beier, R., Doppelfeld, A. (1999). Spatial interpolation and representativeness of air quality data - an intuitive approach. Presented at International Conference Air Quality in Europe: Challenges for the 2000s, Venice, Italy, 19th-21th May 1999.

Bilonick, R. A. (1985). The space-time distribution of sulfate deposition in the northeastern United States. Atmospheric Enviroment 19, 1829-1845.

Brabec, M. (2002). Assimilation of the results of dispersion model and measured air quality data – overview of approaches, suggestion of solutions, illustration on real data (in Czech). In: Environmental project VaV/740/3/02, Prague.

Briggs, D.J., de Hoogh, C., Gulliver, J., Wills, J., Elliott, P., Kingham, S., and Smallbone, K. (2000). A regression based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. Sci. Total Environ. 253:151-167.

Burrough, P.A. 1986. Principles of Geographical information Systems for Land Resource Management, Oxford, USA.

Bytnerowicz, A., B. Godzik, W. Frączek, K. Grodzińska, M. Krywult, O. Badea, P. Barančok, O. Blum, M. Černý, S. Godzik (2002). Distribution of ozone and other air pollutants in forests of the Carpathian Mountains in central Europe. Environmental Pollution 116, 3-25.

Casado, L. S., Rouhani, S., Cardelino, C. A., Ferrier, A. J. (1994). Geostatistical analysis and visualization of hourly ozone data. Atmos. Environ. 28, 2105-2118.

Christakos, G., Vyas, V. M. (1998). A composite space/time approach to studying ozone distribution over eastern United States. Atmos. Environ. 32, 2845-2857.

Christakos, G., Serre, M. L. (2000). BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmos. Environ. 34, 3393-3406.

Collins, W.D. et al. (2001) Simulating aerosols using a chemical transport model with assimilation of satellite aerosol retrievals: methodology for INDOEX. J. Geophys. Res. 106, 7313-7336.

Coyle, M., Smith, R. I., Stedman, J. R., Weston, K. J., Fowler, D. (2002). Quantifying the spatial distribution of surface ozone concentration in the UK. Atmospheric Environment 36, 1013-2002.

Cressie, N. (1993). Statistics for spatial data. Wiley series, New York.

Denby B. and B. Flicstein (2005) Air quality now-cast system for the township of Haifa, Israel. Proceedings of the Fifth international conference on Urban Air Quality, Valencia**.** Published by University of Hertfordshire, CD-ROM, R.S. Sokhi, M.M. Milan and N. Moussiopoulos (eds.)

Dittmann, E., Szalai, S. et al. (1999). Spatial interpolation of climatological parameters. Deutscher Wetterdienst Forschung und Entwicklung Nr. 57, Offenbach.

Eben, K., Juruš, P., Resler, J., Belda, M., Pelikán, E., Krüger, B. C., Keder, J. (2005). Towards an operational air quality forecast with data assimilation for the Czech Republic. Paper presented at 4th WMO International Symposium on Assimilation of Observations in Meteorology and Oceanography, 18-22 April 2005, Prague, Czech Republic.

EC Working group for ambient air quality directives (2000). Guidance on assessment under the EU air quality directives - final draft. (http://europa.eu.int/comm/environment/air/pdf/guidanceunderairquality.pdf)
Elbern, H., Schmidt, H., Ebel, A. (1997) Variational data assimilation for tropospheric chemistry modelling, J. Geophys. Res., 102, 15967-15985.
Elbern, H., and H. Schmidt (1999) A four-dimensional variational chemistry data assimilation scheme for Eulerian chemistry transport modelling. J. Geophys. Res., 104, 18583-18598.
Elbern, H., Schmidt, H., Talagrand, O., Ebel, A. (2000) 4D-variational data assimilation with an adjoint air quality model for emission analysis, Environ. Mod. Softw., 15, 539-548
Evensen, G. (1994) Sequential data assimilation with a non-linear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. J. Geophys. Res., 99 (C5), 10143-10162.
Falke, S.R. and Husar, R.B. (1998a) Declustering in the Spatial Interpolation of Air Quality Data. Presented at the 90th Air and Waste Management association annual meeting, CAPITA Paper No. 98-A927
Falke, S.R. and Husar, R.B. (1998b). Maps of PM2.5 over the U.S. Derived from Regional PM2.5 and Surrogate Visibility and PM10 Monitoring Data; Paper No. 98-A918 by (CAPITA report) (http://capita.wustl.edu/CAPITA/Awma98/HTTP/98_A918.htm)
Fedorov, V. V. (1989). Kriging and other estimators of spatial field characteristics with special reference to environmental studies. Atmospheric Environment 23, 175-184.
Fiala, J., Livorová H., Srněnský, R. et al. (1995). Zdokonalení metod územního a časového hodnocení kvality ovzduší. (Improvement of Methods for Spatial and Temporal Assessment of Ambient Air Quality.) Environmental project, Praha.
Fiala, J., Srněnský, R., Livorová, H., Novotný, P., Riedl, P., Beňo, P. (2000). Postup asimilace měřených a modelových dat při tvorbě polí charakteristik znečištění ovzduší. (Methods of assimilation of measured and modelled data in the production of fields of air quality characteristics.) In: Environmental project VaV/740/2/00. CHMI, Praha.
Gandin, L. S. (1963). Objective Analysis of Meteorological Fields. Leningrad, Gidrometeoizdat (translated by Israel Program for Scientific Translations, Jerusalem, 1965).
Gelman, A. B., Carlin, J. S., Stern, H. S., Rubin, D. B. (1997). Bayesian data analysis. Capman and Hall / CRC, London.
Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. Journal of Hydrology, Volume 228, 113-129.
Hanea, R. G., G. J. M. Velders, and A. Heemink (2004) Data assimilation of ground-level ozone in Europe with a Kalman filter and chemistry transport model, J. Geophys. Res., 109, D10302, doi:10.1029/2003JD004283.
Heemink, A.W., M. Verlaan, and A.J. Segers (2001) Variance reduced ensemble Kalman filtering. J. Mon. Wea. Rev. 129, 1718-1728.
Hjellbrekke, A.G. (2000). Ozone measurements 1998. Kjeller, Norwegian Institute for Air Research (EMEP/CCC-Report 5/2000).
Horálek, J., Fiala, J., Brabec, M., Srněnský, R. (2003). Statistical support of air quality mapping using combination of measurement and model (in Czech). In: Environmental project VaV/740/3/02, Prague.
Hrdá, J., Bártová, I. (1981). Objektivní analýza imisních údajů a její využití k racionalizaci staničních sítí. (Objective Analysis of Air Quality Data and its Application for Optimizing of Measuring Networks.) Meteorologické zprávy 34, No. 4-5.
Jerrett, M., M.A. Arain, P. Kanaroglou, B. Beckerman, D. Crouse, N.L. Gilbert, J.R. Brook, N. Finkelstein, and M.M. Finkelstein (2003) Modelling the intra-urban variability of ambient traffic pollution in Toronto, Canada Strategies for Clean Air and Health Conference, November 5-7, 2003, Rome, Italy (http://www.irr-neram.ca/rome/Proceedings/Entire_Document.pdf)
Johnston, K., Ver Hoef, J. M., Krivoruchko K., Lucas N., 2001. Using ArcGIS Geostatistical Analyst (manual). ESRI, New York.
Kalman, R. (1960) A new approach to linear filtering and prediction problems. Trans. ASME, Ser. D, J. Basic Eng. 82, 35-45.
Kalnay, E. (2003) Atmospheric Modelling, Data Assimilation and Predictability, Cambridge University Press, Cambridge UK.

Kopecká, Š., Fiala, J., Livorová, H. (1995). Vypracování metodiky statistického postupu pro kategorizaci území. (Developing of Methodology of Statistical Procedure for Area Categorization.) Environmental project, Praha.
Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of Chemical, Metallurgical and Mining Society of South Africa 52, 119-139.
Květoň, V. , Srněnský, R., Rett, T. (1998). Contemporary method of computerized spatial analysis of precipitation used in the Czech Hydrometeorological Institute. [Poster]. Map meeting 1998. Chamomix/Font Glanc, France, 17-19 June, 1998.
Lebret, E., Briggs, D., Van Reeuwijk, H., Fischer, P., Smallbone, K., and Harssema, H. (2000). Small area variations in ambient NO_2 concentrations in four European areas. Atmos. Environ. 34:177-185.
Leeuwen, E. P. van, Draaijers, G. P., Erisman J. W. (1996). Mapping wet deposition of acidifying components and base cations over Europe using measurements. Atmos. Environ. 30, 2495-2511.
Lefohn, A. S., Knudsen, H. P., McEvoy, L. R. Jr. (1988). The use of kriging to estimate monthly ozone exposure parameters for the Southeastern United States. Environmental Pollution 53, 27-42.
Lloyd, C. D., Atkinson, P. M. (2004). Increased accuracy of geostatistical prediction of nitrogen dioxide in the United Kingdom with secondary data. Applied Earth Observation and Geoinformation 5, 293-305.
Loibl W., J. Züger, P. Kutschera (2000). Verbesserung des Modells zur Generierung von stündlichen Karten der Ozonkonzentration für Österreich Seibersdorf Research Report, OEFZS--S-0063, BV, 64 pp., März 2000
Lorenc, A. (1986) Analysis methods for numerical weather prediction. Quart. J. Roy. Meteor. Soc. 112, 1177-1194.
Matheron, G. (1963). Principles of Geostatistics. Economic Geology 58, 1246-1266.
Nalder, I. A., Wein, R. W. (1998). Spatial interpolation of climatic normals: test of a new method in the Canadian forest. Agricultural and forest meteorology 92, 211-225.
Pardo-Igúzquiza, E. (1998). Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. Intern. Journal of Climatology 18, 1031-1047.
Pham, D.T., J. Verron, and L. Gourdeau (1998) Singular evolutive Kalman filters for data assimilation in oceanography. C. R. Acad. Sci. Paris, 326, 255-260.
Rouhani, S., Ebrahimpour, M. R., Yaqub, I., Gianella, E. (1992). Multivariate geostatistic trend detection and network evaluation of space-time acid deposition data - I. Methodology. Atmos. Environ. 26A, 2603-2614.
Sarigiannis D., Sifakis N., Soulakellis N., Schafer K., Tombrou M. (2003) A new approach to environmental date fusion for integrated assessment of particulate matter loading and its effect on health in the urban environment. IEEE International Geoscience & Remote Sensing Symposium (IGARSS) 2003, 21-25 July 2003, Toulouse, France, Proc. of IGARSS vol. VII:4579-4581.
Segers, A.J. (2002) Data assimilation in atmospheric chemistry models using Kalman filtering. Ph. D. thesis, Technical Un. Delft, www.library.tudelft.nl/dissertations.
Segers, A. et al. (1997) Data-assimilation applied to non-linear atmospheric chemistry models: first experiences. Proc.1st GLOREAM workshop, Aachen, Germany.
Segers, A., A. Heemink, M. Verlaan, and M. van Loon (2000) A modified RRSQRT-filter for assimilating data in atmospheric chemistry models, Environ. Model. Software, 15, 663– 671.
Şen, Z. (1995). Regional air pollution assesment by cumulative semivariogram technique. Atmosph. Environ. 29, 543-548.
Şen, Z. (1998). An application of a regional air pollution estimation model over Istanbul urban area. Atmosph. Environ. 32, 3425-3433.
Stedman, J. R., Vincent, K. J., Campbell, G. W., Goodwin, J. W. L. and Downing, C. E. H. (1997). New high resolution maps of estimated background ambient NO_x and NO_2 concentrations in the UK. Atmos. Environ. 31, 3591-3602.
Stedman, J. R. (1998). Revised high resolution maps of background air pollutant concentrations in the UK: 1996. AEA Technology, National Environmental Technology Centre. Report AEAT-3133 20008001/006.
Stedman, J. R., Dore, Ch. J. (1998). An empirical model for quantifying current and future benzene and 1,3-butadiene concentrations at the roadside. AEA Technology, National Environmental Technology Centre. Report AEAT-4301.

Stedman, J. R., Bush, T. J., Murrels, T. P., King, K. (2001). Baseline PM10 and NO_x projections for PM_{10} objective analysis. AEA Technology, National Environmental Technology Centre. Report AEAT/ENV/R/0726.

Stedman, J. R., Bush, T. J., Vincent, K. J., Kent, A. J., Grice S., Abbot J. (2005). UK air quality modelling for annual reporting 2003 on ambient air quality assessment under Council Directive 96/62/EC, 1999/30/EC and 2000/69/EC. AEA Technology. Report AEAT/ENV/R/1790.

SURFER for Windows, User's Guide (1994). Golden Software, Inc. 1994.

Tarrasón, L., Semb, A., Hjellbrekke, A.-G., Tsyro, S., Schaug, J., Batnicki, J. and Solberg, S.: (1998), Geographical distribution of sulphur and nitrogen compounds in Europe derived both from modelled and observed concentrations. The Norwegian Meteorological Institute, Oslo, Norway (EMEP/MSC-W Note 4/98).

Van Loon, M., P. J. H. Builtjes, and A. Segers (2000) Data assimilation of ozone in the atmospheric transport chemistry model LOTOS. Env. Mod. and Softw., 15, 603-609.

Van Loon, M, M.Roemer and P.Builtjes (2004) " Model intercomparison in the framework of the review of the Unified EMEP model" TNO-Rep. R 2004/282

Verlaan, M. and A.W. Heemink (1997) Tidal flow forecasting using reduced rank square root filters. Stochastic Hydrology and Hydraulics 11, 346-368.

Verver, G.H.L., J.S. Henzing, G. de Leeuw, C. Robles-Gonzalez, and P.F.J. van Velthoven (2002). Aerosol retrieval and assimilation (ARIA). Final report to NIVR. KNMI-publication 200, December 2002, ISBN 90-369-2223-2 .

Weber, C., Hirsch, J., Perron, G., Kleinpeter, J., Ranchin, T., Ung, A. and Wald, L. (2001). Urban Morphology, Remote Sensing and Pollutants Distribution: An Application To The City of Strasbourg, France. International Union of Air Pollution Prevention and Environmental Protection Associations (IUAPPA) Symposium and Korean Society for Atmospheric Environment (KOSAE) Symposium, 12th World Clean Air & Environment Congress, Greening the New Millennium, 26 – 31 August 2001, Seoul, Korea.

Wiegand, G. and Diegmann, V. (2000). FLADIS - A system for extending air pollution point data to continuous spatial information (according to EU Council Directive 96/62/EC); In: Air Pollution VIII, WIT Press, pp. 191-200, ISBN 1-85312-822-8.