Prog Artif Intell (2012) 1:235–243
DOI 10.1007/s13748-012-0018-6

REGULAR PAPER

Valid predictions with confidence estimation in an air pollution problem

Olga Ivina · Ilia Nouretdinov · Alex Gammerman

Received: 28 November 2011 / Accepted: 13 May 2012 / Published online: 17 June 2012

Abstract The present study is aimed to evaluate levels of air pollution for the Barcelona Metropolitan Region. For this purpose, a newly developed approach called *conformal predictors* is considered, and, in particular, use is made of the *ridge regression confidence machine* (RRCM). The hallmark of this method is that it gives valid estimates, i.e. for a given level of significance of prediction, the probability of error does not exceed this level. Moreover, the chosen specification of the RRCM predictor does not place any requirements on data distribution, apart from being independent and identically distributed. A linear ridge regression conformal predictor has been applied to the data. It has allowed to obtain valid interval estimates of annual nitrogen dioxide concentrations with 95 % confidence. The model has provided good results, but to further increase the efficiency of prediction, the RBF kernel has been used. The data for this study have been provided by the XVPCA (Network for Monitoring and Forecasting of Air Pollution) of the Generalitat of Catalonia. The pollutant considered in this paper is nitrogen dioxide. Its values are represented by annual average concentrations within the period from 1998 to 2009. This paper also describes an application of ordinary kriging, and its results have been compared to those of ridge regression conformal predictor.

Keywords Conformal predictors · Ridge regression · GIS · Kriging · Air pollution

1 Introduction

In our modern environment, air pollution is a very typical problem. It is covered in numerous research works, and those focus on both evaluation of pollution concentrations, and determination of hazardous impact of pollution on people's health and wellbeing. It has been established that each year more than 2 million premature deaths in the world are associated with air pollution [11]. The need for evaluation of contamination levels in particular places has motivated scientists all over the world. They have developed the methods for air pollution assessment since the 1950s [24]. Those methods include geostatistical models, like kriging [10] or inverse distance weighting [23], and land-use regression [9], dispersion models [4] etc. The developed models give quite good estimates of pollution levels. However, those are usually point estimates, and they might lack confidence. Those techniques can additionally provide the estimation error, but the validity of prediction cannot be guaranteed.

Use of conformal predictors solves this problem, because they are always valid [19]. With a given level of confidence, their prediction is correct. An output of a conformal predictor is a prediction set. This set is not necessarily an interval, but often it is. In case of air pollution, an interval estimate given by a conformal predictor can be of greater use compared to a regular point estimate. A valid interval, i.e. an interval that withholds the actual value of pollution with a

O. Ivina (✉)
Research Group on Statistics, Applied Economics and Health (GRECS), Universitat de Girona, Girona, Spain
e-mail: lyolya@gmail.com

O. Ivina
CIBER of Epidemiology and Public Health (CIBERESP), Granada, Spain

I. Nouretdinov · A. Gammerman
Computer Learning Research Centre, Royal Holloway University of London, Surrey, UK
e-mail: ilia@cs.rhul.ac.uk

A. Gammerman
e-mail: alex@cs.rhul.ac.uk

given probability, can be compared to some critical values of pollution established in clinical research.

Regarding the fact that air pollution is a problem of growing concern all over the world, spatial analysis is being carried out to establish the concentrations of pollutants. However, ascertaining the valid intervals of predicted pollution for a given spatial region is a good alternative to point estimates. The World Health Organization has established and published guideline values for contaminative substances suspended in the air [11], and thus valid prediction intervals can be compared to those values to conclude whether the region of interest is "safe" or not pollution wise.

The most frequent contaminant that is being studied now is nitrogen dioxide. It has been shown that nitrogen dioxide has an adverse impact on human health, both in long-term and short-term exposures. Animal toxicological studies also indicate that long-term exposure to NO_2 affects animal health in a hazardous way [11]. In people, nitrogen dioxide is mostly associated with respiratory diseases, but also with cardiovascular illnesses. It has been established that NO_2 is associated with both morbidity and mortality [14]. In high concentrations of over 200 $\mu g/m^3$, NO_2 is a toxic gas, and WHO sets up a guideline value of 200 $\mu g/m^3$ as a 1-h mean concentration [11]. The guideline values are based on expert estimates of air pollution in the world, including both developed and developing countries, and they are aimed to reduce contamination. Those values are outlined for both long-term and short-term exposure. As far as long-term concentrations are concerned, the annual mean guideline value for NO_2 makes up 40 $\mu g/m^3$. The last but not the least of the perilous effects caused by NO_2 is that in the presence of hydrocarbons and ultraviolet light, it is the main contributor to ground level ozone forming. The major source of nitrogen dioxide is road traffic, but it also comes from other combustion sources.

Barcelona is a huge and vibrant city with over 1.6 million inhabitants and over 2000 years of history. It is the heart of the Barcelona Metropolitan Region (BMR) that counts over 5 million of people living there. Furthermore, the area attracts numerous tourists, commerce etc. All those factors make the traffic in the zone busy, which cannot but contribute to traffic-related air pollution. The present study investigates the levels of nitrogen dioxide in the BMR. Its concentrations have been measured at 49 stations across the BMR (see Fig. 1) during the period of time from 1998 to 2009, with the exclusion of 2003. The data have been provided by XVPCA (Network for Monitoring and Forecasting of Air Pollution) of the Generalitat of Catalonia [1].

The data set is represented by the concentrations of nitrogen dioxide together with the geographical coordinates of the measurement spots. Those concentrations are annual averages, although measurements have been taken hourly. The data set consists of 269 observations in total, due to the fact that the data were not available for every station and year and pollutant. The number of valid observations for each year is shown in Table 1.

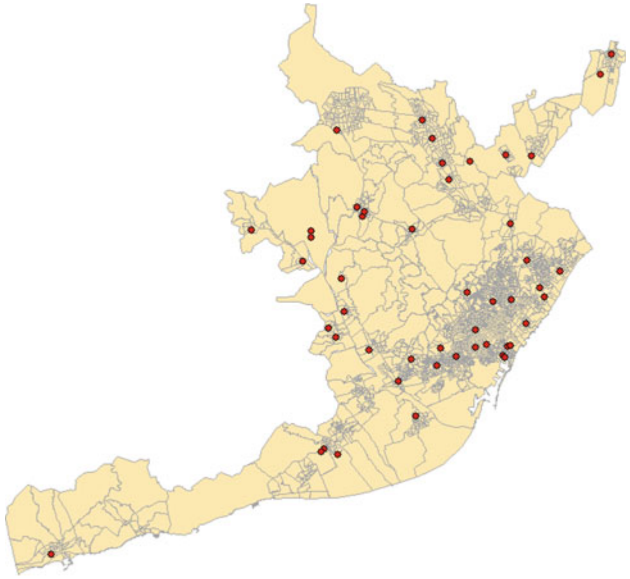

Fig. 1 The location of the stations over the Barcelona Metropolitan Region

Table 1 Available observations for each year

1998	1999	2000	2001	2002	2004
24	25	25	25	25	24
2005	2006	2007	2008	2009	
22	24	25	25	25	

Two methods have been applied in this research to model the concentrations of nitrogen dioxide: ordinary kriging (OK) [21] and ridge regression confidence machine (RRCM) [19]. To compare both methods, the concentrations have been predicted for each year at the stations where the observations were not available, so that for each year the data set would be completed up to 49 points. Of course, the estimates could be expanded on a grid to cover up the whole study region. This is planned to be done in the future when more data will be available, so that the predictive models, both kriging and RRCM, can be properly adjusted.

2 Methods

2.1 Conformal prediction and RRCM

Conformal prediction is a relatively new technique. Its main characteristic feature is that it can be used with any existing machine learning method. Conformal predictors have two major hallmarks: they are valid and effective. Validity here means that in the long run, the frequency of prediction errors does not exceed the given level of significance. An output of a conformal predictor is a prediction set. Efficiency here means that prediction sets should to be as small as possible. Validity is guaranteed, while efficiency is a quantitative value depending on the appropriate selection of an underlying method.

The particular specification of conformal prediction used in this study is called RRCM [19]. Ridge regression may be treated as an advance of the classical least squares approach aimed to deal with "ill-conditioned" situations when the correlations between independent variables in a model lead to necessity to invert an almost singular matrix. Ridge regression procedure suggests the introduction of a small *ridge coefficient* to the regression equation. The method was proposed in 1960s by Hoerl [6]. RRCM, together with the nearest neighbors regression (KNN), is the most standard regression algorithms in machine learning.

Let us briefly describe the method. Suppose there is $\mathbf{X} = \mathbb{R}^d$ called the *object* space, $Y = \mathbb{R}$ called the *label* space, and $\mathbf{Z} = \mathbf{X} \times Y$ is the *example* space. In other words, $\mathbf{z_i} = (\mathbf{x_i}, y_i)$. Here, the object space would describe the independent variables, or regressors, the label space would consist of the dependent variables, and the example set would be made up of the observations. Suppose, the observed data sequence is incomplete, i.e. there is a necessity to predict the missing data. Ridge regression computes the following minimization expression:

$$a \cdot ||\boldsymbol{\omega}||^2 + \sum_{i=1}^{n} (y_i - \boldsymbol{\omega} \cdot x_i)^2 \to \min, \tag{1}$$

with the aim to establish the optimal values of the weights $\boldsymbol{\omega}$. Here, $a \geq 0$ is a *ridge factor*. The least squares minimization algorithm is a special case of the ridge regression when the parameter a is equal to zero.

So far, there is no analytical procedure that would allow calculating the optimal value of a ridge factor. Thus, a feasible way to obtain such a value is simple brute force search. The nonconformity measure for the RRCM is the absolute value of the residuals, i.e. $\alpha_i := |e_i| = |y_i - \hat{y}_i|$. Let $\mathbf{X_n}$ be the $n \times p$ matrix of regressors, and $\mathbf{Y_n}$ be the vector of labels (dependent variables), then the minimization expression will take the form:

$$a||\boldsymbol{\omega}||^2 + ||\mathbf{Y_n} - \mathbf{X_n}\boldsymbol{\omega}||^2 \to \min_{\omega}, \tag{2}$$

or

$$\mathbf{Y_n'Y_n} - 2\boldsymbol{\omega}'\mathbf{X_n'Y_n} + \boldsymbol{\omega}'(\mathbf{X_n'X_n} + a\mathbf{I_p})\boldsymbol{\omega} \to \min_{\omega}. \tag{3}$$

Solving the equation, the optimal weights can be obtained:

$$\boldsymbol{\omega}^* = (\mathbf{X_n'X_n} + a\mathbf{I_p})^{-1}\mathbf{X_n'Y_n}. \tag{4}$$

The ridge regression approximation is provided by the expression:

$$\widehat{\mathbf{Y_n}} := (\hat{y}_1, \ldots, \hat{y}_n)' = \mathbf{X_n}(\mathbf{X_n'X_n} + a\mathbf{I_p})^{-1}\mathbf{X_n'Y_n}, \tag{5}$$

where the matrix

$$\mathbf{H_n} := \mathbf{X_n}(\mathbf{X_n'X_n} + a\mathbf{I_p})^{-1}\mathbf{X_n'} \tag{6}$$

is called the *hat matrix* (because it transforms Y_n into $\widehat{Y_n}$). Therefore, the vector of nonconformity scores is represented as follows: $\alpha_i = |e_i|$. Then, considering that an incomplete data set is given, the label y for x_n is unknown. The vector $Y_n = (y_1, \ldots, y_{n-1}, y)'$ can be split into two:

$$\mathbf{Y_n} = (y_1, \ldots, y_{n-1}, 0)' + (0, \ldots, 0, y)'. \tag{7}$$

Then,

$$\mathbf{A_n} = (\mathbf{I_n} - \mathbf{H_n})(y_1, \ldots, y_{n-1}, 0)', \tag{8}$$

and

$$\mathbf{B_n} = (\mathbf{I_n} - \mathbf{H_n})(0, \ldots, 0, y)'. \tag{9}$$

For each y, the expression $\alpha(y_i) - \alpha(y_n)$ changes sign only at some points for each $i = 1, \ldots, n$. The set of such points can be calculated instead of testing all possible y. The set:

$$\begin{aligned} S_n &:= \{y : \alpha_i(y) \geq \alpha_n(y)\} \\ &= \{y : |a_i + b_i y| \geq |a_n + b_n y|\} \end{aligned} \tag{10}$$

is the prediction set for each $i = 1, \ldots, n$. S_n can either be a ray, a union of two rays, an interval, a point, empty, or represented by the whole real line. With no loss of generality, we could assume $b_i \geq 0$. Then, if $b_i \neq b_n$, α_i and α_n are equal at two points:

$$-\frac{a_i - a_n}{b_i - b_n} \quad \text{and} \quad -\frac{a_i + a_n}{b_i + b_n} \tag{11}$$

then S_n is an interval, or a union of two rays.

The RRCM setting described above is referred to as the *primary setting*. However, ridge regression can only deal with situations when the number of parameters is relatively small. This algorithm implies inverting a $p \times p$ matrix which can be computationally difficult. In case of a high-dimensional problem, the so-called "kernel trick" [19] is used. For using this technique, it is essential to represent the ridge regression equation in the *dual form*. The duality is based on the following matrix equation:

$$\mathbf{X_n}(\mathbf{X_n'X_n} + a\mathbf{I_p})^{-1} = (\mathbf{X_nX_n'} + a\mathbf{I_n})^{-1}\mathbf{X_n}. \tag{12}$$

The prediction can thus be rewritten as:

$$\hat{y} = \boldsymbol{\omega} \cdot \mathbf{x} = \mathbf{Y_n'}(\mathbf{X_nX_n'} + a\mathbf{I_n})^{-1}\mathbf{X_n x}. \tag{13}$$

The main aspect of this representation is that the prediction depends on the objects $x_1, \ldots, x_n$ only through scalar product between them. The hat matrix for the dual representation will take the form:

$$\mathbf{H_n} = (\mathbf{X_n X_n'} + a\mathbf{I_n})^{-1}\mathbf{X_n X_n'} \tag{14}$$

or

$$\mathbf{H_n} = (\mathbf{K_n} + a\mathbf{I_n})^{-1}\mathbf{K_n}. \tag{15}$$

Here, $\mathbf{K_n}$ is the matrix with the elements $(\mathcal{K}_n)_{(i,j)}$. Scalar product in this equation can be substituted by any suitable *kernel*:

$$\mathcal{K}(x^{(1)}, x^{(2)}) = F(x^{(1)}) \cdot F(x^{(2)}), \tag{16}$$

where F is a function defined as:

$$F : \mathbb{R}^2 \rightarrow \mathcal{H}, \tag{17}$$

and $\mathcal{H}$ is referred to as *feature space*. For more information, please see [15].

In this research, the iid specification of the RRCM model has been used [20]. This specification implies that the data are independent and identically distributed. This model is a very convenient option for air pollution data, as it does not require any prior knowledge on the data distribution, apart from that the observations $(\mathbf{x_n}, y_n)$ should be iid. Both a standard (linear) model and a model with a non-linear kernel have been fitted. Here, the number of features is equal to 2, and there is no need to convert big matrices. However, there might be a need to try different approaches for an empirical variogram of spatial data. For those purposes, "kernel trick" can be used. A kernel that has been implemented in this work is the Gaussian radial basis function (RBF kernel) [15]:

$$\mathcal{K}(x^{(1)}, x^{(2)}) = \exp -\frac{||x^{(1)} - x^{(2)}||^2}{2a^2}, \tag{18}$$

where a is a scale parameter.

2.2 Kriging

Kriging is a classical interpolation method aimed to assess geographical data. It serves for conversion of spatial data into an estimate of a random field together with a measure of error or uncertainty [17]. The measure of uncertainty here is provided with the so-called *kriging variance*. This method was first introduced by a South African engineer D. G. Krige in his master thesis devotes to statistical methods to estimation of a mineral ore body in the 1950s [5]. Since then, kriging has been studied and broadened, and nowadays it serves as a generic name for a set of methods: simple kriging, ordinary and universal kriging, cokriging, etc. as well as a Bayesian approach to kriging [8]. Bayesian kriging can yield a smaller error variance than the "traditional" ordinary kriging. However, this gain in precision depends on whether the Bayesian specification of the model is actually reasonable. In practice, there are two major obstacles to implementation of Bayesian kriging: a correct specification of the prior distribution, and a computational complexity, although the latter problem can be solved by implementation of Monte Carlo methods [5]. The specification of the prior distribution can be avoided with the use of parametric bootstrap, like the trans-Gaussian Bayesian kriging model suggests [16]. This method can help relax the Gaussian assumption which is common for kriging, though not always appropriate.

Ordinary kriging is a convenient method that must be used when dealing with raw data with unknown mean [3]. This specification has been used in the present research. Suppose $x_1, \ldots, x_n$ are points of the spatial domain $\mathcal{D}$, and x_0 is an unobserved location. $Z(x)$ is a realization of a second-order stationary isotropic stochastic process with a known variogram. The term "variogram" is the cornerstone of kriging. It is defined as one-half the expected value of the squared difference between random variables $Z(x)$, $Z(x+h)$ at two different locations x and $x+h$, where h is a distance between two points [17]:

$$\mathrm{var}[Z(x+h) - Z(x)] = 2\gamma(h). \tag{19}$$

The actual value of $Z(x)$ is unobserved at the point x_0. Then:

$$Z^*_{OK}(x_0) = \sum_{i=1}^{n} \omega_i Z(x_i), \tag{20}$$

where ω_i are the *kriging weights* [21]. The estimates $Z^*_{OK}(x_0)$ are unbiased, which implies that the estimation error is nil on the average. The unbiasedness of the estimates is guaranteed by the condition:

$$\sum_{i=1}^{n} \omega_i = 1. \tag{21}$$

The unbiasedness constraint of ordinary kriging allows the variance (known as *kriging variance*) to be expressed in terms of a variogram [17]:

$$\begin{aligned}\sigma_E^2 &= \mathrm{var}(Z^*(x_0) - Z(x_0)) \\ &= -\gamma(x_0 - x_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} \omega_i\omega_j\gamma(x_i - x_j) \\ &\quad +2\sum_{i=1}^{n} \omega_i\gamma(x_i - x_0).\end{aligned} \tag{22}$$

For a more detailed specification, please see [21]. Kriging variance is an analog of a regression mean squared error, provided the variogram model is specified correctly [22].

Error variance for an unbiased linear kriging estimator can be expressed in terms of covariances between different $Z(x_i)$, $Z(x+h)$ at different spatial points x_i, (x_i+h) [17]. A covariance function is defined on the assumption the field is second-order stationary [21]:

$$E[Z(x) \cdot Z(x+h)] - m^2 = C(h), \tag{23}$$

where $E[Z(x)] = m$. A variogram can be derived from a covariance function as follows:

$$\gamma(h) = C(0) - C(h) \tag{24}$$

Covariance functions can be of different form, and they are extremely useful for spatial modeling, because they allow to consider spatial dependence and specific features of data distribution over a spatial region. Different types of covariance functions are being used in practical research. Two covariance functions are used here. The first one is the exponential covariance function:

$$C(h) = b \exp\left(-\frac{h}{\sigma}\right), \tag{25}$$

where b is the value at the origin, σ is a range parameter, and h is the distance between two spatial points. Another one is the Gaussian covariance function:

$$C(h) = b \exp\left(-\left(\frac{h}{\sigma}\right)^2\right). \tag{26}$$

Both of these functions belong to the Matérn family [5].

The situation when a variable is discontinuous at the origin of a varoigram is called "nugget effect". This means that the values of the variable change abruptly at a very small scale [21]. In practice, when sampling design implies one single measure at each of the locations, the nugget effect is usually attributed to measurement error or spatial variation at very small distances: smaller than the smallest separation between two sampling locations, or to a combination of the two [5]. The discontinuity at zero separation is usually introduced to parametric covariance models [17].

Error variances at unobserved locations are computed from chosen covariance models, and model parameters are obtained from the observed data [17]. Covariance and variogram estimates are thus random values under the assumed model, and so the error variance is also an estimate, and its properties depend on the model.

3 Results

For computational purposes, R [18] has been used. Kriging has been performed with the use of **geoR** package [13], in particular, employing its function *krige.conv*. As for conformal prediction models, the **PredictiveRegression** [20] package has been availed of. For RRCM modeling in the iid setting, use has been made of the function *iidpred* [20] from this package. For executing RRCM models with RBF kernel, an additional function has been created on the basis of the *iidpred* function. The modification implied rewriting the ridge regression procedure in the "dual form" and applying the "kernel trick" as introducing a relevant kernel.

At first, ordinary kriging with the exponential covariance function has been applied to the data. Models have been derived for each year, with the exception of 2003 for which the data were not available, so there were 11 models in total. Then, an RRCM model in the iid basic specification has been executed for the same data. Confidence level for prediction has been set to 95 %. Ridge factor has been set to 0.01. It is noteworthy that ridge regression suggests scaling and normalization of the independent variables [6], which is though not always necessary. However, here the data have been scaled and normalized. It has been done for purely computational purposes. For precise comparison of the two approaches, ordinary kriging and RRCM, scaled and normalized data have been used for both of them.

For each year, RRCM output intervals contained the kriging predicted values. Naturally, it would be desirable to contrast the predictions to actual observations. However, observed values were not available for the points at which the predictions were made. To track validity of predictions, leave-one-out cross-validation was further performed for the observed points. A mean predicted concentration was considered as a measure for evaluation of the kriging estimate. Also, taking into account the feature of kriging to output the estimation variance together with the result itself, mean variance was also tracked. As for RRCM predictions, they are sets in the form of intervals. These have been evaluated in terms of efficiency, i.e. their size: the narrower the better. This comparison aimed to conclude the following: whether the intervals for a given year withheld the kriging predictions or not. A data set must count just over 20 observations to assure the confidence of prediction equal to 95 %. A data set must count $1/\epsilon$ observations or more, where ϵ is the level of significance of prediction, to provide than an iid model would yield informative prediction intervals [20]. As the chosen confidence level was equal to 95 %, 5 % of errors was allowed. It is possible that there were no such errors: all of the RRCM intervals contained the kriging predictions. The results of the comparison are depicted in Fig. 2. There, the central curve represents mean ordinary kriging predicted concentrations for each year. The upper curve shows mean values of the upper bounds of prediction sets for each year. The lower curve, on the other hand, represents averages of the lower bounds of prediction intervals for each year.

Assuming that the Gaussian model could probably be a good fit for the given data, kriging with a Gaussian covariance function has been performed. Figure 3 represents how exponential and Gaussian variomodels approach the empirical variogram for 1998 data. Such plots have been processed for all the years of study.

An RRCM model with the Gaussian RBF kernel has been also used, and the results of both methods have been compared. The aim of testing various RRCM approaches was to find one that would give the most effective intervals.

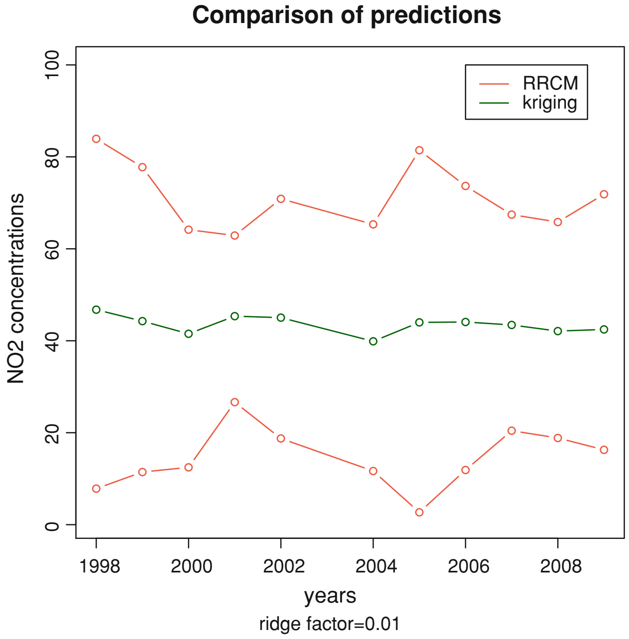

Fig. 2 Comparison of mean predictions: OK with exponential covariance and RRCM in default setting

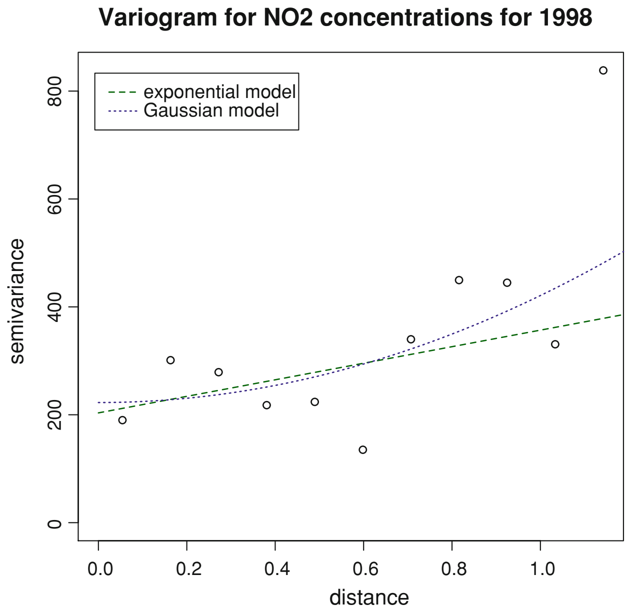

Fig. 3 Comparison of fit of exponential and Gaussian variomodels for 1998

One factor to vary here is the kernel, while another one is the ridge factor. The latter is explained in detail below. The results of prediction with kriging and RRCM with the RBF kernel are shown in Fig. 4. For each year, the central curve represents mean kriging predicted concentrations, while the upper and the lower curves show average upper and lower bounds of RRCM RBF intervals. Ridge factor and confidence level were equal to 0.01 and 95 %, respectively. In these models, since the data have been scaled and normalized, range parameter in the RBF kernel has been set to 1.

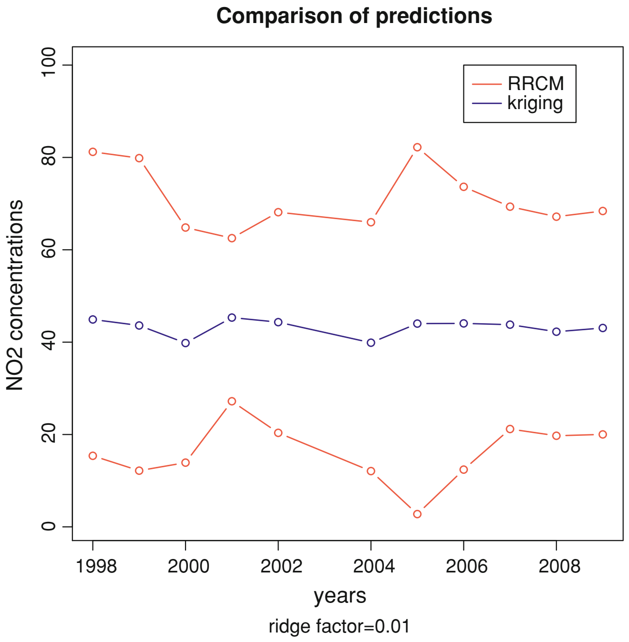

Fig. 4 Comparison of mean predictions: OK with Gaussian covariance and RRCM with RBF kernel

Table 2 Mean RRCM interval width

Year	Dot product	RBF
1998	76.08	65.82
1999	66.31	67.68
2000	51.69	50.91
2001	36.24	35.32
2002	52.12	47.78
2004	53.65	53.89
2005	78.75	79.44
2006	61.78	61.24
2007	47.00	48.15
2008	46.96	47.44
2009	55.59	48.38

Table 2 shows mean widths of RRCM prediction intervals. The results are shown for both standard linear model and the one with the RBF kernel. It is seen that the implementation of the Gaussian RBF kernel provides a slightly better result.

As for kriging variances, they are huge. Those values are estimates, and due to small number of observations available, it is hard to come up with an accurate variogram model. Table 3 depicts the estimated ordinary kriging variances for both models taking up an exponential covariance function and a Gaussian one.

Some words should be said on the role of a ridge factor in RRCM modeling. There is no general rule of choice for it.

Table 3 Mean kriging variances

Year	Exponential	Gaussian
1998	242.02	239.81
1999	260.13	249.73
2000	66.73	88.35
2001	26.12	52.12
2002	73.05	97.63
2004	129.87	129.87
2005	171.83	173.41
2006	136.64	140.81
2007	86.30	95.87
2008	71.18	86.96
2009	64.28	88.19

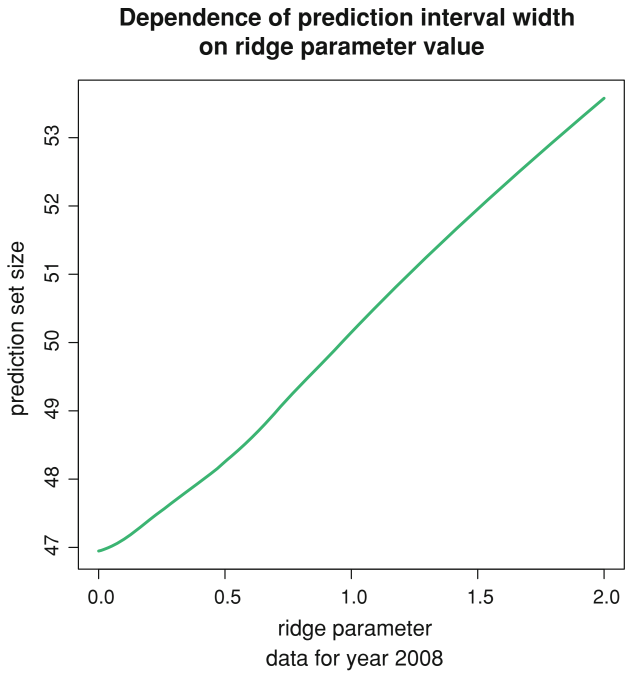

Fig. 5 Mean width of prediction set for a given ridge factor

In both of the models demonstrated here, ridge factor has been set to 0.01. In fact, the value of ridge factor influences the efficiency of RRCM prediction. For some values, prediction sets are smaller than for the other ones. Plotting ridge regression parameter estimates against ridge factor values is called *ridge trace* [6]. Figure 5 demonstrates an example of how an output of a RRCM predictor varies for different values of ridge factor. Mean annual concentrations of nitrogen dioxide have been modeled for 2008, for a sequence of ridge factors from 0 to 2 with a step equal to 0.01. It is noteworthy that here it is not the parameter estimates that have been plotted against the ridge factor values, but the actual output given by mean size of prediction intervals. The latter one is the average difference between upper and lower bounds of the prediction sets for each given spatial point. It is evident from the plot that the optimal value of ridge factor for these data is zero, i.e. the best predictor is the least squares predictor. However, optimal values are different for different years. It is a general practice to use the brute force method to choose the optimal value of a ridge factor for each study data set upon which RRCM predictor is planned to be executed.

Table 4 Comparison of models with various ridge factors

Ridge	0	0.01	0.1	1	1.5	2
Dot prod.	56.93	56.93	56.94	58.23	59.25	60.31
RBF	–	55.10	55.46	57.12	58.29	59.39

For all the years in the data set, Table 4 depicts the results of modeling for the following ridge factors: 0, 0.01, 0.1, 1.5, 1, 2. The results are provided for both of the RRCM settings used in this research: iid and Gaussian RBF. For every model, each cell in the Table 4 stands for mean width of prediction intervals for all the years for a given value of ridge factor. In other words, for every chosen ridge factor, mean prediction interval width for each of the years has been evaluated, and then those values have been further averaged for all of the years. This has been done to come up with the mean prediction interval width for each ridge factor for the whole data set. It is noteworthy that the computation of a least squares estimate (i.e., with a ridge factor equal to 0) is not possible for RRCM with the RBF kernel, as it implies inversion of a singular matrix. Taking into consideration the last notion, it is seen from Table 4 that an almost least squares estimate (i.e., the one with a ridge factor almost equal to 0) is on the average the best one for both models.

This research compares two techniques aimed to predict the data at unobserved locations. As the actual values are unavailable for the locations, goodness of fit of the models cannot be fully explored. To test the models on the observed data, leave-one-out (LOO) cross-validation [2] has been performed for each year. Table 5 shows an example of application of LOO cross-validation for a kriging model (with an exponential covariance function) for the data referring to 1999.

Table 6 shows the results of LOO cross-validation for a standard RRCM iid linear model for the same 1999 data. Ridge factor has been taken equal to 2. Cross-validation of RRCM procedure for this particular data set has revealed one error: for one observation (object), the actual value of the variable of interest (mean annual concentration of NO2) does not fall within the prediction interval. Out of 25 observations, 1 incorrect prediction makes up 4 % of errors, which is acceptable, since the chosen confidence level is 95 %.

Cross-validation for ordinary kriging models with a Gaussian covariance function and RRCM models with the RBF kernel were performed as well for each year.

Table 5 OK leave-one-out cross-validation

Observed	Predicted	Absolute error
17.00	47.42	30.42
20.00	45.42	25.42
22.00	44.38	22.38
25.00	48.10	23.10
28.00	47.25	19.25
30.00	47.44	17.44
32.00	45.29	13.29
33.00	44.46	11.46
37.00	43.92	6.92
38.00	42.19	4.19
38.00	47.02	9.02
44.00	44.95	0.95
45.00	44.86	0.14
48.00	44.25	3.75
50.00	44.09	5.91
51.00	43.00	8.00
52.00	43.75	8.25
53.00	44.42	8.58
57.00	43.68	13.32
61.00	44.14	16.86
61.00	44.10	16.90
64.00	43.58	20.42
65.00	41.37	23.63
68.00	43.70	24.30
71.00	43.77	27.23

Table 6 RRCM leave-one-out cross-validation

Observed value	Lower bound	Upper bound	Interval width
17.00	12.96	71.47	58.51
20.00	−1.74	75.54	77.28
22.00	8.95	69.69	60.74
25.00	10.22	69.93	59.71
28.00	11.26	70.94	59.68
30.00	11.41	71.68	60.27
32.00	11.09	71.29	60.20
33.00	10.20	70.18	59.98
37.00	9.16	69.96	60.80
38.00	6.52	69.51	62.99
38.00	12.21	75.82	63.61
44.00	7.72	69.55	61.82
45.00	10.13	71.36	61.23
48.00	10.62	75.86	65.23
50.00	9.23	71.58	62.35
51.00	7.47	70.56	63.09
52.00	5.13	69.12	63.99
53.00	9.98	72.07	62.09
57.00	9.02	71.02	62.00
61.00	7.03	70.20	63.17
61.00	8.26	71.84	63.58
64.00	9.03	72.25	63.22
65.00	4.74	68.18	63.44
68.00	5.80	70.12	64.32
71.00	8.62	68.87	60.25

4 Conclusion

Nowadays, a growing traffic burden together with other related factors imposes the problem of air pollution assessment. Since geographical data are involved, using geostatistical methods is a regular practice, and kriging is among the most widely used geostatistical techniques. It is a well-developed algorithm. However, some limitations of kriging are known and described elsewhere. One of them is that kriging predictions lack confidence. The method is known for providing a measure of uncertainty, i.e., kriging variance, together with the estimates at unobserved locations. However, kriging variance is an estimate itself. A new approach of conformal predictors, and, in particular, the RRCM, can be treated as a good alternative to kriging. Having a similar regression underlying algorithm, it yields valid prediction intervals. Also, RRCM implies confidence prediction with no prior assumption on data distribution, apart from being iid.

This research has taken up both kriging and conformal predictors. Kriging's main adjustment tool is the covariance function. It allows to take into account covariances between observations at different spatial points. Choosing the right shape of covariance function for the data is important, as model parameters are obtained from it. In RRCM, a kernel can be seen as an instrument similar to covariance function. In this research, a Gaussian kernel approach has been used. It has been guessed, since the data are geographical that the character of spatial dependence is probably not linear. Also, the given data set represents annual concentrations of a substance suspended in the air and carried with the wind. It is very hard to derive a fitting distribution for such a data set. It is important to mention that the data set is small, so the task of an effective prediction gets very complex. However, the implementation of the described methods gives good results with these data, so they would most probably work well with wider data sets.

RRCM intervals could be narrowed, so an increase in the efficiency of prediction is the aim. Regarding that the data set is small and not so easy to process, the RBF kernel has been used without particular justification. There should be better kernels available which can be established with more trials, and/or more data.

The results of kriging and RRCM coincide in terms of RRCM intervals withholding kriging predictions. Generally, for each specification of kriging the results should match, provided that the kriging covariance function is taken up as a kernel in RRCM and vice versa, and provided the ridge factor is the same.

5 Discussion

This research has two major points. First is that machine learning methods, and, in particular, RRCM can be used for air pollution assessment, and, as a consequence, for relevant clinical research. As confidence of the prediction means a lot in epidemiological studies, because those are dealing with health and well-being of people, the use of those newly developed methods can yield valid and effective estimates. Second is that in geostatistics, RRCM can be seen as a complementary method to the classical kriging. Kriging methods are known to be largely based on the assumption that the underlying stochastic process is Gaussian [12]. If a relaxation of the Gaussianity assumption is the aim, an RRCM iid predictor might come in handy as it only assumes that the data are iid.

This study marks some directions for future research. First of all, models should be tested on larger data sets, which it is planned to do. Second, implementing of other covariance functions and kernels can be tested. Third, a Bayesian conformal predictor can be considered and compared to Bayesian kriging. Finally, as said before, a conformal predictor can be built upon almost any underlying algorithm. Using the underlying algorithm to obtain a non-conformity measure (or strangeness measure), a conformal predictor delivers confidence to the prediction. This means that not only kriging can be used as an underlying algorithm for this particular research, but any suitable regression technique, such as land-use regression [7,9], can serve this purpose.

Acknowledgments We would like to thank Dr. Evgeny Ivin from Lomonosov Moscow State University for his helpful advice on regression related issues, Matthias Birk from Helmholtz Zentrum München for the well-developed GIS map of the study region, Dr. Alexey Chervonenkis from Royal Holloway University of London for his counseling on geostatistics and Dr. Marc Saez from the University of Girona for his valuable input regarding air pollution assessment.

References

1. Barceló, M., Saez, M., Saurina, C.: Spatial variability in mortality inequalities, socioeconomic deprivation, and air pollution in small areas of the Barcelona Metropolitan Region, Spain. Sci. Total Environ. **407**(21), 5501–5523 (2009)
2. Cawley, G., Talbot, N.: Fast exact leave-one-out cross-validation of sparse least-squares support vector machines. Neural Netw. **17**(10), 1467–1475 (2004)
3. Chilès, J., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty. Wiley Series in Probability and Statistics. Wiley, New York (2009)
4. Cyrys, J., Hochadel, M., Gehring, U., Hoek, G., Diegmann, V., Brunekreef, B., Heinrich, J.: GIS-based estimation of exposure to particulate matter and NO2 in an urban area: stochastic versus dispersion modeling. Environ. Health Perspect. **113**(8), 987–992 (2005)
5. Diggle, P., Ribeiro, P. Jr.: Model-Based Geostatistics. Springer, Berlin (2007)
6. Draper, N., Smith, H.: Applied Regression Analysis. Wiley, New York (1981)
7. Gilbert, N., Goldberg, M., Beckerman, B., Brook, J., Jerrett, M.: Assessing spatial variability of ambient nitrogen dioxide in Montreál, Canada, with a land-use regression model. J. Air Waste Manag. Assoc. **55**(8), 1059–1063 (2005)
8. Handcock, M.S., Stein, M.L.: A Bayesian analysis of kriging. Technometrics **35**(4), 403–410 (1993)
9. Jerrett, M., Arain, A., Kanaroglou, P., Beckerman, B., Potoglou, D., Sahsuvaroglu, T., Morrison, J., Giovis, C.: A review and evaluation of intraurban air pollution exposure models. J. Expo. Anal. Environ. Epidemiol. **15**, 185–204 (2005)
10. Lee, P., Talbott, E., Roberts, J., Catov, J., Sharma, R., Ritz, B.: Particulate air pollution exposure and C-reactive protein during early pregnancy. Epidemiology **22**(4), 524–531 (2011)
11. Organization, W.H.: Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide. EURO Nonserial Publication. World Health Organization (2006)
12. Pilz, J., Spöck, G.: Why do we need and how should we implement Bayesian kriging methods. Stoch. Environ. Res. Risk Assess. **22**(5), 621–632 (2008)
13. Ribeiro, P. Jr., Diggle, P.: geoR: a package for geostatistical analysis. R-NEWS **1**(2) (2001)
14. Samoli, E., Aga, E., Touloumi, G., Nisiotis, K., Forsberg, B., Lefranc, A., Pekkanen, J., Wojtyniak, B., Schindler, C., Niciu, E., Brunstein, R., Dodic Fikfak, M., Schwartz, J., Katsouyanni, K.: Short-term effects of nitrogen dioxide on mortality: an analysis within the APHEA project. Eur. Respir. J. **27**(6), 1129–1138 (2006)
15. Schölkopf, B., Smola, A.: Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
16. Spöck G., Kazianka, H., Pilz, J.: Bayesian trans-Gaussian kriging with log-log transformed skew data. In: Pilz, J. (ed.) Interfacing Geostatistics and GIS, pp. 29–43. Springer, Berlin (2009)
17. Switzer, P.: Kriging. In: Encyclopedia of Environmetrics. John Wiley & Sons, Ltd (2006)
18. Team, R.D.C.: R: A language and environment for statistical computing (2011). http://www.R-project.org
19. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer, Berlin (2005)
20. Vovk, V., Nouretdinov, I., Gammerman, A.: On-line predictive linear regression. Ann. Stat. **37**(3), 1566–1590 (2009)
21. Wackernagel, H.: Multivariate Geostatistics: An Introduction With Applications. Springer, Berlin (2003)
22. Webster, R., Oliver, M.: Geostatistics for Environmental Scientists. Wiley, New York (2007)
23. Zimmerman, D., Pavlik, C., Ruggles, A., Armstrong, M.: An experimental comparison of ordinary and universal kriging and inverse distance weighting. Math. Geol. **31**, 375–390 (1999)
24. Zou, B., Wilson, J.G., Zhan, F.B., Zeng, Y.: Air pollution exposure assessment methods utilized in epidemiological studies. J. Environ. Monit. **11**, 475–490 (2009)