The choice of optimal lag for Kriging interpolation of NWP model forecast

Lesya Mykolaivna Katsalova, Vitalii Mykhailovych Shpyg
Ukrainian Hydrometeorological Institute, Prospekt Nauki 37, Kyiv, Ukraine, e-mail: vitold82@i.ua

Abstract. In this paper a Kriging method is reviewed and a way of its application in numerical weather prediction is proposed. The basic principles of the Kriging are shown; the main advantage is its accuracy, but at the same time a disadvantage is its large computational complexity. The construction stage of the variographical model is highlighted, as it is the most important stage and has a significant impact on the accuracy of interpolation. The algorithm for the construction of the variographical model is described. Special attention is paid to averaging an experimental variogram by introducing a special interval, called "lag". Precisely this issue, according to the authors has a significant impact on the effectiveness of the practical application of Kriging for the interpolation of meteorological parameters. The advantages of averaging an experimental variogram by the administration of lag are presented, and the error that arises in this case is estimated. A theoretical study for the determination of the optimal lag was conducted. The lag proposed for the determination is guided by the criteria of accuracy and the economy of computer time. The two-criteria problem is solved, and the formula, which makes it possible to determine the optimal lag on these criteria, is received. An example shown here is the application of the obtained results for solving the applied task associated with meteorological parameters forecast by the COSMO model.

Keywords: Kriging, interpolation, variography, experimental variogram, lag

Submitted 22 October 2015, **revised** 29 March 2016, **accepted** 13 July 2016

1. Introduce

Interpolation – a method for the estimation of intermediate values of spatially distributed variables on a discrete set of known values. There are many interpolation methods (Samarsky, Gulin 1989; Samarsky 1982; Roache 1985; Stein 1999; Will et al. 2015), but their effectiveness is largely dependent on the assigned task and distribution of the input data (Gunes et al. 2006; ArcGIS Resources 2013). For the best evaluation in terms of statistics, the Kriging method was used– its evaluation has the minimum variation of error (Matheron 1967; Oliver 1990). An important property of this method is the reproduction of the exact values in the measured nodes.

The initial and most important stage of Kriging is variography, which means the selection of the variogram on a set of known values (Kanevskiy et al. 1999). A variogram is a theoretical model that defines the distribution of squared differences of values (correlation) depending on their relative position, but not depending on their absolute position. In the first variographical stage, an experimental variogram is built according to known values, that is, a sequence of pairs: the distance between two points, and the value of the variogram for this distance. Because of the irregular placement or very large number of the input nodes, the experimental variogram may contain a very large number of such pairs, leading to significant complications with the selection of the theoretical model. A large number of points on the experimental variogram may also increase the interpolation error.

The conventional way to solve this problem is the introduction of a certain interval, known as lag (Koshel, Musin 2000; Katsalova, Shpyg 2013). Lag divides the axis of distances into new intervals, on which, are the averaged values of the experimental variogram.

The size of the lag is generally determined on the basis of numerical experiments, depending on the criteria governing the solution of specific problems. Among the scientific literature on the Kriging method, the authors have not come across work in which the choice of the interval was structurally substantiated.

In this work the authors present a study of the selection of the interval according to the criteria of the accuracy of the experimental variogram and the optimisation of the determination of the time of a mathematical model for the next phase of Kriging interpolation.

2. Kriging – basic principles

Suppose we have the value of a variable $v(x)$ in the points $x_1, x_2,\ldots, x_N$: $v(x_1)$, $v(x_2),\ldots, v(x_N)$. It is necessary to assess the value of v at a point x^*. We apply an ordinary Kriging method.

The estimation of the value v at point x^* defined as a linear combination of the known values of the weights $w_1, w_2,\ldots, w_N$ is:

$$v\left(x^*\right)=\sum_{i=1}^{N} w_i v(x_i) \tag{1}$$

Ratios w_i are calculated from equation:

$$\begin{pmatrix} w_1 \\ \ldots \\ w_N \\ \mu \end{pmatrix} = A^{-1} \begin{pmatrix} D\left(x^*, x_1\right) \\ \\ D\left(x^*, x_N\right) \\ 1 \end{pmatrix} \tag{2}$$

where:

$$A=\begin{pmatrix} D(x_1,x_1) & \ldots & D(x_1,x_N) & 1 \\ \ldots & \ldots & \ldots & \ldots \\ D(x_N,x_1) & \ldots & D(x_N,x_N) & 1 \\ 1 & \ldots & 1 & 0 \end{pmatrix};$$

μ is a Lagrange multiplier; $D(x_i,x_j) = D(\rho_\kappa)$; $\rho_\kappa = |x_i - x_j|$, $i,j \in \{1, 2,\ldots, N\}$, $\kappa \in \{1, 2,\ldots, 0.5N(N + 1)\}$ is a theoretical variogram, that is, a continuous function describing the spatial continuity of spatially distributed data.

From this system are obtained (2) weights $w_1, w_2,\ldots, w_n$ which substitute in the equation (1) and produce the evaluation $v\left(x^*\right)$.

3. Using the lag in the construction of the experimental variogram

As mentioned above, the important role the accuracy of Kriging plays in the variographical model (DeMers 1999) reflects how the mutual influence of the values of the variable at the nodes changes depending on the changes of distance between them.

The variographical model is built on the basis of an experimental variogram which is a sequence of pairs $(\rho,\gamma(\rho))$, ρ is the distance between the two input nodes (Kanevskiy et al. 1999):

$$\gamma(\rho)=\frac{\sum\limits_{i,j:|x_i-x_j|=\rho}((v(x_i)-v(x_j))^2)}{K} \tag{3}$$

where K is the number of pairs of input points, and the distance between them is ρ. Often there are problems in which the number K is very large; in this case the variogram modeling is a complex and resource-intensive task (Armstrong 1984).

That is why the concept LAG, an interval along the axis distances ρ, is introduced, and why the variogram is calculated in the following way:

- $\rho_{max} = \max\limits_{i,\,j \in 1,\ldots,\,N} |x_i - x_j|$ divided at equal intervals of points h_m:$h_0 = 0$, $h_m = h_{m-1} + LAG$, $m \in \{1, 2,\ldots, M\}$;
- at each point h_m the value of the smoothed experimental variogram is calculated by the equation:

$$\gamma(h_m)=\frac{\sum\limits_{i,j:h_m-LAG<|x_i-x_j|\le h_m}((v(x_i)-v(x_j))^2)}{k_m}$$

for all $i,j \in \{1, 2,\ldots, N\}$, $h_{m-1} < |x_i - x_j| \le h_m$, k_m is the number of pairs of points, the distance between them falls into the interval distances $(h_{m-1}, h_m]$.

Obviously, the smoothed experimental variogram obtained by the described principle will be smoother, but simpler for modeling. This variogram will be called the smoothed experimental variogram.

The advantage of this approach is easy to see, defining the computational complexity of calculating the coefficients of theoretical models by least squares.

Suppose we have N input points and the distance between them is unique, then we get $0.5{*}N(N+1)$ pairs $(\rho, \gamma(\rho))$ in the experimental variogram. The number of machine operations that must be carried out to find the coefficients of the model by least squares equals $0.5(6+A_1)$ $N(N+1)$, where A_1 is the number of operations of the calculation of values of theoretical variogram, and $2 \le A_1 \le 26$ for models considered by the authors (Katsalova, Shpyg 2014). Obviously, the computational complexity of calculating the coefficients of the model will be $O(N^2)$.

Let us introduce $LAG = h_{min}$, where $h_{min} = \min\limits_{i,\,j \in 1,\ldots,\,N} |x_i - x_j|$. The smoothed variogram will be presented in sequence pairs $(h_m, \gamma(h_m))$, $m \in \{1, 2,\ldots, M\}$, $M \le N$. Then the number of machine operations in the calculation of coefficients of the model of the least squares method equals $(6+A_1)$ M and computational complexity will be not more than $O(N)$. Obviously, finding the coefficients of the theoretical model, which is based on the smoothed experimental variogram, is an order of magnitude faster than it is based on the original.

Explore the error of averaging of the experimental variogram.

Let the initial experimental variogram be represented by a set of pairs $(\rho_k, \gamma(\rho_k))$, $\rho_k \in [0, \rho_{max}]$, $k = 0, 1,\ldots, 0.5N(N+1)$. Suppose that the experimental model is described by some continuous and differentiable function f, that is $f(\rho_k) = \gamma(\rho_k)$ and its derivative are bounded on $[0, \rho_{max}]$. We introduce into the computational domain $[0, \rho_{max}]$ the grid:

$$\omega=\left\{h_m : h_o=0,\ h_m=h_{m-1}+LAG,\ m=1, 2,\ldots, M,\ M=\left[\frac{\rho_{max}}{LAG}\right]_{round}\right\}$$

When building the smoothed models, the function $f(\rho)$ was replaced by piecewise linear function $P(\rho)$:

$$P(\rho)=\begin{cases} f(0),\ \rho=0, \\ P(h_m)=\dfrac{1}{k_m}\sum f(\rho_k),\rho_k\in[h_{m-1},h_m],\mathrm{k}=1,2,...,\ k_m,\sum k_m=0.5N(N+1) \\ P(\rho)=P(h_{m-1})+\dfrac{P(h_m)-P(h_{m-1})}{h_m-h_{m-1}}(\rho-h_{m-1}),\ h_{m-1}<\rho<h_m \end{cases}$$

Then $f(\rho) = P(\rho)+R(\rho)$, $R(\rho)$, is the smoothing error. Since we are interested in estimating the error R only in the nodes of the grid ω, it is enough to consider:

$$R(h_m)=f(h_m)-\frac{1}{k_m}\sum_{k=1}^{k_m}f(\rho_k)=\frac{1}{k_m}\sum_{k=1}^{k_m}\left(f(h_m)-f(\rho_k)\right)$$

According to Lagrange's theorem:

$$R(h_m)=\frac{1}{k_m}\sum_{k=1}^{k_m}f'(\rho)(h_m-\rho_k)\leq\frac{\max\limits_{\rho\in(h_{m-1},h_m]}|f'(\rho)|}{k_m}\sum_{k=1}^{k_m}(h_m-\rho_k)\leq \max_{\rho\in(h_{m-1},h_m]}|f'(\rho)|\cdot LAG \quad \forall h_m\in\omega$$

Summing the last inequality to the entire domain, we obtain the estimation:

$$R(\rho)\leq\bar{R}=\max_{[0,\rho_{max}]}|f'(\rho)|LAG=A_2LAG \tag{4}$$

In which $f(\rho_k)=\gamma(\rho_k)$:

$$A_2=\max_{[0,\rho_{max}]}|f'(\rho)|\approx\max_{\rho_k\in[0,\rho_{max}]}\frac{|\gamma(\rho_{k+1})-\gamma(\rho_k)|}{(\rho_{k+1}-\rho_k)} \tag{5}$$

The assessment (4) indicates that the error of the averaging is limited and depends on the averaging interval.

The next numerical experiment illustrates the obtained mathematical evaluations.

Consider the COSMO model forecast for the estimated area from 44.5° to 52.5° north and from 22° to 40° east with a spatial step of 0.5. Interpolation of the forecast of temperature on the grid of Ukrainian weather stations by the Kriging method based on Gaussian variogram was performed. The theoretical model is obtained on the basis of the appropriate experimental variogram and on the basis of the smoothed experimental variograms with different averaging intervals.

The experimental variogram of temperature forecast in the said computational domain consists of 410 pairs $(\rho,\gamma(\rho))$ and is shown in Fig. 1a. The same figure shows the Gaussian model, built by the least squares, based on experimental variogram of forecast data for 03 UTC 08 April 2012.

For comparison, it is worth mentioning that experimental values of the variogram of the forecast of meteorological parameters obtained in UHMI using the COSMO model consist of 10972 pairs $(\rho,\gamma(\rho))$. It is clear that the construction of the variographical model by the least squares on such a large array of data takes a lot of computer time.

Based on the initial experimental variogram, the smoothed experimental variograms with lags of 0.5°, 1°, 2°, 4° respectively were constructed. These variograms were used for obtaining Gaussian models for each lag. In Fig. 1b are shown: the smoothed variogram for *LAG* = 0.5°, the Gaussian model that is based on this experimental variogram.

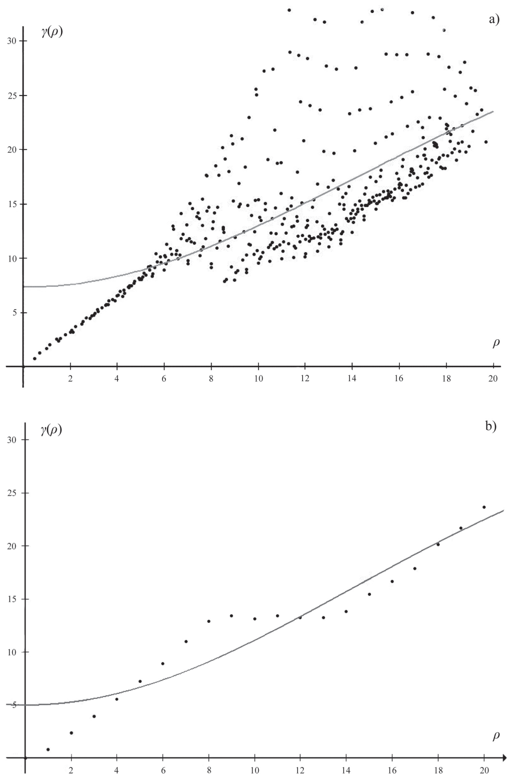

Fig. 1. a) Experimental variogram of COSMO forecast for 03 UTC 08 April 2012; Gaussian model is based on this experimental variogram; b) Smoothed experimental variogram (*LAG* = 1°); Gaussian model is based on this experimental variogram

Mean-square errors of the variographical models based on primary and smoothed experimental variograms shown in Table 1. Mean-square errors between observed and calculated temperature (°C) obtained through Kriging interpolation of the COSMO model forecast from a model (regular) grid to the Ukrainian meteorological stations network (irregular grid) are also represented in Table 1 (see σ of interpolation).

From Table 1, we can see that the mean-square errors for the initial experimental variogram are somewhat larger than the mean-square errors for the smoothed experimental variogram with *LAG* = 0.5°, 1°, 2°. This effect occurs because a large number of points leads to increased errors of variography, and consequently interpolation. For the smoothed experimental variograms, the mean-square

Table 1. Comparison of variography (Gaussian model) and Kriging interpolation that are conducted on the basis of the initial experimental variogram and smoothed experimental variograms with different steps of averaging

	Without averaged	$LAG = 0.5°$	$LAG = 1°$	$LAG = 2°$	$LAG = 4°$
number of pairs $(\rho, \gamma(\rho))$	410	41	21	11	5
σ of variography	4.4406	2.2498	2.2763	2.2886	5.6307
σ of interpolation	3.3016	2.1907	2.1905	2.1960	2.23578

errors of variography and Kriging interpolation increase with the increase in *LAG*, but the increase in interpolation error is less significant than the variography error. At the same time, the number of nodes of the smoothed experimental variogram is significantly less than the number of nodes of the initial experimental variogram which, of course, also significantly reduces the computing time of interpolation. The numerical experiment fully confirmed the theoretical research.

4. Determining the optimal lag

Consider the problem of the choice of lag which would provide the optimal ratio of accuracy of correlation data and cost of computing time.

Suppose $LAG = m \cdot h_{min}$, where $m = 1, 2,\dots, M$:

$$M = \left[\frac{\rho_{max}}{h_{min}}\right]_{round} \tag{6}$$

$[]_{round}$ is rounded to an integer.

It is shown above that the time required to obtain the coefficients of the theoretical variogram will be equal to $(6 + A_1)M\tau$ at $m = 1$, where τ is the time required by computers to perform one operation.

Introduce a function that will reflect changing the computing time needed for the calculation of coefficients depending on the change of lag:

$$T(m) = (6 + A_1) \cdot \frac{M}{m} \tag{7}$$

The function of the change of error smoothing in the experimental variogram depending on the value of lag is written using (4):

$$\overline{R}(m) = A_2 h_{min} m \tag{8}$$

The value m that allows the determination of the optimal *LAG* is obtained from a two-criteria problem:

$$T(m) \rightarrow min \tag{9}$$

$$\overline{R}(m) \rightarrow min \tag{10}$$

$$m = 1, 2,..., M$$

Solving the problem (9), (10) is as follows:

- replace two criteria with one:

$$F(m) = \alpha T(m) + (1-\alpha)\,\overline{R}(m) \rightarrow min \qquad m = 1, 2,..., M \tag{11}$$

where α, $\alpha \in (0,1)$ is a parameter that lets you set the priority of criteria (9), (10) in determining the optimal lag;

- find the minimum point of the function $F(m)$:

$$F(m) = \alpha(6 + A_1) \cdot \frac{M}{m} + (1-\alpha) A_2 h_{min} m$$

$$F'(m) = -\alpha(6 + A_1) \cdot \frac{M}{m^2} + (1-\alpha) A_2 h_{min} = 0$$

$$F''(m) = \alpha(6 + A_1) \cdot \frac{2M}{m^3} > 0$$

wherefrom:

$$m_{opt} = \sqrt{\frac{\alpha(6 + A_1)M}{(1-\alpha) A_2 h_{min}}} \tag{12}$$

Then, without limiting the generality, we can state that the optimal lag is found from the equation:

$$LAG_{opt} = \left[\sqrt{\frac{\alpha(6 + A_1)M}{(1-\alpha) A_2 h_{min}}}\right]_{round} h_{min} \tag{13}$$

where: h_{min} is the minimum distance between input points; $M = \left[\frac{\rho_{max}}{h_{min}}\right]_{round}$; ρ_{max} is the maximum distance between input points; A_1 is the number of operations in the model for which ratios are calculated; A_2 is calculated from equation (5).

Thus equation (13) makes it possible to calculate lag, which would provide the optimal ratio of accuracy of the spatial distribution of data and the cost of computer time to determine the coefficients of the variographical model. In determining the optimal lag, the priority of criteria (9), (10) can change depending on the conditions and objectives of a specific task.

5. Determining the optimal lag for data forecast model COSMO

Let us show how the eq. (13) operates with a practical example. Consider the interpolation of the meteorological parameters of the COSMO model forecast (Doms 2013) on the grid of Ukrainian meteorological stations by the Kriging method.

The estimated grid of the COSMO model for the territory of Ukraine is an array of 21 109 points (De Morsier et al. 2015), in which every 3 hours forecasts are obtained for 17 meteorological parameters with a forecast period of 78 hours. It is clear that the processing of such a significant amount of data requires considerable computer time. The reduction of the time for the post-processing of the model is an extremely important issue. This is why the need for a smooth experimental variogram for during Kriging interpolation of data from COSMO forecasts is apparent, and, therefore, there is a problem in determining the optimal lag for a smooth variogram.

Thus, we have a set of input data, which is a set of numeric values of some meteorological parameter in 21 109 nodes of a two-dimensional computational grid. The distance between nodes $h = h_{min} = 0.125°$. The grid covers the area from 42.5° to 55° north, and 17° to 43°east, i.e. the maximum distance between two points. Hence, according to (6), $M = 231$.

For Kriging interpolation of various parameters, different theoretical models are used (Katsalova, Shpyg 2014). To illustrate the determination of the optimal lag, the surface pressure data of the COSMO forecast model for February 2014 will be used. In (Katsalova, Shpyg 2014) it is shown that linear ($A_1 = 2$) and Gauss ($A_1 = 8$) models reproduce the best correlation structure of the pressure field. From (13) it is clear that LAG_{opt} will vary depending not only on the choice of the theoretical model, but with each set of data, as A_1 changed for each new array of pressure values. The value A_2 is calculated with the equation (5).

Let us also consider two cases of the priority of criteria. In the first case, assume that the criteria of accuracy and the criteria of time have equal priority ($\alpha = 0.5$). In the second case the criteria of accuracy is considered much more important than saving machine time ($\alpha = 0.1$).

Table 2. The calculation results LAG_{opt}

Date	A_2	Linear model: $A_1 = 2$		Gauss model: $A_1 = 6$	
		$\alpha = 0.5$	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 0.1$
1 February 2014	29.31	2.75	0.875	3.5	1.125
2 February 2014	31.98	2.625	0.875	3.25	1.00
3 February 2014	31.58	2.75	0.875	3.25	1.00
4 February 2014	26.77	2.875	0.875	3.5	1.125
5 February 2014	30.92	2.625	0.875	3.25	1.00
6 February 2014	23.47	3.125	1.00	3.875	1.25
7 February 2014	30.27	2.75	0.875	3.375	1.125
8 February 2014	23.11	3.125	1.00	3.875	1.25
9 February 2014	30.14	2.75	0.875	3.375	1.125
10 February 2014	30.52	2.75	0.875	3.375	1.125

Given all of the above, a series of numerical experiments were performed. Their results are given in Table 2.

According to the data presented in Table 2, LAG_{opt} (in the common case LAG has the same units as ρ, here – degree of latitude/longitude) changes significantly with a change in the priority of criteria. For models with a large number of operations the interval is naturally larger than for models with a small number of operations. LAG_{opt} varies for each set of input data, but this change is limited, and we can talk about the definition of average LAG_{opt} for this task of pressure data interpolation.

It should be noted that when $\alpha = 0$ (maximal accuracy is priority, calculation time is not important) or $\alpha = 1$ (a low level of accuracy is acceptable, the calculation time is important), in such cases problems (9) and (10) should be solved the separately, and equation (13) does not give the correct solution.

6. Conclusions

The research of smoothing of an experimental variogram by introducing lag was conducted. The advantages and disadvantages associated with smoothing were indicated. The saving of computer time and the simplification of the definition of the theoretical variogram are significant, and this transition is prevalent in solving problems with large amounts of input data. At the same time, the smoothing of the variogram produces an error. Its evaluation is obtained in this work.

The research, the purpose of which was to establish the size of the optimal interval, based on the criteria of accuracy and economy of computing time for solving the problem, was conducted. A two-criteria problem, making it possible to determine LAG_{opt}, taking into account the priority of each criterion, was solved. It is established that the value of the optimal interval depends on the prioritisation of criteria as well as on the type of variogram and correlation structure of input data.

The definition of LAG_{opt} was provided with the example of the problem of interpolation data from the COSMO forecast model of a pressure field.

The results may be useful when using Kriging for problems related to the interpolation of meteorological parameters and for other problems of data interpolation.

Bibliography

ArcGIS Resources, 2013, Help of ArcGIS 10.1, http://resources.arcgis.com (data access 17.08.2016)

Armstrong M., 1984, Common problems seen in variograms, Mathematical Geology, 16 (3), 305-313

DeMers M.N., 1999, Fundamentals of Geographic Information System, John Wiley and Bong Inc., 512 pp.

De Morsier G., Fuhrer O., Kaufman P., Schubiger F., 2015, Developing a 1.1 km Model Setup at MeteoSwiss: Impact of changing the boundary conditions, [in:] COSMO/CLM/ART User Seminar 2015, Book of Abstracts, Offenbach, Germany, 3

Doms G., 2013, A description of the Nonhydrostatic Regional COSMO-Model, www.cosmo-model.org (data access 17.08.2016)

Gunes H., Sirisup S., Karniadakis G.E., 2006, Gappy data: to Krig or not to Krig?, Journal of Computational Physics, 212 (1), 358-382, DOI: 10.1016/j.jcp.2005.06.023

Kanevskiy M.F., Demyanov V.V., Savelyev E.A., Chernov S.Y., Timonin V.A., 1999, An elementary introduction to geostatics (in Russian), VINITI, Series Problems of the Environment and Natural Resources, 11, 136 pp.

Katsalova L.M., V.M. Shpyg, 2013, Kriging-interpolation in weather forecast, (in Ukrainian), Scientific Papers of UHMI, 264, 3-9

Katsalova L.M., V.M. Shpyg, 2014, Variographic models of meteorological parameters distribution on the territory of Ukraine for Kriging-interpolation, (in Ukrainian), Scientific Papers of UHMI, 266, 20-26

Koshel S.M., Musin O.R., 2000, Digital simulation methods: Kriging and radial interpolation, (in Russian), Newsletter GIS Association, 4-5 (26-27), 32-33

Matheron G., 1967, Kriging or polynomial interpolation procedures?, Mathematical Geology Transactions, 70, 240-244

Oliver M.A., 1990, Kriging: a method of interpolation for Geographical Information Systems, International Journal of Geographic Information Systems, 4 (3), 313-332, DOI: 10.1080/02693799008941549

Roache P.J., 1985, Computational fluid dynamics, (in Russian), Albuquerque: Hermosa Publishers, 616 pp.

Samarsky A.A., Gulin A.V., 1989, Numerical methods, (in Russian), Nauka, 432 pp.

Samarsky A.A., 1982, Introduction to numerical methods, (in Russian), Nauka, 552 pp.

Stein M.L., 1999, Interpolation of spatial data: some theory for kriging, Springer Series in Statistics, Springer-Verlag, New York, USA, 249 pp.

Will A., Weiher S., Blahak U., 2015, Overview of the interpolation methods. Analysis of the main error sources and first improvements in int2lm, [in:] COSMO/CLM/ART User Seminar 2015, Book of Abstracts, Offenbach, Germany, 2