Liet. matem. rink., T. 43, spec. nr., 504–508

Applying of kriging and cokriging methods for prediction of Curonian lagoon data

Ingrida KRŪMINIENĖ, Kęstutis DUČINSKAS, Rolandas GARŠKA (KU)
e-mail: duce@gmf.ku.lt

1. Introduction

Geostatistics (spatial statistics) is a branch of applied statistics focusing on the characterization of the geospatial dependence on one or more attributes whose values vary over space (in 1-D, 2-D, or 3-D); and the use of that spatial dependence to predict (model) values at unsampled locations.

Any of a number of methods to produce estimates of a variable at unsampled locations is based on values at discrete points. Examples include: tesselation (Theissen polygons, triangular irregular network, Delauney triangulation, etc.), moving average, inverse distance weighting, spline functions, trend surfaces. The geostatistical equivalent of best linear unbiased predictor is kriging predictor.

2. Spatial data analysis

Geostatistical spatial data customarily refer to measurements on several attributes at the points where spatial locations are referred as $s_1, s_2, \ldots, s_n$, in region D.

In the univariate case we observe $Y(s_i)$ at a site s_i arising from an underlying random spatial process $\{Z(s)\colon s \in D\}$, where D is a fixed subset of R^d with positive measure. That is, the spatial index s varies continuously throughout the region D and realization of the process is a random surface above D.

Let $(Z(s_1), Z(s_2), \ldots, Z(s_n))'$ to be a vector of the observed values at locations $s_1, s_2, \ldots, s_n$. The objective is to predict the unobserved value $Z(s_0)$ at a location s_0 which is not one of $s_1, s_2, \ldots, s_n$. These data may involve spatial correlation which cannot be ignored.

Kriging, a term introduced by Matheron (1963), is a very popular method to solve the problem of spatial prediction. It was first used in mining data. It assumes a random field expressed through a variogram or covariance function, and correct estimation of the variogram (or covariance function) is crucial.

The model assumption (see [1]) is $Z(s) = \mu + E(s)$ where $E(s)$ is a zero mean stochastic term with variogram $2\gamma(\cdot)$. If we assume intrinsic stationarity then $E(Z(s+h)Z(s)) = 0$ and the variogram is defined as

$$2\gamma(h) = Var[Z(s+h) - Z(x)]. \tag{1.1}$$

This can be written as $Var(Z(s+h)Z(s)) = E(Z(s+h)Z(s))^2$ and thus the method of moments estimator for the variogram can be used (also called the classical estimator, [1])

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{k=1}^{N(h)} [Z(s_k + h) - Z(s_k)]^2, \tag{1.2}$$

where $N(h)$ is the number of data pairs within a given class of distance. When we have n observations the number of pairs becomes $N(h) = \frac{n(n-1)}{2}$.

The spatial variability between two correlated random variables is described by the cross semivariogram. An estimator of the cross semivariogram is

$$\hat{\gamma}_{ij}(h) = \frac{1}{2N(h)} \sum_{k=1}^{N(h)} [Z_i(s_k + h) - Z_i(s_k))(Z_j(s_k + h) - Z_j(s_k)], \tag{1.3}$$

where $Z_i(.)$ and $Z_j(.)$ denote two different variables, $i \neq j$ $i, j = 1, \ldots, q$ and $N(h)$ is the number of pairs of observations separated by the distance h.

Kriging method is known to be the Best Linear Unbiased Prediction, because it minimizes the variance error between the model and the predictor.

Linear predictor of the value $Z_j(s_0)$ of the data at the unsampled site s_0 from the data $Z(s_1), \ldots, Z(s_n)$ at the sampled sites $s_1, s_2, \ldots, s_n$ is

$$\widehat{Z}_j(s_0) = \sum_{k=1}^{n} w_{kj} Z_j(s_k), \quad j = 1, \ldots, q, \tag{1.4}$$

where w_{kj} is the weight for the jth variable of observation at location s_k and n is the number of point observations.

w_{kj} are chosen to minimize the mean squared error $MSE = E([\widehat{Z}_j(s_0) - Z_j(s_0)]^2$. $\widehat{Z}(s_0)$ is unbiased for $Z(s_0)$ if and only if $\sum_{k=1}^{n} w_{kj} = 1$.

Ordinary kriging gives the optimal predictions under the assumption that the mean value is constant (but unknown) across the whole area under study.

The ordinary kriging variance for Z_j is given by

$$\sigma^2_{ok} = \sum_{k=1}^{n} w_{kj} \gamma(s_k - s_0) + m, \tag{1.5}$$

where m is a Lagrange multiplier ([1], p. 122).

Cokriging is prediction of a primary variable using additional information from a secondary variable. This method is used in data sets containing more than one regionalized variables which are correlated with one another.

Suppose that $q = 2$. The prediction of $\widehat{Z}_1$ is done, not only on the basis of Z_1, but also on measurements of Z_2.

Cokriging involves the prediction of $Z_j(s_0)$ at an unsampled site s_0 from the data $Z(s_1), Z(s_2), \ldots, Z(s_n)$ (where $Z(s)^T = (Z_1(s), Z_2(s))$ from all variables at the sampled sites $s_1, s_2, \ldots, s_n$. The linear prediction of cokriging is

$$\widehat{Z}_1(s_0) = \sum_{k=1}^{n} v_1 Z_1(s_k) + \sum_{k=1}^{n} v_2 Z_2(s_k). \tag{1.6}$$

To obtain an unbiased estimate the following constraints are needed $\sum_{k=1}^{n} v_1^k = 1$ and $\sum_{k=1}^{m} v_2^k = 0$.

Similary as (1.5) the variance of cokriging the prediction can be written as

$$\sigma_{cok}^2 = \sum_{k=1}^{n} v_1^k \gamma_{k1}(s_k - s_0) + \sum_{k=1}^{n} v_2^k \gamma_{k2}(s_k - s_0) + m_1. \tag{1.7}$$

Cross-validation is a method of evaluating the aptness of spatial correlation model using only data from the sample. This method is especially useful for pointing out which specific areas in a region are difficult to estimate from the observed data.

Cross-validation procedure:

1. For location s_i temporaly exclude $Z_j(s_i)$ from the data set temporarily.
2. Estimate $Z_j(s_i)$ from the remaining points.
3. Compare $\widehat{Z}_j(s_i)$ to $Z_j(s_i)$ (compute square difference).
4. Do steps (1) to (3) for all $i = 1, \ldots, n$ points in the sample.
5. Compute summary statistics.

Summary statistics:

1. $\frac{1}{n} PRESS = \frac{1}{n} \sum_{i=1}^{n} (\widehat{Z}_j(s_i) - Z_j(s_{-i}))^2$ where $\widehat{Z}_j(s_{-i})$ indicates the prediction of $Z_j(s_i)$ from the rest of the data. This quantity should be small if the model fits well.

2. Mean of standartized PRESS residuals $\frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{Z}_j(s_i) - Z_j(s_{-i})}{\tilde{\sigma}_{R(-i)}}$, where $\tilde{\sigma}_{R(-i)}$ is the mean squared prediction error for predicting $Z_j(s_i)$ from the rest. This quantity should be close to *zero* if the model fits well. We would like the prediction errors to be small.

3. Root mean squared prediction residuals (standardized)

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\widehat{Z_j}(s_i) - Z_j(s_{-i})}{\tilde{\sigma}_{R(-i)}} \right)^2}.$$

This quantity should be close to *one* if the model fits well. The variance of the cross-validation errors is an empirical estimate of the prediction variance.

3. Results

The Curonian lagoon (also known as Kuršių marios, Kurshskij zaliv, Kurische Haff) is a large (length 95 km, width up to 48 km) shallow (mean depth of 3.8 m, the maximum

5.8 m) coastal water body in the south-eastern part of the Baltic Sea. The outlet of the lagoon to the Baltic Sea, Klaipėda Strait, is artificially deepened down to 12 m.

Data have been measured in 1990 year by S. Gulbinskas. They consist of bed sediments and soundings of the Curonian Lagoon. Sediments where measured in 213 locations, depth was measured in 263 locations. Their x coordinates values are between 278199 and 333376 and y coordinates values are between 6088178 and 6172784.

Sediments have been divided into 7 groups depending on median diameter (Md) (in mm): (1) more than 0.5, (2) 0.5–0.25, (3) 0.25–0.125, (4) 0.125–0.063, (5) 0.063–0.01, (6) 0.01–0.004, (7) less than 0.004.

In order to apply the above statistical methods for data analysis (modelling and fitting (cross) semivariogram, evaluating kriging and cokriging, comparing spatial statistics) we have chosen free available software R (package Gstat). R provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, ...).

Statistical methods for data on bed fractions percentage and soundings have been described and applied. The methods are general, but in this paper they have been applied only to measurements of the Curonian lagoon.

To check which one of the kriging and cokriging methods predict the true data best first we calculated semivariogram (cross variogram), then used cross-validation method and calculated summary statistics.

In Gstat variogram models are coded as the sum of one or more simple models (and optionally an anisotropy structure). A simple variogram model is denoted by $cMod(a)$ with c the vertical (variance) scaling factor, Mod the model type, and a the range (horizontal, distance scaling factor) of this simple model.

Semivariogram models of kriging method have been made using percentage of all fractions. Semivariogram model for (1) fraction is 16.62548 Nug(0) + 31.57597 Sph (256783.5), where Sph represents model types, nugget effect equals 16.62548, when $\frac{1}{2}\gamma(h) = 0$, sill is 31.57597 and range is 256783.5.

Semivariograms models of kriging method for all fractions are:

(1) fraction: 16.62548 Nug(0) + 31.57597 Sph (256783.5);

(2) fraction: 285.8347 Nug(0) + 176.9284 Sph (30712.18);

(3) fraction: 411.6521 Nug(0) + 578.6877 Sph (11884.07);

(4) fraction: 265.04047 Nug(0) + 62.43803 Sph (15553.02);

(5) fraction: 227.0382 Nug(0) + 345.0651 Sph (18237.69);

(6) fraction: 17.07367 Nug(0) + 531.80577 Sph (532720);

(7) fraction: 15.54634 Nug(0) + 52.93388 Sph (20565.16).

The spherical model had the best-fit to all (cross) semi-variograms.

In cokriging method case cross semivariograms models were changed in each "leave one out point" step.

The best model is the one that has the smallest root-mean-squared prediction error and the standardized root-mean-squared prediction error nearest to one, see [1].

Summary statistics (root mean squared prediction residuals) for all fractions are:

Fraction	Kriging	Cokriging
(1)	0.2537	0.2088
(2)	0.0589	0.0601
(3)	0.0400	0.0414
(4)	0.0536	0.0590
(5)	0.0624	0.0673
(6)	0.2479	0.2808
(7)	0.2487	0.2939

We can see, that the root mean squared prediction residuals of cokriging method is much closer to one.

So the results of this research show that the cokriging method for prediction the data of most fractions is better than kriging.

References

[1] N. Cressie, *Statistics for Spatial Data*, John Wiley, New York (1993).
[2] B.D. Ripley, *Spatial Statistics*, John Wiley, New York (1981).
[3] H. Wackernagel, *Multivariate Geostatistics*, Springer, New York (1995).
[4] K. Krivoruchko, A. Gribov, J. Ver Hoef, Predicting exact, filtered, and new values using kriging, *Stochastic Modeling and Geostatistics*, **2** (2000).
[5] K. Krivoruchko, Using linear and non-linear kriging interpolators to produce probability maps, *Annual Conference of the International Association for Mathematical Geology*, Cancun, Mexico, September (2001).
[6] A. Gribov, K. Krivoruchko, J. Ver Hoef, Modified weighted least squares semivariogram and covariance model fitting algorithm, *Stochastic Modeling and Geostatistics*, **2** (2000).
[7] L.S. Richard, *Environmental Statistics*, University of North Carolina Chapel Nill, NC 27599-3260 (2001).
[8] N.L. Sören, *Reconstruction of Data from the Aquatic Environment*, LYNGBY (2001).
[9] A.A. Nielsen, *Geostatistics and Kriging* (1995).
[10] J. Welhan, *G606 Geostatistics and Spatial Modeling* (2001).

Apie krigingo ir kokrigingo modelių taikymą prognozuojant Kuršių marių duomenis

I. Krūminienė, K. Dučinskas, R. Garška

Straipsnyje aprašomi nežinomų erdvės duomenų prognozavimo metodai krigingas ir kokrigingas. Taip pat nurodomos sąlygos, kurias reikia įvykdyti norint atlikti prognozę. Siekiant nustatyti, kuris iš metodų yra efektyvesnis, Kuršių marių duomenims buvo pritaikytas kryžminės validacijos metodas. Rezultatai parodė, kad tikslesni prognozės rezultatai gaunami duomenis prognozuojant kokrigingu.