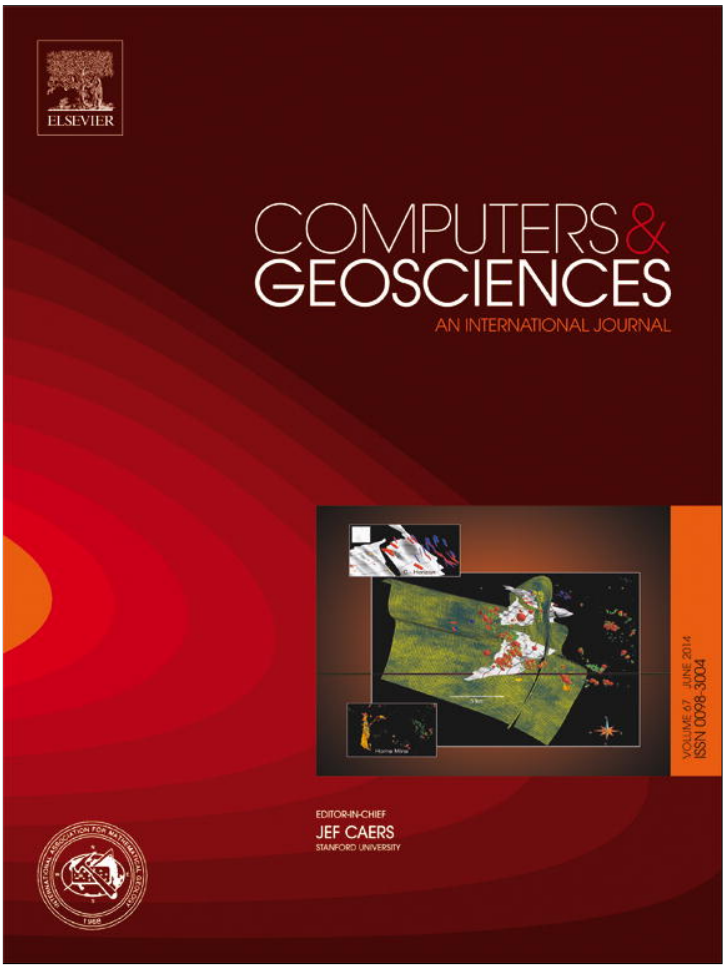

Computers & Geosciences 67 (2014) 180–190

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

rtop: An R package for interpolation of data with a variable spatial support, with an example from river networks

J.O. Skøien [a,*], G. Blöschl [b], G. Laaha [c], E. Pebesma [d], J. Parajka [b], A. Viglione [b]

[a] Institute of Environment and Sustainability/Joint Research Centre/European Commission, Italy
[b] Institute for Engineering Hydrology and Water Resources Management/Vienna University of Technology, Austria
[c] Institute of Applied Statistics and Computing/University of Natural Resources and Life Sciences, BOKU Vienna, Austria
[d] Institute for Geoinformatics/Westfälische Wilhelms-Universität Münster, Germany

ARTICLE INFO

Article history:
Received 13 May 2013
Received in revised form
17 February 2014
Accepted 28 February 2014
Available online 14 March 2014

Keywords:
Geostatistics
Support problem
Hydrology
River networks
Interpolation

ABSTRACT

Geostatistical methods have been applied only to a limited extent for spatial interpolation in applications where the observations have an irregular support, such as runoff characteristics along a river network and population health data. Several studies have shown the potential of such methods, but these developments have so far not led to easily accessible, versatile, easy to apply and open source software. Based on the top-kriging approach suggested by Skøien et al. (2006), we will here present the package *rtop*, which has been implemented in the statistical environment R (R Core Team, 2013). Taking advantage of the existing methods in R for analysis of spatial objects (Bivand et al., 2013), and the extensive possibilities for visualizing the results, *rtop* makes it easy to apply geostatistical interpolation methods when observations have a non-point spatial support. The package also offers integration with the intamap package for automatic interpolation and the possibility to run *rtop* through a Web Service.

1. Introduction

For many applications where a process has been observed, it is also necessary to have predictions of the process at locations without observations; this could be for visualization purposes, for input in a different model, or for further analysis. There exists an appealing set of methods for such predictions, based on a model of spatial correlation between random variables defined at observation and prediction locations. These are commonly referred to as geostatistical methods (Isaaks and Srivastava, 1989) due to the background from mining, although similar methods were also developed within meteorology (Gandin, 1963) and referred to as objective methods. Kriging is defined as the best linear unbiased estimator for spatial data, i.e. a linear interpolation method where the expected bias is zero and the expected interpolation error is minimized. One major advantage of these methods is that they can give an estimate of the prediction uncertainty in addition to the prediction itself.

Whereas geostatistical and similar methods mostly have been developed for and applied to observations with point support or a regular support (e.g. pixels from satellite images), and also to make predictions for areas or volumes (Journel and Huijbregts, 1978), many data sources have a more irregular support in time and/or space, such as runoff related data and aggregated health data. The literature shows some different approaches to solve the interpolation problem for observations with non-point support (Goovaerts, 2006, 2008; Gottschalk, 1993; Gottschalk et al., 2006; Gotway and Young, 2002; Kyriakidis, 2004; Sauquet et al., 2000; Skøien et al., 2006). Although all these studies make similar assumptions and solve the problem in a similar way, there has not yet been any easily accessible, easily applicable and open source software that is able to interpolate such data.

Based on the top-kriging approach suggested in Skøien et al. (2006), and extended with suggestions by Gottschalk (1993) and Gottschalk et al. (2011), we present the package *rtop*, which has been implemented in the open source statistical environment R (R Core Team, 2013). The R environment includes a large range of tools for data analysis and visualization and is a platform where it is easy to extend the existing system with new methods in a package system. Several packages already deal with spatial data within R (Bivand et al., 2013), but none of these can do geostatistical interpolation with a variable spatial support of observations taken into account. *rtop* therefore makes it considerably easier to do such interpolation and to use visualize results, in comparison to former implementations of the method. And although the top-kriging method was originally developed for interpolation of runoff characteristics, we also suggest that the package can be useful for a number of other analyses such as health statistics from administrative regions, aggregated forest data and remote sensing images.

* Corresponding author. Tel.: +39 0332 789205; fax: +39 0332 789960.
E-mail address: jon.skoien@gmail.com (J.O. Skøien).

http://dx.doi.org/10.1016/j.cageo.2014.02.009
0098-3004/

There are several other R-packages that can do some similar operations. The *dissever*-package (Malone et al., 2012) can for example do downscaling of remote sensing images, but it is based on regular areas and is not available on CRAN. The *SSN*-package (VerHoef and Peterson, 2010) is able to do interpolation along stream networks, but this is based on a different theory which does not take the areal support into account. It also requires preprocessing of the stream network with the STARS geoprocessing toolset running under ArcGIS.

The *constrainedKriging*-package (Hofer and Papritz, 2011) can do interpolation to irregular blocks, but not from observations with irregular supports, and does not include variogram fitting tools comparable to *rtop*. It is also possible to make extensions to the *gstat*-package (Pebesma, 2004) for interpolation between different supports, but this is currently not in the code itself, and it is not possible to fit variograms for such observations. It is also likely that the framework of the *INLA*-package (Rue et al., 2013) can deal with the same type of problems as *rtop*, but such functionality has, to our knowledge, not yet been implemented. Also the C-code of the *psgp*-package (Barillec et al., 2011) can in theory interpolate observations and predictions with a spatial support, but this is not available from the R-package.

rtop is currently the only package which provides all necessary functionality for interpolating data with variable spatial support. In this paper we demonstrate the package with an example from hydrology: the interpolation of streamflow along river networks, where it was earlier necessary to combine the use of spreadsheets, analyses tools, graphical tools and GIS to perform the interpolation. We show how *rtop* can be applied to carry out this chain of analyses within R, to get an easy solution for the river network problem and other problems of spatial interpolation with a variable support.

2. Theory

The *rtop* package is based on the top-kriging method presented by Skøien et al. (2006), but also includes simplifications suggested by Gottschalk (1993). The theoretical background of top-kriging is briefly described below, with some additions regarding validity.

2.1. Assumptions

The methods implemented in the *rtop* package are mainly based on two assumptions. The first is that the observed variable of interest can be seen as the output from a continuous process in space and/or time. The process is usually not observable at a local (point) scale, but as an integrate or average over some larger area or time interval, referred to as spatial or temporal support. This process can, for example, be the runoff generating process in hydrology, the probability of a single person contracting a certain disease in health statistics or spatial analyses of plots of forested areas. The observed values can then in most cases be seen as the aggregates (linear averages) of local realizations of the process over the support, such as runoff per unit area (specific runoff) and the percentage of a population contracting the disease. The assumption of a linearly aggregating process can in general be somewhat relaxed in the sense that the process needs to aggregate linearly within the range of supports of the observations and prediction locations. In the case of runoff generation, this means that we assume that the specific runoff from catchments and basins can be seen as the average of the runoff generated in subcatchments, but that the connection between runoff generation and runoff might be more complicated and intermittent for smaller scales.

The second assumption is the general stationarity assumption for geostatistical methods, i.e., that expected variance between observations is a function of separation distance. The implementation of the method does not take non-stationarity of the mean into account, as commonly done through universal kriging although it is possible to add an external drift (Laaha et al., 2013). The standard implementation, however, is a local kriging approach where only the observations with the highest modeled correlations are used for interpolation. The examples shown in Skøien et al. (2008) indicate that predictions can be good despite violations of the stationarity assumptions, whereas estimates of prediction uncertainty may be less reliable. This is similar to ordinary kriging, and in accordance with general knowledge that the quality of the predictions is relatively insensitive to the choice of the variogram, at least as long as there are several observations within the range of the variogram (e.g. Lark, 2000).

2.2. Kriging

Kriging is an interpolation method where the value of a spatial variable at locations without observations is predicted as a weighted average of the observations at the surrounding locations. By assuming that the expected variance of the process measured at different locations is only a function of distance, it is possible to find the weights by solving a set of kriging equations (Cressie, 1991). The prediction $\hat{z}(\vec{x}_0)$ of the variable z at position $\vec{x}_0$ (i.e. the target position) is

$$\hat{z}(\vec{x}_0) = \sum_{i=1}^{n} \lambda_i z(\vec{x}_i) \tag{1}$$

where λ_i is the interpolation weight of the measurement at position $\vec{x}_i$ and n is the number of neighboring measurements used for interpolation. The weights λ_i can be found by solving the kriging system:

$$\begin{aligned} &\sum_{j=1}^{n} \lambda_j \gamma_{ij} + \mu = \gamma_{0i}, \quad i = 1, \ldots, n \\ &\sum_{j=1}^{n} \lambda_j = 1 \end{aligned} \tag{2}$$

where the gamma value γ_{ij} is the expected semivariance between two measurements i and j, which can be modeled through a theoretical variogram model. μ is the Lagrange parameter, which is a parameter necessary for the unbiasedness constraint $\hat{z}(\vec{x}_0) = z(\vec{x}_0)$. We can also, as described in Skøien et al. (2006), take the uncertainty of the observations into account, by introducing a term for the measurement error variance σ_i^2 of the observations in the first line in Eq. (2) above, assuming uncorrelated measurement errors. The interpolation method can then be referred to as kriging with uncertain data (KUD) (de Marsily, 1986), and similar to factorial kriging with data containing measurement errors.

2.2.1. Taking area into account

In the methods used here, the spatial variable $Z(A)$ is assumed to be representative for an area with a non-zero spatial support A:

$$Z(A) = \frac{1}{|A|} \int_A z(\vec{x})\, d\vec{x} \tag{3}$$

where $z(\vec{x})$ is the value at location $\vec{x}$ and $|A|$ refers to the size of A. We assume a constant mean, as mentioned above, meaning that $E(Z(A_i) - Z(A_j)) = 0$. If a non-zero support A is taken into account, the kriging system remains the same, but the semivariances between the measurements must be integrated over the supports (Cressie, 1991, p. 66), as was done in Skøien et al. (2006). We can assume the existence of a point variogram γ_p, which is valid at the

point scale even if the process itself might not be observable at point scale. Our interest is rather that this point variogram can describe the scaling of the variability between areas as a function of their spatial support. The semivariance between two observations with supports A_i and A_j can then be found through regularization of the variogram (Skøien et al., 2006; Cressie, 1991):

$$\gamma_{ij} = 0.5 \times \mathrm{Var}(Z(A_i) - Z(A_j)) = \frac{1}{|A_i||A_j|}\int_{A_i}\int_{A_j}\gamma_p(|\vec{x}_i - \vec{x}_j|)\,d\vec{x}_i\,d\vec{x}_j - 0.5 \times \left[\frac{1}{|A_i|^2}\int_{A_i}\int_{A_i}\gamma_p(|\vec{x}_i - \vec{x}_j|)\,d\vec{x}_i\,d\vec{x}_j + \frac{1}{|A_j|^2}\int_{A_j}\int_{A_j}\gamma_p(|\vec{x}_i - \vec{x}_j|)\,d\vec{x}_i\,d\vec{x}_j\right] \quad (4)$$

where $\vec{x}_i$ and $\vec{x}_j$ are position vectors within each area used for the integration. The first part of this expression integrates the semivariance between the two areas, while the integrated semivariance within the areas are subtracted in the second part, causing the semivariance of $Z(A)$ to decrease with increasing support. The semivariances are inserted into the kriging matrix Eq. (2) and the kriging system can be solved in the normal way to calculate the weights λ_i for Eq. (1). Eq. (4) always gives a positive semivariance.

The nugget effect needs a special attention when we are dealing with observations with a support. Traditionally, this is the discontinuity that is often seen in a variogram at a distance infinitesimally larger than zero, caused by measurement errors and variability at small distances.

The idea behind the regularization of the nugget effect in *rtop* is the following. If we assume that small scale variability gives a nugget variance equal to C_{0p} for a unit area, the nugget variance of a larger area will be $C_{0p}/(A/A_0)$, where A is the size of the larger area and A_0 is the size of the unit area. This is similar to the variance of the mean of a set of observations, and follows the description by Chilés and Delfiner (1999) and Skøien et al. (2006). For two areas of different sizes, A_i and A_j, the nugget variance will then be the average of the nugget variances of these two areas, if they are not overlapping. If they are overlapping, it will also be necessary to subtract the part of the nugget variance that is common to the two areas, as the small scale variability will be equal for the part that is overlapping. The combined nugget effect for two areas (A_i, A_j) of different sizes can be then be generalized as

$$C_0(A_i, A_j) = 0.5\left(\frac{C_{0p}}{|A_i|/|A_0|} + \frac{C_{0p}}{|A_j|/|A_0|} - \frac{2C_{0p}\cdot \mathrm{Meas}(A_i \cap A_j)}{|A_i||A_j|/|A_0|}\right) \quad (5)$$

where $\mathrm{Meas}(A_i \cap A_j)$ represents the area shared by the two areas A_i and A_j. If they are overlapping, then this will be the area of the intersection of the two areas, if they are not then this will be zero. For runoff related variables, the shared area is equal to the smallest area if they are overlapping. The point nugget effect in Eq. (5) will depend on the size of the unit area A_0, and can therefore appear unreasonably large if the unit area is small. However, we are, as mentioned above, not interested in the validity of Eq. (5) at point scale, but the ability to reproduce area-dependent scaling of the variability at the size of the observation areas.

2.2.2. A simplification for regularization

It is computationally expensive to compute and integrate the semivariogram values for all different pairs of distances between the different areas. When fitting the variogram model to a variogram cloud or to binned variograms through an iterative optimization procedure, all semivariogram values between points within the different areas have to be recomputed for each iteration step. Gottschalk (1993) and Gottschalk et al. (2011) suggested to simplify this calculation for regularization of covariances, by applying the covariance model on the averaged distance, d^*, between areas instead of integrating the covariance function for all distances between the areas, as is usually done when regularizing:

$$d^*_{ij} = \frac{1}{|A_i||A_j|}\int_{A_i}\int_{A_j}(|\vec{x}_i - \vec{x}_j|)\,d\vec{x}_i\,d\vec{x}_j \quad (6)$$

We can also use d^*_{ij} for semivariograms, and find the regularized semivariance between the two areas as

$$\gamma^*_{ij} = \gamma_p(d^*_{ij}) - 0.5 \times [\gamma_p(d^*_{ii}) + \gamma_p(d^*_{jj})] \quad (7)$$

where d^*_{ii} represents the average distances within A_i. We will later refer to this distance as the Ghosh (1951)-distance.

Averaging distances is mathematically simpler and computationally faster than calculating the semivariogram values for all distances between two areas and it only has to be done once, whereas traditional regularization implies that the regularized semivariogram has to be recomputed for all candidate variogram models. The effect of the approximation depends on the variogram and the configuration of areas, but Gottschalk et al. (2011) indicate that the approximation will usually give a good result for different types of variograms, although the best results can be expected for variograms which are close to linear shape, particularly near the origin.

3. The *rtop* package

The new package *rtop* is based on a reimplementation of the top-kriging method presented by Skøien et al. (2006) in the statistical language/environment R (R Core Team, 2013). The original implementation was in FORTRAN, whereas the *rtop* package is almost entirely implemented as R code. Some computationally demanding functions have been kept in FORTRAN for faster computation. The main functions have been implemented using the S3 object handling of R, making it easier to use the same functions for different types of objects. The package uses spatial objects as defined in the *sp*-package (Bivand et al., 2013). It is therefore straightforward to use existing functionality for import/export of data and for visualization and analysis.

3.1. Creating an object for interpolation

The easiest interface to the methods in *rtop* is to store all variables (such as observations, prediction locations and parameters) in an *rtop*-object, which is created by a call to *createRtopObject*. The different functions will take this object as an argument and add their results as new elements of it. The only necessary object in the call to *createRtopObject* is a *SpatialPolygonsDataFrame* with the observed values (divided by support areas) and observation support, although it would also be common to add a *SpatialPolygons*-object with the prediction locations. Shapefiles or other standard formats with the polygons can be imported with tools in *rgdal* (Keitt et al., 2012) or other packages. It is also possible to use files describing the coordinates of the polygons.

The user can set a range of parameters, all of them with default values. A list of the most important parameters with their default values is given in Table 1, a complete list is given in the help file of *getRtopParams*. Some of the parameters will be further explained in the application section.

The integration in *rtop* is based on discretized areas. The discretization of the polygons is done once and the coordinates of the discretization points are added to the rtop-object. It is

Table 1
Some of the most important parameters of *rtop* and their default values.

model = "Ex1"	Variogram model type, other models are Exp, Gau, Ga1, Sph, Sp1, Fra, where the 1 denotes a fractal version of the variogram. The fractal version is obtained by multiplying the standard variogram by d^b, as in Skøien et al. (2006). It can be shown that this is positive definite for small b (Laaha et al., 2014)
nugget = FALSE	Logical; if point nugget effect should be estimated
unc = TRUE	Logical; if TRUE it will look for observations errors in column *unc*
rresol = 25	Minimum number of discretization points in each element (area or line)
cloud = FALSE	logical; if true use the variogram cloud
amul = 2	Defines the number of areal bins within one order of magnitude
dmul = 3	Defines the number of distance bins within one order of magnitude
fit.method = 9	Defines the type of Least Square method for fitting of variogram. The methods 1–7 correspond to the similar methods in *gstat*. Methods 8 and 9 are derived from the weighted least squares method suggested by Cressie (1985)
gDist = TRUE	Use Ghosh-distance (Gottschalk et al., 2011) when fitting variograms and computing covariance matrices (also possible to set individually with *gDistEst* and *gDistPred*)
maxdist = Inf	Maximum distance between prediction location and observation locations
nmax = 10	For local kriging: the number of nearest observations to be used for kriging prediction, where nearest is defined as the observations with the lowest modeled semivariance
hstype = "regular"	Sampling type for binned variograms
rstype = "rtop"	Sampling type for the elements, method "rtop" as described in Section 2.2.1
cv = FALSE	Logical; for cross-validation of observations

possible to choose a random discretization, but experiments indicate that regularly gridded discretization give numerically better results for the same number of discretization points. As the size of areas can differ by several orders of magnitude, we use an adaptive grid where the resolution scales with the area. Starting with a coarse grid covering the region of interest, this grid will for a certain area be refined until a requested minimum number of points is within the area. This method also assures that points used to discretize a large support will be reused when discretizing smaller supports within the large one, e.g. subcatchments within larger catchments. Fig. 1 gives an example where one area (gray) is overlapping another (white). The areas are discretized into 21 and 31 points, small points indicating the smallest area, large points indicating the largest area. The points from the largest area within the smallest area coincide with some, but not all the points within the smallest area.

3.2. Variogram modeling and interpolation

The function *rtopVariogram* will calculate a sample variogram based on the observations. The package supports both binned variograms and variogram clouds. The variogram cloud is almost identical to the variogram cloud from *gstat* (Pebesma, 2004), with columns for the area sizes added. The binned variogram is not only based on distance, but also on combinations of support areas. It is therefore a function of three independent variables (distance, smallest support, largest support) instead of the traditional variogram which is only a function of distance. This can also be understood as having a set of distance based variograms for different combinations of support areas. The bins for distances and support are created with regular spacing in the log_{10}-domain. The distance between the centers of gravity for the two areas is used for the distance-axis.

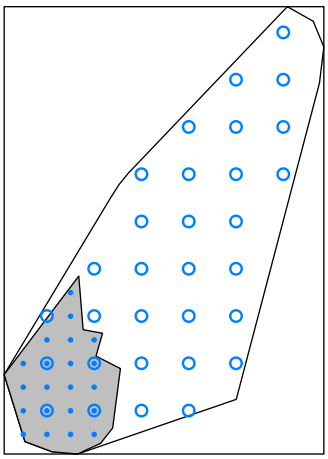

Fig. 1. Discretization of two different areas, small dots in smallest area, large points in largest area.

The theoretical point variogram cannot be fitted directly from the sample data, due to the different support of the observations. Instead, the function *fitVariogram* optimizes a point variogram whose regularized semivariogram values yield best fit to the sample variogram values. This can be referred to as a back-calculation of the sample variogram (Skøien et al., 2006), similar to the deconvolution method described by Goovaerts (2008). Regularized semivariogram values are found from the variogram model for all the combinations of support sizes and distances of pairs of observations, or from the averaged distances and areas of the binned variogram. If binned observations are used, the support is approximated by squares, which allows the areas to partly overlap. Different variogram models are available and also different least squares methods for the fitting. The optimization procedure is based on the Shuffle Complex Evolution Method (Duan et al., 1992).

The interpolation function (*rtopKrige*) solves the kriging system based on the regularized semivariances. These are computed in a separate regularization function, and are stored in the *rtop*-object if it is necessary to redo parts of the analysis, as this is the computationally expensive part of the interpolation.

The kriging result is added to the *rtop*-object as a *SpatialPolygonsDataFrame* of name *predictions*. This element will be the same as *predictionLocations*, with the predictions and prediction errors added, using notation for a single interpolation variable from *gstat* (Pebesma, 2004).

3.3. Interfacing the INTAMAP package

The *rtop*-package has also been developed so that it is possible to use it through the *intamap*-package (Pebesma et al., 2011) which is an *R* package for automatic interpolation. The *intamap*-package is developed to run as the computational back-end of a Web Service, and integration with this package thus makes it easy also to access *rtop* through a Web Service.

The linking is possible due to a similar structure between the two packages, and some adaptions of the *rtop*-package. Both packages use a single object for passing data to and from functions. Most of the

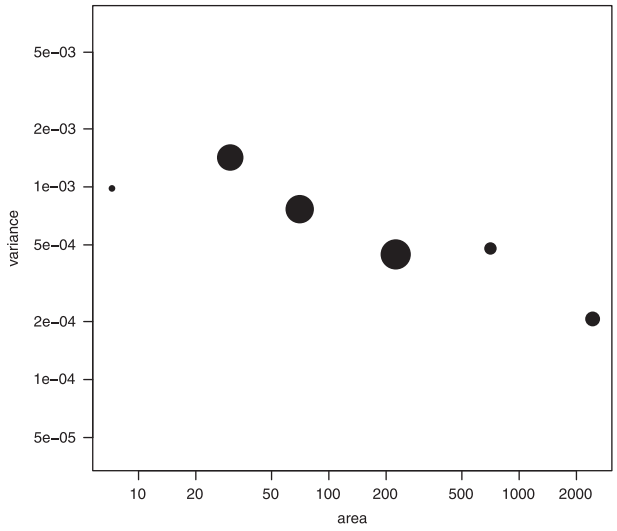

Fig. 2. Relationship between spatial variance of observations and area size, size of circles is proportional to the number of observations in each bin.

element names within an object are equal, such as *observations*, *predictionLocations* and *params*. The resulting predictions are also similar to the prediction-format *gstat* (Pebesma, 2004). Both packages modify and use parameters in a similar way, so that the parameters of one package can easily be extended with the parameters of the other package.

The *rtop*-package adds new methods for two of the main functions of *intamap*: *estimateParameters.rtop* and *spatialPredict.rtop*; these are wrapper functions around *rtopFitVariogram* and *rtopKrige*. It is necessary to call the function *useRtopWithIntamap* to be able to use the two packages together.

4. Example

In the example application we use top-kriging for prediction of a hydrologic characteristic (average summer runoff) along river networks in the federal country of Upper Austria. The average summer runoff is interpolated from the gauge locations to locations on the river where runoff measurements were absent. The top-kriging approach is applicable if we assume that the average runoff is a linear aggregate of all runoffs generated within the upstream contributing area. Runoff prediction at ungauged locations is a fundamental issue in hydrology and has been the objective of a 10 years initiative of the International Association of Hydrological Sciences (IAHS) just concluded in 2013 (Blöschl et al., 2013).

4.1. Data

The example application uses average summer runoff from 134 runoff gauges in the federal country of Upper Austria, including their catchment polygons. This is a subset of an extensive data set similar to the one analyzed by Skøien et al. (2006) and Skøien and Blöschl (2007). For this example we will also use a polygon shapefile of 542 catchments for predictions, and a line shapefile of 720 river segments matching the network IDs of either observations or prediction locations. A subset of this limited data set is also included in the rtop package, but to be able to reproduce the example in this paper it is necessary to download the example data from a web site at the Vienna University of Technology,[1] or conveniently with the function *downloadRtopExampleData*.

[1] http://www.hydro.tuwien.ac.at/downloads.html

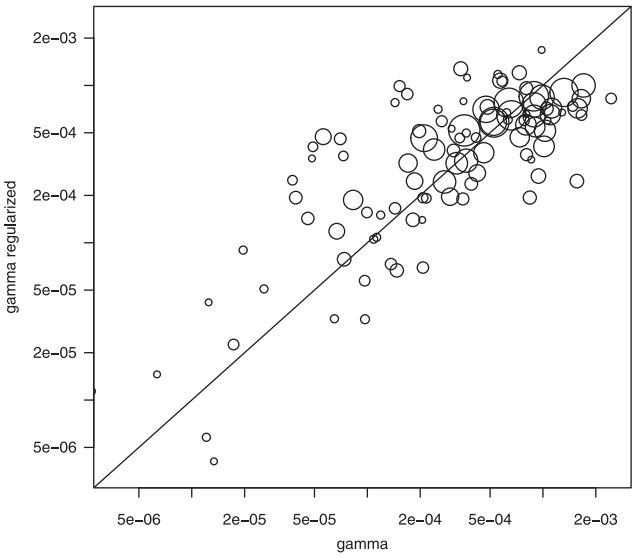

Fig. 3. Comparison of observed and regularized semivariogram values. Size of circle is relative to the number of pairs in each bin.

Different runs can give some deviations of the results due to the use of random numbers in some functions, but set.seed(1) in the start of the script assures that the results are reproduced exactly.

4.2. Starting the process

The polygons of the catchments are read from shapefiles, and have the observations attached. In the example, the runoff has been given as runoff per time unit, $m^3\ s^{-1}$, which is a function of catchment area. However, top-kriging requires the average runoff per unit area (specific runoff in $m^3/s/km^2$).

```
library(rtop)
library(rgdal)
set.seed(1)
downloadRtopExampleData()
rpath = system.file("extdata", package = "rtop")
observations = readOGR(rpath, "observations")
```

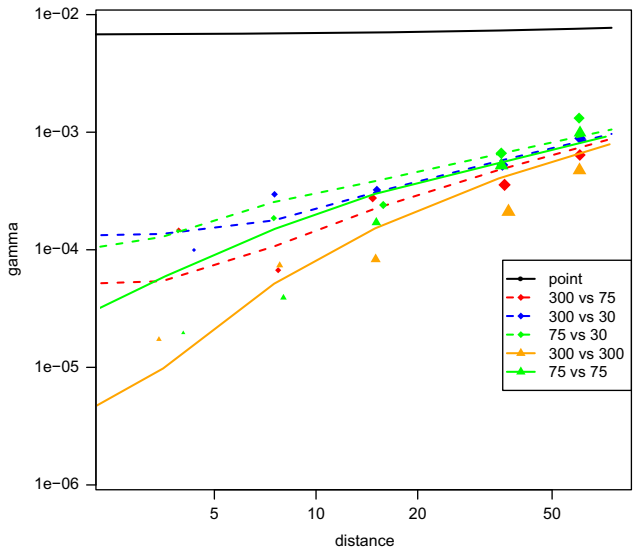

Fig. 4. Sample variogram and regularized semivariograms as a function of distance and area. Solid lines represent regularized semivariograms of equally sized catchments and dotted lines represent combinations of catchments sizes. Point size is relative to the number of pairs in each bin.

```
predictionLocations=readOGR(rpath,
   "predictionLocations")
observations$obs = observations$QSUMMER_OB/
   observations$AREASQKM
```

The easiest way to use the functionality in the *rtop*-package is by creating a particular *rtop*-object. Building the object can take some time, as the function also will start building some of the elements necessary at a later stage. We use the option *gDist=TRUE* (the simplification presented in Section 2.2.2) which will reduce computation time.

```
rtopObj = createRtopObject(observations,
          predictionLocations,
          formulaString = obs~1, params = list
          (gDist=TRUE, rresol = 25))
```

To allow for more control over the different steps, it is also possible to directly call the functions below by using the required elements of the rtop object as arguments.

4.3. Variogram model

A sample variogram can be estimated as a variogram cloud or a binned variogram, as explained in Section 3.2, and we can fit a point variogram model with *rtopFitVariogram*. The fitting function will call the function that creates the sample variogram (*rtopVariogram*) if this is not a part of the *rtopObject* that is given as an argument.

```
rtopObj = rtopFitVariogram(rtopObj)
```

The output to the console indicates the convergence of the optimization routine. “Best” refers to the best value of the error function so far, “function convergence” gives the relative improvement for the last *kstop* (usually 5) iteration cycles, and “parameter convergence” gives the variability of the current set of parameters relative to a stop criterion. Both criteria have to be below 1 for the optimization procedure to stop.

The function *checkVario* can create up to four figures. The first two explore the data before variogram fitting and interpolation, whereas the last two show the correspondence between the sample variogram and the fitted variogram. Three of these plots are shown in Figs. 2–4.

```
rtopObj = checkVario(rtopObj, cloud = TRUE,
          identify = TRUE,
          acor = 0.000001)
```

The extra parameters to this function is *cloud*, a logical defining whether to compare the individual semivariances with the regularized ones (as a variogram cloud) and *identify*, a logical that makes it possible to identify catchment pairs for which the semivariance is particularly large, and might not follow the homogeneity assumption. *acor* is a variable that can be used to transform the unit of the area label used in figures, e.g. from square meters to square kilometers (giving labels as 1 km^2 instead of 1,000,000 m^2 if the projection of the data sets is based on meters).

Fig. 2 shows the relationship between area and dispersion variance. The observations have been ordered in classes according to their area, and the dispersion variance has been calculated from the observations from each area class. One of the assumptions of top-kriging is that the dispersion variance decreases with increasing area, what should be visible from this figure. The sizes of the dots are relative to the number of observations in each area class.

The function also produces a figure (not included in this article) showing the sample variogram cloud of the observations. It is normal that some pairs of catchments exhibit quite large semivariances also for small distances, this can particularly be the case for combinations of small and large catchments. However, large differences could indicate deviations from the stationarity assumption and might need further examination. In the example in this paper, many of the high values are caused by catchment 27 (14.0 km^2), which has higher floods than its neighbors. We do not have any other information that give reason to exclude this catchment, and will continue using it in the further analysis.

Fig. 3 shows a log–log scatter plot of regularized semivariogram values plotted against the sample variogram values from a fitted variogram model. The diagonal line is the 1:1 line, which would represent a perfect fit. We can see that most of the points are centered around the diagonal line, although there is quite a large scatter. Such deviations can still be expected, similar to the difference between a cloud variogram and a fitted variogram model. The size of the dots is relative to the number of pairs in each bin for binned variograms, showing that the largest outliers are mainly bins with few pairs.

The last plot produced by *checkVario* (Fig. 4) is a comparison between some selected bins of the sample variogram and the regularized semivariance for those bins. The numbers in the key refer to the combinations of areas, e.g., 300 vs 75 means the regularized semivariogram as a function of distance for two catchments of sizes 300 and 75 km^2, respectively. The regularized semivariances for the variogram model in this figure are based on quadratic catchments, whereas natural catchments usually are more stretched. We can see that the regularization is not fully able to reproduce the variance reduction as a function of area particularly for large distances. A part of the reason for this is the (automatic) choice of areas for comparison. The fit appears slightly better for a different set of areas, such as

```
rtopObj = checkVario(rtopObj, acor = 0.000001,
acomp = data.frame (acl1 = c(2,2,2,2,3,3,3,4,4),
acl2 =c(2,3,4,5,3,4,5,4,5)))
```

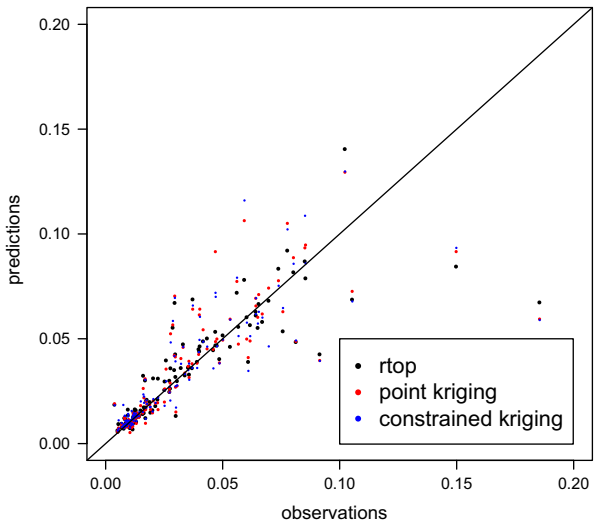

Fig. 5. Scatter plots of observations and cross-validation predictions from the two interpolation methods.

where *acomp* is used to select the combination of area bins. The fit still appears to be suboptimal, which could be a result of a suboptimal optimization result or because the real variance reduction is more complex than what we try to model with the methods in this paper. However, the general tendency is correct that smaller catchments have higher semivariances for large distances than the larger catchments.

From the figure we can also note that variogram values do not approach zero for small distances for pairs of catchments with different areas. This is because two catchments can share the same centroid, but the smallest will then be a tributary of the larger catchments, and there will be a variance between the observations of these catchments.

4.4. Interpolation

In this subsection we present the results from interpolating the mean summer runoff (using Eq. (1)) to new locations along the river network, as well as a cross-validation of the ability to predict at the observation locations.

Table 2
Some statistics (mean, R^2, mean absolute error and mean error) from cross-validation of the example data set with top-kriging, point kriging and constrained kriging.

Summary statistic	Top-kriging	Constrained kriging	Point kriging
Mean	0.03	0.031	0.031
R^2	0.72	0.66	0.65
ME	−0.00038	0.00053	0.00031
MAE	0.0068	0.0078	0.008

4.4.1. Cross-validation

Cross-validation in *rtop* can be done easily by calling *rtopKrige* with the parameter *cv=TRUE*. See Section 4.4.2 for further options. The result is in the *predictions* element of *rtopObj*, with column names similar to those from cross-validation with *krige* in *gstat*.

```
rtopObj = rtopKrige(rtopObj, cv=TRUE)
predictions = rtopObj$predictions
sstot = sum((predictions$obs - mean(predictions
  $obs))^2)
rtopsserr = sum((predictions$obs - predictions
  $var1.pred)^2)
rtoprsq = 1 - rtopsserr/sstot
summary(predictions)
```

Fig. 5 presents the results from a leave-one-out-cross-validation carried out for the catchments above in the form of a scatter plot with the observations of mean summer runoff on one axis and the predictions on the other axis. We have also added the results for the same catchments based on ordinary point kriging using the center-of-gravity of each catchment and from the constrained kriging approach of Hofer and Papritz (2011), using the point

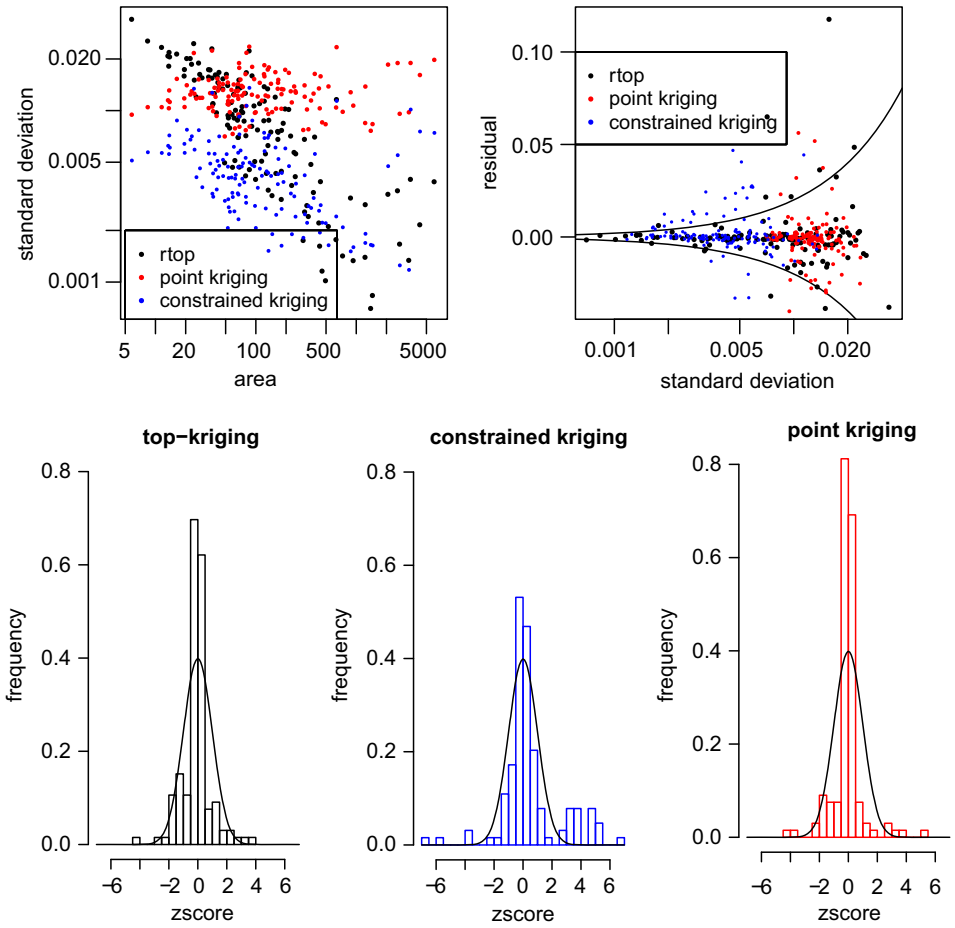

Fig. 6. Top left: Relationship between kriging standard error and area. Top right: Relationship between residual and kriging standard errors. Lines represent +/- two times the kriging standard error. Bottom left: Histogram of *z*-score (residuals/kriging standard error) for top-kriging. Bottom center: Histogram of *z*-score for constrained kriging. Bottom right: Histogram of *z*-score for point kriging. The lines represent the normal distribution for all bottom panels.

variogram estimated by *rtop*. Point kriging was done with the *automap*-package (Hiemstra et al., 2008), which is a wrapper for automatic interpolation based on *gstat* (Pebesma, 2004).

The figure indicates that all methods perform reasonably well, although there is a slightly larger scatter for the point kriging. All methods fail in predicting the two highest observations correctly, whereas point kriging overestimates some catchments, most likely as a result of giving too large weights to the two small catchments with high observations.

Table 2 shows some summary statistics of the cross-validation from the three methods. Top-kriging performs slightly better for the coefficient of determination (R^2) and mean average error (MAE). The small MAEs show that all methods are practically unbiased, but (R^2) increases from 0.65 to 0.71 for this example, whereas the mean absolute error is slightly reduced from 25–26 to 22 percent of the mean of the observations (0.031 $m^3/s/km^2$).

Geostatistical interpolation methods will also give an estimate of the prediction error in addition to the prediction itself in the form of an estimated kriging variance or a kriging standard error (square root of the kriging variance) . Fig. 6 shows the prediction errors from the three methods in different ways.

The top left panel shows the relationship between the area and the kriging standard error. Whereas the kriging standard errors from top-kriging and constrained kriging decrease with increasing catchment area, the kriging standard errors from point kriging are independent of the catchment area. The decreasing trend from top-kriging corresponds to our expectations, as stream flow is an integrating process, so the specific runoff tends to vary more smoothly along the stream for large catchments, and should thus also be easier to interpolate. The decreasing trend is stronger for top-kriging than for constrained kriging, although some of the difference might depend on parameter settings in constrained kriging.

The top right panel shows the relationship between the cross-validation residuals and the kriging standard error. Here we can notice that all the catchments with a low kriging standard error from top-kriging also have low residuals. The low kriging standard errors therefore seem justified. The two lines correspond to $+/-$ two times the kriging standard error, and the top-kriging residuals are in most cases within these lines. The kriging standard errors from point kriging are always high, whereas the kriging standard errors from constrained kriging are somewhat smaller. There are, however, more points from constrained kriging which are plotted outside the lines.

The two panels on the bottom show the *z*-score (residual divided by kriging standard error) of the three methods (2, 6 and 1 *z*-scores have absolute values above 7 for top-kriging, constrained kriging and point kriging, respectively). This *z*-score should ideally have a $N(0,1)$ distribution, which is plotted with a line in the figure. We can see that neither of the methods gives a *z*-score with a normal distribution, but top-kriging and constrained kriging deviates less from this distribution than point kriging.

4.4.2. Prediction

The aim of top-kriging is to predict for locations without measurements. Predictions can be made by a call to *rtopKrige* as for the cross-validation:

```
rtopObj = rtopKrige(rtopObj)
```

The function will first create variance matrices between the observations catchments and between observation catchments and prediction catchments if these are not already a part of *rtopObj*. The result will be in the element called *prediction* of *rtopObj*, accessible as *rtopObj$predictions*.

Some optional parameters to this function (also possible in the cross-validation above) are *nmax*, the maximum number of neighbors in a kriging neighborhood, *maxdist*, the maximum distance to neighbors and *wlim*, which is an upper limit to the sum of the absolute weights. Whereas the first two are for local kriging, partly to avoid wrong results if the assumption of a constant mean is wrong, and for reduction of computation time, the last parameter is a method to limit the effect of possible numerical problems that cause large weights, both positive and negative.

To plot the predictions on the river network, it is necessary to join the predicted values with the river network shape file. In the example in this paper, we match the code *EZGID* of the prediction (and observation) catchments with the code *EZGA* of the river network.

```
rnet = readOGR(rpath, "riverNetwork")
pred = rtopObj$predictions
rnet$pred = predvar1.pred[match(rnet$EZGA, pred
  $EZGID)]
spplot(rnet, "pred", col.regions = bpy.colors())
```

At this point, the river network only contains the segments with predictions, and there will be some missing lines for all the segments that represent observations. We can therefore add these observations to the network, and also create a dot with a color code equal to the river network at the gaging stations to see where observations have been made:

```
at = seq(0, max(rnet$pred, na.rm =TRUE), 0.01)
cols = bpy.colors(length(at))
```

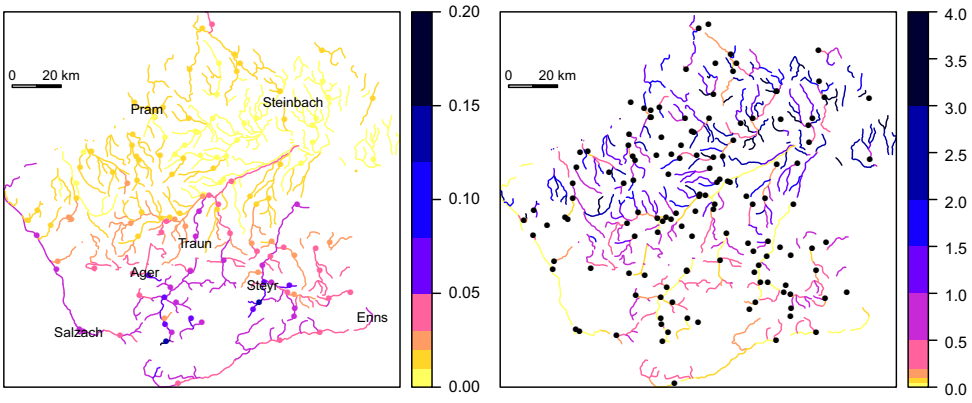

Fig. 7. Left: Predictions of specific average summer runoff ($m^3/s/km^2$) plotted on river segments together with observations as points at the stations, including names of some river segments. Right: estimated uncertainty presented as a coefficient of variation (kriging standard deviation divided by the prediction itself).

```
cobs = observations@data[, c("XSTATION",
   "YSTATION", "obs")]
names(cobs) = c("x", "y", "obs")
coordinates(cobs) = ~ x + y
cobs$class = findInterval(cobs$obs, at)
```

```
spplot(rnet, "pred", col.regions = bpy.colors(),
   at = at,
   panel = function(x, y, ...){
    panel.polygonsplot(x, y, ...)
    sp.points(cobs[, "obs"], cex = 1, pch = 16,
              col = cols[cobs$class]))})
```

This makes it easy to visualize how the predictions match the observations, as seen in the left panel of Fig. 7. One can easily see that the observations are fairly consistent with the observations. The deviations are mostly in areas where there are large deviations between adjacent stations.

We can also look at the uncertainty of the predictions by plotting the coefficient of variation (CV: kriging standard deviation divided by the prediction itself). This is shown in the right panel of Fig. 7, where the river segments are colored according to the CV and the stations are shown as dots. We can notice two things. First of all the CV is consistently higher for the segments representing small catchments, the larger catchments have smaller CV. We can also note that the CV is slightly higher in the northern part of the figure, something that comes from the lower observations and predictions in this region. Note that we have clipped some of the lowest values of the variance, some being negative due to the numerical issues, and some of the highest values of the CV, for better contrasts.

After visualizing the result, it is also easy to write the result in any vector format, such as a shapefile

```
writeOGR(rnet, dsn, layer, "ESRI Shapefile")
```

where *dsn* and *layer* refer to the directory and the name of the shapefile to be created, respectively.

4.5. Prediction efficiency and limitations

The computation time will, in addition to the number of observation and prediction locations, depend on the parameter settings, such as the choice between Ghosh-distance or full integration of the variogram, and on the number of discretization points. The optimal choice of the latter will depend on the correlation structure and on the distribution of observation locations. However, below we will show some analyses of different parameter settings for illustrational purposes.

The computation time particularly depends on the choice between Ghosh-distance and full integration of the semivariogram. Fig. 8 shows the computation time for doing different tasks of the interpolation process, depending on the number of observations and predictions, and of different parameter settings. There is first of all a large difference in computation time depending on the use of Ghosh-distance (dashed lines) or integration of the variogram (solid lines). Second, the total computation time appears

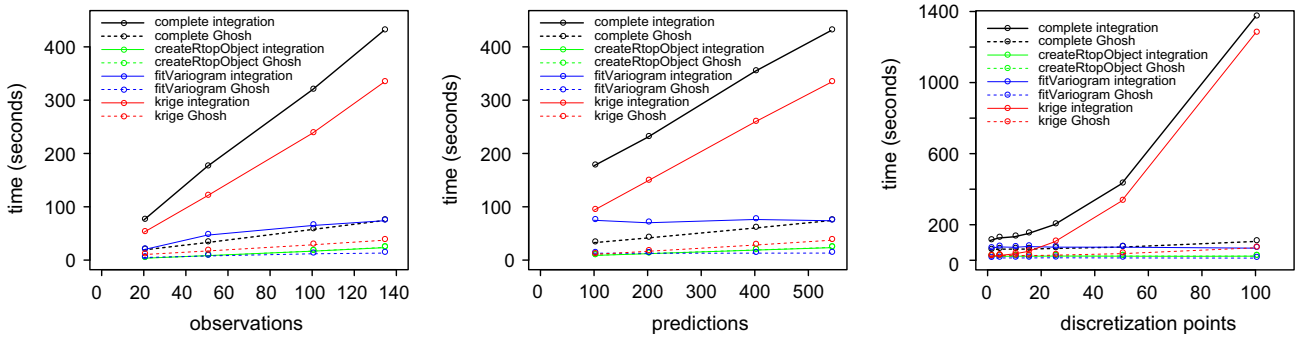

Fig. 8. Computation time for different parts of the interpolation process as a function of observations, prediction locations and resolution. The number of observations, prediction locations and discretization points are 134, 542 and 50, respectively, for the cases where they are not varied.

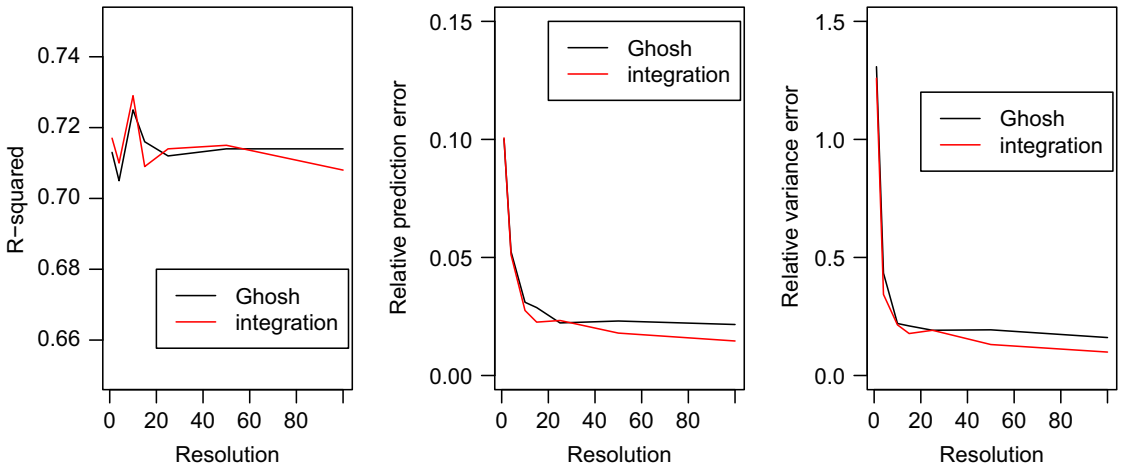

Fig. 9. Prediction efficiency as a function of resolution. Panels from the left: *R*-squared, relative prediction error (compared to an average of predictions with high resolution interpolation) and relative kriging variance error (compared to the same predictions).

linearly dependent on the number of observations and the number of prediction locations. The resolution seems on the other hand to have an O^2 effect on the computation time for full integration, whereas the effect is less pronounced when using the Ghosh-distance; there is only a small increase in computation time with increasing resolution.

The effect of different parameters on the computation errors will depend on the spatial correlation and on the distribution of observation and prediction locations. However, for the example in this paper, we can in the left panel of Fig. 9 see that good results (high R^2) can be achieved with relatively few discretization points. This figure is based on an average of 10 cross-validation runs (the variogram refitted every time) with different resolution, and the highest R^2 is observed when the number of discretization points is set to 10, which means that there can be 10–40 discretization points per catchment. There is only a small difference between the results using Ghosh-distance and full integration.

The two next panels are also based on a comparison with the average of the results from 10 cross-validations per resolution. However, here we compare with an average of the cross-validation results using full integration and 100 discretization points, as these are the results that correspond to what should in theory be the best prediction. The absolute prediction error (kriging variance error) for each pixel has been found by subtracting and dividing by the prediction (kriging variance) of the average described above, and then taking the norm. These relative errors reach a relatively stable level for 10–15 discretization points for the predictions and 15–25 discretization points for the kriging variance. It is therefore in many cases not necessary to use a large number of discretization points, not even for a good kriging variance.

5. Conclusions

An R package for interpolating observations with a non-point support has been developed. This package is easy to use, is an open source, and is developed within the R environment, which can handle a large range of formats of input and output, and simplifies creation of graphical output and diagnostics. The versatility of R makes it possible to make most, if not all, of the analyses within one framework, where it was earlier necessary to exchange data between spreadsheets, visualization tools, analysis tools and GIS. Some other R packages have some of the functionalities in *rtop*, but none of them are able to fit variograms and make predictions from a set of shapefiles without preprocessing.

The coefficients of determination (R^2) of the predictions indicate that *rtop* performs better than the other two methods, although the differences in the visual comparisons were minor in the example in this paper. We did not fully explore all possible parameter settings of the *constrainedKriging*-package, so that it is possible that better results could be obtained with further experimentation with these. A possible reason for the small visual differences is that large rivers in Upper Austria are draining a range of different geological formations, which contributes to the large differences in the observed values of specific discharge. The diversity of hydrogeological conditions then violates the stationarity assumption of kriging, and an alternative approach could be to split the study area into homogeneous regions.

However, we will claim that the method is based on a conceptually better theory than ordinary point kriging for the same data, and publishing an R package for this method allows for using geostatistical methods within a field where such methods have only been used in a limited set of applications. The method also estimates kriging standard errors that can be used as a measure of the prediction uncertainty, which cannot be correctly estimated with ordinary kriging from this type of data.

The methods provided by this package are optimal in the least squares sense as long as the main underlying assumption, the intrinsic hypothesis, is valid. Particularly when used for streamflow prediction, the data set will often consist of a combination of highly variable runoff with large uncertainty and short spatial correlation in and close to mountains, and lower specific runoff with lower uncertainty and longer distance correlation in flatter regions. Further developments should probably focus on methods which can improve the predictions for these cases. A second limitation is that the method is not likely to perform well when the size of the prediction areas is very different from the size of the observed areas.

The prediction efficiency of top-kriging will to a large degree depend on the density of gauges. Although not tested here, it is likely that the difference between point kriging and top-kriging will be larger for higher gauge density. Top-kriging explicitly takes the nested structure of the catchments into account and will therefore give better predictions for highly nested datasets. The difference is likely to be smaller where few of the gauges are nested. For non-overlapping data (such as some administrative records), the advantage will depend on the configuration of administrative boundaries. Estimates of prediction uncertainty will in all cases be better with a top-kriging approach, as point kriging does not take the lower sampling uncertainty for larger supports into account.

There are several choices to be made by the users of this package. Our results indicate that the use of Ghosh-distance achieves similar results as the integration of the variogram. The small differences observed in Fig. 9 are most likely not significant. And although we only looked at one particular example in this paper, it seems like the number of discretization points can be kept fairly low. Use of Ghosh-distance and the number of discretization points are currently the two most important choices regarding computational speed. Liu and Journel (2009) have suggested an alternative method for integration of the variogram, based on a fast fourier transform (FFT), a method which has been parallelized by Guan et al. (2011).

We have not properly analysed the memory consumption, as it will depend on more than the parameter settings. However, for this example, the largest memory consumption is caused by the shapefiles of the catchments. With more regular polygons, and a larger number of prediction locations or observation locations, the covariance matrices can be even larger. These are matrices of size $nObs \times nObs$ and $nObs \times nPred$ where *nObs* and *nPred* are the number or observation locations and prediction locations, respectively. This will, depending on the available memory, usually not be an issue before the number of observations or predictions is in the order of several thousands. A large number of prediction locations are not a problem as long as the number of observation locations is not too high (much more than 1000).

Some tests (not shown in this paper) also indicate that it is preferable, both from a computational and a numerical point of view to work with binned variograms when the number of observations is in the order of 30–40 or higher, similar to what is done for similar cases in ordinary kriging. The use of a nugget effect in regularization is rather uncommon. When redoing the examples of this paper with a separate regularization of the nugget effect, this effect will for many pairs of catchments be of the same magnitude as the regularized semivariance. An alternative could be to use a nested variogram with a short range (1 km in the example of this paper) for the small scale variability.

Acknowledgements

We would like to thank two anonymous reviewers for their constructive comments and suggestions which helped to improve the quality of this manuscript.

References

Barillec, R., Ingram, B., Cornford, D., Csató, L., 2011. Projected sequential Gaussian processes: a c++ tool for interpolation of large datasets with heterogeneous noise. Comput. Geosci. 37 (3), 295–309.

Bivand, R.S., Pebesma, E., Gómez-Rubio, V., 2013. Applied Spatial Data Analysis with R, Use R, 2nd Ed., vol. 10, Springer, New York, USA.

Blöschl, G., Sivapalan, M., Wagener, T., Viglione, A., Savenjie, H., 2013. Runoff Prediction in Ungauged Basins—Synthesis Across Processes, Places and Scales. Cambridge University Press.

Chilés, J.-P., Delfiner, P., 1999. Geostatistics. Modelling Spatial Uncertainty. Wiley, New York

Cressie, N., 1985. Fitting variogram models by weighted least squares. Math. Geol. 17 (5), 563–586.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, a Wiley-Interscience Publication, New York, NY

de Marsily, G., 1986. Quantitative Hydrogeology. Academic Press Inc., London

Duan, Q., Sorooshian, S., Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour. Res. 28 (4), 1015–1031.

Gandin, L.S., 1963. Objective Analysis of Meteorological Fields (in Russian). Gidrometeorologicheskoe Izdate'stvo, Leningrad, U.S.S.R.

Ghosh, B., 1951. Random distances within a rectangle and between two rectangles. Bull. Calcutta Math. Soc. 43, 17–24.

Goovaerts, P., 2006. Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point poisson kriging. Int. J. Health Geograph. 5 (art. no. 52).

Goovaerts, P., 2008. Kriging and semivariogram deconvolution in the presence of irregular geographical units. Math. Geosci. 40 (1), 101–128.

Gottschalk, L., 1993. Interpolation of runoff applying objective methods. Stoch. Hydrol. Hydraul. 7, 269–281.

Gottschalk, L., Krasovskaia, I., Leblois, E., Sauquet, E., 2006. Mapping mean and variance of runoff in a river basin. Hydrol. Earth Syst. Sci. 10, 469–484.

Gottschalk, L., Leblois, E., Skøien, J.O., 2011. Distance measures for hydrological data having a support. J. Hydrol. 402 (3–4), 415–421.

Gotway, C.A., Young, L.J., 2002. Combining incompatible spatial data. J. Am. Stat. Assoc. 97 (458), 632–648.

Guan, Q., Kyriakidis, P.C., Goodchild, M.F., 2011. A parallel computing approach to fast geostatistical areal interpolation. Int. J. Geograph. Inf. Sci. 25 (8), 1241–1267.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J., Heuvelink, G.B., 2008. Real-time automatic interpolation of ambient gamma dose rates from the dutch radioactivity monitoring network. Comput. Geosci. 35 (8), 1711–1721.

Hofer, C., Papritz, A., 2011. Constrainedkriging: An R-package for customary, constrained and covariance-matching constrained point or block kriging. Comput. Geosci. 37 (10), 1562–1569.

Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford University Press, New York

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, London, UK

Keitt, T.H., Bivand, R., Pebesma, E., Rowlingson, B., 2012. rgdal: Bindings for the Geospatial Data Abstraction Library. R Package Version 0.7-12. URL ⟨http://CRAN.R-project.org/package=rgdal⟩.

Kyriakidis, P., 2004. A geostatistical framework for area-to-point spatial interpolation. Geograph. Anal. 36 (3), 259–289.

Laaha, G., Skøien, J.O., Blöschl, G., 2014. Spatial prediction on river networks: comparison of top-kriging with regional regression. Hydrol. Processes 28 (2), 315–324.

Laaha, G., Skøien, J.O., Nobilis, F., Blöschl, G., 2013. Spatial prediction of stream temperatures using top-kriging with an external drift. Environ. Model. Assess. 18 (6), 671–683.

Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51 (1), 137–157.

Liu, Y., Journel, A.G., 2009. A package for geostatistical integration of coarse and fine scale data. Comput. Geosci. 35 (3), 527–547.

Malone, B.P., McBratney, A.B., Minasny, B., Wheeler, I., 2012. A general method for downscaling earth resource information. Comput. Geosci. 41, 119–125.

Pebesma, E., Cornford, D., Dubois, G., Heuvelink, G.B.M., Hristopulos, D., Pilz, J., Stöhlker, U., Morin, G., Skøien, J.O., 2011. Intamap: the design and implementation of an interoperable automated interpolation web service. Comput. Geosci. 37 (3), 343–352.

Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL ⟨http://www.R-project.org/⟩.

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., 2013. INLA: Functions Which Allow to Perform Full Bayesian Analysis of Latent Gaussian Models Using Integrated Nested Laplace Approximaxion. R Package Version 0.0-1379326596.

Sauquet, E., Gottschalk, L., Leblois, E., 2000. Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. Hydrol. Sci. J. 45 (6), 799–815.

Skøien, J.O., Blöschl, G., 2007. Spatio-temporal top-kriging of runoff time series. Water Resour. Res. 43, W09419.

Skøien, J.O., Merz, R., Blöschl, G., 2006. Top-kriging—geostatistics on stream networks. Hydrol. Earth Syst. Sci. 10, 277–287.

Skøien, J.O., Pebesma, E.J., Blöschl, G., 2008. Geostatistics for automatic estimation of environmental variables—some simple solutions. Georisk 2 (4), 257–270.

VerHoef, J.M., Peterson, E.E., 2010. A moving average approach for spatial statistical models of stream networks (with discussion). J. Am. Stat. Assoc. 105, 6–18.