NONPARAMETRIC GEOSTATISTICS WITH THE NPSP PACKAGE

Rubén Fernández-Casal
Universidade da Coruña (Spain), Centro de investigación CITIC
ruben.fcasal@udc.es

Abstract

In this work the `R` package `npsp` (Nonparametric spatial statistics) is presented. This package implements nonparametric methods for inference on multidimensional geostatistical processes, avoiding the misspecification problems that may arise when using parametric models. The spatial process can be either stationary or show a non-constant trend. Joint estimation of the trend and the semivariogram can be performed automatically, by using the function `np.fitgeo`, or by a step-by-step approach. .

Introduction

- We will assume that $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$ is a spatial process that can be modeled as:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

where $\mu(\cdot)$ is the trend function and the error term ε, is a second order stationary process with zero mean and covariogram $C(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, with $\mathbf{u} \in D$.

- In this framework, given n observed values $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^t$, the first step consists in estimate the trend $\mu(\mathbf{x})$ and the semivariogram $\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{u})$.

- There are several `R` packages, such as `gstat` or `geoR`, which implement the traditional geostatistical techniques to approximate these functions. Nevertheless, as these approaches usually assume parametric models, they can present misspecification problems.

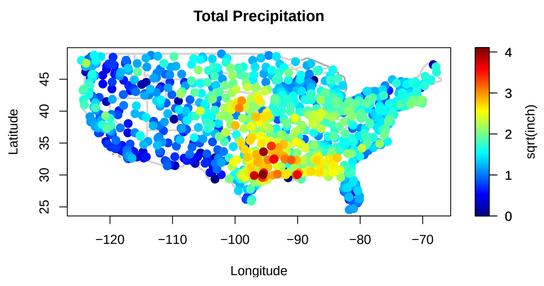
- The aim of the `npsp` package is to provide nonparametric tools for geostatistical modeling and interpolation, under the general spatial model (1), and without assuming any specific form (parametric model) for the trend and the variogram of the process.

```
library(npsp)
```

- The `precipitation` data set, supplied with the `npsp` package, will be used in the examples in this work. The data consist of total precipitations (square-root of rainfall inches) during March 2016 recorded over 1053 locations on the continental part of USA.

- For instance, `spoints()` or `scattersplot()` functions may be used to perform a descriptive analysis of the data.

```
spoints(precipitation)
```



Nonparametric estimation

Local polynomial trend estimation

- The local polynomial trend estimator $\hat{\mu}_{\mathbf{H}}(\mathbf{x})$ (e.g. [5]), obtained by polynomial smoothing of $\{(\mathbf{x}_i, Y(\mathbf{x}_i)) : i = 1, \ldots, n\}$, is the solution for β_0 to the least squares minimization problem

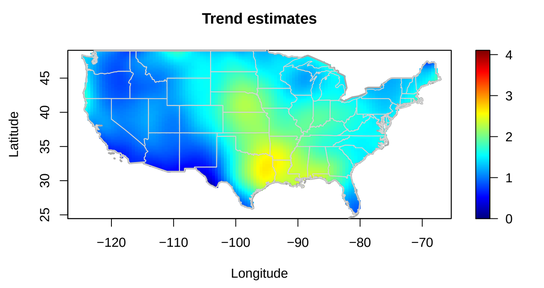
$$\min_{\beta_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p} \sum_{i=1}^{n} \Big\{ Y(\mathbf{x}_i) - \beta_0 - \boldsymbol{\beta}_1^t (\mathbf{x}_i - \mathbf{x}) - \ldots - \boldsymbol{\beta}_p^t (\mathbf{x}_i - \mathbf{x})^p \Big\}^2 K_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x}), \tag{2}$$

where $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix; K is a d-dimensional kernel and $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$.

- The bandwidth matrix $\mathbf{H}$ controls the shape and size of the local neighborhood used to estimate $\mu(\mathbf{x})$.

- The local trend estimates can be computed in practice with the `locpol()` function. A full bandwidth matrix and a multiplicative triweight kernel is used to compute the weights. Main calculations are performed in Fortran using the LAPACK library.

```
x <- coordinates(precipitation)
y <- precipitation$y
lp <- locpol(x, y, nbin = c(120,120), h = diag(c(5,5)))
```



- The smoothing procedures in `npsp` use linear binning to discretize the data. Usually two grids will be considered, a low resolution one to speed up computations during the modeling (e.g. by using the default values), and a higher resolution one to obtaing the final results.

Bandwidth selection

- Traditional bandwidth selectors, such as cross validation (CV) or generalized cross validation (GCV), do not have a good performance for dependent data [9], since they tend to undersmooth the trend function.

- The modified cross-validation criteria (MCV), proposed in [2] for time series, can be adapted for spatial data, by ignoring observations in a neighbourhood $N(i)$ around $\mathbf{x}_i$. Note that the ordinary CV approach is a particular case with $N(i) = \{\mathbf{x}_i\}$.

- An alternative is the corrected generalized cross-validation criterion (CGCV), proposed in [8], that takes the spatial dependence into account.

- In practice, the bandwidth can be selected through the `h.cv()` function.

```
bin <- binning(x, y)
lp0.h <- h.cv(bin)$h
lp0 <- locpol(bin, h = lp0.h, hat.bin = TRUE)
```

Variogram estimation

- When the process is assumed stationary, the pilot local linear estimate $\hat{\gamma}(\mathbf{u})$ is obtained by the linear smoothing of $\left(\mathbf{x}_i - \mathbf{x}_j, (Y(\mathbf{x}_i) - Y(\mathbf{x}_j))^2/2\right)$. The corresponding bandwidth can be selected, for instance, by minimizing the cross-validation relative squared error.

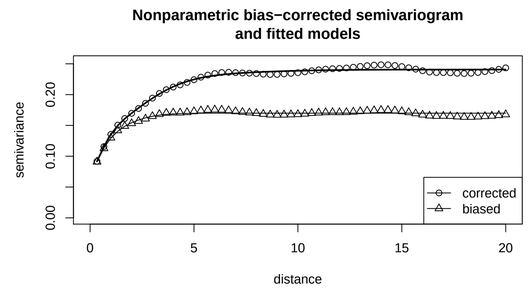
- In the general setting (1), the natural approach consists in removing the trend and estimating the variogram from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{SY}$. Nevertheless, the residuals variability may be very different to that of the true errors (see e.g. [3], Section 3.4.3, for the case of the linear trend estimator). As the bias due to the direct use of residuals in variogram estimation may have a significant impact on inference, a similar approach to that described in [6] could be considered.

- In practice, the local linear variogram estimates can be computed with functions `np.svar()` and `np.svar.corr()`. The generic function `h.cv()` may be used to select the corresponding bandwidth.

```
svar.bin <- svariso(x, residuals(lp0), nlags = 60,
                    maxlag = 20)
svar.h <- h.cv(svar.bin)$h
svar.np <- np.svar(svar.bin, h = svar.h)
svar.np2 <- np.svariso.corr(lp0, nlags = 60, maxlag = 20,
                    h = svar.h, plot = FALSE)
```

- The final variogram estimate is obtained by fitting a "nonparametric" isotropic Shapiro-Botha variogram model [10], to the nonparametric pilot estimate, by using the function `fitsvar.sb.iso()`.

```
svm0 <- fitsvar.sb.iso(svar.np2, dk = 0)
```

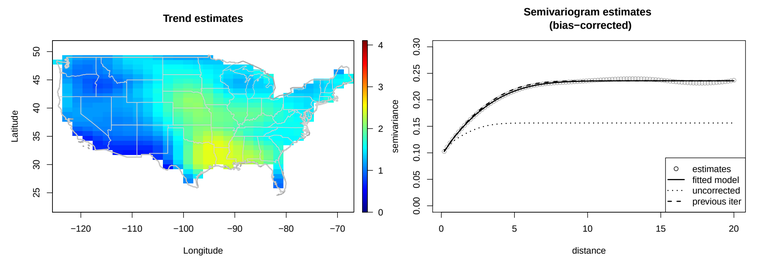


Automatic modeling algorithm

- Function `np.fitgeo()` implements an automatic procedure for the joint estimation of the trend and the semivariogram. The algorithm is as follows:

1. Select an initial bandwidth to estimate the trend using the MCV criterion.
2. Compute the local linear trend estimate (2).
3. Obtain a bias-corrected pilot semivariogram estimate from the corresponding residuals (similar to that described in [6]).
4. Fit a valid Shapiro-Botha model [10] to the pilot semivariogram estimates.
5. Select a new bandwidth matrix for trend estimation with the CGCV criterion, using the fitted variogram model to approximate the data covariance matrix.
6. Repeat steps 2-4 until convergence (in practice, only two iterations are usually needed).

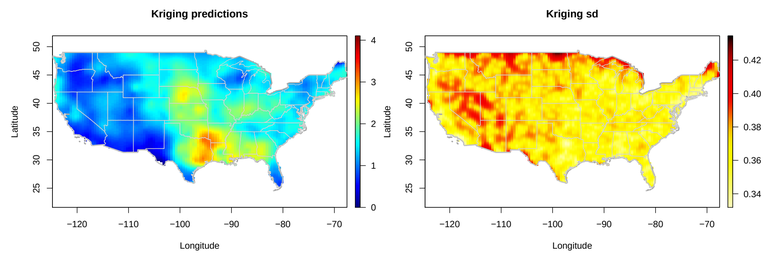
```
geomod <- np.fitgeo(x, y, nbin = c(30, 30), maxlag = 20,
                    svm.resid = TRUE)
```



Kriging

- The final trend and variogram estimates could be employed for spatial prediction, by using method `np.kriging()`.

```
krig.grid <- np.kriging(geomod, ngrid = c(120, 120))
```



- Currently, only global residual kriging is implemented. Users are encouraged to use `gstat::krige()` (or `krige.cv()`) together with `as.vgm()` for local kriging.

Conclusions and future work

- Unlike traditional geostatistical methods, the nonparametric procedures avoid problems due to model misspecification.

- The bias due to the direct use of residuals, may have a significant impact on variogram estimation (and consequently on the selected bandwidth with the CGCV criterion). Therefore, the use of the bias-corrected nonparametric variogram estimator would be recommended.

- Currently, only isotropic semivariogram estimation is supported, but the intention is to extend this approach to anisotropy in two components [7], which could be suitable for modeling spatio-temporal dependency.

- The nonparametric trend and variogram estimates also allow to perform inferences about other characteristics of interest of the process, for instance, by using the bootstrap algorithm described in [1], which will be implemented in the next version of the package.

Acknowledgments

This research has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the European Regional Development Fund (ERDF).

References

[1] Castillo-Páez, S., Fernández-Casal, R., García-Soidán, P., 2019. A nonparametric bootstrap method for spatial data. *Comput. Stat. Data An.*, **137**, 1–15.

[2] Chu, C. K. and Marron, J. S. (1991). Comparison of Two Bandwidth Selectors with Dependent Errors. *The Annals of Statistics* **19**, 1906–1918.

[3] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

[4] Fernández-Casal, R. (2019). npsp: Nonparametric Spatial Statistics. R package version 0.7-5. *http://github.com/rubenfcasal/npsp*.

[5] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

[6] Fernández-Casal, R. and Francisco-Fernández, M. (2014). Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment* **28**, 1247–1259.

[7] Fernández-Casal, R., González Manteiga, W. and Febrero-Bande, M. (2003). Flexible Spatio-Temporal Stationary Variogram Models. *Statistics and Computing* **13**, 127–136.

[8] Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[9] Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

[10] Shapiro, A. and Botha, J.D. (1991). Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.