EnvirometriX

Consultancy, Research and Innovation

Machine Learning (ranger package) as a framework for spatial and spatiotemporal prediction

tom.hengl@enviometrix.net

@tom_hengl

thengl

http://envirometrix.net

Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables

Research article | Biogeography | Soil Science | Computational Science | Data Mining and Machine Learning | Spatial and Geographic Information Science

Tomislav Hengl[1], Madlene Nussbaum[2], Marvin N Wright[3], Gerard B.M. Heuvelink[4]

March 14, 2018

Author and article information

Abstract

Random forest and similar Machine Learning techniques are already used to generate spatial predictions, but spatial location of points (geography) is often ignored in the modeling process. Spatial auto-correlation, especially if still existent in the cross-validation residuals, indicates that the predictions are maybe biased, and this is suboptimal. This paper presents a random forest for spatial predictions framework (RFsp) where buffer distances from observation points are used as explanatory variables, thus incorporating geographical proximity effects into the prediction process. The RFsp framework is illustrated with examples that use textbook datasets and apply

README.md

RFsp — Random Forest for spatial data (R tutorial)

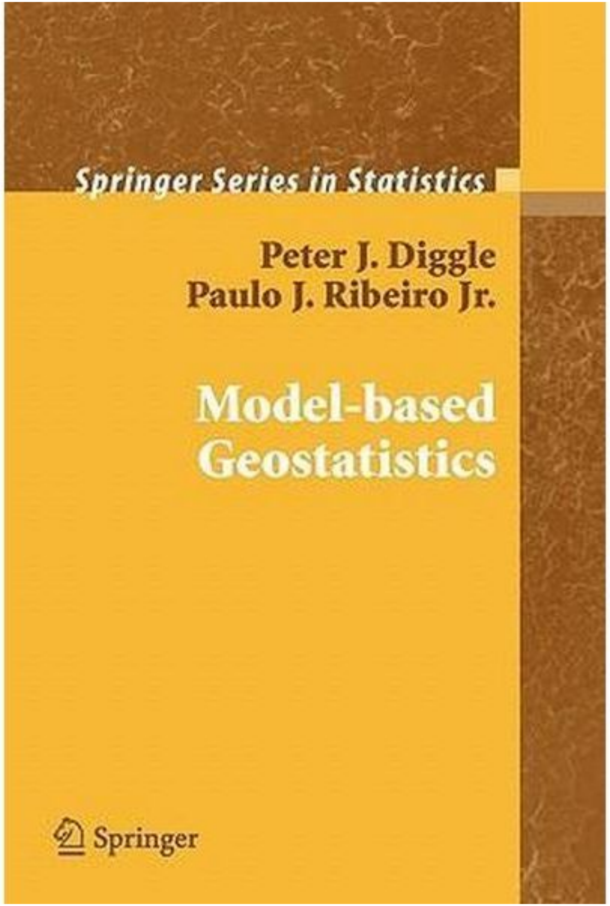
Hengl, T., Nussbaum, M., and Wright, M.N.

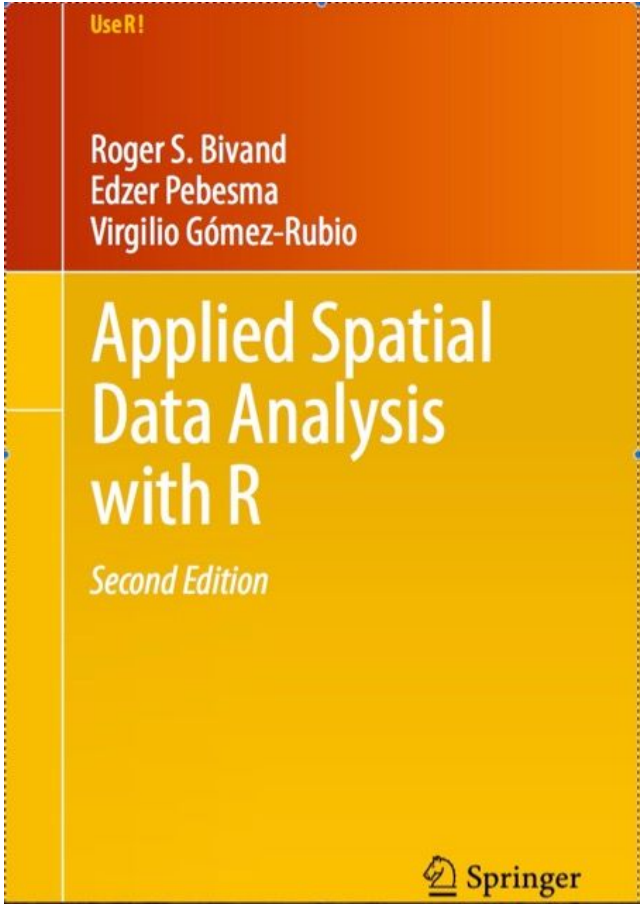


https://github.com/thengl/GeoMLA

Abstract: This tutorial explains how to use Random Forest to generate spatial and spatiotemporal predictions (i.e. to make maps from point observations using Random Forest). Spatial auto-correlation, especially if still existent in the cross-validation residuals, indicates that the predictions are maybe biased, and this is suboptimal. To account for this, we use Random Forest (as implemented in the ranger package) in combination with geographical distances to sampling locations to fit models and predict values. We describe eight typical situations of interest to spatial prediction applications: (1) prediction of 2D

Kriging has been a synonym for geostatistics since 1960s





Standard steps

1. Determine distribution of the target variable and appropriate transformation (normal, log-normal, zero-inflated, Gamma, Poissonic ...)

2. Fit variogram (WLS, REML, ...), deal with multicolinearity (PCA?), non-stationary properties, support size, mixed effects…

3. Predict (mean values and uncertainty)

4. Validate predictions (mapping accuracy)

Variogram modeling and predictions (kriging)

```
R> zinc.vgm <- likfit(zinc.geo, lambda = 0,
ini=c(var(log1p(zinc.geo$data)), 500), cov.model
= "exponential")

R> zinc.ok <- krige.conv(zinc.geo, locations =
locs, krige = krige.control(obj.m = zinc.vgm))
```

krige.conv: model with constant mean
krige.conv: performing the Box-Cox data transformation
krige.conv: back-transforming the predicted mean and variance
krige.conv: Kriging performed using global neighbourhood

Web Images More...

Google regression-kriging

Scholar About 3,930 results (**0.06** sec) My Citat

Articles

Case law

My library

Any time

Since 2017

Since 2016

Since 2013

Custom range...

Sort by relevance

Sort by date

include patents

include citations

Create alert

A generic framework for spatial prediction of soil variables based on **regression-kriging** [PDF] researchgate.net

T Hengl, GBM Heuvelink, A Stein - Geoderma, 2004 - Elsevier

A methodological framework for spatial prediction based on **regression-kriging** is described and compared with ordinary kriging and plain regression. The data are first transformed using logit transformation for target variables and factor analysis for continuous predictors

Cited by 673 Related articles All 16 versions Web of Science: 352 Cite Saved

Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and **regression-kriging**

IOA Odeh, AB McBratney, DJ Chittleborough - Geoderma, 1995 - Elsevier

Several methods involving spatial prediction of soil properties from landform attributes are compared using carefully designed validation procedures. The methods, tested against ordinary kriging and universal kriging of the target variables, include multi-linear regression,

Cited by 511 Related articles All 9 versions Web of Science: 303 Cite Saved

About **regression-kriging**: from equations to case studies [PDF] researchgate.net

T Hengl, GBM Heuvelink, DG Rossiter - Computers & geosciences, 2007 - Elsevier

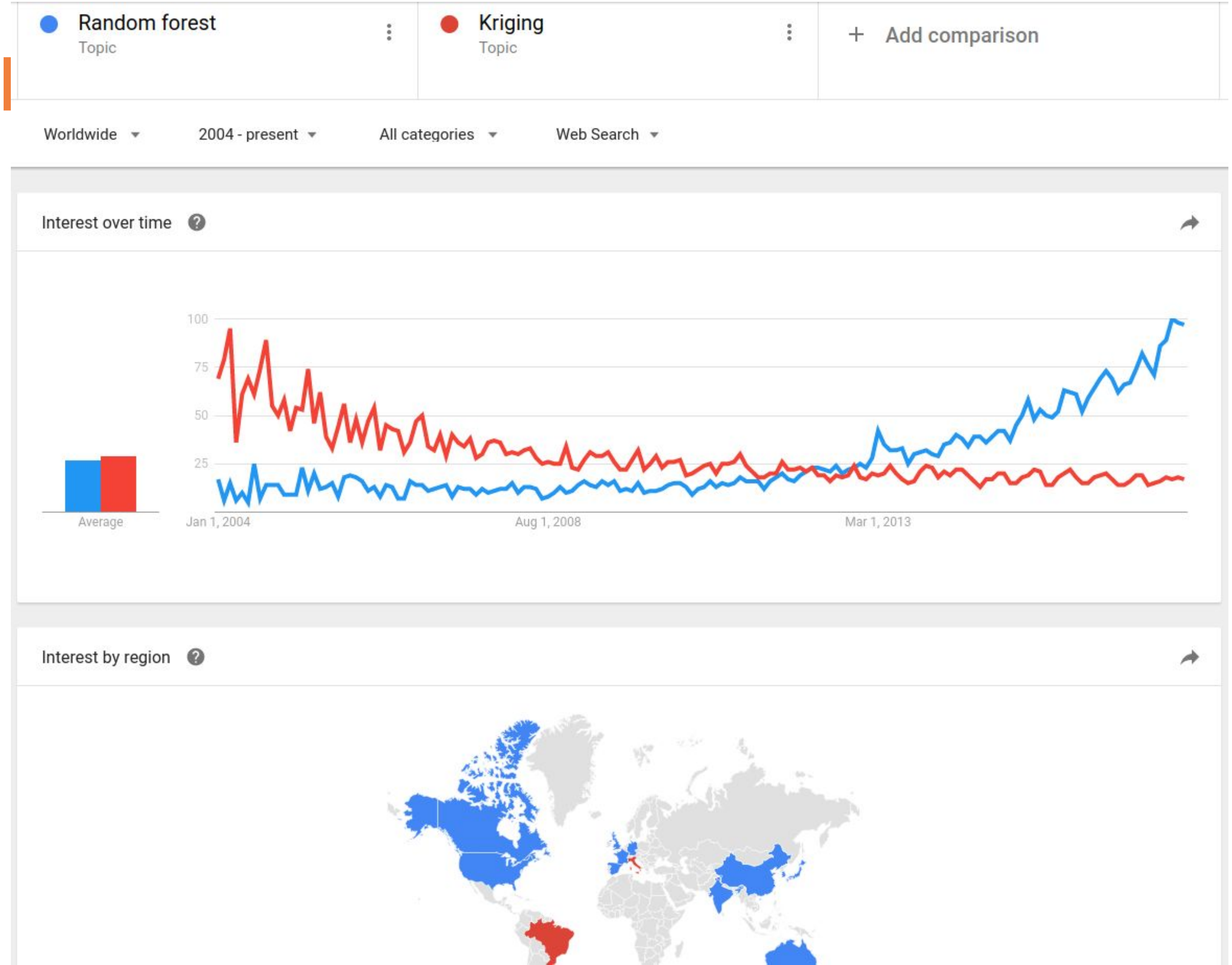
This paper discusses the characteristics of **regression-kriging** (RK), its strengths and limitations, and illustrates these with a simple example and three case studies. RK is a spatial interpolation technique that combines a regression of the dependent variable on

Cited by 435 Related articles All 10 versions Web of Science: 259 Cite Saved

[BOOK] Index [PDF] academia.edu

R Webster, MA Oliver - 1999 - Wiley Online Library

... 159–160 kriging with trend 195–211 E-BLUP 202 kriging with external drift 203–205 universal kriging 196–203 lognormal kriging 184–185 mapping 173–174, 181–191 ordinary kriging


Random forest
Topic
Kriging
Topic
+ Add comparison
Worldwide
2004 - present
All categories
Web Search
Interest over time
100
75
50
25
Average
Jan 1, 2004
Aug 1, 2008
Mar 1, 2013
Interest by region

MLA is interesting for generating sp, however…

Spatial auto-correlation, especially if still existent in the cross-validation residuals, indicates that the predictions are maybe biased, and this is suboptimal.

To account for this, we use Random Forest (as implemented in the ranger package) in combination with geographical distances to sampling locations to fit models and predict values.

Solution: RFsp

$$Y(\mathbf{s}) = f(\mathbf{X}_G, \mathbf{X}_R, \mathbf{X}_P) \tag{18}$$

where $\mathbf{X}_G$ are covariates accounting for geographical proximity and spatial relations between observations

$$\mathbf{X}_G = (d_{p1}, d_{p2}, \ldots, d_{pN}) \tag{19}$$

where d_{pi} is the buffer distance (or any other complex proximity upslope/downslope distance, as explained in the next section) to the observed location pi from $\mathbf{s}$ and N is the total number of training points. $\mathbf{X}_R$ are surface reflectance covariates, i.e. usually spectral bands of remote sensing images, and $\mathbf{X}_P$ are process-based covariates.

Geographical distances (proximity)

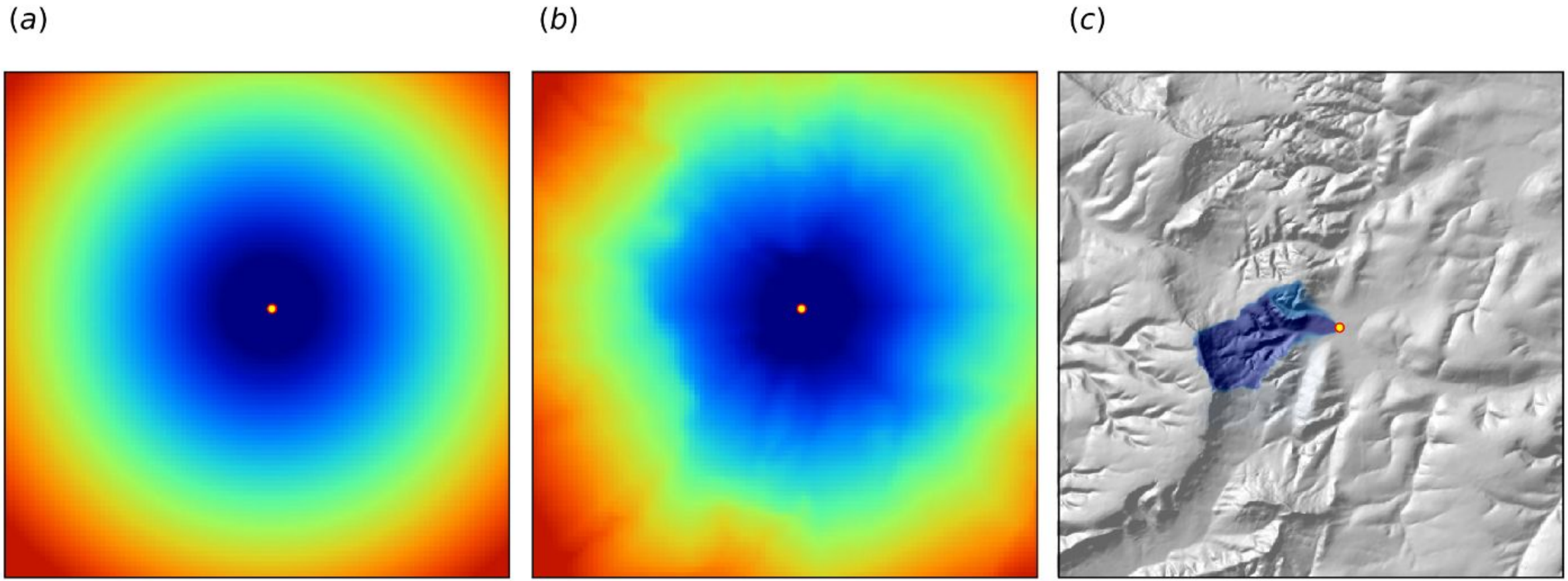
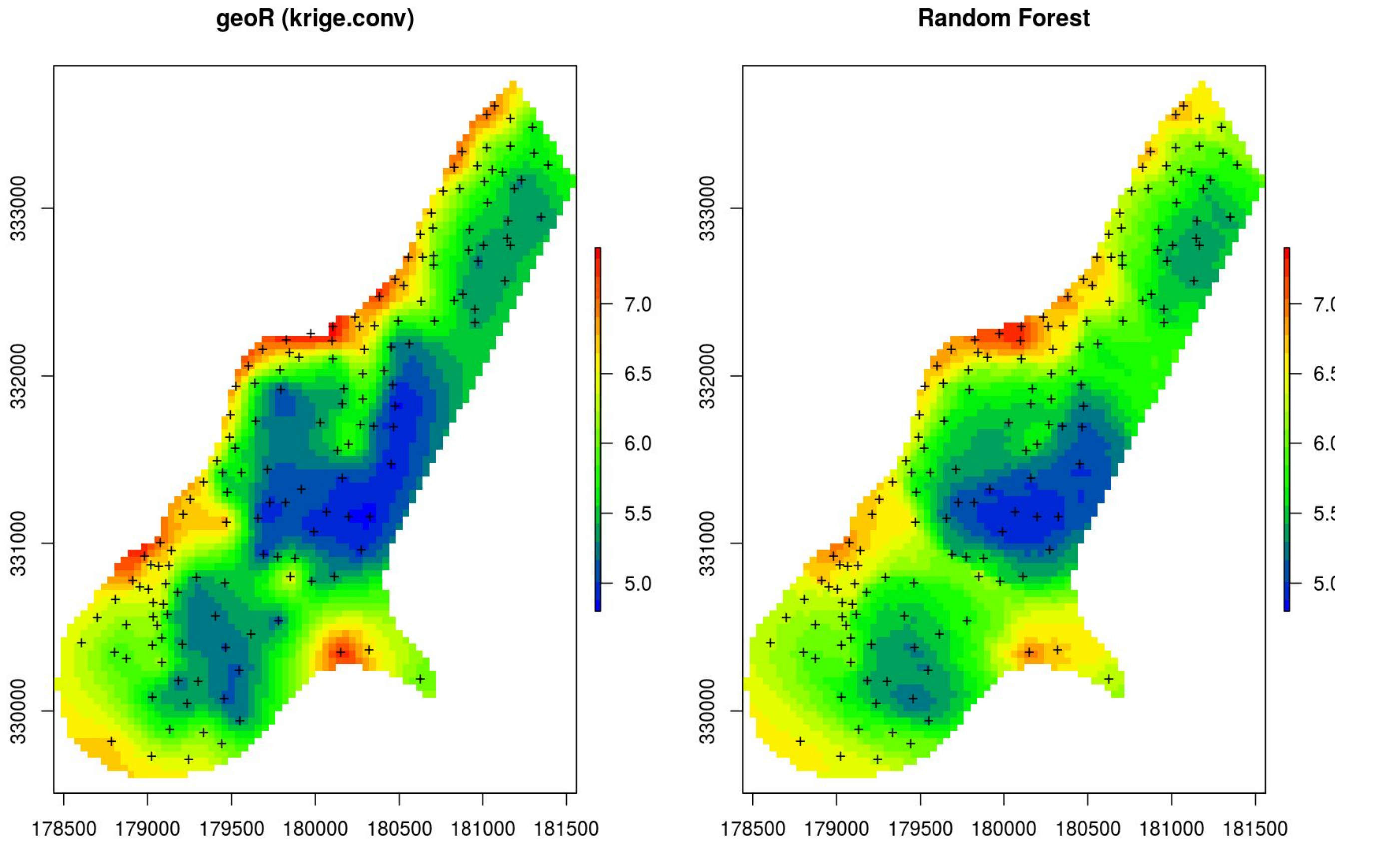


Figure 2. Examples of distance maps to some location in space (yellow dot) based on different derivation algorithms: (a) simple Euclidean distances, (b) complex speed-based distances based on the gdistance package and Digital Elevation Model (DEM) (van Etten, 2017), and (c) upslope area derived based on the DEM in SAGA GIS (Conrad et al., 2015). Case study: Ebergötzen (Böhner et al., 2006).

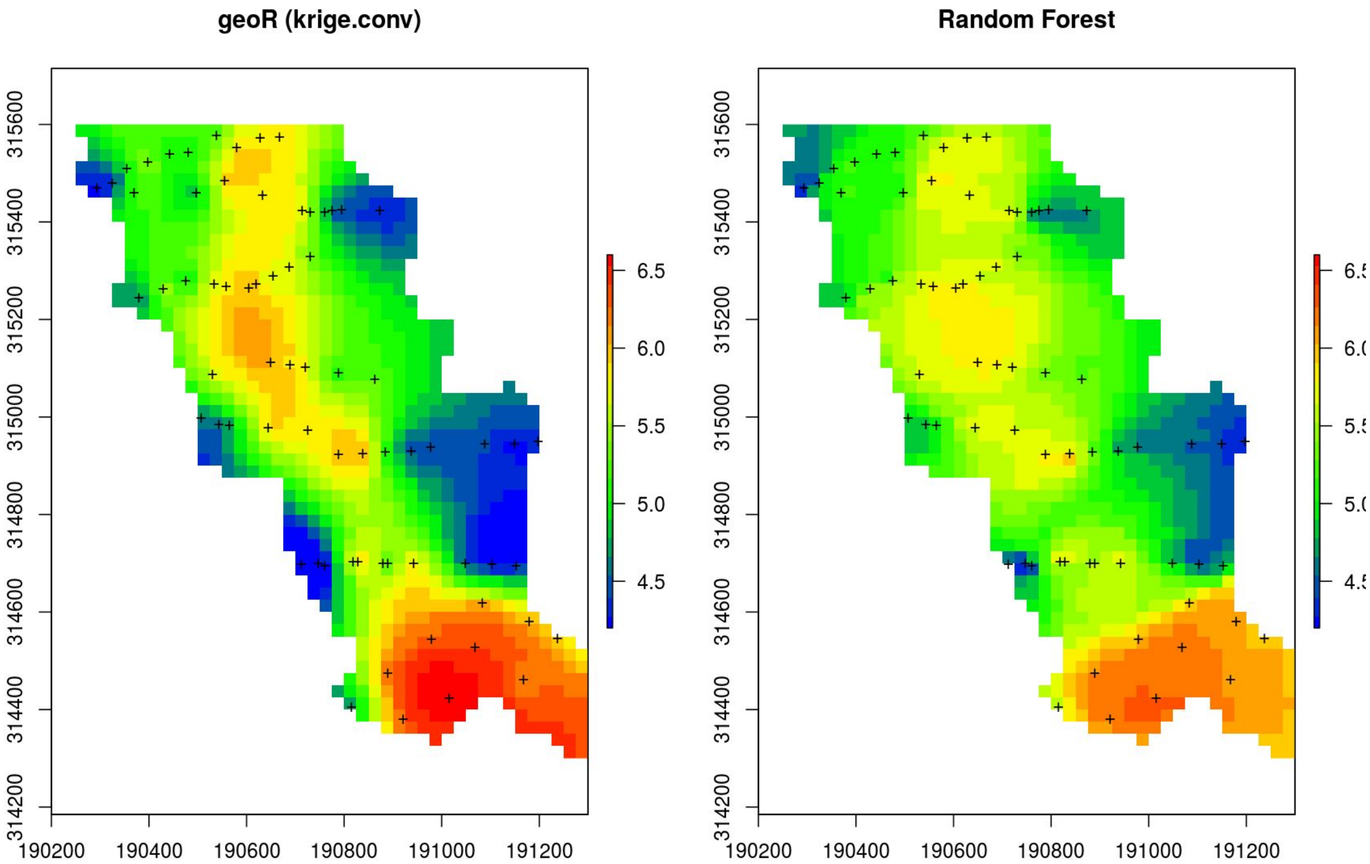
Variogram modeling and predictions (kriging)

```
R> grid.dist0 <- buffer.dist(meuse["zinc"],
meuse.grid[1], as.factor(1:nrow(meuse)))

R> ov.zinc <- over(meuse["zinc"], grid.dist0)

R> m.zinc <- ranger(as.formula(paste("zinc ~",
paste(names(grid.dist0), collapse="+")),
cbind(meuse@data["zinc"], ov.zinc))

R> zinc.rfd <- predict(m.zinc, grid.dist0@data)
```

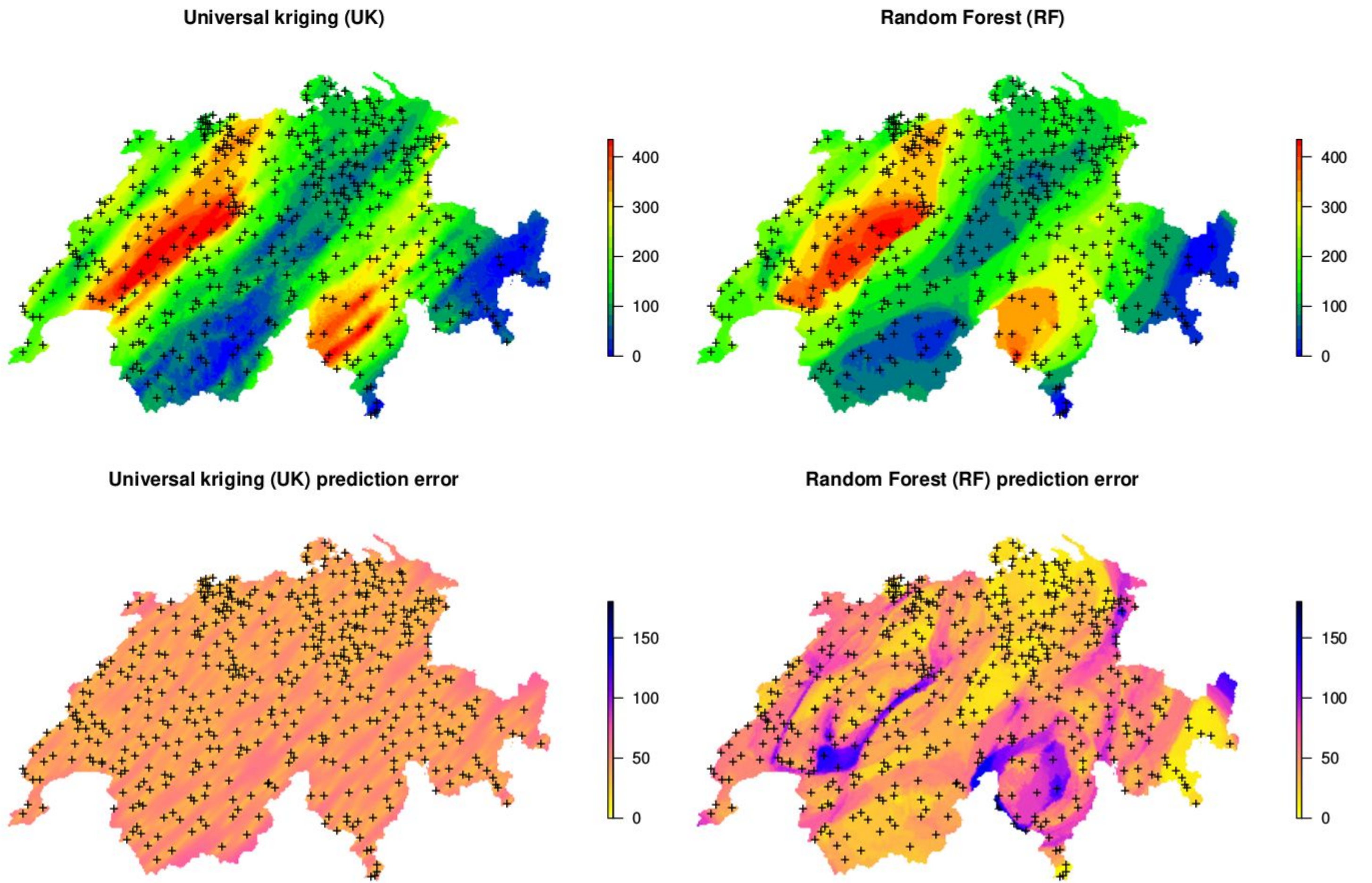
Meuse data set



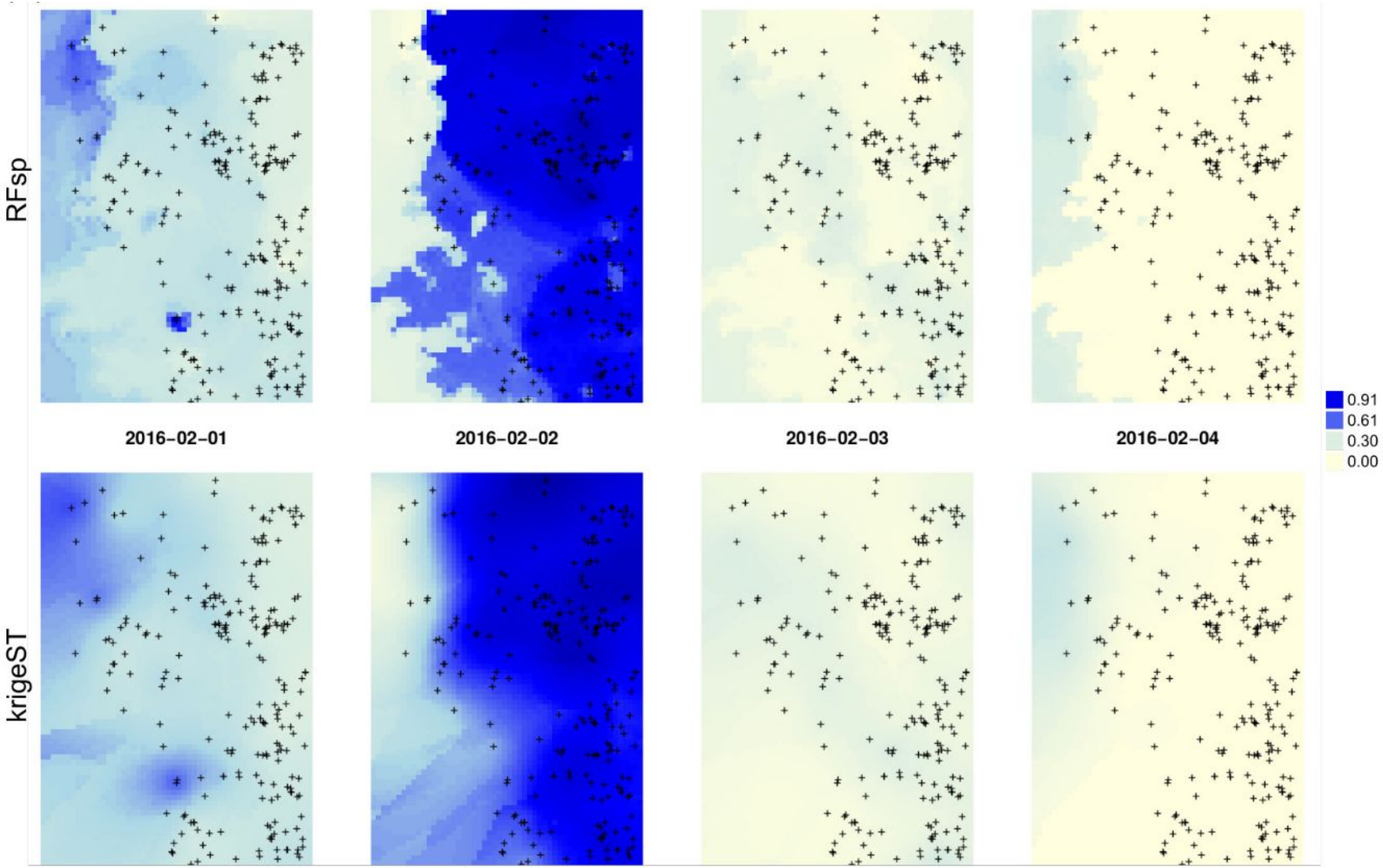
Geul data set



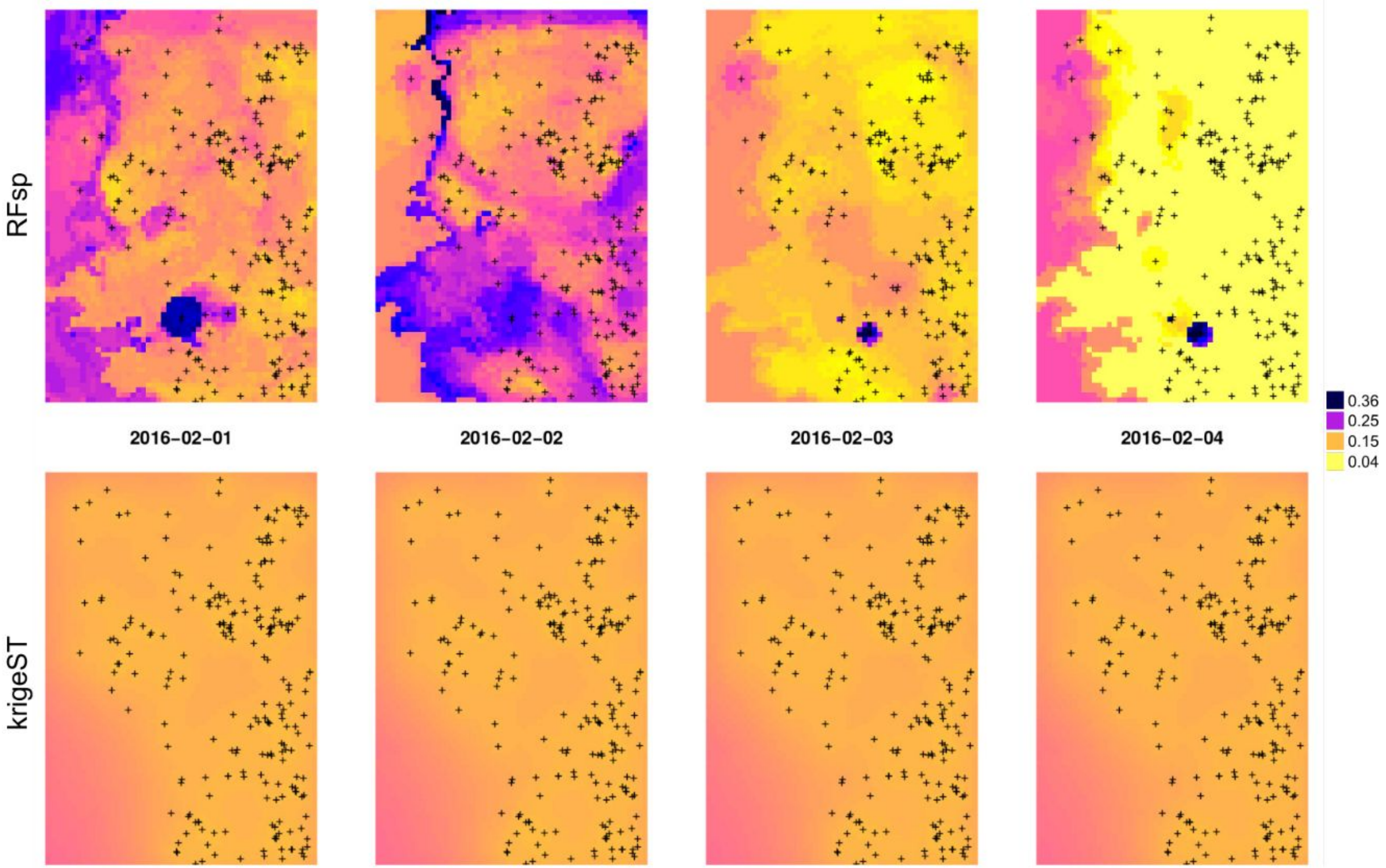
SIC97 data set



Daily precipitation (spatiotemporal)



Daily precipitation (spatiotemporal) prediction error maps



Summary results

Our results indicate that RFsp can produce comparable results to model-based geostatistics. The advantage of RFsp over model-based geostatistics is that RFsp requires much less statistical assumptions and is easier to automate (and scale up through parallelization). For smaller data sets with linear relationships model-based geostatistics could still a better choice.

RFsp is still an experimental method and application with large data sets (>>1000 points) is not recommended.

Advantages of RFsp vs kriging

- ★ No stationarity requirements.
- ★ No Normal distribution requirements.
- ★ No problems with choosing the right variogram (in fact, there is no need for a vgm at all!!).
- ★ No (serious) problems with hot-spots.
- ★ More complex distances can be added.

Problems to solve

1. Extrapolation problems (quality of spatial sampling)
2. Computation intensity very high
3. Validation with spatial declustering (over-fitting problems)
4. Match geostatistical simulations, co-kriging etc.

RF is not a good idea for extrapolation

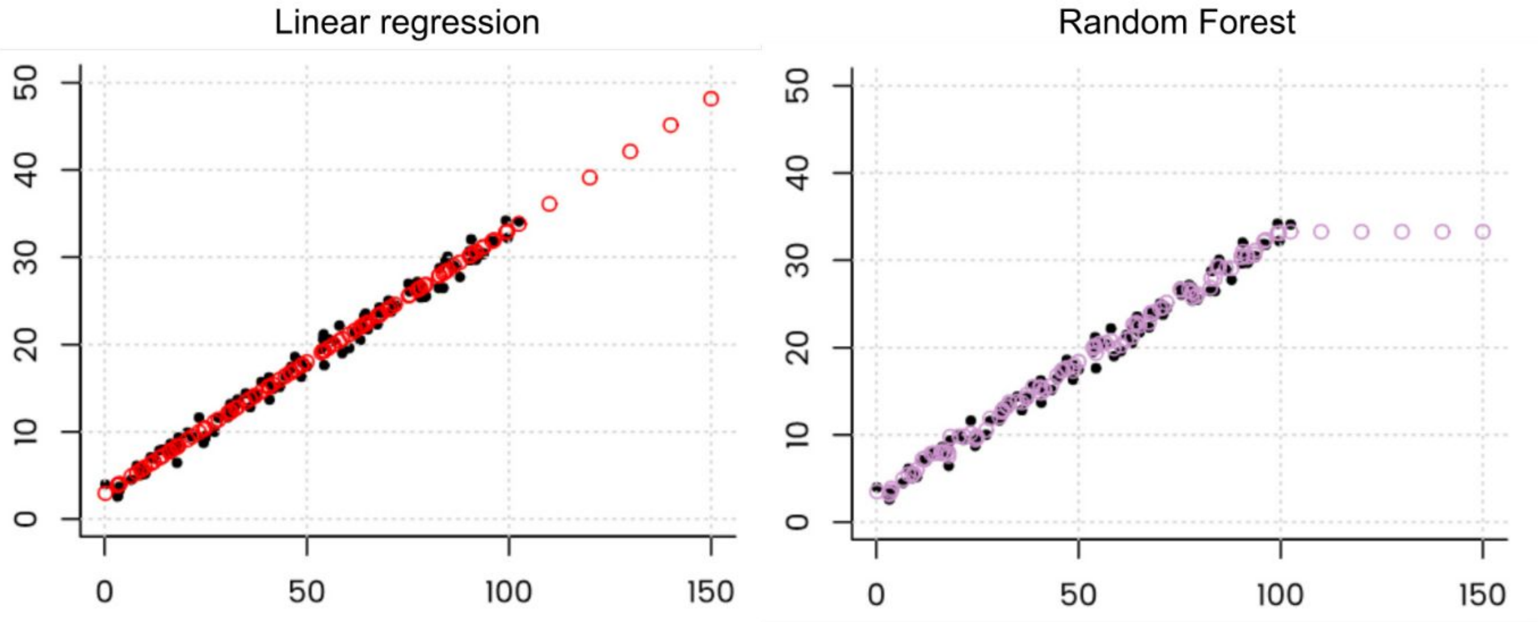


Figure 14. Illustration of the extrapolation problem of Random Forest based on the code examples from Peter Ellis (http://freerangestats.info). Even though Random Forest is more generic than linear regression and can be used also to fit complex non-linear problems, it can lead to completely nonsensical predictions if applied to extrapolation domains.

EnvirometriX