Machine Learning Algorithms for Spatial Prediction problems

tom.hengl@enviometrix.net
@tom_hengl
thengl
http://envirometrix.net

Agenda

- ➔ 9:00–10:30: Theoretical introduction
- ➔ 11:00–12:30: Step-by-step tutorial
- ➔ 12:30–13:30: Lunch
- ➔ 13:30–15:00: Step-by-step tutorial (advanced)
- ➔ 15:30–17:00: Q&A

Preprint

View 6 tweets

NOT PEER-REVIEWED

"*PeerJ Preprints*" is a venue for early communication or feedback before peer review. Data may be preliminary.
Learn more about preprints or browse peer-reviewed articles instead.

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

Mapping Prediction Uncertainty using RFsp

Hengl, T. and Wright, M.N.

https://github.com/thengl/GeoMLA/

There is an increasing interest in using Machine Learning techniques for the purpose of generating spatial predictions, and for mining environmental data in general. Machine Learning algorithms, such as random forests, artificial neural networks and support vector machines have already shown predictive potential for various environmental applications (Biau & Scornet, 2016; Nussbaum et al., 2018; Prasad, Iverson, & Liaw, 2006). This tutorial explains how to use Machine Learning to efficiently generate spatial predictions and derive associated uncertainty. Our focus is on using Random Forest as

Objectives

- ➔ Expose you to some new ideas and new software (a brand new way to do geostatistics!)
- ➔ Point you to literature / tutorials that can help you master MLA
- ➔ Provoke!

Some basic assumptions

- ➔ You work with spatial data and have interest in MLA’s
- ➔ You are familiar with R (but maybe not advanced developer)
- ➔ **You have produced spatial interpolations (maps) from point samples before**

MLA's of interest (CART-based)

- ➔ Random forest
- ➔ Quantile regression random forest
- ➔ Gradient boosting (CART-based)
- ➔ H2o framework for ML
- ➔ Neural nets / deep learning

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

Standard steps in mb geostatistics

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)

Vgm 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

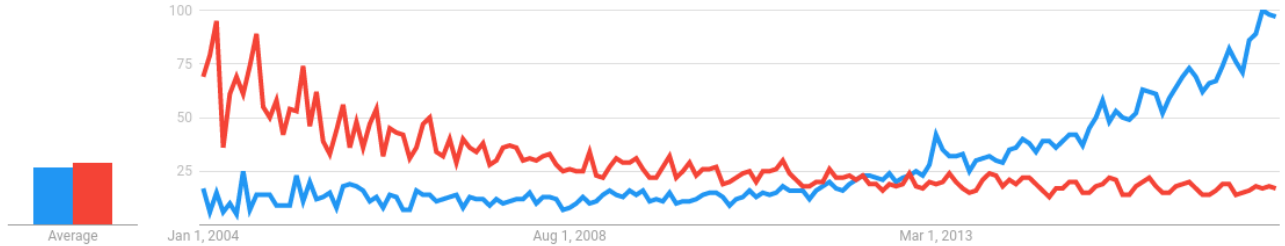
Random forest
Topic

Kriging
Topic

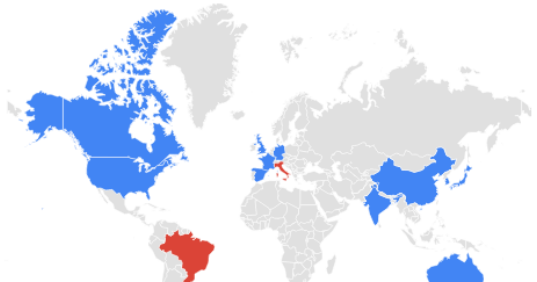
+ Add comparison

Worldwide | 2004 - present | All categories | Web Search

Interest over time



Interest by region



Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Computing

The Extraordinary Link Between Deep Neural Networks and the Nature of the Universe

Nobody understands why deep neural networks are so good at solving complex problems. Now physicists say the secret is buried in the laws of physics.

by Emerging Technology from the arXiv September 9, 2016

In the last couple of years, deep learning techniques have transformed the

COMPUTING

How the Computer Beat the Go Master

As a leading go player falls to a machine, artificial intelligence takes a decisive step on the road to overtaking the natural variety

By Christof Koch on March 19, 2016 أعرض هذا باللغة العربية

South Korean professional Go player Lee Sedol is seen on a TV screen during the

READ THIS NEXT


Computer Beats Go Champion for the First Time


Go Players React to Computer Defeat

AI Software Teaches Itself

ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R

Marvin N. Wright
Universität zu Lübeck

Andreas Ziegler
Universität zu Lübeck,
University of KwaZulu-Natal

Abstract

We introduce the C++ application and R package **ranger**. The software is a fast implementation of random forests for high dimensional data. Ensembles of classification, regression and survival trees are supported. We describe the implementation, provide examples, validate the package with a reference implementation, and compare runtime and memory usage with other implementations. The new software proves to scale best with the number of features, samples, trees, and features tried for splitting. Finally, we show that **ranger** is the fastest and most memory efficient implementation of random forests to analyze data on the scale of a genome-wide association study.

Keywords: C++, classification, machine learning, R, random forests, **Rcpp**, recursive partitioning, survival analysis.

A Random Forest Guided Tour

Gérard Biau
Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
& Institut universitaire de France
gerard.biau@upmc.fr

Erwan Scornet
Sorbonne Universités, UPMC Univ Paris 06, F-75005, Paris, France
erwan.scornet@upmc.fr

Abstract

The random forest algorithm, proposed by L. Breiman in 2001, has

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)

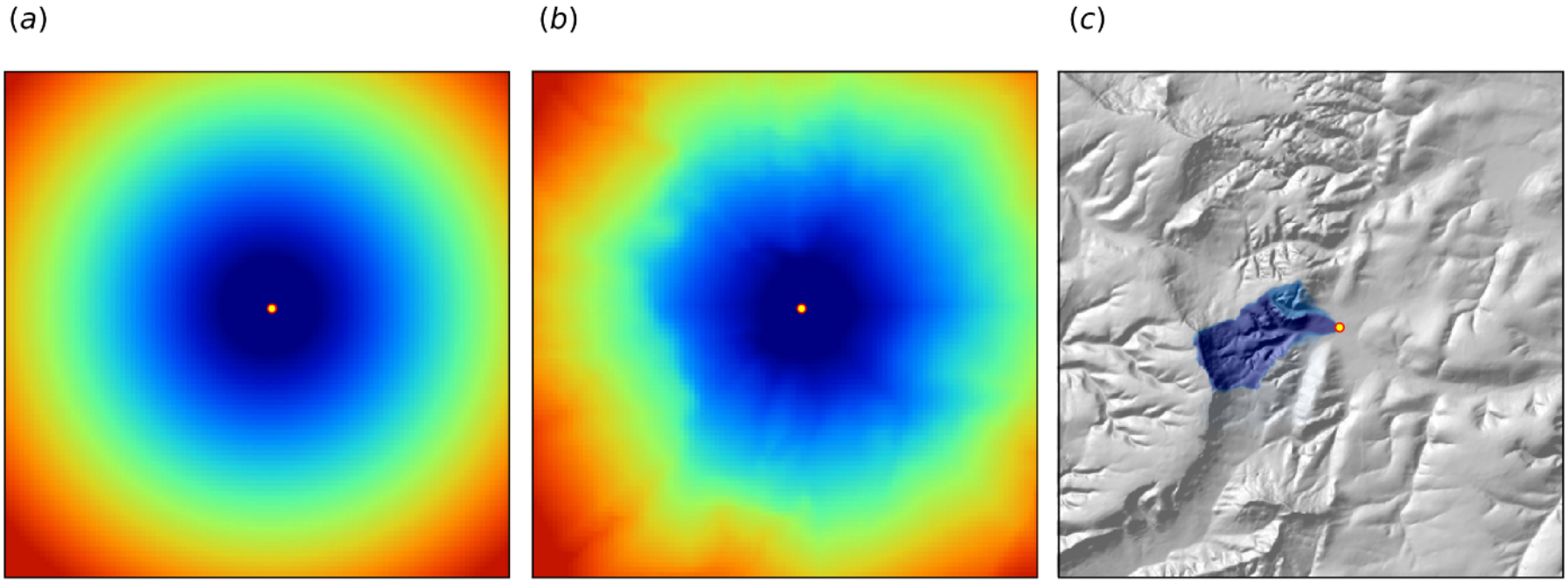
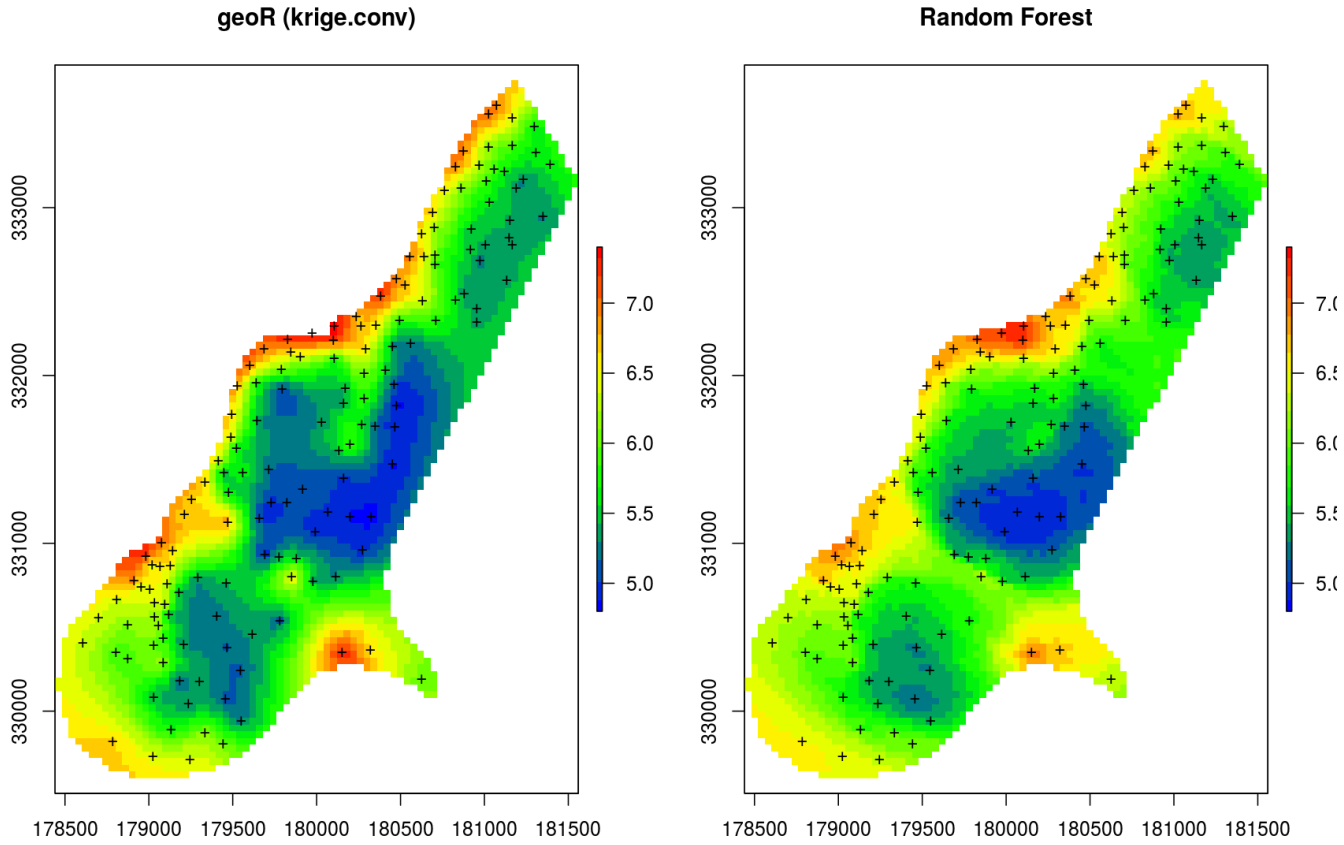


Figure 2. Examples of distance maps to some location in space (yellow dot) based on different derivation algorithms: (a) simple Euclidean distances, (b) complex speed-based distances based on the gdistance package and Digital Elevation Model (DEM) (van Etten, 2017), and (c) upslope area derived based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

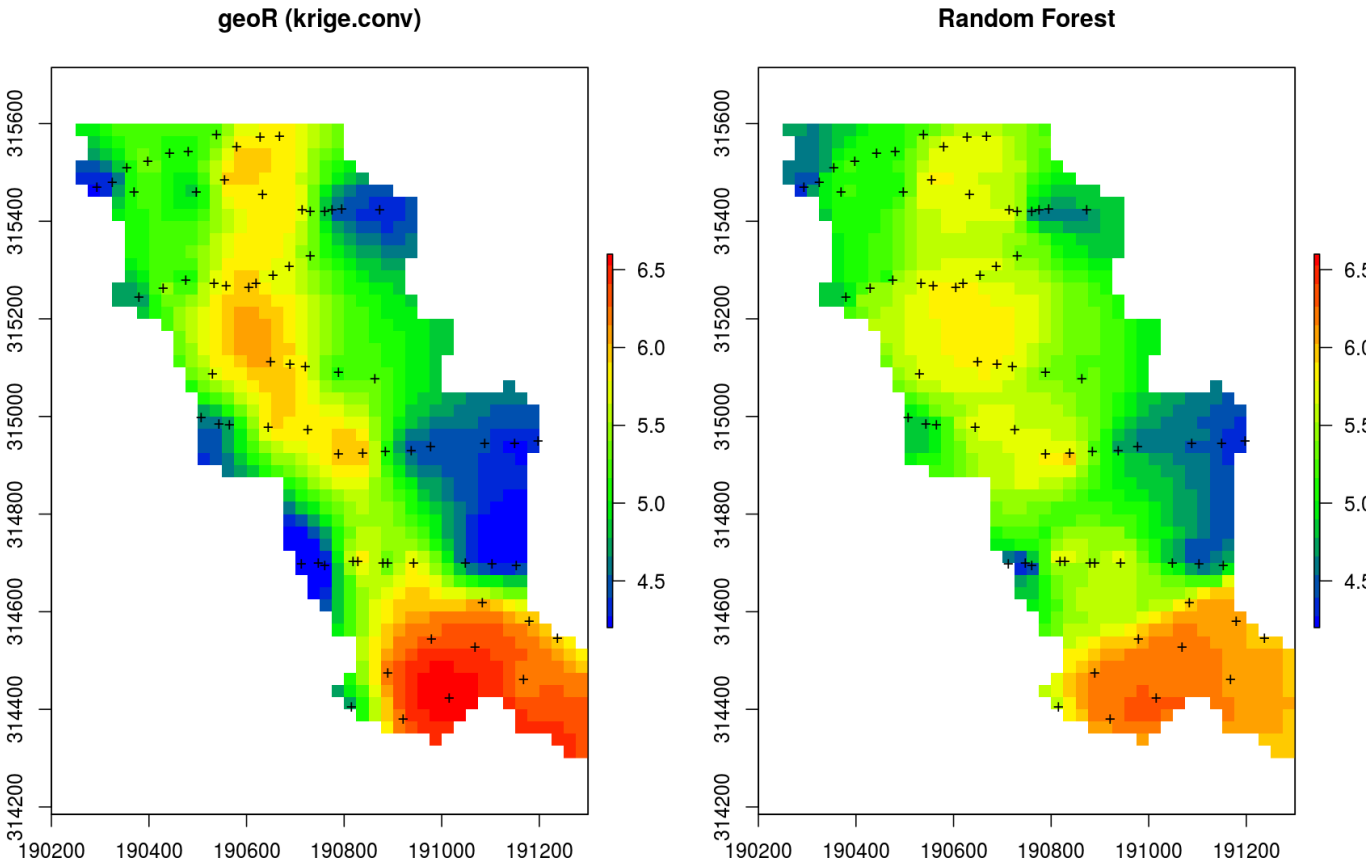
Vgm modeling and predictions (kriging)

```
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)
```

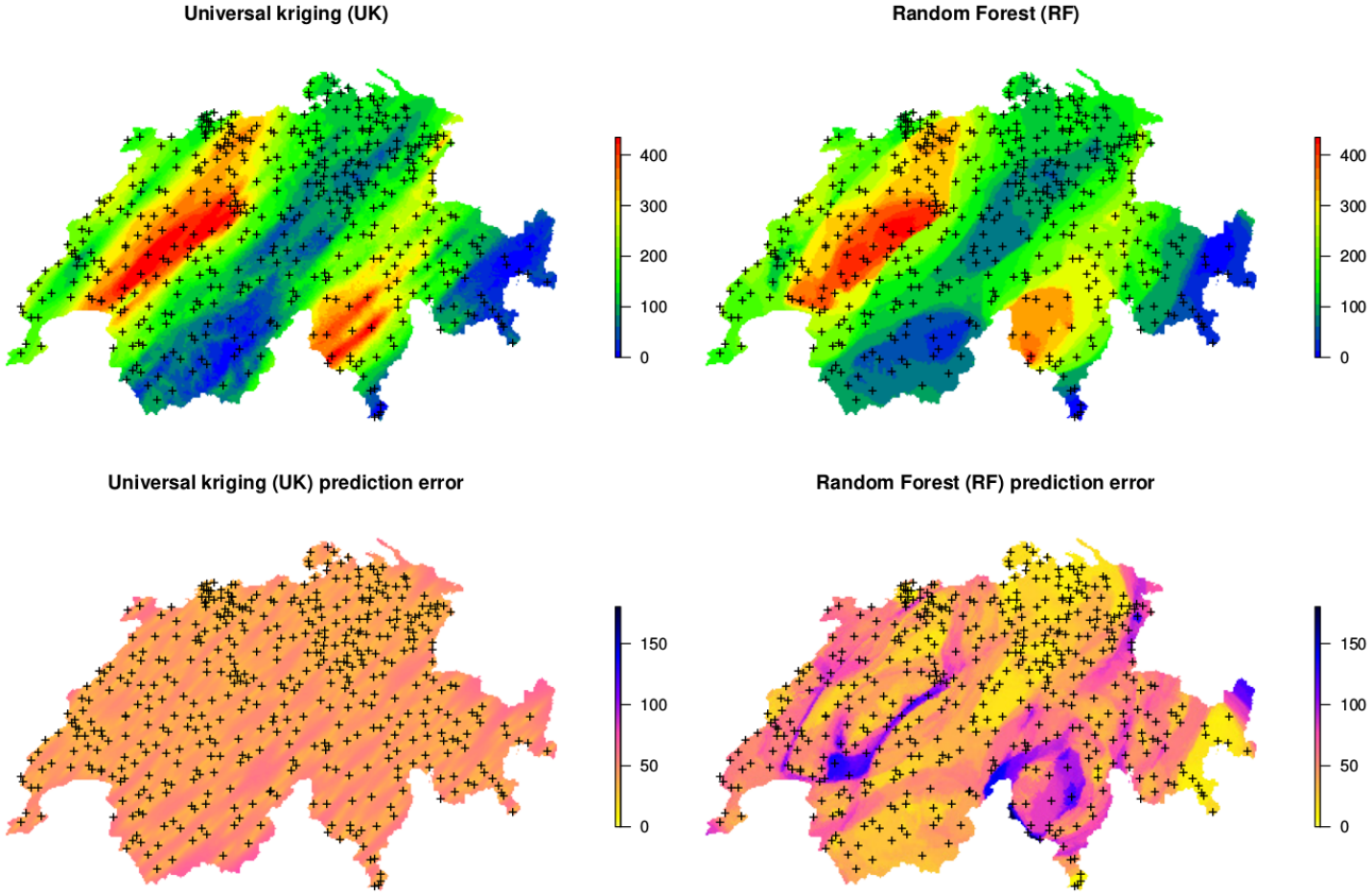
Meuse data set

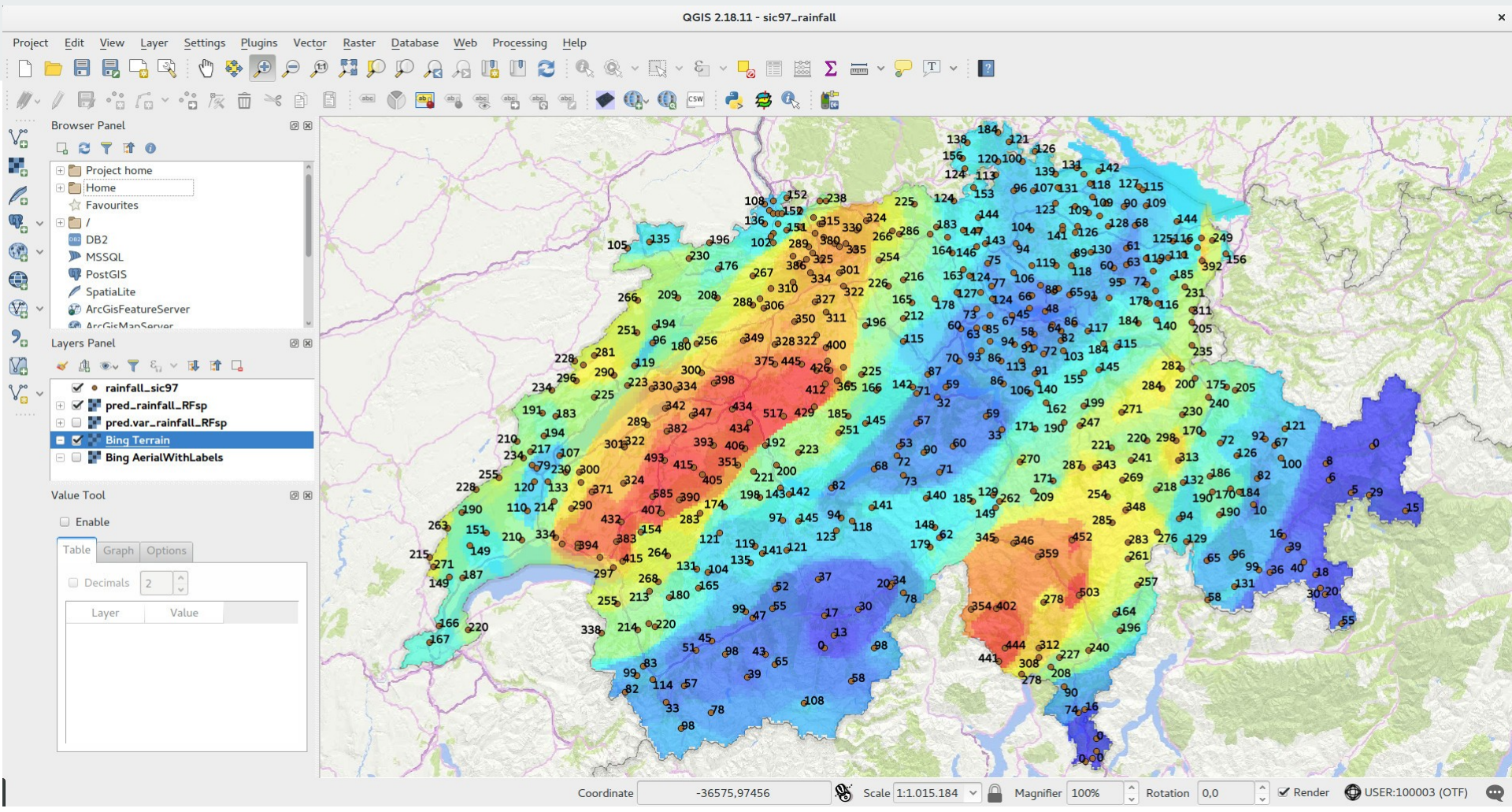


Geul data set

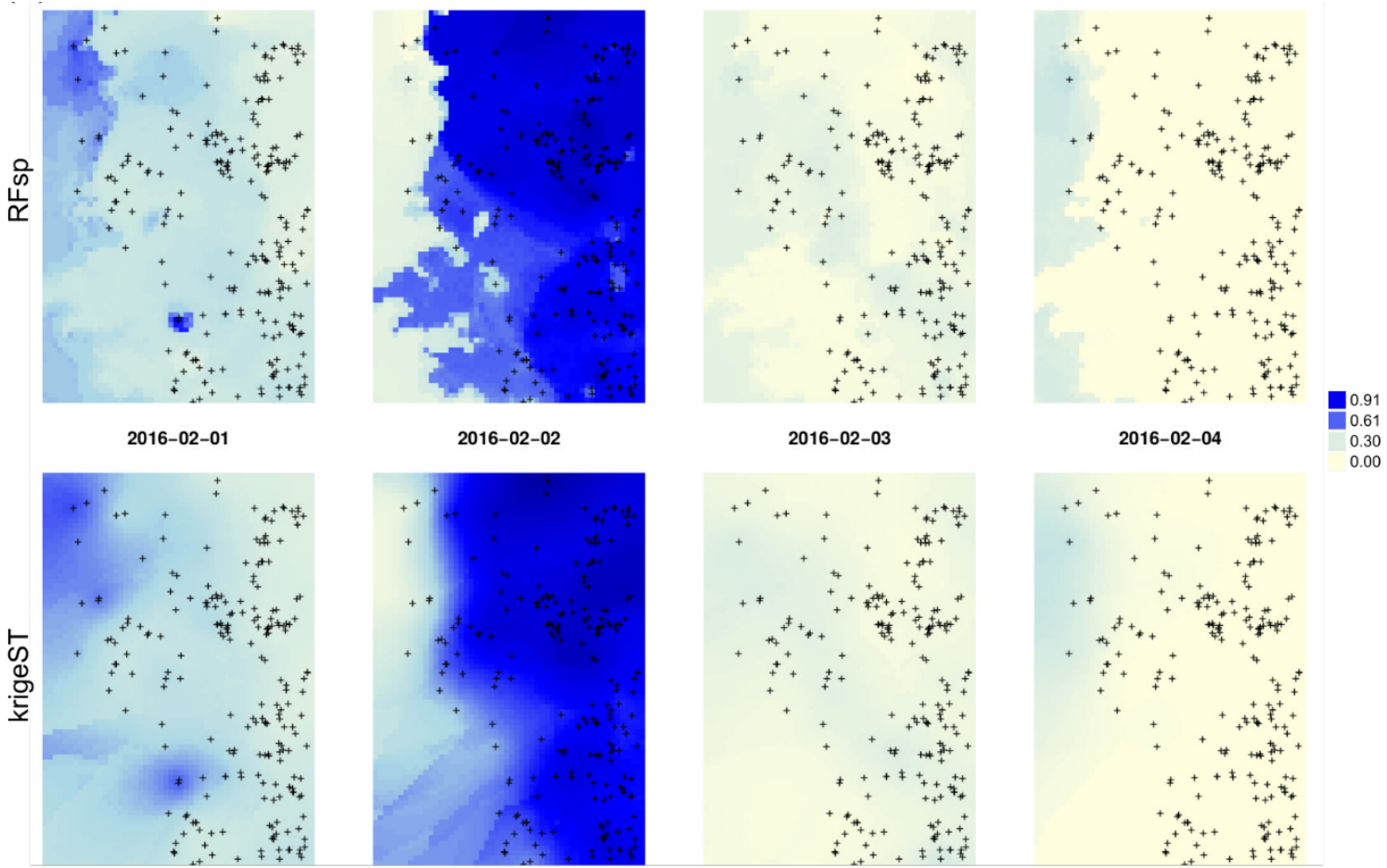


SIC97 data set

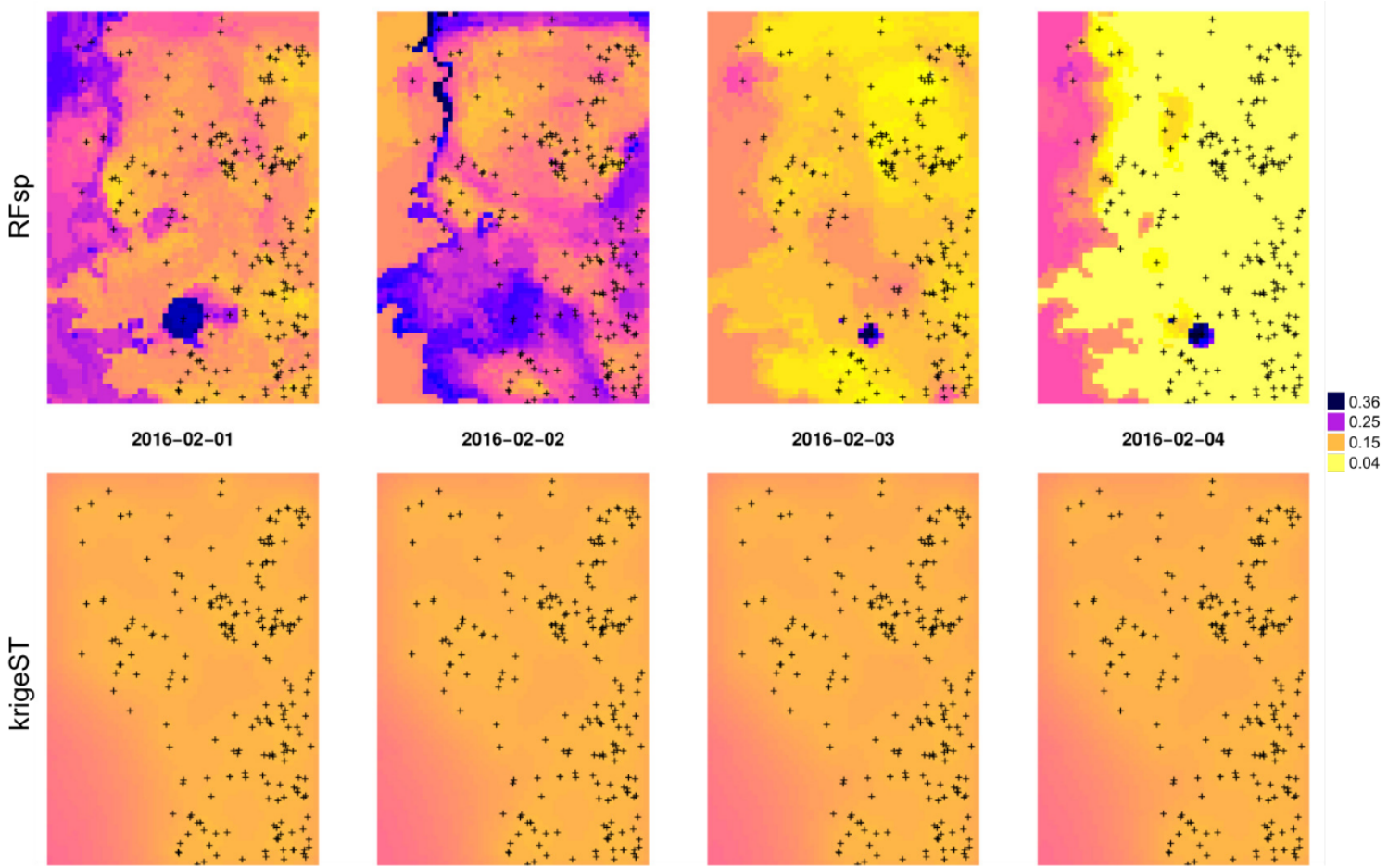



QGIS 2.18.11 - sic97_rainfall
Project Edit View Layer Settings Plugins Vector Raster Database Web Processing Help
Browser Panel
Project home
Home
Favourites
/
DB2
MSSQL
PostGIS
SpatiaLite
ArcGisFeatureServer
Layers Panel
rainfall_sic97
pred_rainfall_RFsp
pred.var_rainfall_RFsp
Bing Terrain
Bing AerialWithLabels
Value Tool
Enable
Table
Graph
Options
Decimals
Layer
Value
Coordinate
-36575,97456
Scale
1:1.015.184
Magnifier
100%
Rotation
0,0
Render
USER:100003 (OTF)

Daily precipitation (spatiotemporal)



Daily precipitation (st) prediction error maps

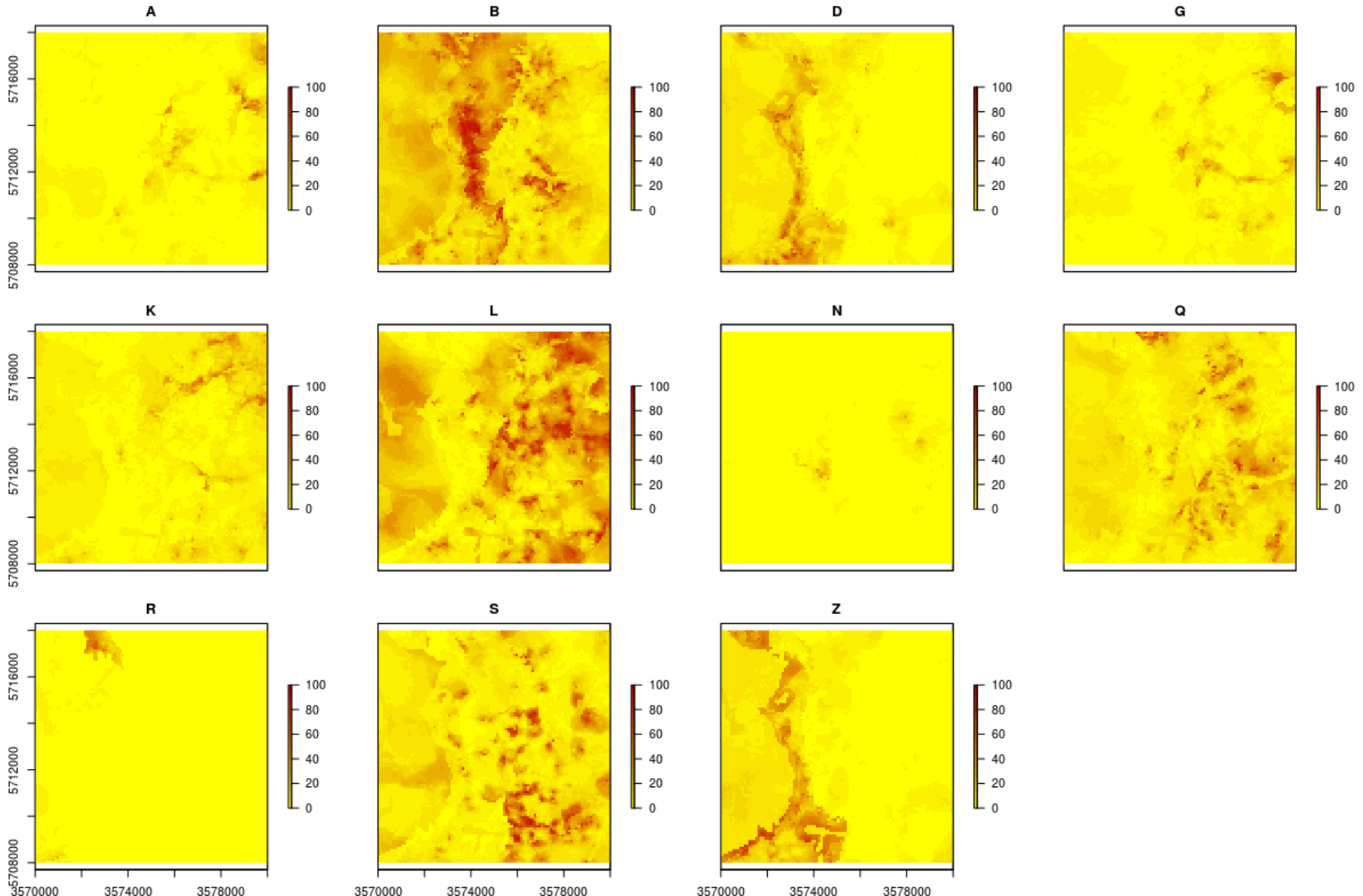


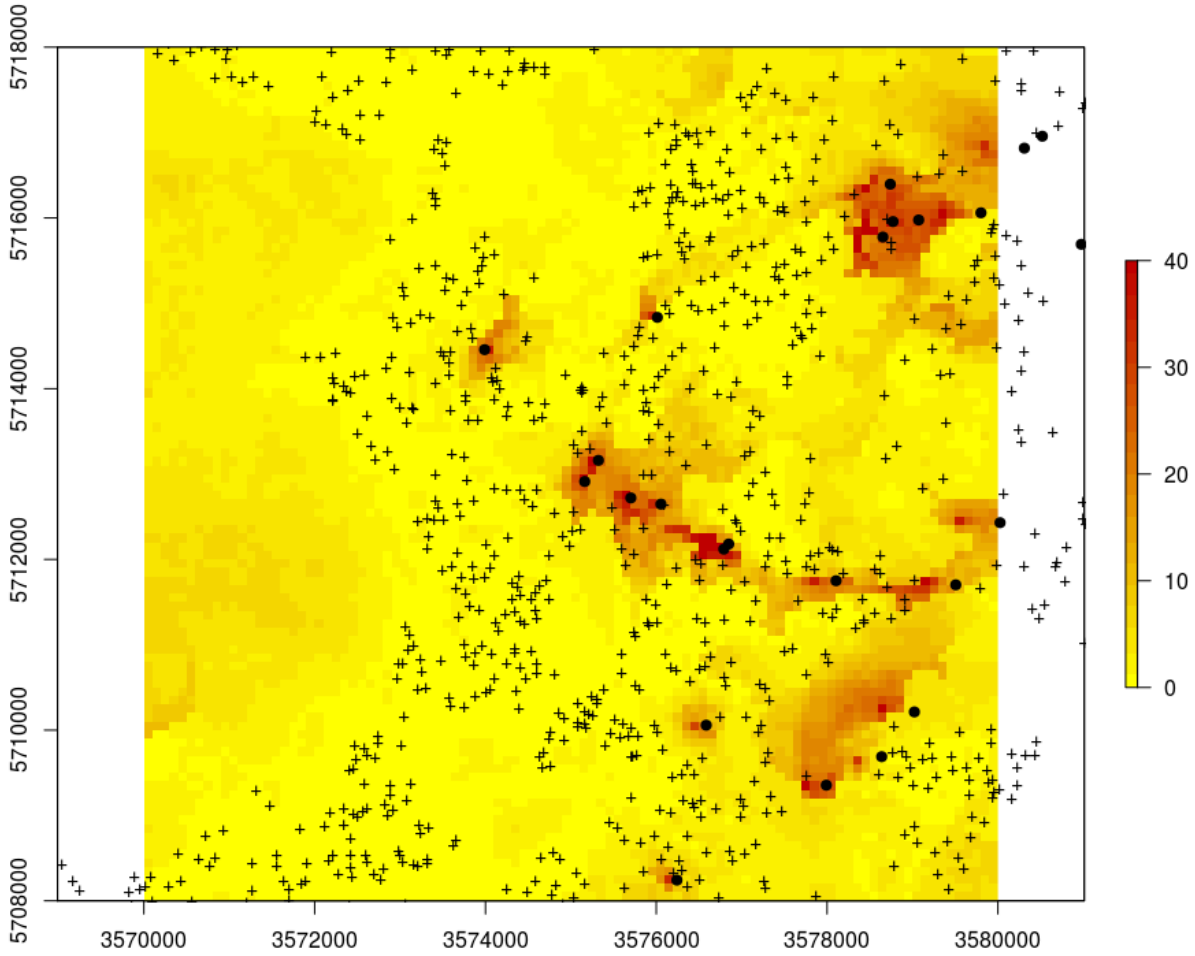
Summary points

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.

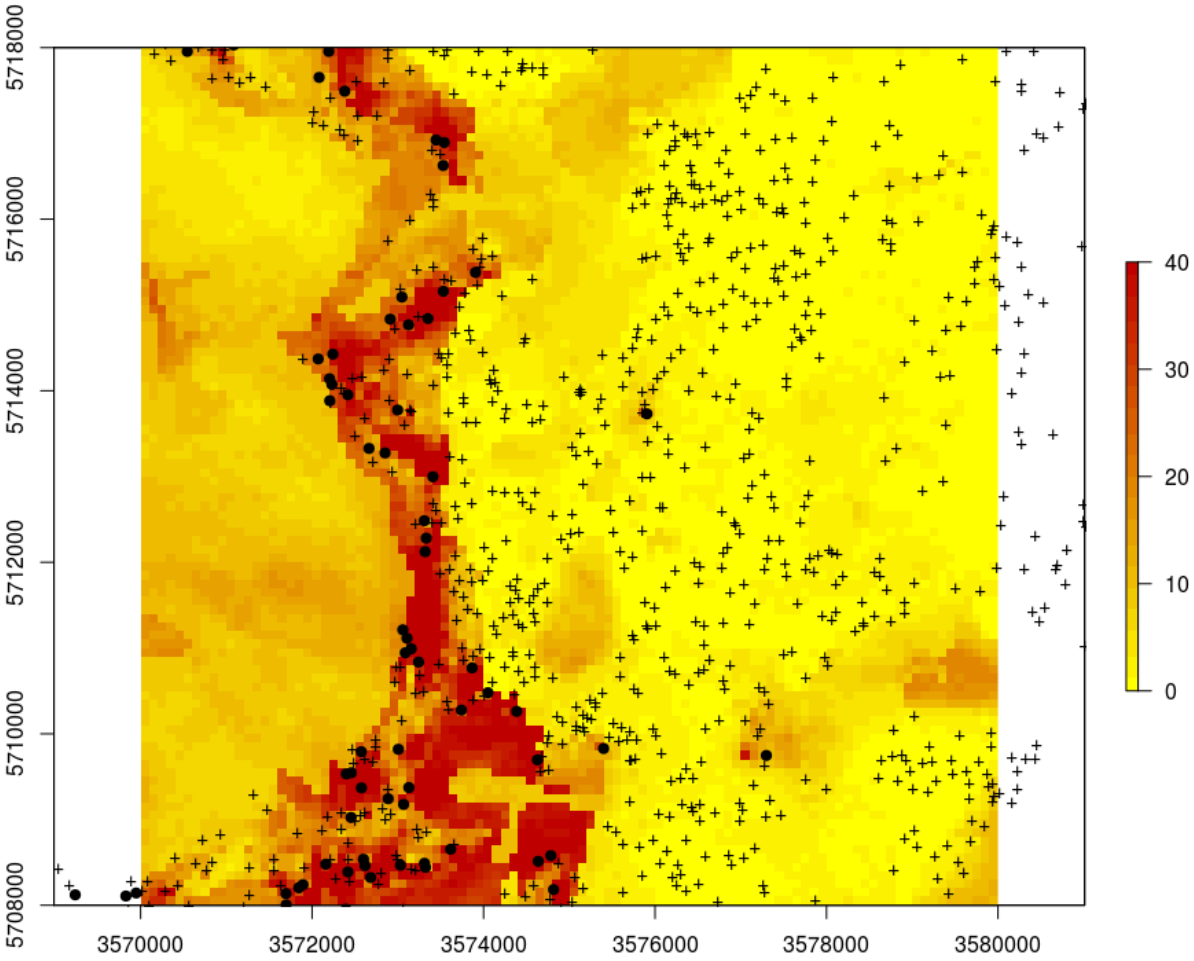
RFsp is still an experimental method and application with large data sets (>>1000 points) is not recommended.

Spatial prediction of categorical/factor variables is even easier


A
B
D
G
K
L
N
Q
R
S
Z
5716000
5712000
5708000
3570000
3574000
3578000
100
80
60
40
20
0

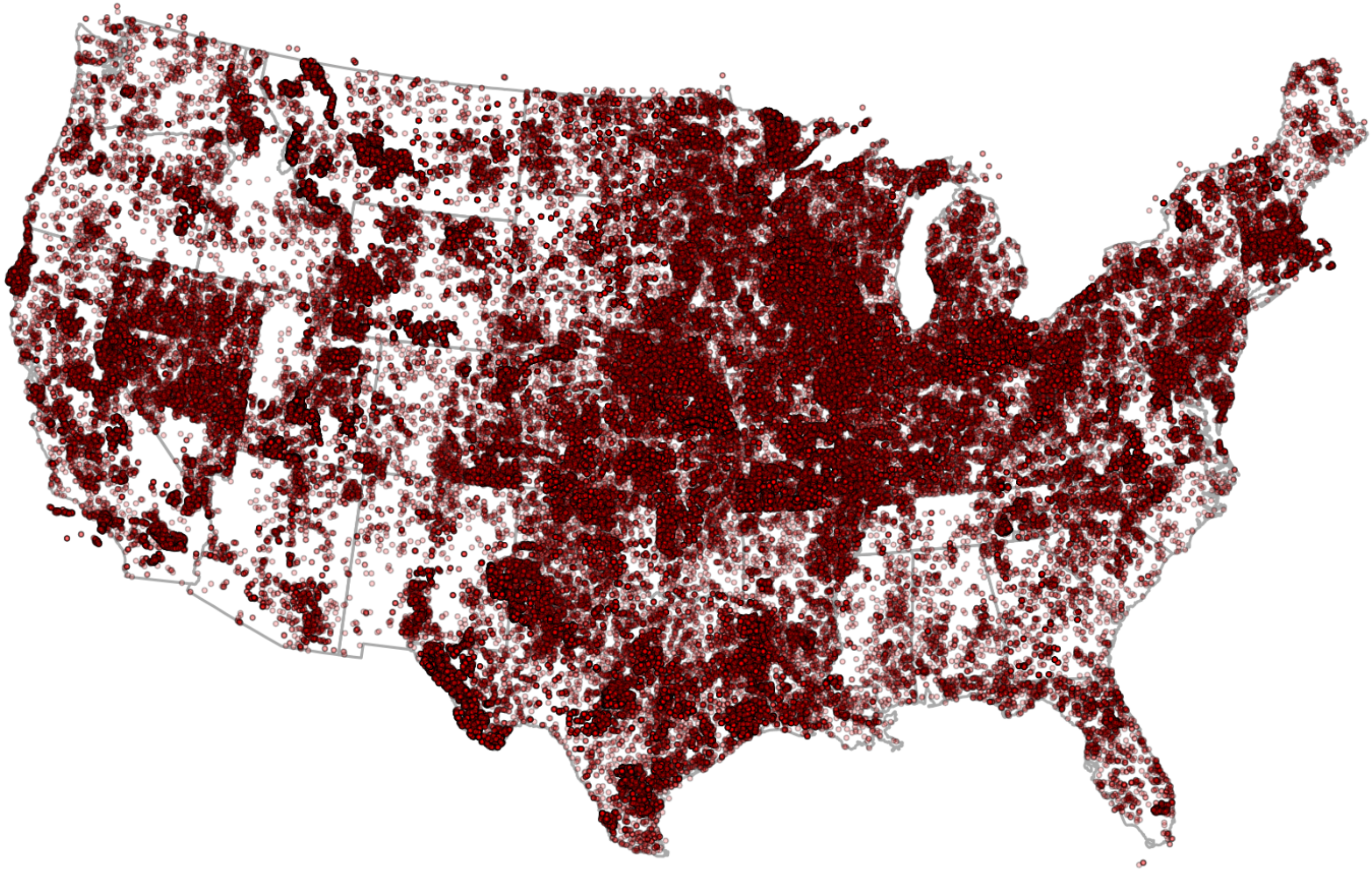

5718000
5716000
5714000
5712000
5710000
5708000
3570000
3572000
3574000
3576000
3578000
3580000
40
30
20
10
0

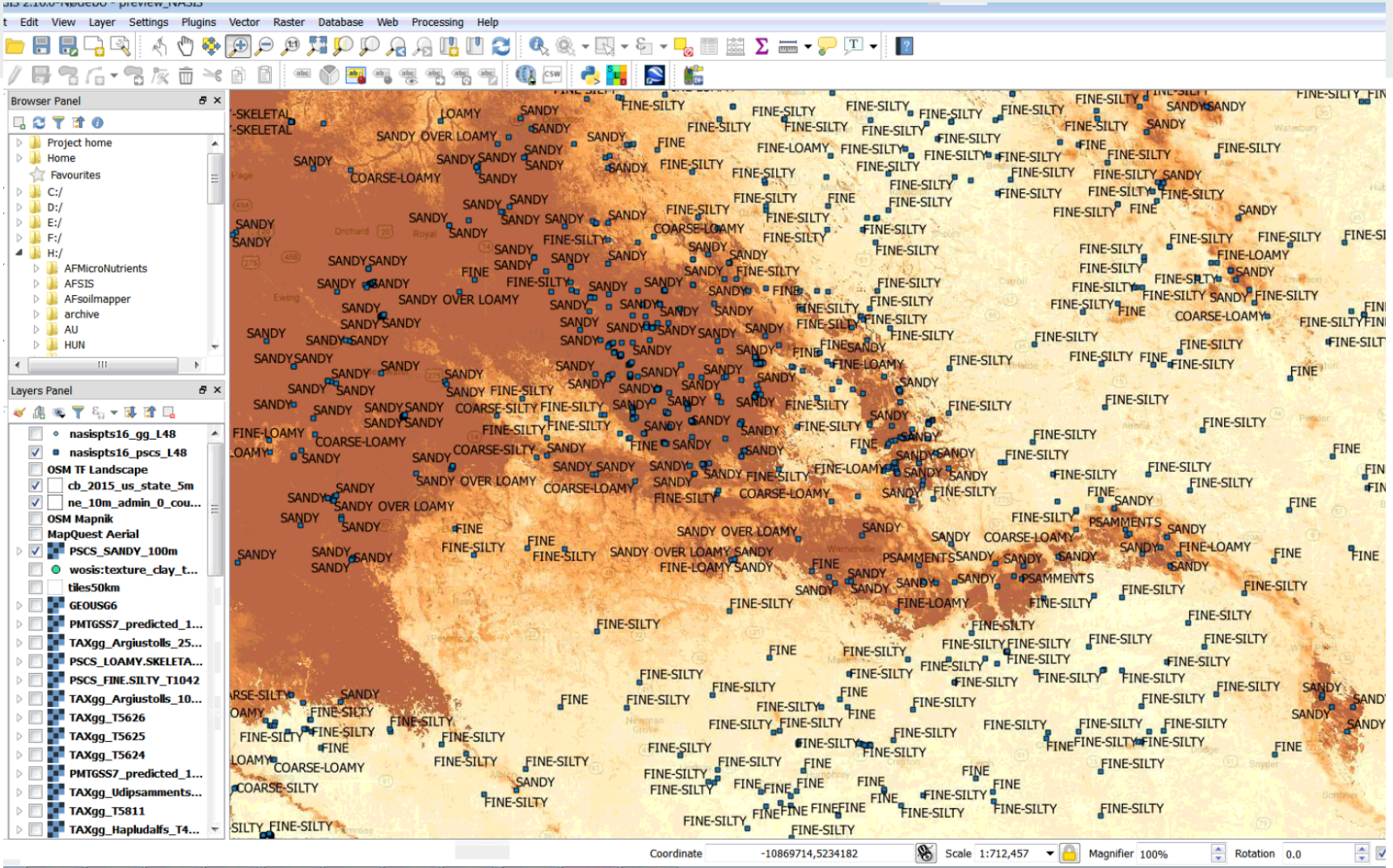
Class G

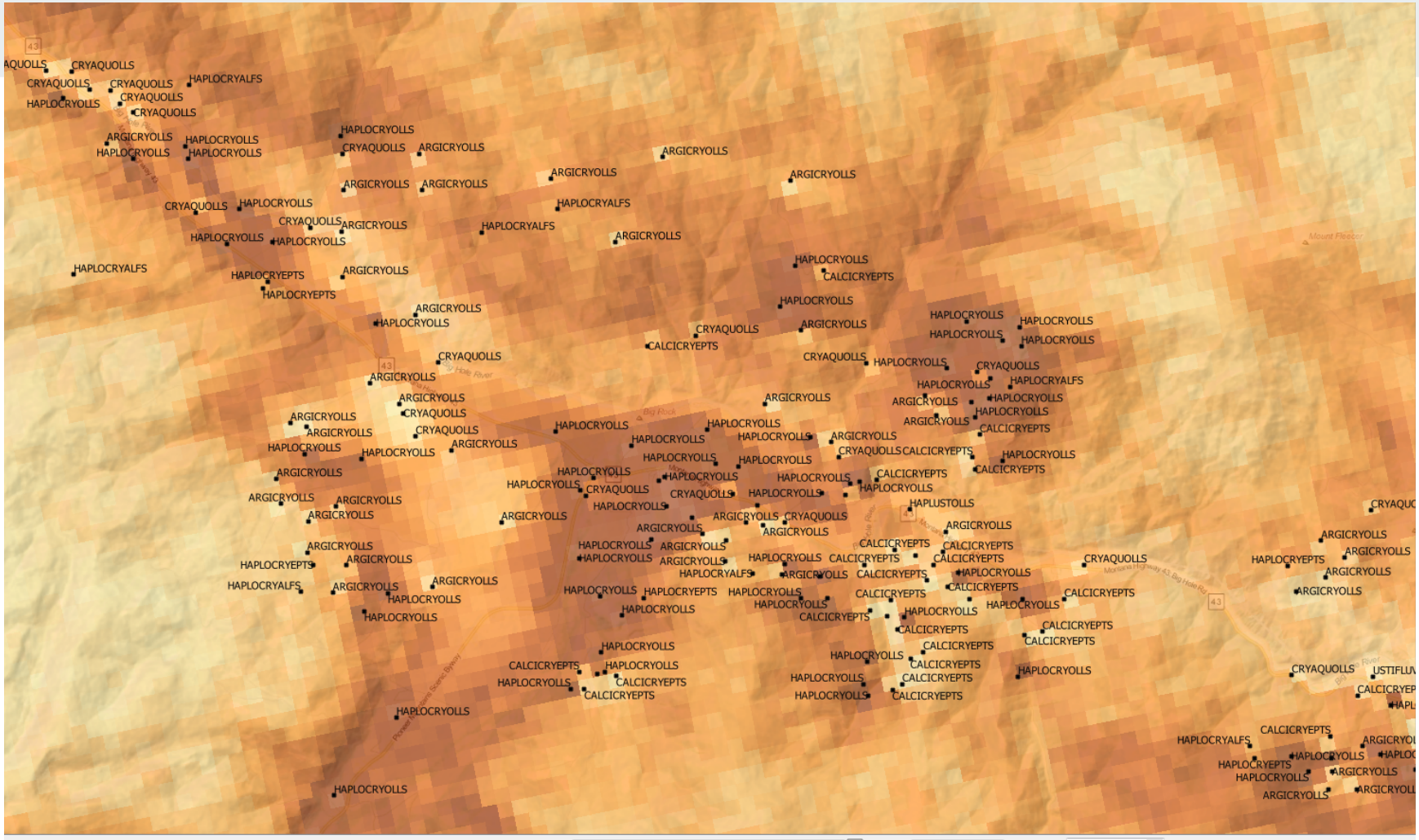

5718000
5716000
5714000
5712000
5710000
5708000
3570000
3572000
3574000
3576000
3578000
3580000
40
30
20
10
0

Class D

NASIS points (N=327,041)



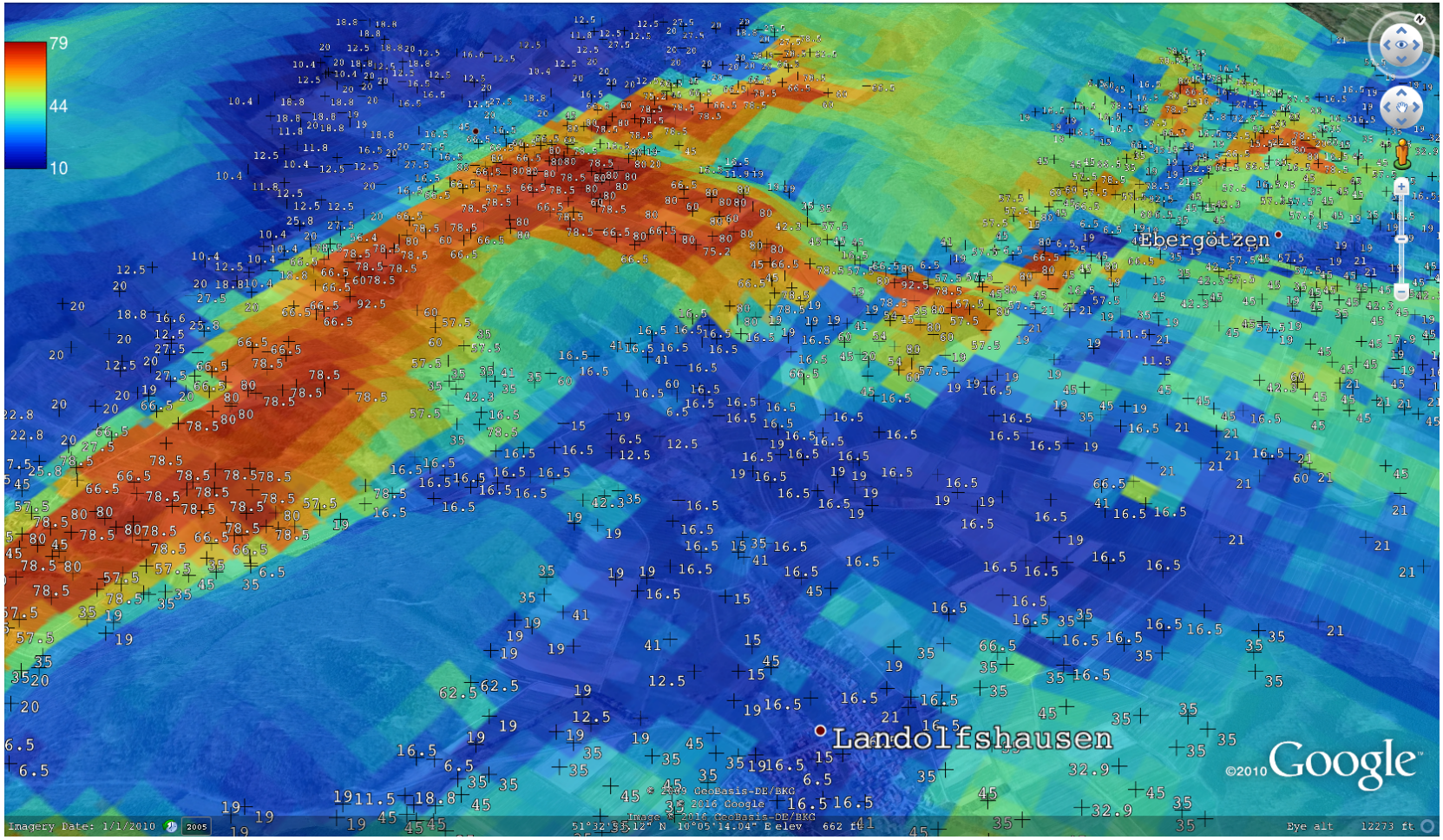

Edit View Layer Settings Plugins Vector Raster Database Web Processing Help
Browser Panel
Project home
Home
Favourites
C:/
D:/
E:/
F:/
H:/
AFMicroNutrients
AFSIS
AFsoilmapper
archive
AU
HUN
Layers Panel
nasispts16_gg_L48
nasispts16_pscs_L48
OSM TF Landscape
cb_2015_us_state_5m
ne_10m_admin_0_cou...
OSM Mapnik
MapQuest Aerial
PSCS_SANDY_100m
wosis:texture_clay_t...
tiles50km
GEOUSG6
PMTGSS7_predicted_1...
TAXgg_Argiustolls_25...
PSCS_LOAMY.SKELETA...
PSCS_FINE.SILTY_T1042
TAXgg_Argiustolls_10...
TAXgg_T5626
TAXgg_T5625
TAXgg_T5624
PMTGSS7_predicted_1...
TAXgg_Udipsamments...
TAXgg_T5811
TAXgg_Hapludalfs_T4...
Coordinate -10869714,5234182
Scale 1:712,457
Magnifier 100%
Rotation 0.0



Coordinate -12571418,5758361 Scale 1:107,419 Rotation 0.0 Render EPSG:3857 (OTF)

If machine learning is so efficient in generating spatial predictions, do we still need kriging?

Autopredict = *Deus ex machina*



Model-based (geo)statistics (traditional)

Specialists x hours

Sampling, detection of artifacts (outliers), derivation of covariates ...

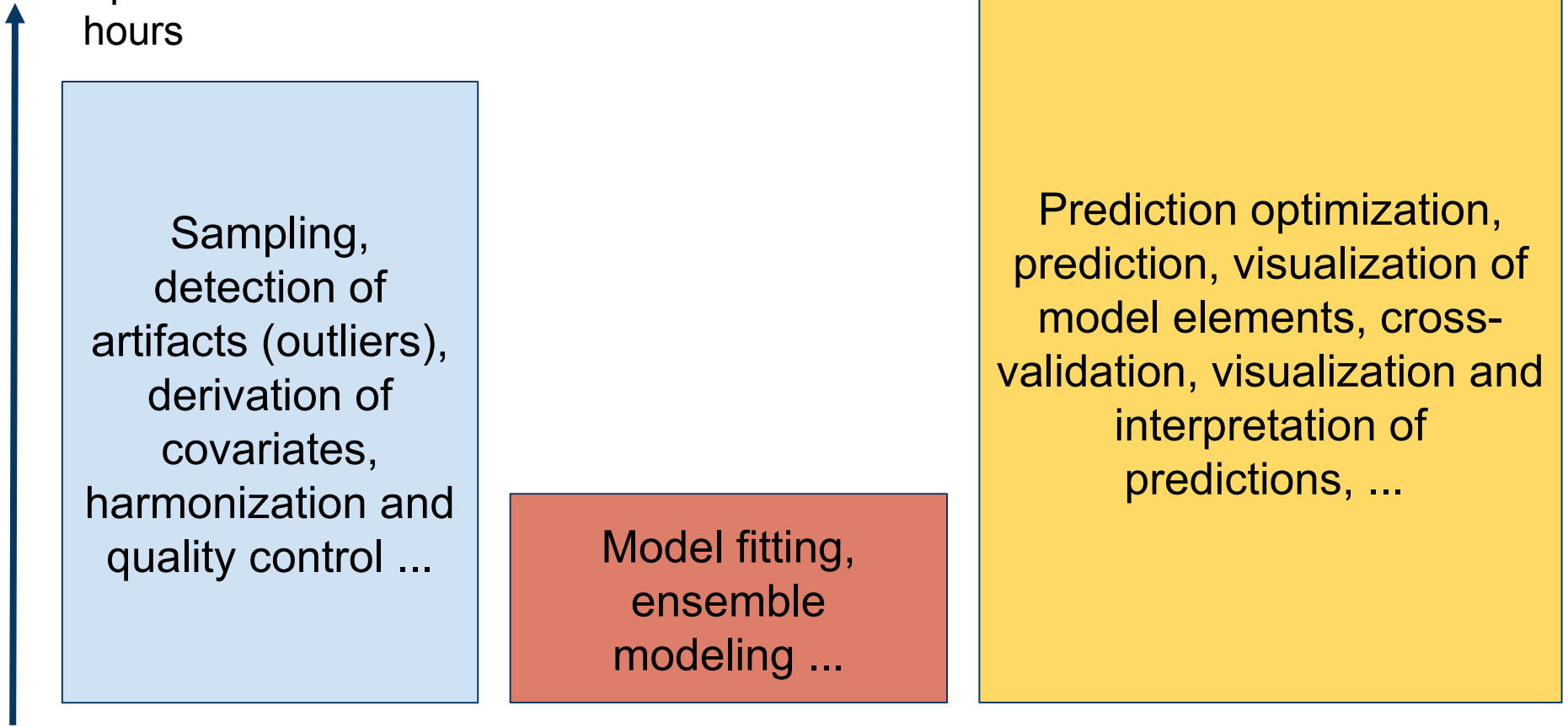
Model selection, testing of model requirements, model fitting (regression, variogram), step-wise selection of optimal set of covariates, model interpretation ...

Prediction, cross-validation, model visualization ...

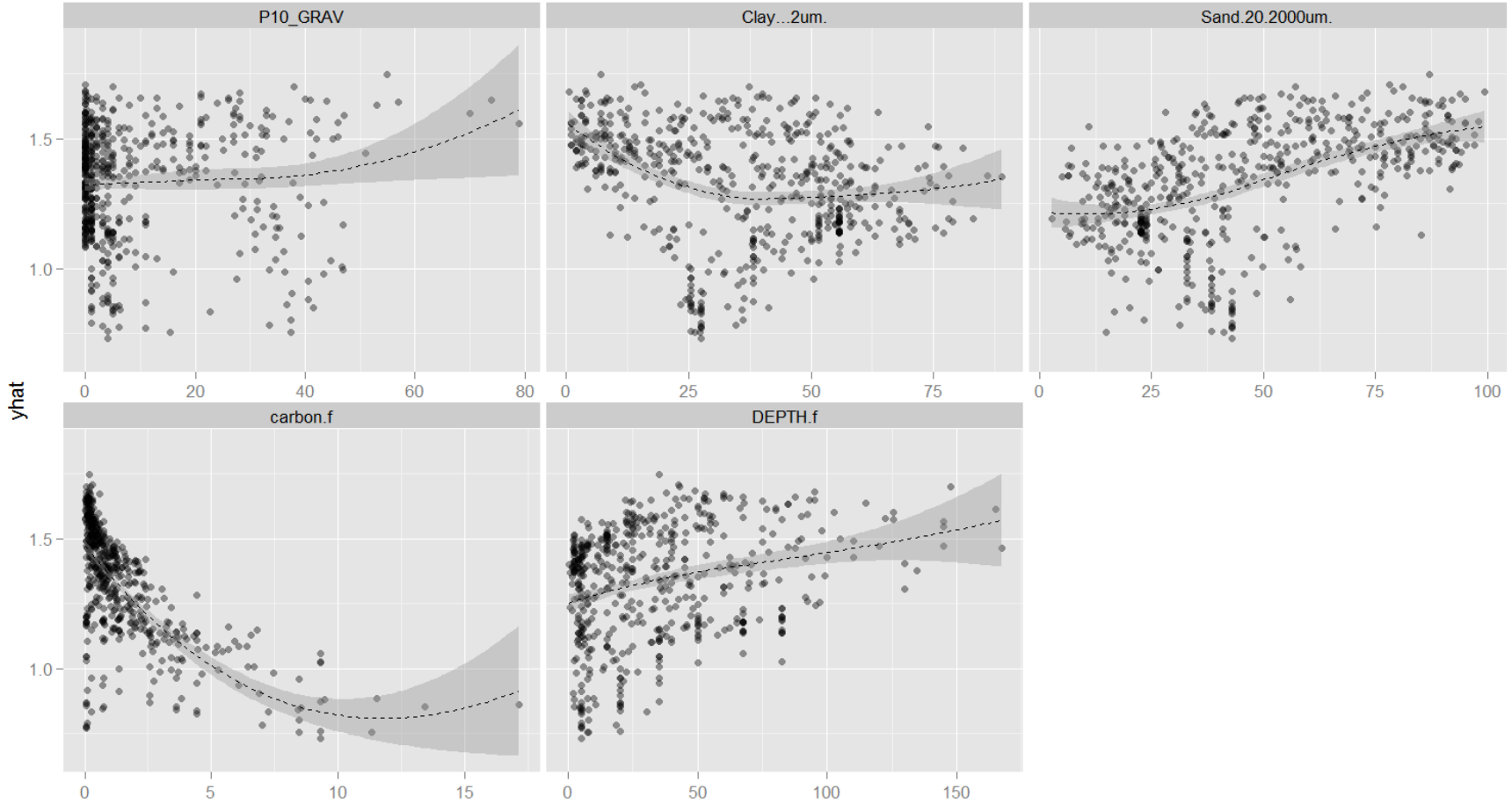
However!

- ➔ Computing intensity of MLA can be **MASSIVE**
- ➔ MLA usually **very very sensitive to artifacts in the input data** (even few typos can shift all predictions)
- ➔ There are still many things unsolved (how to plugin uncertainty? how to account for spatial clustering? spacetime distances?)

Data-driven (algorithmic) modeling



RandomForestSRC package



But what about (prediction) uncertainty?

First time I questioned this

Tomislav Hengl-4 ➤Jun 23, 2013; 5:51pm **Prediction variance (map) for predictions derived using RandomForest pac** Reply | Threaded | More

153 posts

Dear list,

I have a question about the randomForest models. I'm trying to figure out a way to estimate the prediction variance (spatially) for the randomForest function (http://cran.r-project.org/web/packages/randomForest/).

If I run a GLM I can also derive the prediction variance using:

```
 > demo(meuse, echo=FALSE)
 > meuse.ov <- over(meuse, meuse.grid)
 > meuse.ov <- cbind(meuse.ov, meuse@data)
 > omm0 <- glm(log1p(om)~dist+ffreq, meuse.ov, family=gaussian())
 > om.glm <- predict.glm(omm0, meuse.grid, se.fit=TRUE)
 > str(om.glm)
List of 3
  $ fit          : Named num [1:3103] 2.34 2.34 2.32 2.29 2.34 ...
   ..- attr(*, "names")= chr [1:3103] "1" "2" "3" "4" ...
  $ se.fit       : Named num [1:3103] 0.0491 0.0491 0.0481 0.046 0.0491 ...
   ..- attr(*, "names")= chr [1:3103] "1" "2" "3" "4" ...
  $ residual.scale: num 0.357
```

when I fit a randomForest model, I do not get any estimate of the model uncertainty (for each pixel) but just the predictions:

```
 > meuse.ov <- meuse.ov[-omm0$na.action,]
```

... and of course R-sig-geo knows it

Tomislav Hengl-4 Jun 25, 2013; 12:44am **Re: Prediction variance (map) for predictions derived using RandomForest** | Reply | Threaded | More

153 posts

In reply to this post by Forrest Stevens

Dear Forrest,

Thanks a lot for your tip. I think quantregForest is what we were looking for. It takes much more time to compute, but the method looks sound (http://jmlr.org/papers/volume7/meinshausen06a/meinshausen06a.pdf). I do simplify everything on the end and assume that I can derive upper and lower confidence limits for +/- 1 s.d. (0.15866, 1-0.15866) and then use this as the prediction variance, but this is probably as good as it goes. Here is the revised code:

https://code.google.com/p/gsif/source/browse/trunk/meuse/RK_vs_RandomForestK.R

Thank you all for your suggestions / opinions (very useful as usual).

cheers,

T. (Tom) Hengl
Url: http://www.wageningenur.nl/en/Persons/dr.-T-Tom-Hengl.htm
Network: http://profiles.google.com/tom.hengl
Publications: http://scholar.google.com/citations?user=2oYU7S8AAAAJ

On 23/06/2013 15:08, Forrest Stevens wrote:

> Hi Tom, I've done something similar in the past to visualize the
> distribution of the predictions attained for each observation across

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

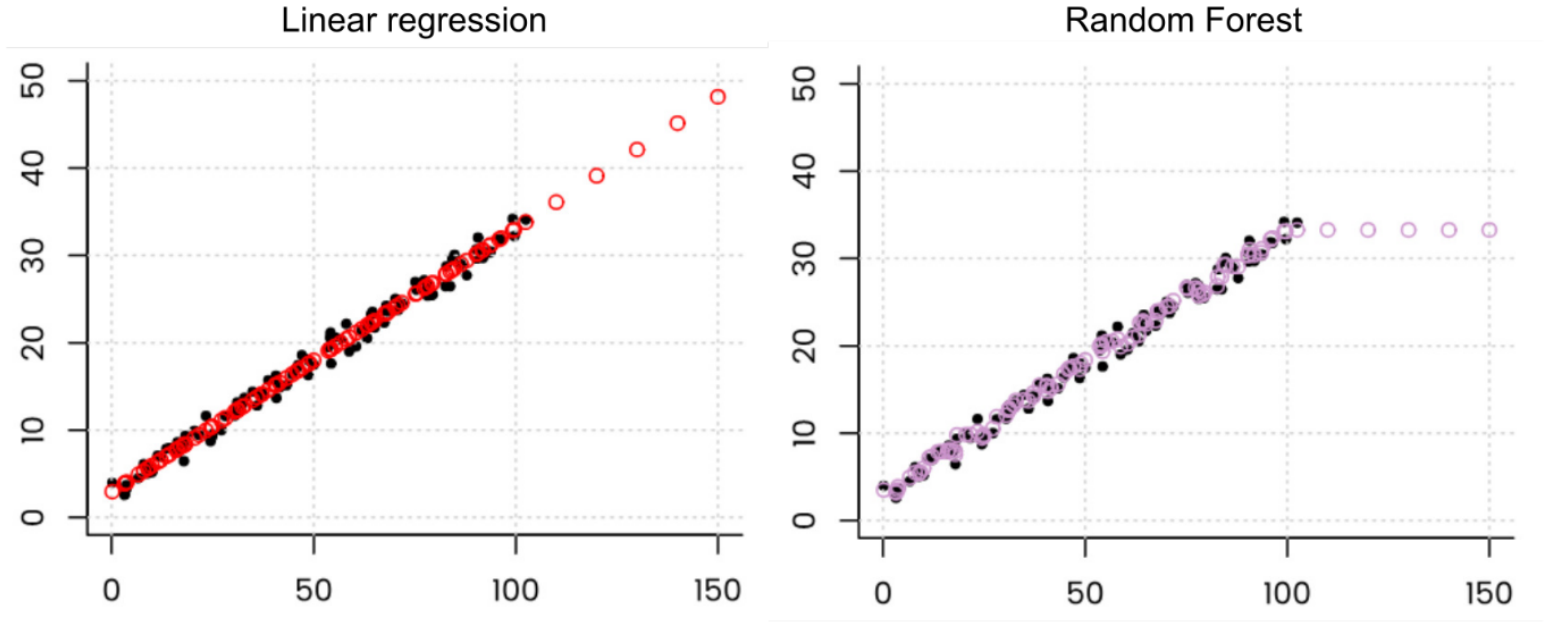
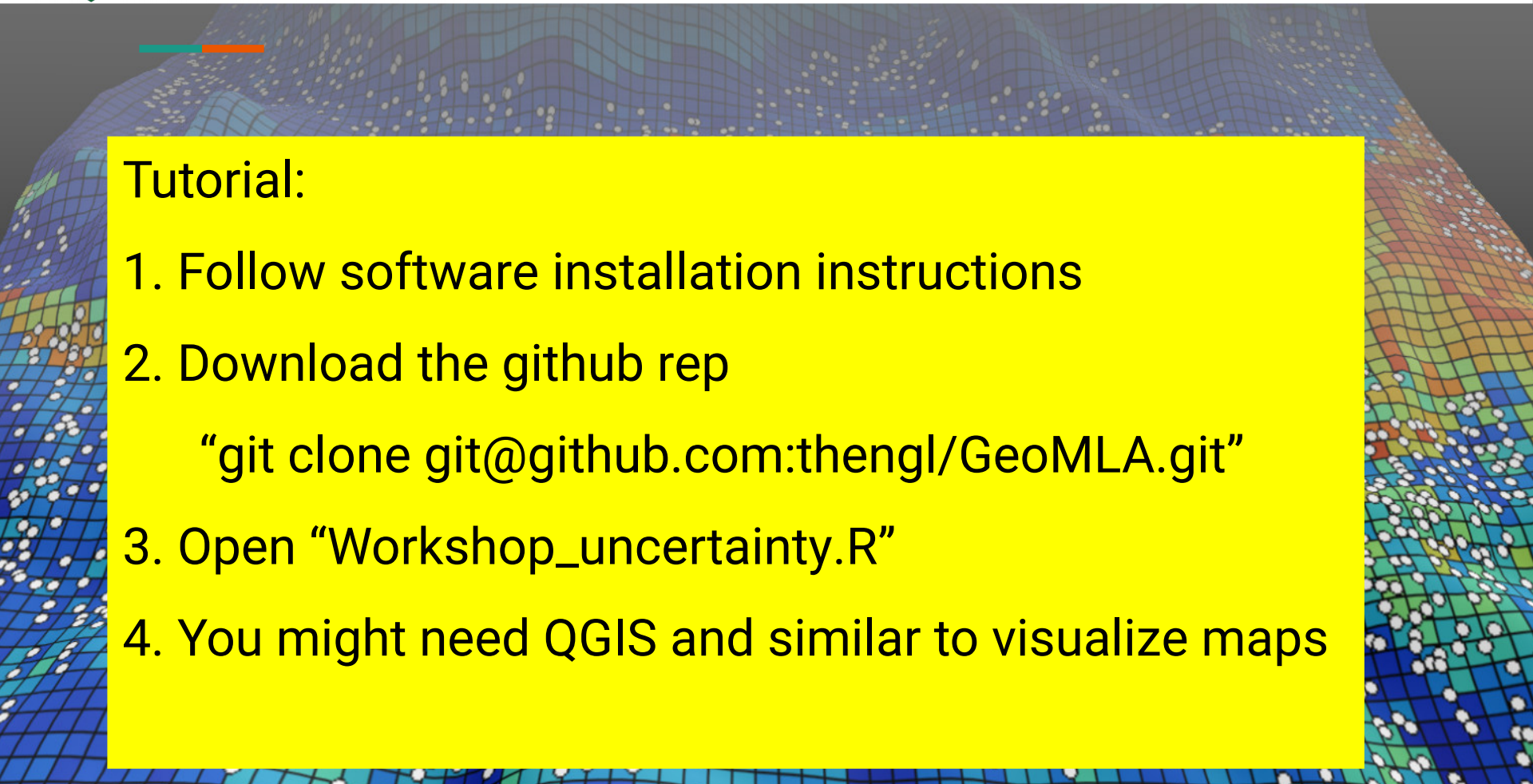


Figure 14. Illustration of the extrapolation problem of Random Forest based on the code examples from Peter Ellis (http://freerangestats.info). Even though Random Forest is more generic than linear regression and can be used also to fit complex non-linear problems, it can lead to completely nonsensical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



Tutorial:

1. Follow software installation instructions
2. Download the github rep

   “git clone git@github.com:thengl/GeoMLA.git”
3. Open “Workshop_uncertainty.R”
4. You might need QGIS and similar to visualize maps