ISRIC
World Soil Information

End of kriging? Or how tree-based Machine Learning Algorithms can be used to generate more accurate spatial predictions with combined geographical and feature space covariates

T. (Tom) Hengl

Death to Kriging?

Journal
International Journal of Remote Sensing >
Volume 32, 2011 - Issue 15

Original Articles

Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment

Robert Gilmore Pontius Jr & Marco Millones

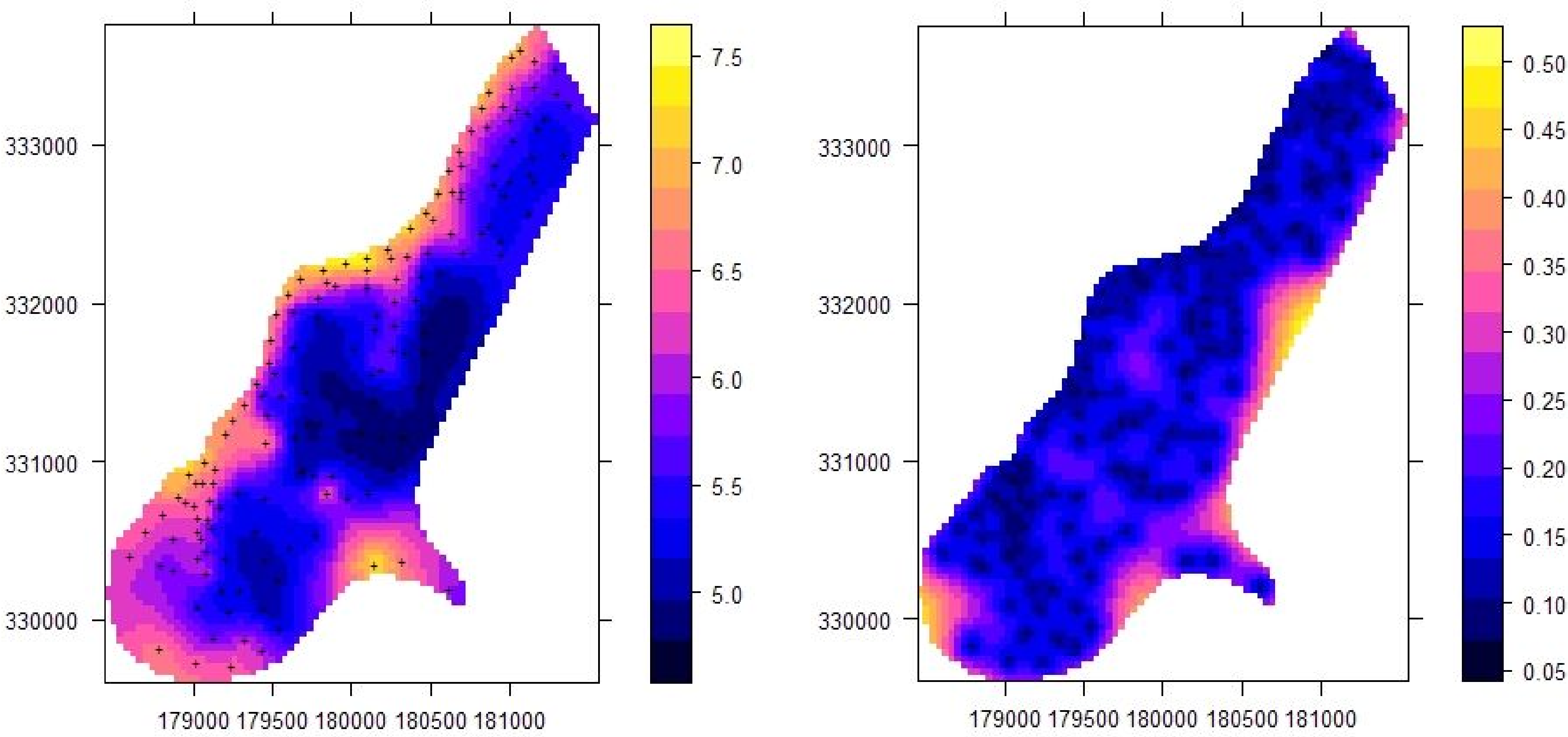
Pages 4407-4429 | Received 27 Aug 2010, Accepted 20 Dec 2010, Published online: 17 Aug 2011
Download citation http://dx.doi.org/10.1080/01431161.2011.552923

Abstract

The family of Kappa indices of agreement claim to compare a map's observed classification accuracy relative to the expected accuracy of baseline maps that can have two types of randomness: (1) random distribution of the quantity of each category and (2) random spatial allocation of the categories. Use of the Kappa indices has become part of the culture in remote sensing and other fields. This article examines five different Kappa indices, some of which were derived by the first author in 2000. We expose the indices' properties mathematically and illustrate their limitations graphically, with emphasis on Kappa's use of randomness as a baseline, and the often-ignored conversion from an observed sample matrix to the estimated population matrix. This article concludes that these Kappa indices are useless, misleading and/or flawed for the practical applications in remote sensing that we have seen. After more than a decade of working with these indices, we recommend that the profession abandon the use of Kappa indices for purposes of accuracy assessment and map comparison, and instead summarize the cross-tabulation matrix with two much simpler summary parameters: quantity disagreement and allocation

Kriging, in essence, is statistically correct / well established and of course well documented, but this talk shows that it could also be potentially replaced.

“Kriging” is basically a synonym for Geostatistics



Tom Hengl (tom.hengl@isric.org)

Model-based geostatistics (state-of-the-art)

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)

Tom Hengl (tom.hengl@isric.org)

Regression-kriging

The Regression-kriging approach:

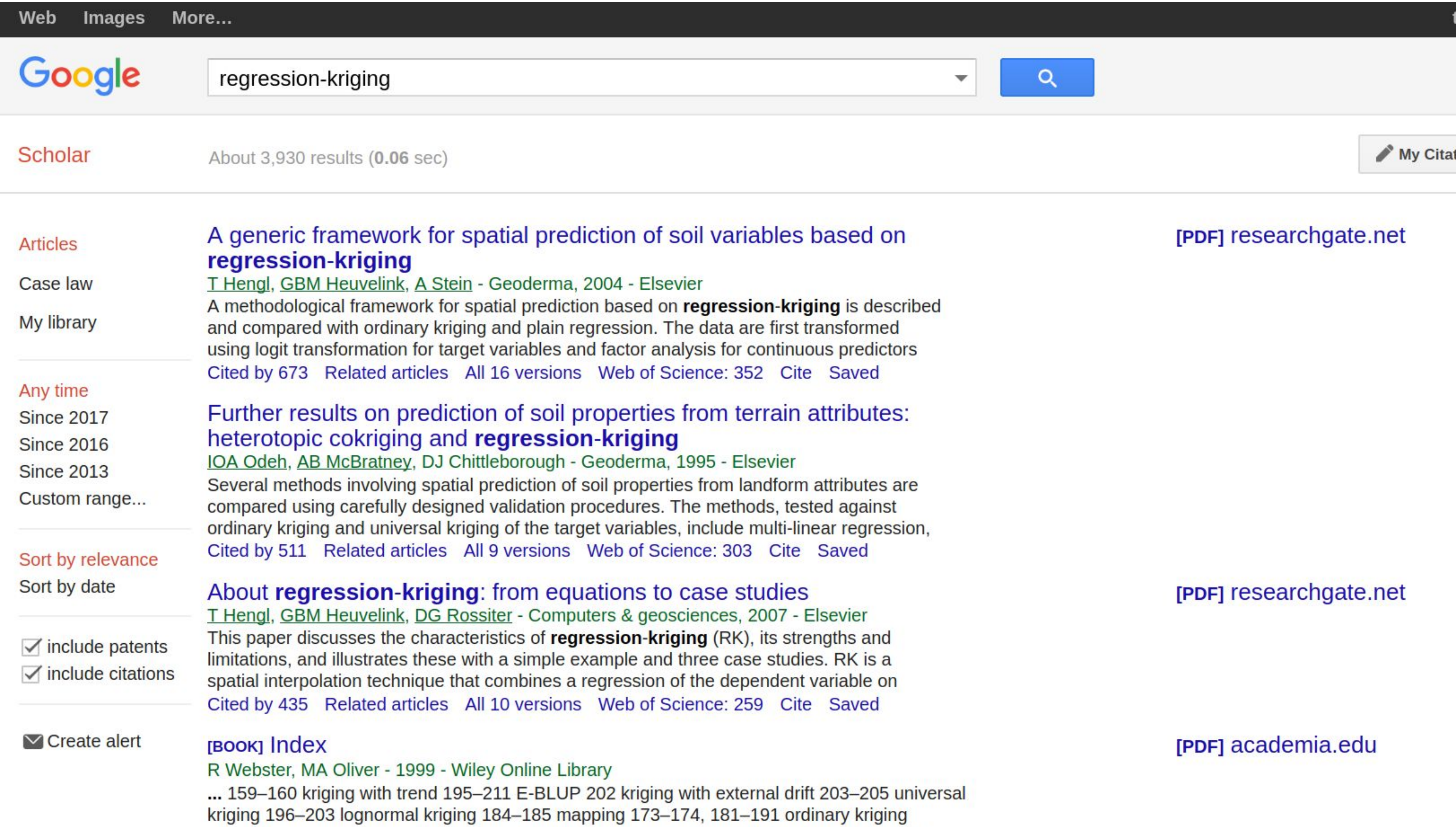
$$Y'(s) = m'(s) + e'(s) + \varepsilon$$

$$m'(s) = f[X(s)]; \quad e'(s) = f[\mathbf{h}(s)|Y]$$

- First and second order stationarity
- Normal distribution (residuals)
- Anysotropy, multicolinearity...

Tom Hengl (tom.hengl@isric.org)

If you search Regression-kriging looks like my name comes high on the list



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

Two cultures: GLM vs RF

```
m = glm( zinc~log1p(dist)+ffreq, meuse,
family=gaussian(link=log))

m = quantregForest(
x=meuse@data[,c("dist","ffreq")],
y=meuse$zinc)
```

Read more in: [Leo Braiman "Two cultures"]()

This repository Search Pull requests Issues Marketplace Gist

ISRICWorldSoil / **GSIF_tutorials**

Unwatch 8 | Star 8 | Fork 13

Code | Issues 0 | Pull requests 0 | Projects 0 | Wiki | Settings | Insights

Branch: master | **GSIF_tutorials** / geul / **random_forest_vs_kriging.R** | Find file | Copy path

root RF vs OK — 68cb598 34 minutes ago

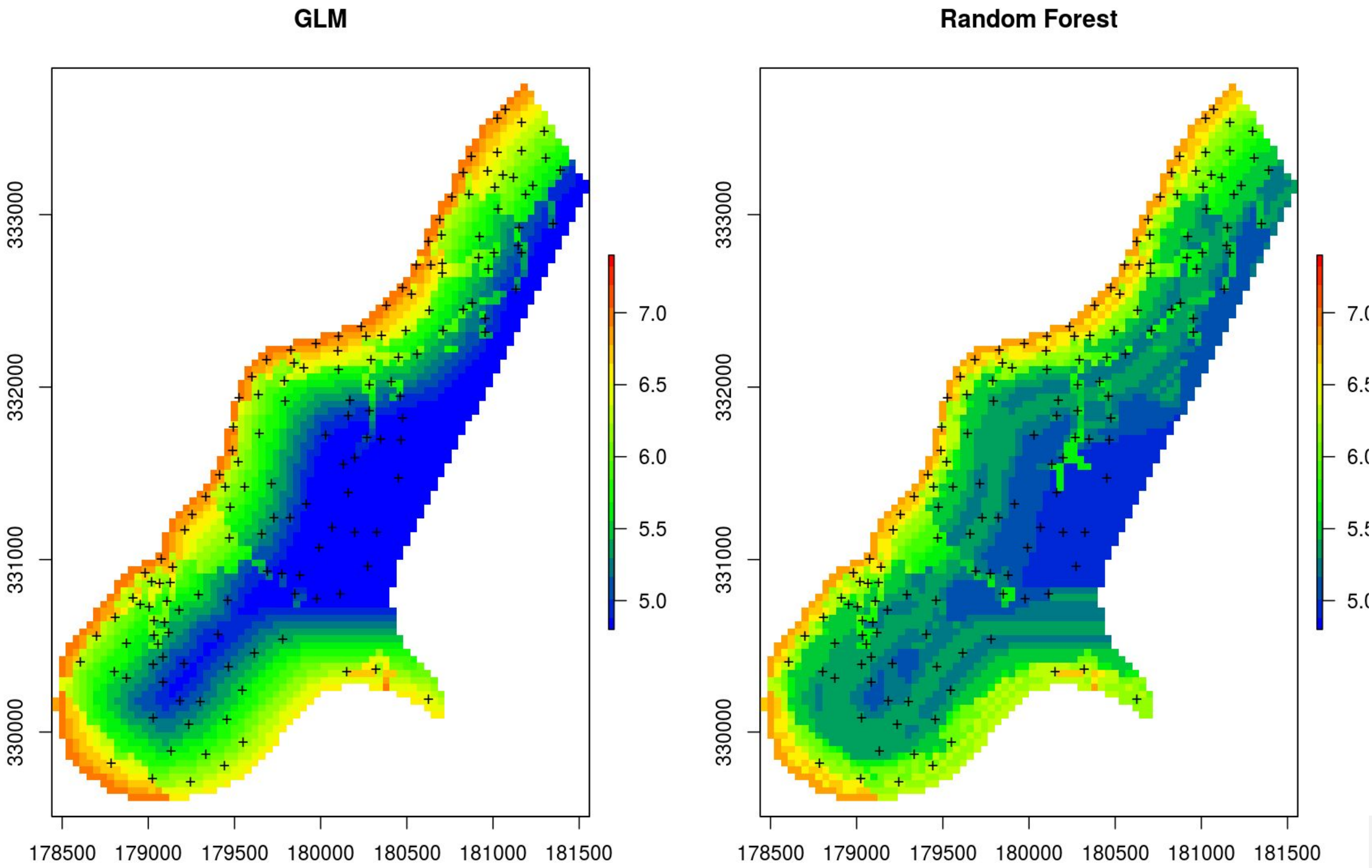
0 contributors

152 lines (139 sloc) | 6.16 KB — Raw | Blame | History

```r
## Comparison RF vs kriging
## tom.hengl@isric.org

library(GSIF)
library(rgdal)
library(raster)
library(gstat)
library(randomForest)
library(quantregForest)
library(plotKML)
library(scales)
library(ranger)
library(RCurl)
#library(geoR)
leg = c("#0000ff", "#0028d7", "#0050af", "#007986", "#00a15e", "#00ca35", "#00f20d", "#1aff00", "#43ff00", "#6bff00", "#94ff00",
## Load the Meuse data set:
demo(meuse, echo=FALSE)

## compare GLM vs RF
m <- glm(zinc~log1p(dist)+ffreq, meuse, family=gaussian(link=log))
plot(m$fitted.values~m$y, asp=1)
abline(0,1)
rf <- quantregForest(x=meuse@data[,c("dist","ffreq")], y=meuse$zinc)
plot(rf$predicted~rf$y, asp=1)
abline(0,1)
meuse.grid$glm.zinc <- predict(m, meuse.grid@data, type="response")
```

All examples shown are available via github

GLM vs RF (zinc meuse)



geoR

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

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
```

Variogram (zinc meuse)

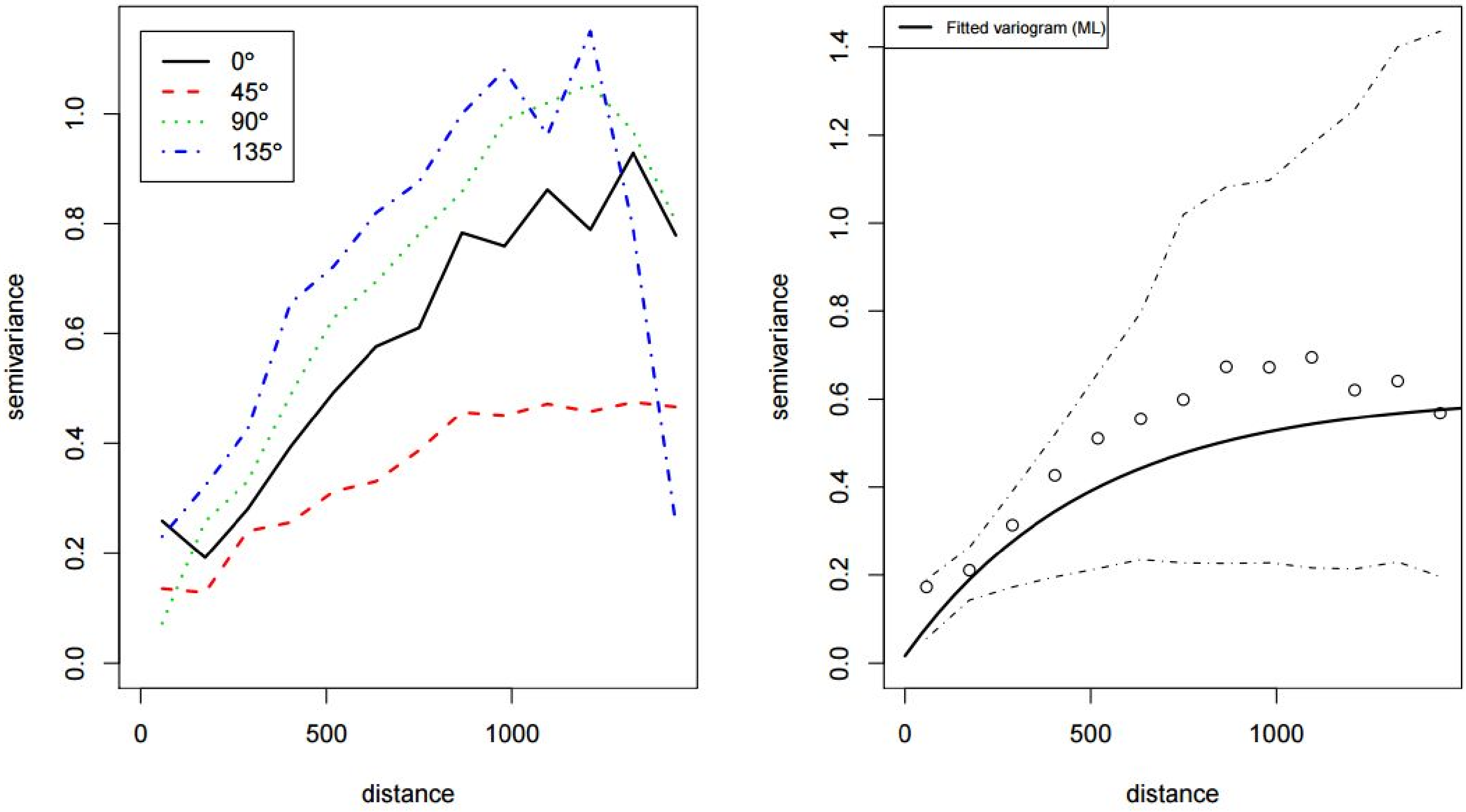
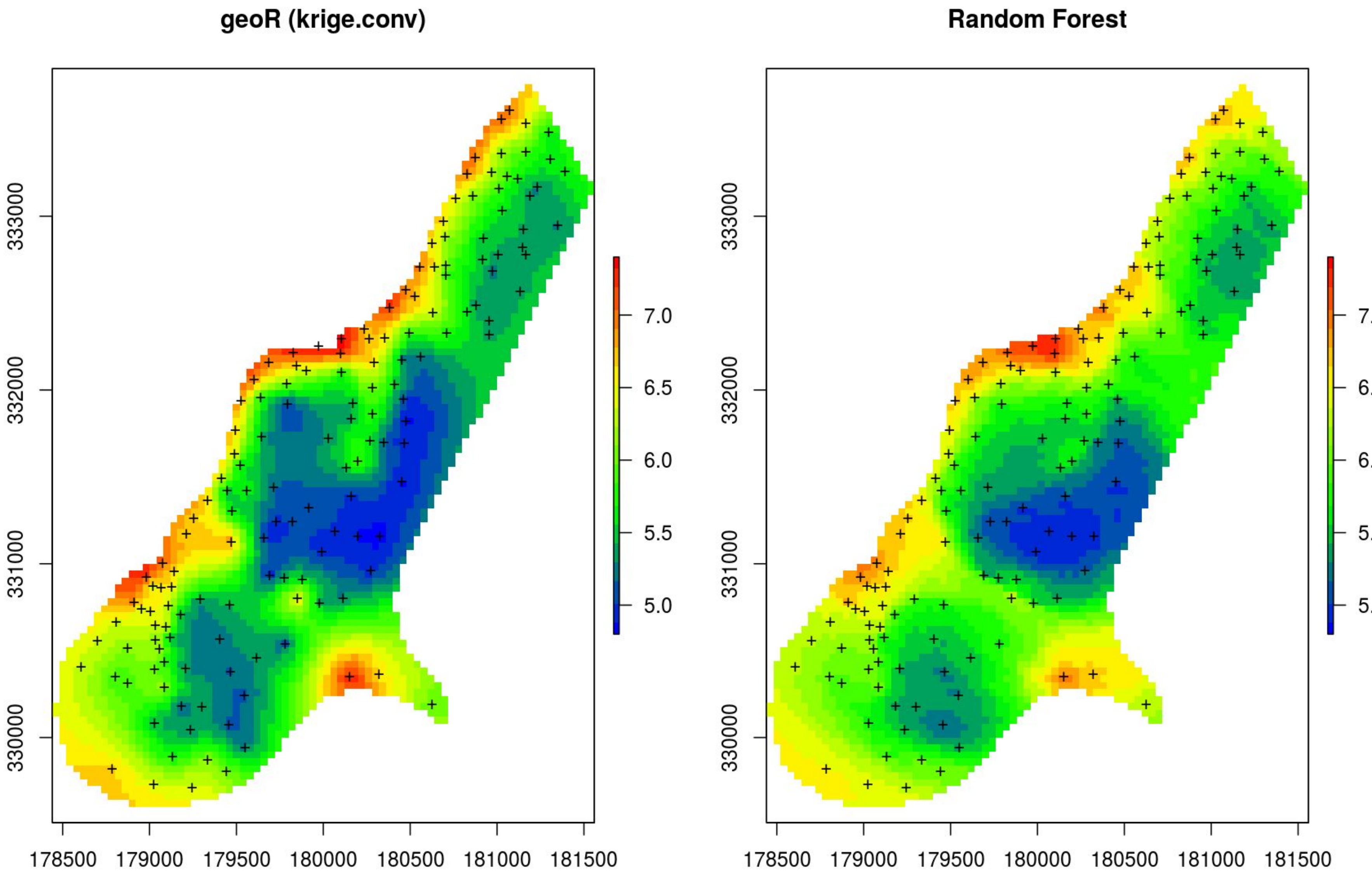


Fig. 5.15: Anisotropy (left) and variogram model fitted using the Maximum Likelihood (ML) method (right). The confidence bands (*envelopes*) show the variability of the sample variogram estimated using simulations from a given set of model parameters.

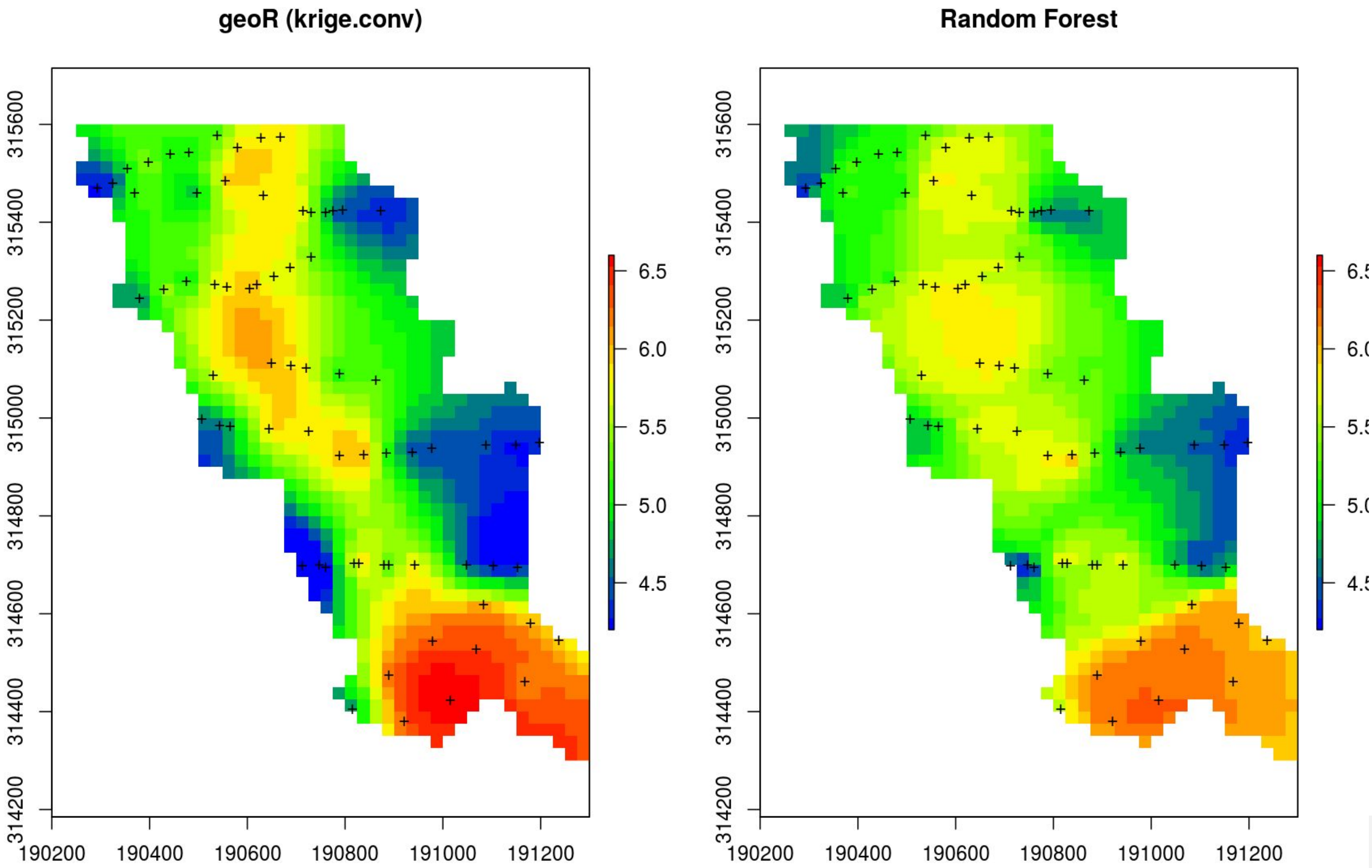
GSIF: correlate values with buffer distances

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

dn0 <- paste(names(grid.dist0), collapse="+")
fm0 <- as.formula(paste("zinc ~", dn0))

m0 <- fit.gstatModel(meuse, fm0, grid.dist0,
method="ranger", rvgm=NULL)

rk.m0 <- predict(m0, grid.dist0)
```

OK vs RF (zinc Meuse)



OK vs RF (Pb Gaul)



sp predictions based on Random Forest

- ✓ 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.

Random Forest complete list

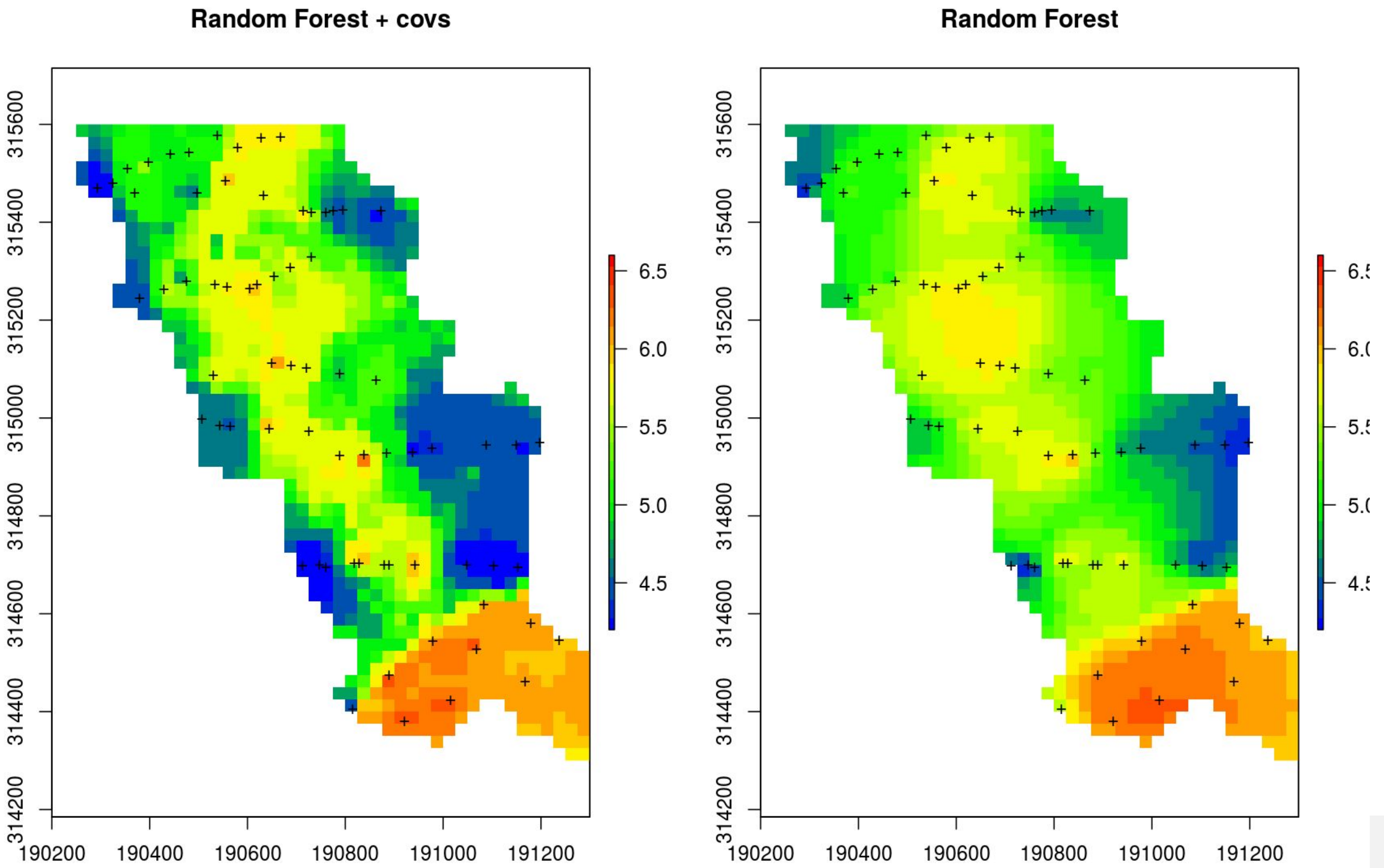
Use both location-buffers + covariates:

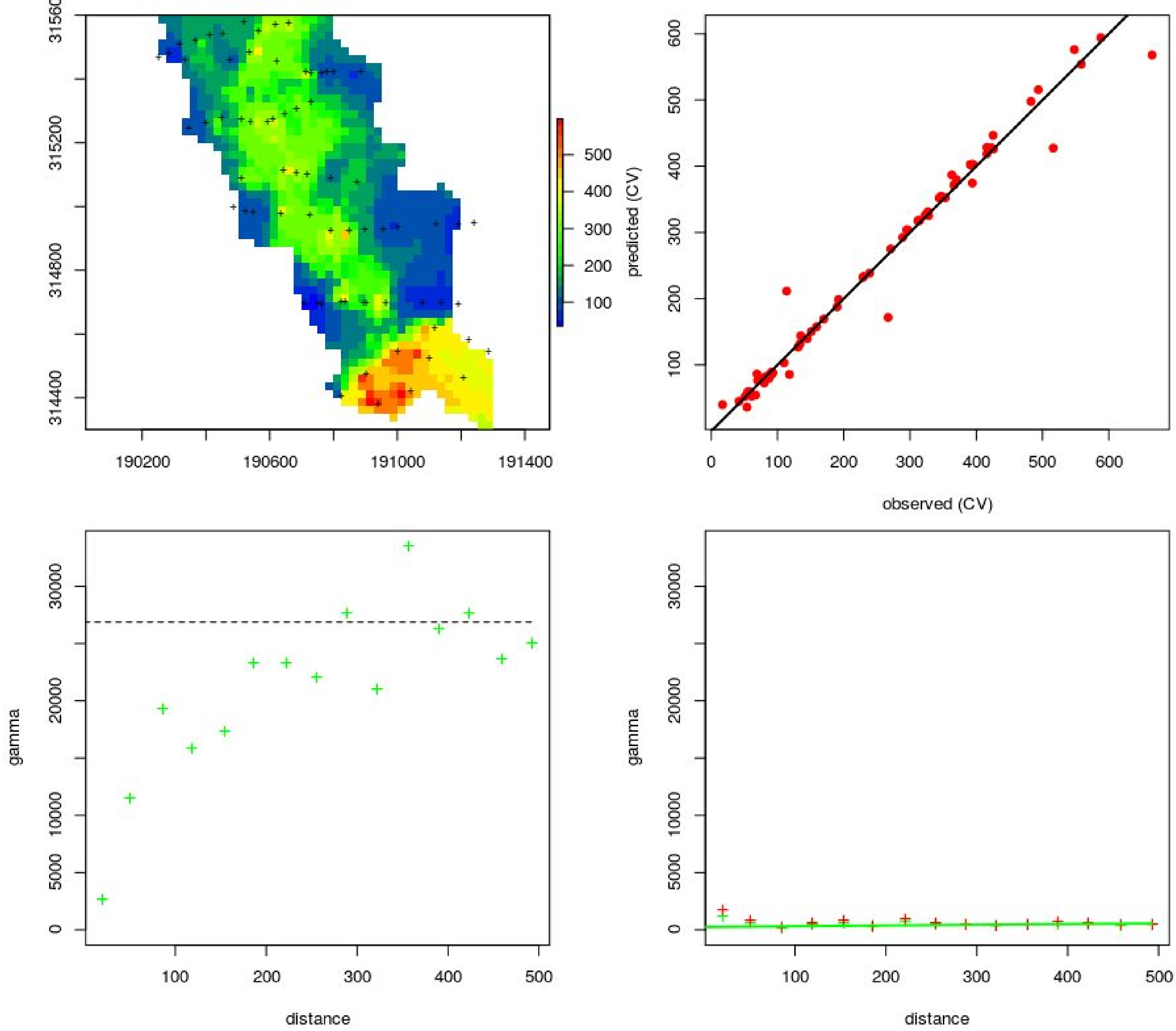
$$Y'(s) = f[\mathbf{h}(s)|Y, X(s)]$$

- $\mathbf{h}(s)|Y$ = buffer distances to points
- $X(s)$ = covariates / grids

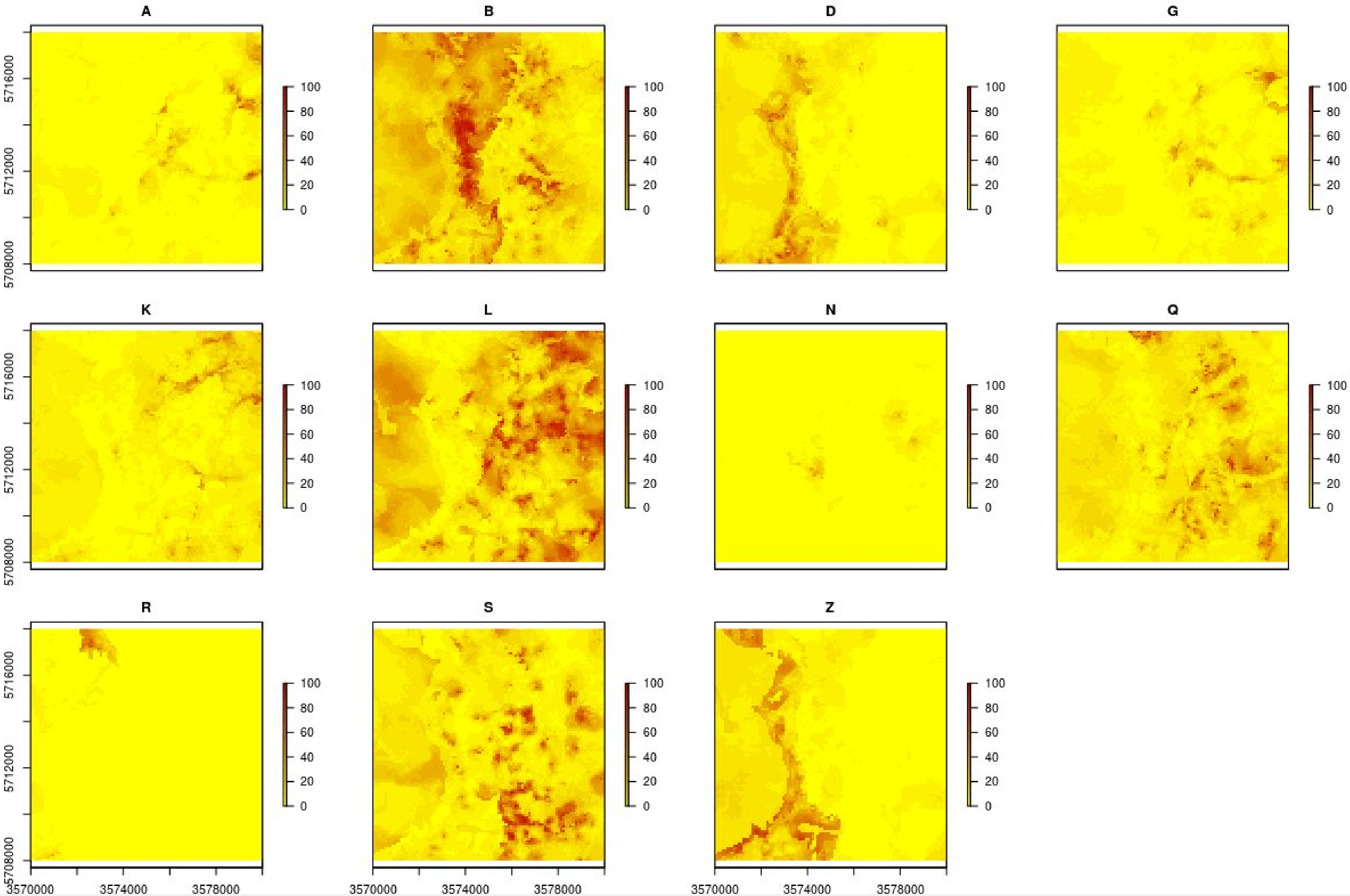
Tom Hengl (tom.hengl@isric.org)

RF with covariates (Pb Gaul)

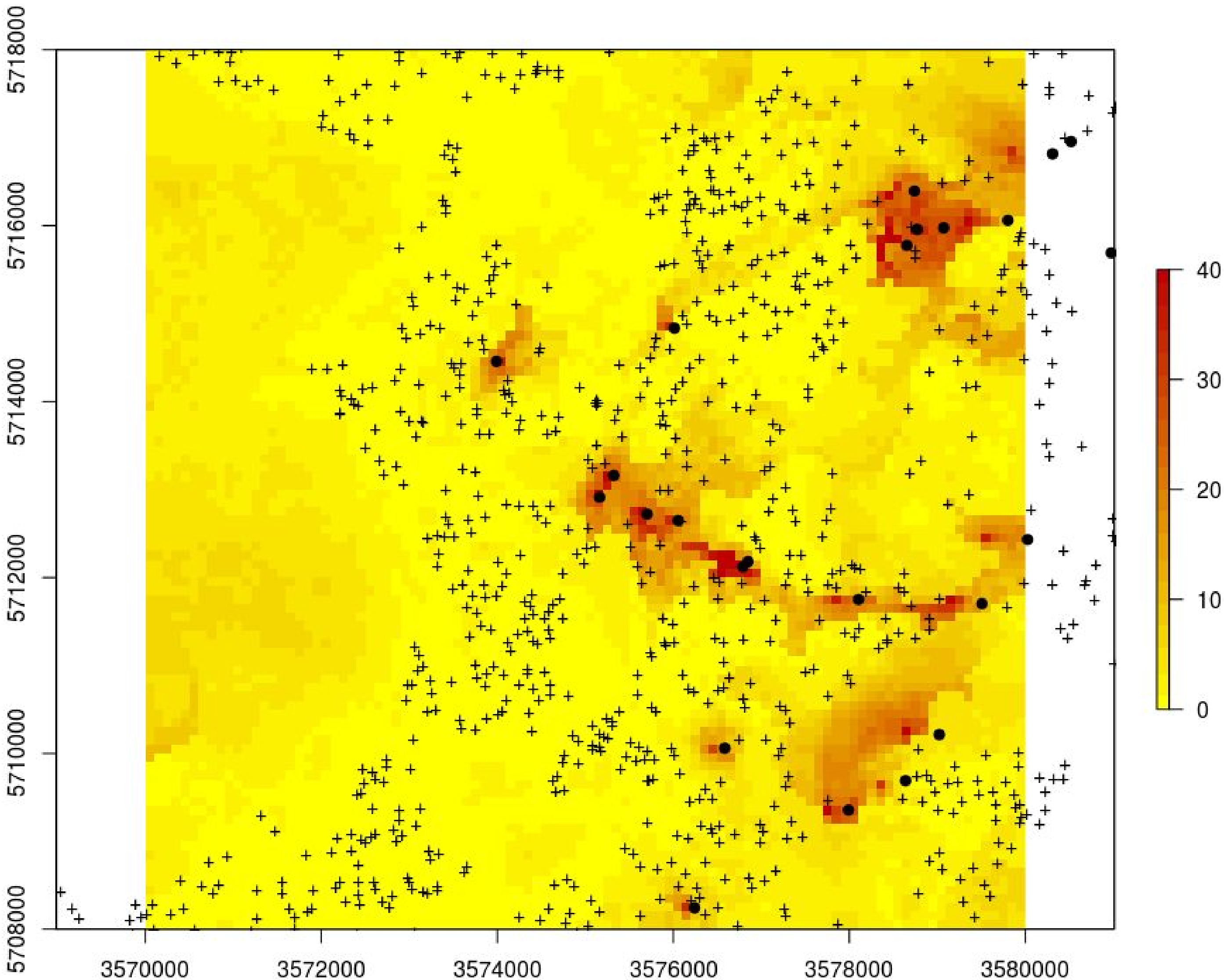




315600
315200
314800
314400
190200
190600
191000
191400
500
400
300
200
100
predicted (CV)
600
500
400
300
200
100
0
100
200
300
400
500
600
observed (CV)
gamma
30000
20000
10000
5000
0
100
200
300
400
500
distance
gamma
30000
20000
10000
5000
0
100
200
300
400
500
distance

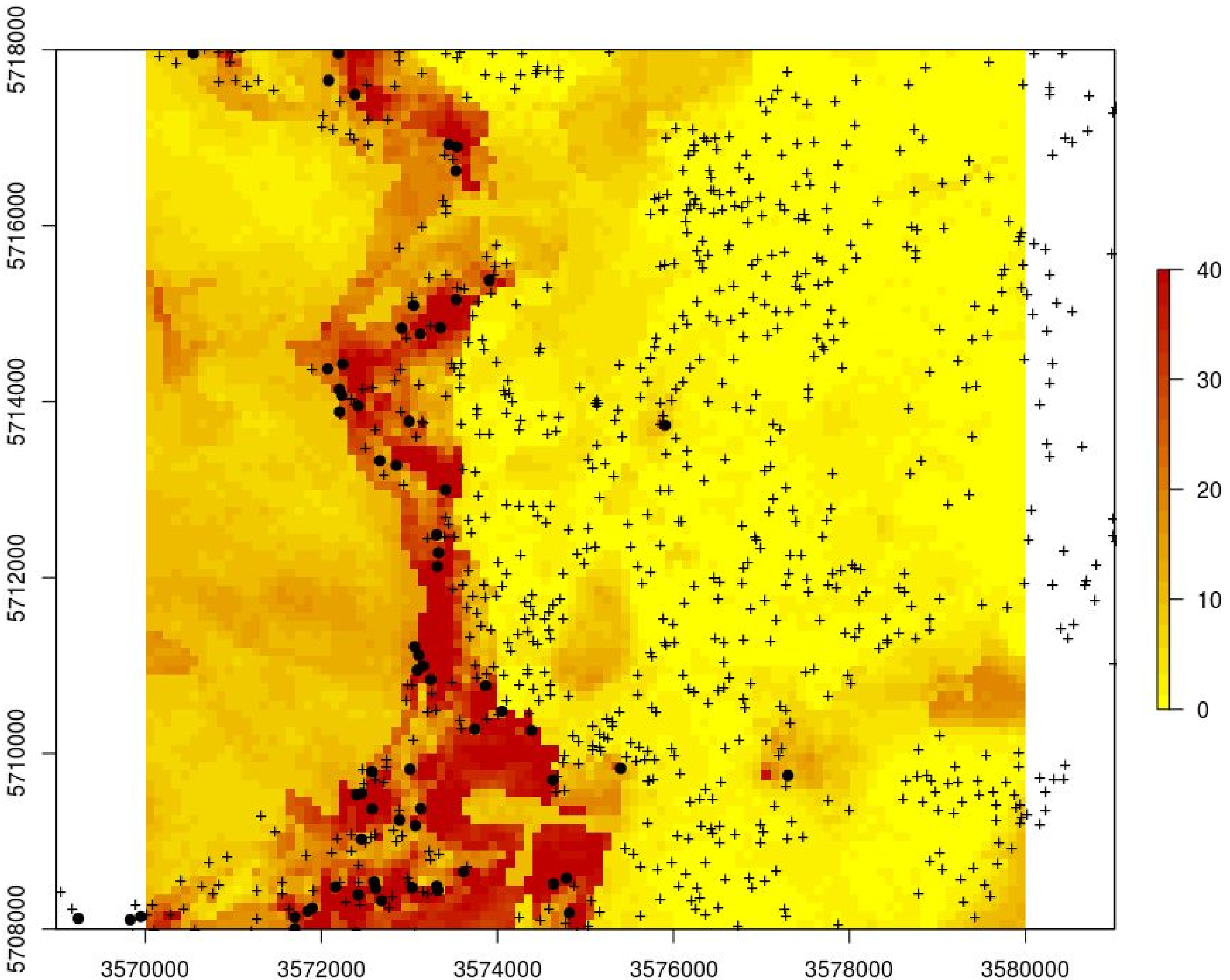

A
B
D
G
K
L
N
Q
R
S
Z
100
80
60
40
20
0
5716000
5712000
5708000
3570000
3574000
3578000

Soil type “G”



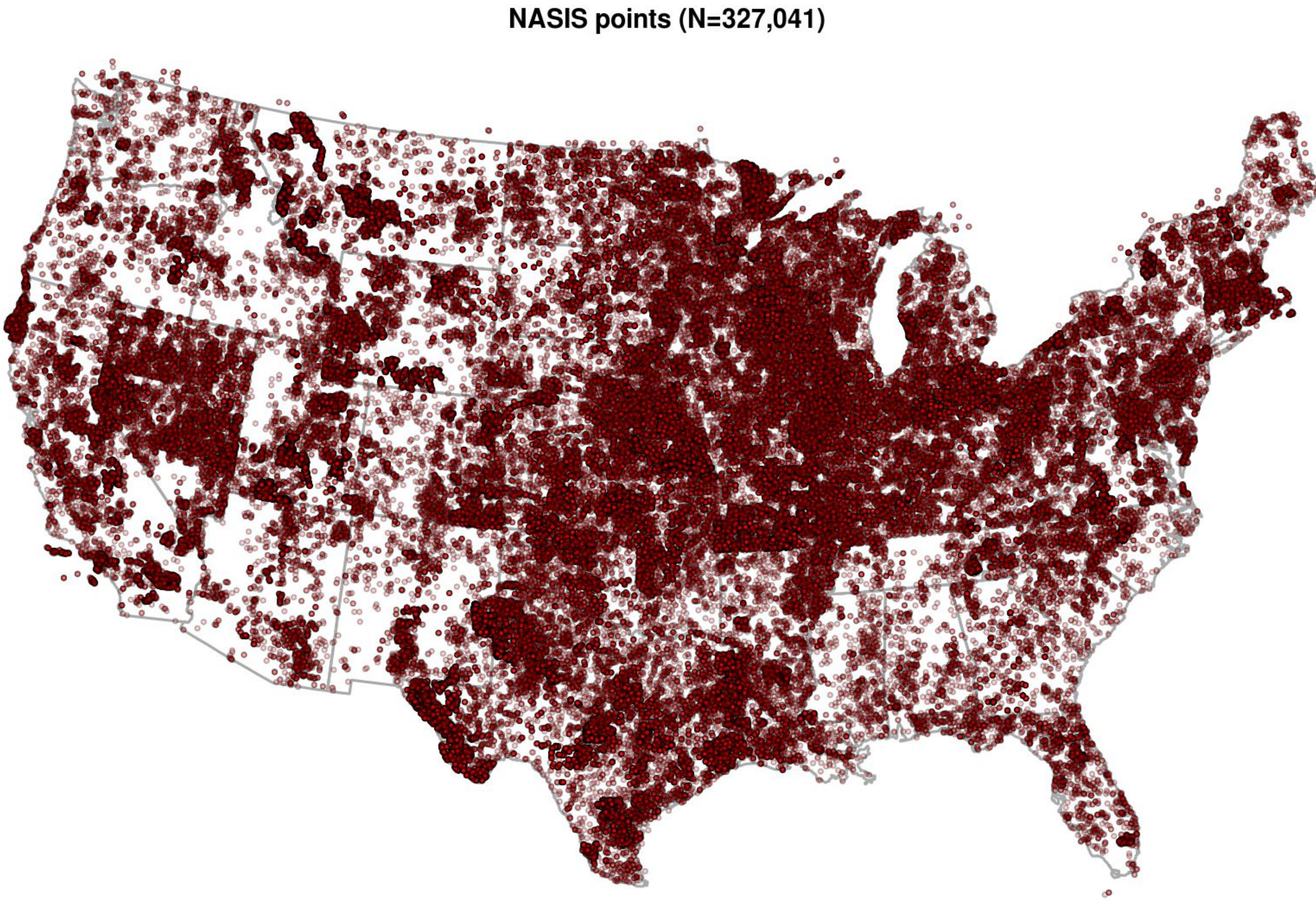
om.hengl@isric.org)

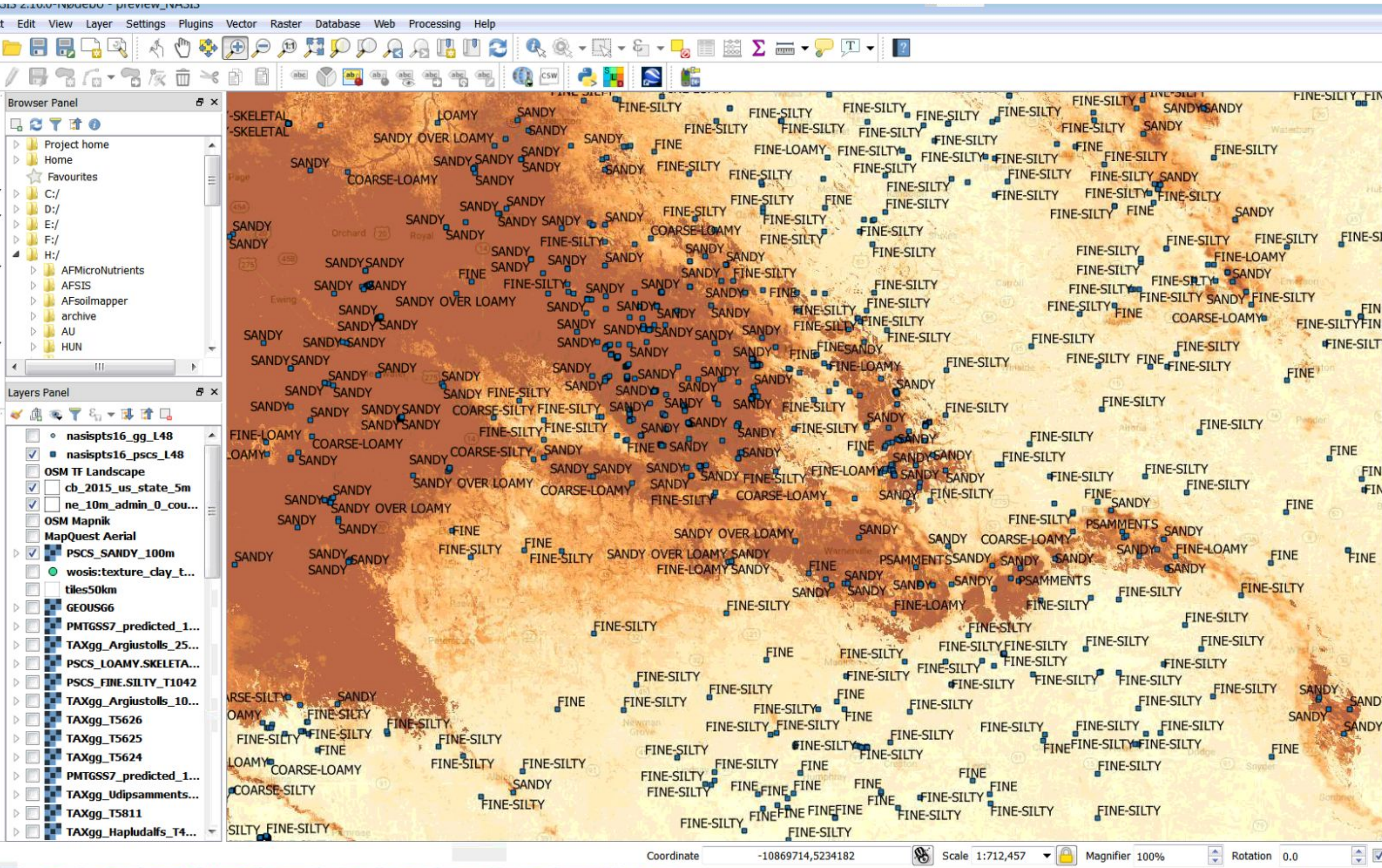
Soil type “D”



om.hengl@isric.org)

USA data

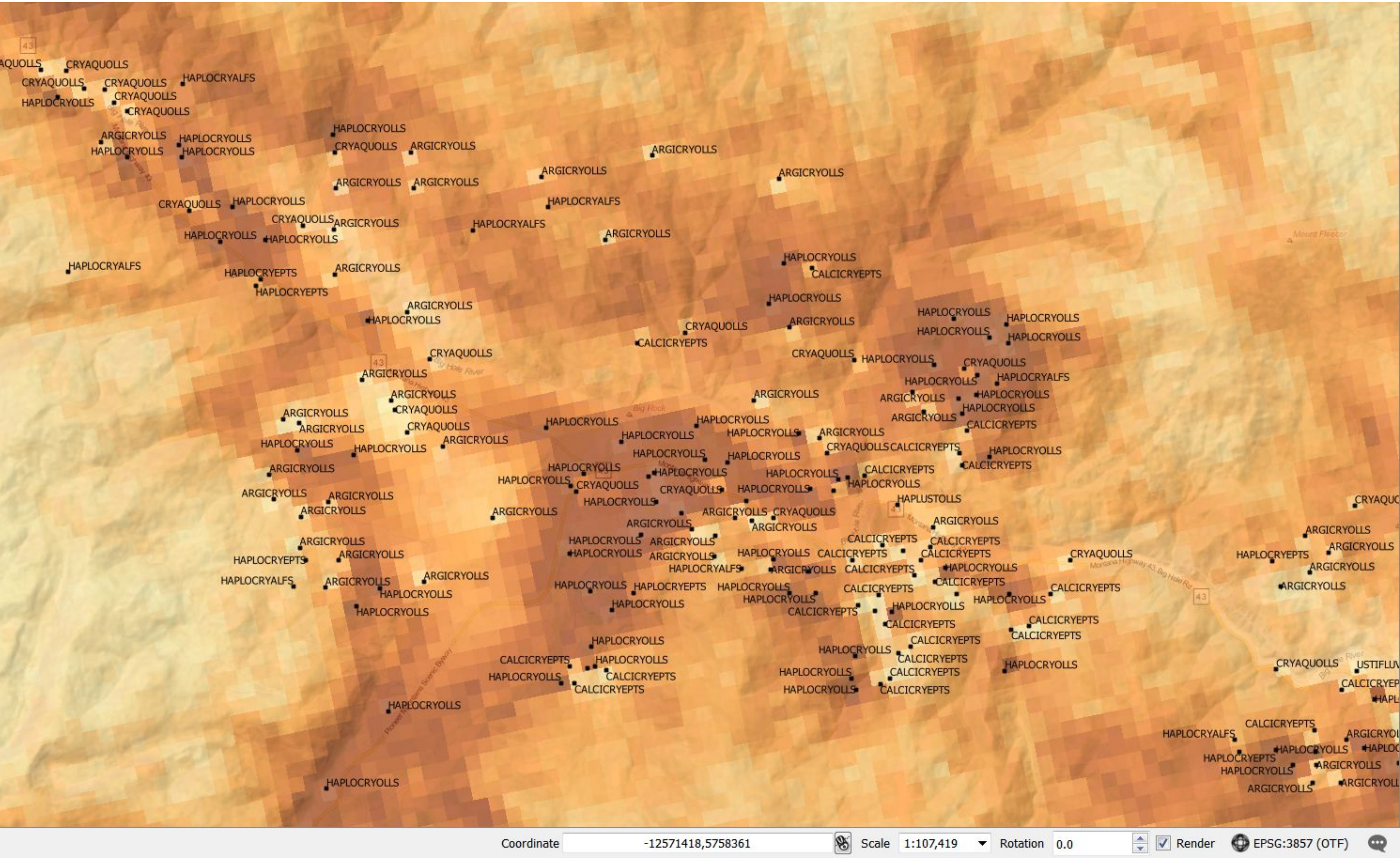




ISRIC World Soil Information

Tom Hengl (tom.hengl@isric.org)

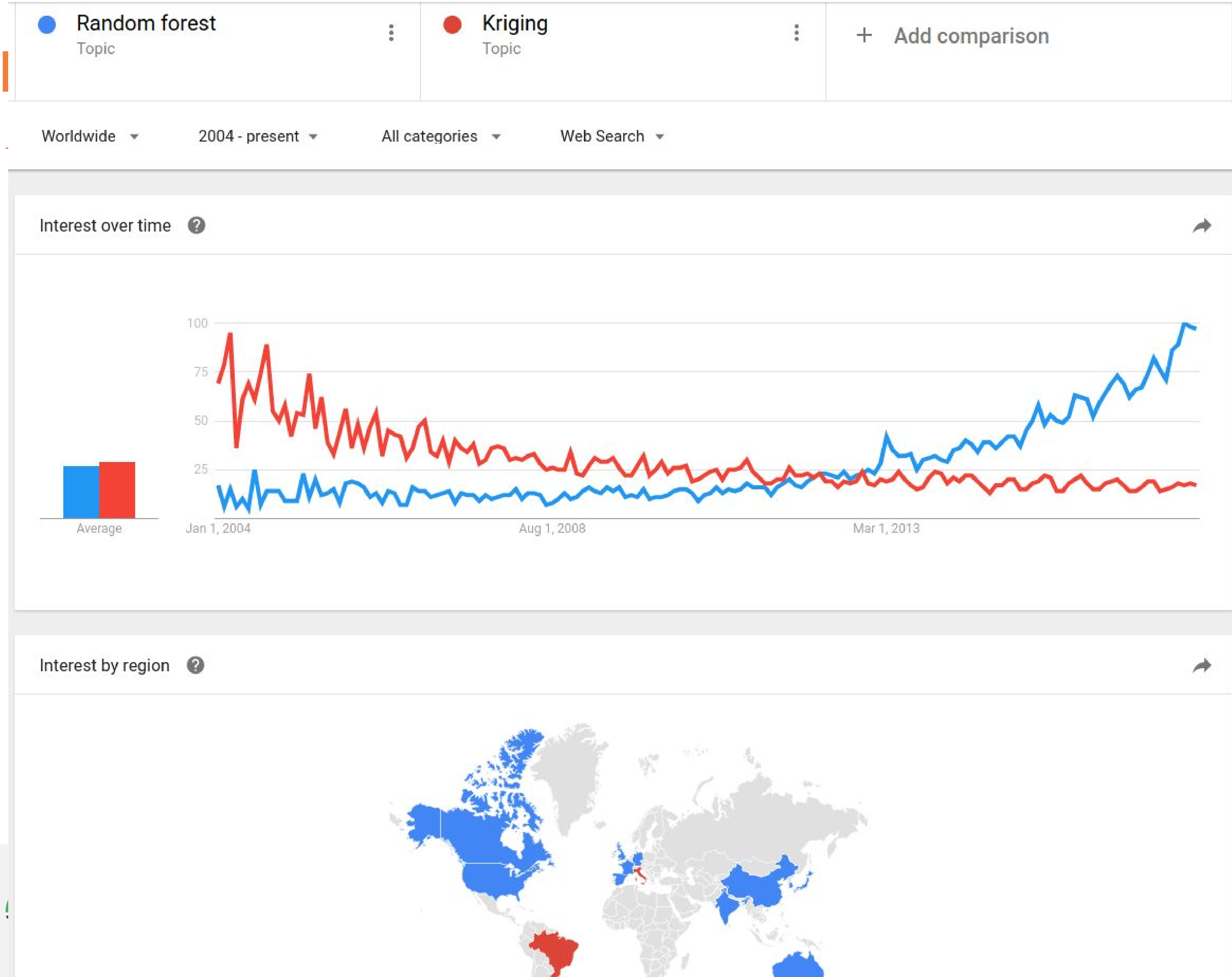
Kriging patterns



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

ISRIC
World Soil Information

Tom Hengl (tom.hengl@isric.org)


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

A generic framework for sp

In essence, three types of covariates:

$$Y'(s) = f[\ \mathbf{h}(s)|Y,\ \ X_R(s),\ \ X_P(s)\ ]$$

- $\mathbf{h}(s)|Y$ = buffer distances to points
- $X_R(s)$ = reflectances
- $X_P(s)$ = physical and chemical processes

Data-driven modeling (MLA)

1. Any target variable distribution is fine.
2. Multicolinearity — no problem.
3. A lot of covariates — no problem.
4. Complex non-linear relationships — great!

MLA = Hyper-parametric non-linear (nested) models, usually requiring intensive computing

Tom Hengl (tom.hengl@isric.org)

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); most importantly **it requires high quality sampling**
- ➔ There are still many things unsolved (spatial simulations? spatial clustering? spacetime distances?)

Tom Hengl (tom.hengl@isric.org)

Conclusions #1

Kriging?

Probably not needed any more.

Tom Hengl (tom.hengl@isric.org)

Conclusions #2

Spatial statistics?

Probably needed more than ever!

Tom Hengl (tom.hengl@isric.org)

Conclusions #3

MLA is magic?

Yes, but also very sensitive on data quality and still many many issues to be solved.

Tom Hengl (tom.hengl@isric.org)

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 world of artificial intelligence. One by one, the abilities and techniques that humans once imagined were uniquely our own have begun to fall to the onslaught of ever more powerful machines. Deep neural networks are now better than humans at tasks such as face recognition and object recognition. They've mastered the ancient game of Go and thrashed the best human players.

But there is a problem. There is no mathematical reason why networks arranged in layers should be so good at these challenges. Mathematicians are flummoxed. Despite the huge success of deep

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

ISRIC
World Soil Information

Thanks for your attention

www.isric.org