OPEN DATA SCIENCE EUROPE

WORKSHOP

Modeling with spatial and spatiotemporal data in R: spatial interpolation

Sept 6, 2021: 11:00 - 12:30

Tom Hengl

tom.hengl@opengeohub.org

https://opengeohub.org

Programme

Day 1 Sept 6, 2021

9:00 - 10:30

Introduction to spatial and spatiotemporal data in R

By **Tomislav Hengl** Spatiotemporal Ensemble ML in R / computing with Cloud-Optimized GeoTIFFs

11:00 - 12:30

Modeling with spatial and spatiatemporal data in R

By **Tomislav Hengl** Spatiotemporal Ensemble ML in R / computing with Cloud-Optimized GeoTIFFs

13:30 - 15:00

Spatiotemporal Ensemble ML in R

By **Tomislav Hengl** Spatiotemporal Ensemble ML in R / computing with Cloud-Optimized GeoTIFFs

15:30 - 17:00

Computing with Cloud-Optimized GeoTIFFs

By **Tomislav Hengl** Spatiotemporal Ensemble ML in R / computing with Cloud-Optimized GeoTIFFs

Day 2 Sept 7, 2021

9:00 - 10:30

Data visualization: from R to Google Earth and QGIS

By **Tomislav Hengl** Spatiotemporal Ensemble ML in R / computing with Cloud-Optimized GeoTIFFs

Outline

- Modeling seasonality effects in spatiotemporal data;
- ML for predictive mapping: standard ML steps;
- Extrapolation and over-fitting problems of ML;
- Predictive mapping using machine learning:
  - Using buffer distances to points to generate predictions;
  - Using distances to neighbors;

Example of seasonality pattern

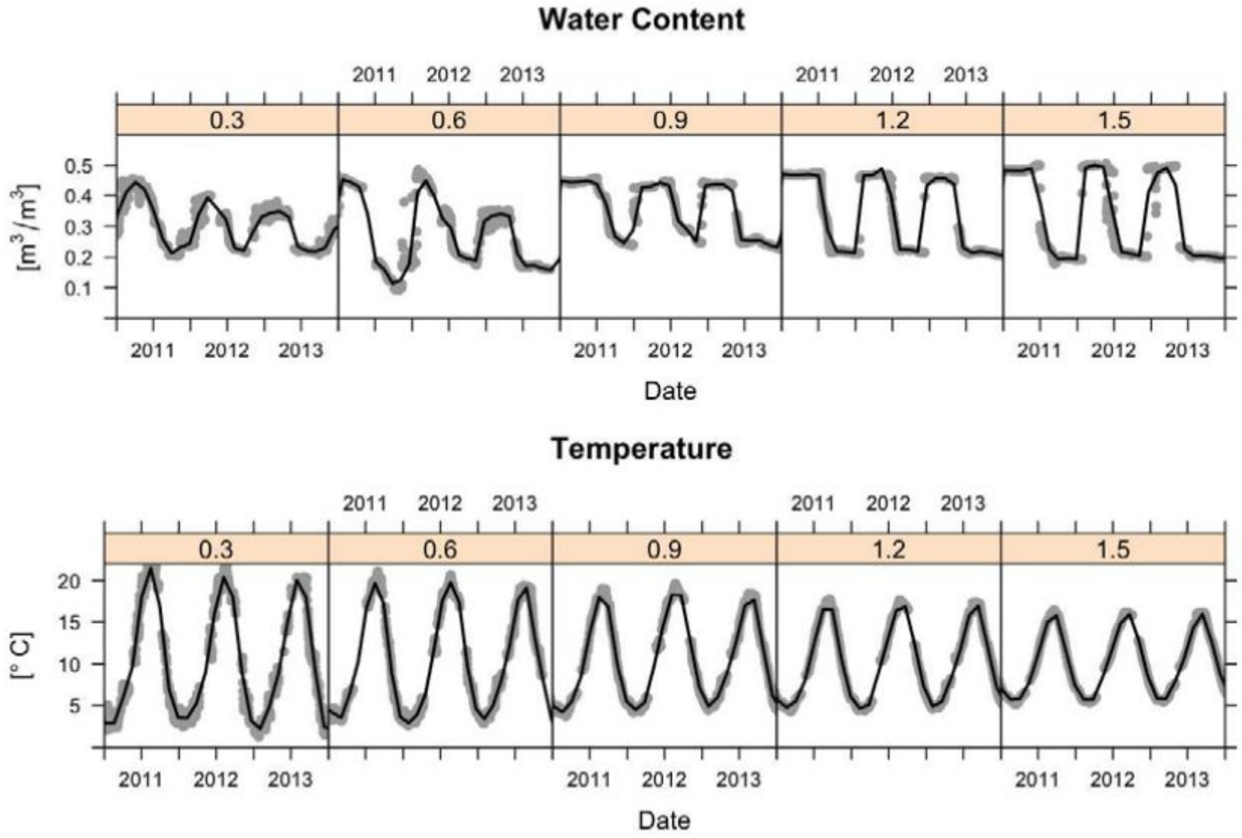


Figure: Sensor values from five depths (0.3, 0.6, 0.9, 1.2, and 1.5 m) at one station at Cook Agronomy Farm from January 2011–January 2014. The black line indicates locally fitted splines ([Gasch et al., 2015](#)).

Daily temperature as a function of Lat, Doy, Alt

AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE

10.1002/2013JD020803

Key Points:
- Global spatio-temporal regression-kriging daily temperature interpolation
- Fitting of global spatio-temporal models for the mean, maximum, and minimum temperatures
- Time series of MODIS 8 day images as explanatory variables in regression part

Correspondence to:
M. Kilibarda,
kili@grf.bg.ac.rs

Citation:
Kilibarda, M., T. Hengl, G. B. M. Heuvelink, B. Gräler, E. Pebesma, M. Perčec Tadić, and B. Bajat (2014), Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution, *J. Geophys. Res. Atmos.*, *119*, 2294–2313,

Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution

Milan Kilibarda[1], Tomislav Hengl[2], Gerard B. M. Heuvelink[3], Benedikt Gräler[4], Edzer Pebesma[4], Melita Perčec Tadić[5], and Branislav Bajat[1]

[1]Department of Geodesy and Geoinformatics, Faculty of Civil Engineering, University of Belgrade, Belgrade, Serbia, [2]ISRIC-World Soil Information, Wageningen, Netherlands, [3]Soil Geography and Landscape Group, Wageningen University, Wageningen, Netherlands, [4]Institute for Geoinformatics, University of Münster, Münster, Germany, [5]Meteorological and Hydrological Service, Zagreb, Croatia

Abstract Combined Global Surface Summary of Day and European Climate Assessment and Dataset daily meteorological data sets (around 9000 stations) were used to build spatio-temporal geostatistical models and predict daily air temperature at ground resolution of 1 km for the global land mass. Predictions in space and time were made for the mean, maximum, and minimum temperatures using spatio-temporal regression-kriging with a time series of Moderate Resolution Imaging Spectroradiometer (MODIS) 8 day images, topographic layers (digital elevation model and topographic wetness index), and a geometric temperature trend as covariates. The accuracy of predicting daily temperatures was assessed using leave-one-out cross validation. To account for geographical point clustering of station data and get a more representative cross-validation accuracy, predicted values were aggregated to blocks of land of size 500 × 500 km. Results show that the average accuracy for predicting mean, maximum, and minimum daily temperatures is root-mean-square error (RMSE) = ±2°C for areas densely covered with stations and between

Daily temperature as a function of Lat, Doy, Alt

perature is a function of geometric position of a particular location on Earth and day of the year. We call this a *geometric temperature trend*. The geometric temperature trend for the mean temperature was modeled as a function of the day of year and latitude (ϕ):

$$t_{\text{geom}} = 30.4 \cos \phi - 15.5(1 - \cos \theta) \sin |\phi|, \tag{3}$$

where θ is derived as

$$\theta = (\text{day} - 18)\frac{2\pi}{365} + 2^{1-\text{sgn}(\phi)}\pi. \tag{4}$$

The number 18 represents the coldest day in the Northern and warmest day in the Southern Hemisphere and was derived empirically by graphical inspection of mean daily temperature plots from stations in the Northern and Southern Hemispheres. The sgn denotes the signum function that extracts the sign of a real number. Parameters 30.4°C and 15.5°C of the geometric temperature trend were calculated by least squares fitting on circa 44 million daily temperature observations from 2000 to 2011. These two numbers are, in fact, similar to the mean annual temperature on the equator and the mean global Earth temperature.

The linear model for the minimum daily temperature uses the same covariates as the linear model for mean daily temperature. The geometric temperature trend for minimum daily temperature was

$$t_{\text{geom}} = 24.2 \cos \phi - 15.7(1 - \cos \theta) \sin |\phi|, \tag{5}$$

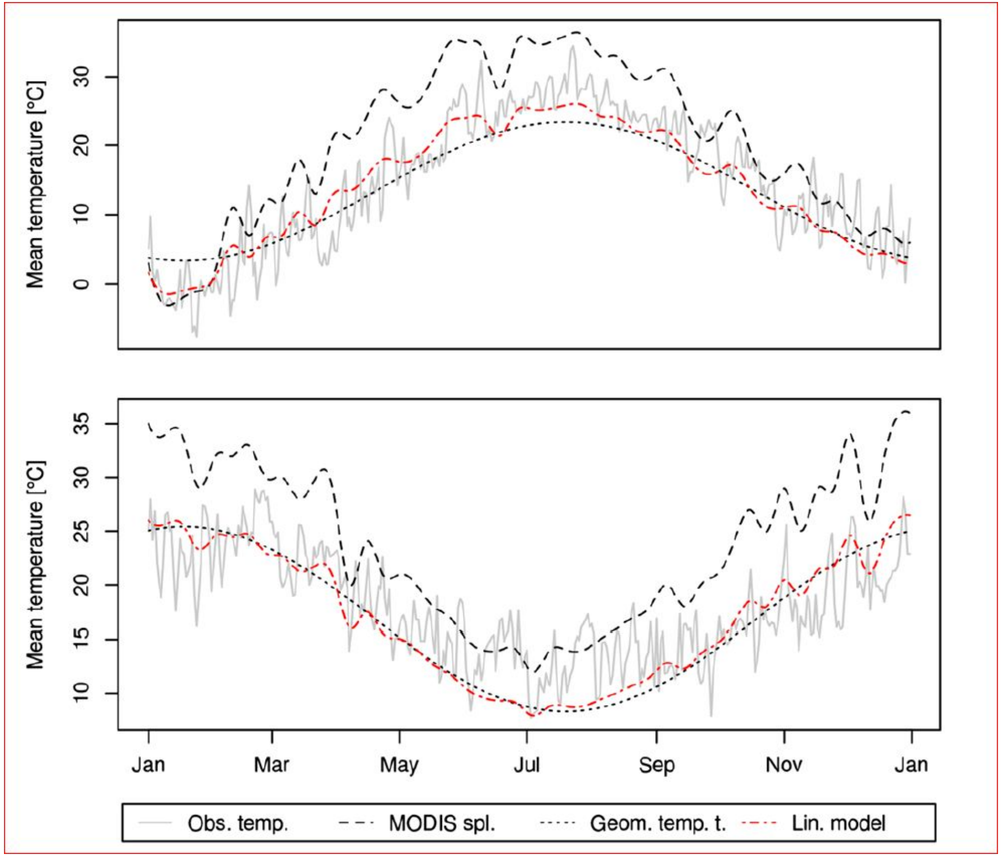
The geometric temperature trend for maximum daily temperature was

$$t_{\text{geom}} = 37 \cos \phi - 15.4(1 - \cos \theta) \sin |\phi|, \tag{6}$$

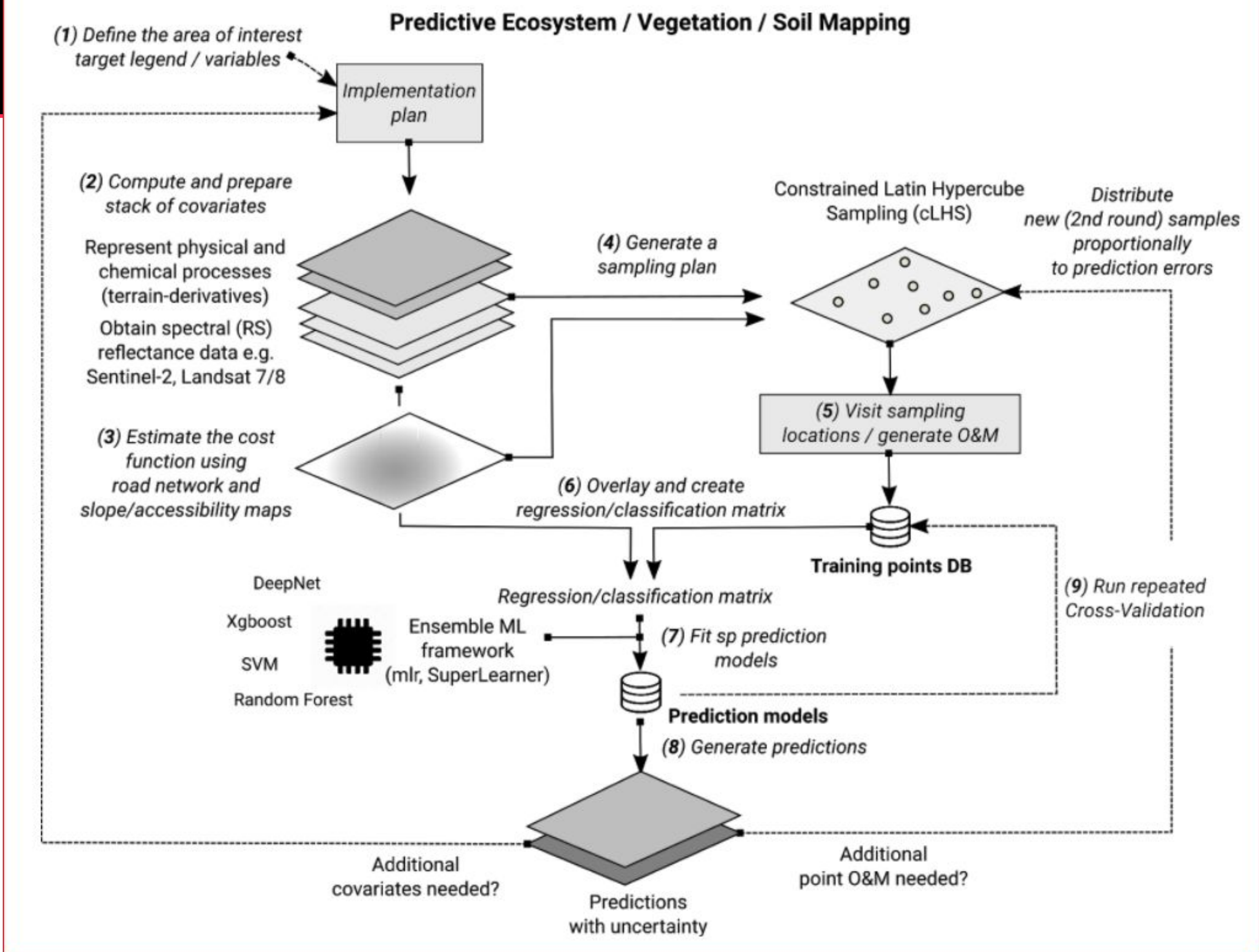
Daily temperature as a function of Lat, Doy, Alt

```
temp.from.geom <- function(fi, day, a=30.419375,
                b=-15.539232, elev=0, t.grad=0.6) {
  f = ifelse(fi==0, 1e-10, fi)
  costeta = cos( (day-18 )*pi/182.5 + 2^(1-sign(fi) ) *pi)
  cosfi = cos(fi*pi/180 )
  A = cosfi
  B = (1-costeta ) * abs(sin(fi*pi/180 ) )
  x = a*A + b*B - t.grad * elev / 100
  return(x)
}
```

Daily temperature as a function of Lat, Doy, Alt



Kilibarda, Milan, Hengl, T., Heuvelink, G. B., Gräler, B., Pebesma, E., Perčec Tadić, M., & Bajat, B. (2014). Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. Journal of Geophysical Research: Atmospheres, 119(5), 2294–2313. doi:[10.1002/2013JD020803](https://doi.org/10.1002/2013JD020803)



https://soilmapper.org & Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6:e5518 https://doi.org/10.7717/peerj.5518

Predictive mapping using ML (2D)

1. Prepare **training (points)** data and data cube with all covariates ideally as **an analysis-ready datacube**.
2. Overlay points and create a **regression-matrix**.
3. Fine-tune initial model, reduce complexity and produce **production-ready prediction model**.
4. **Run mapping accuracy assessment** and determine prediction uncertainty including the per pixel uncertainty.
5. Generate predictions and save as **maps**.
6. Visualize predictions using **web-GIS solutions**.

topepo.github.io/caret/model-training-and-tuning.html#model-trainin...

5 Model Training and Tuning

Contents

The caret package (short for Classification And REgression Training)

2D+T mapping

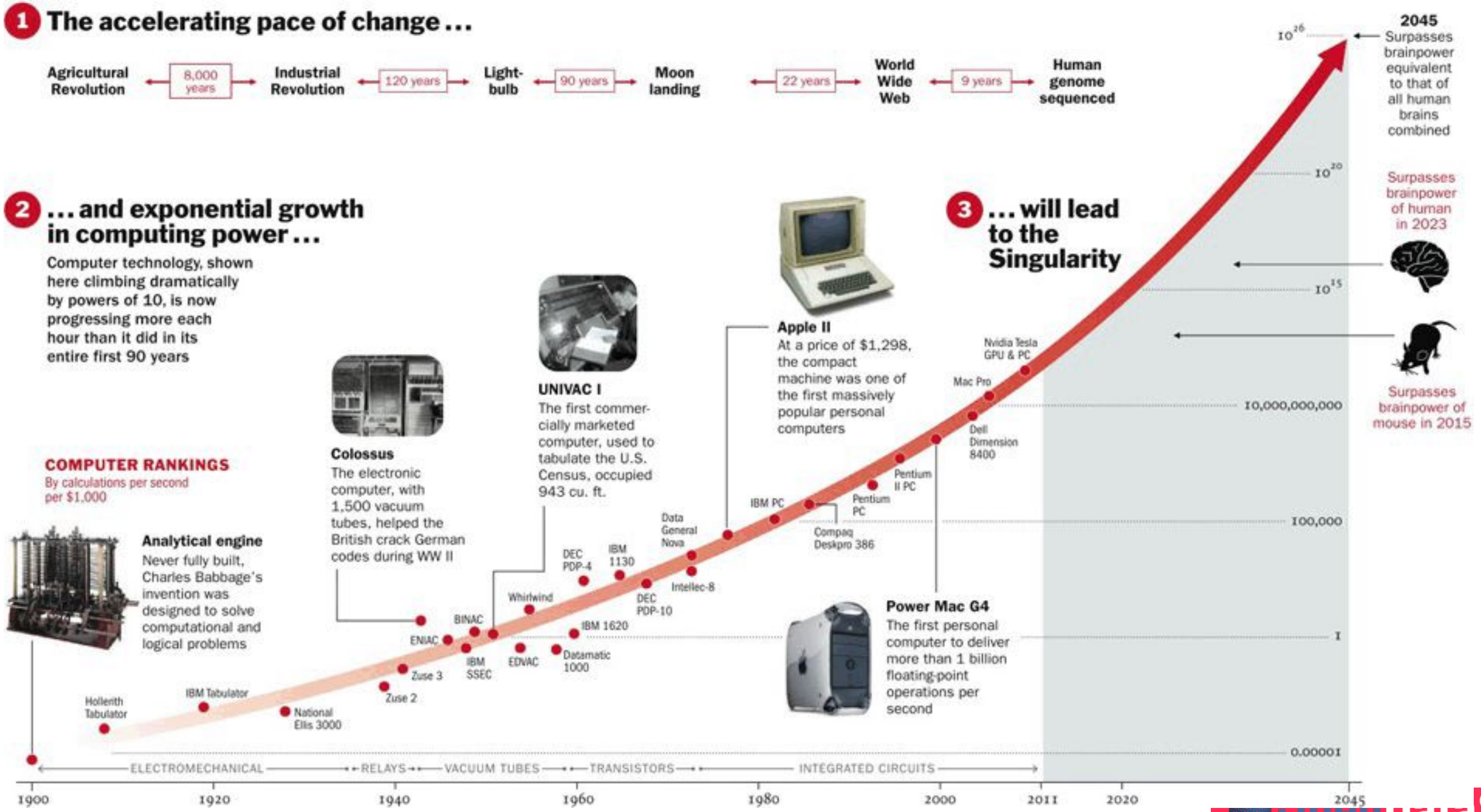
The principle is similar, but there are several additions:

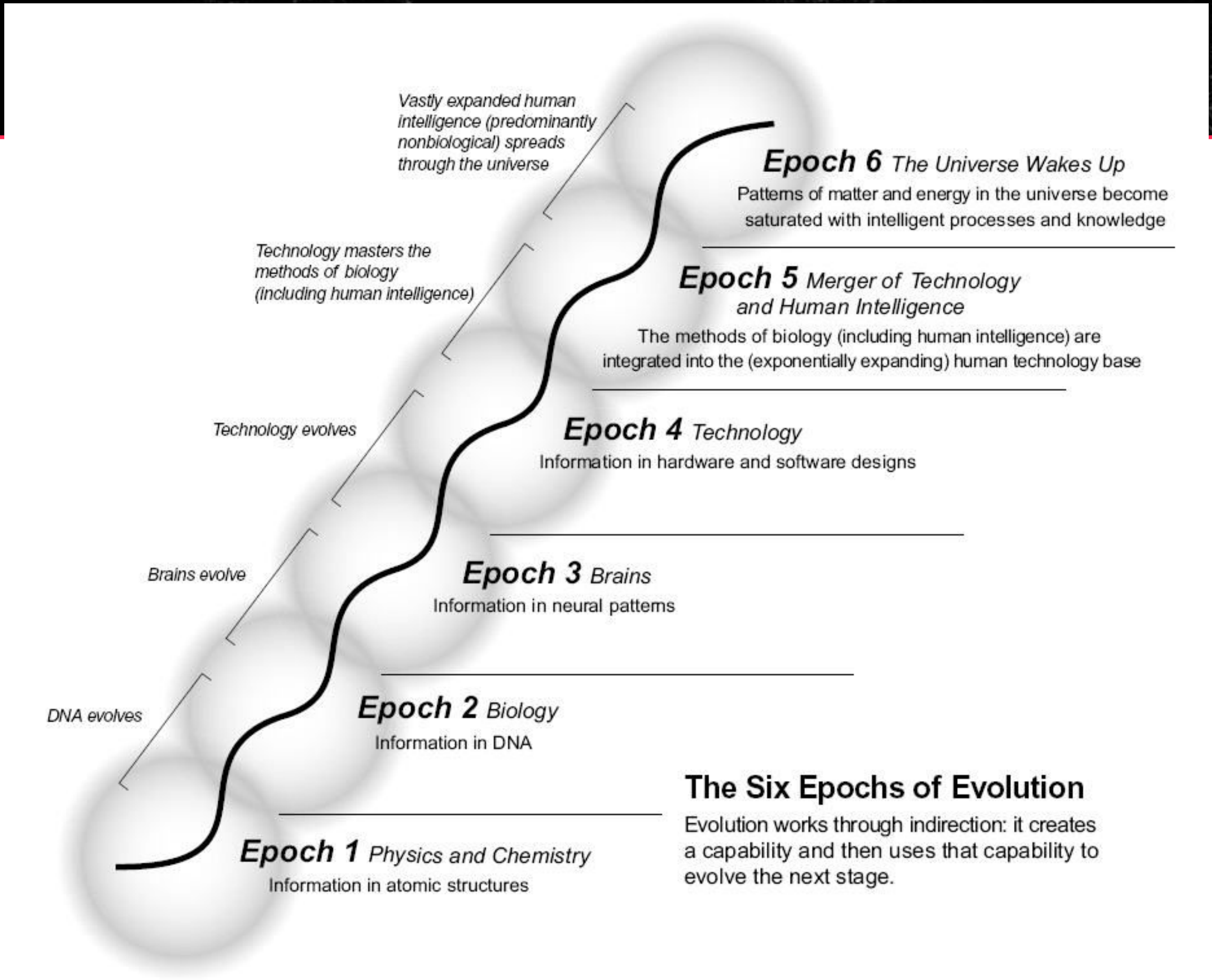
- Covariates are time-series of images;
- Seasonality component needs to be dealt with;
- Instead of simple spatial overlay, we need to do spatiotemporal overlay (can be very computational!);
- The spatiotemporal regression-/classification-matrix can be at the order of magnitude larger;
- Predicting in future is not trivial; predicting while ignoring processes and physical laws can lead to artifacts;

My story

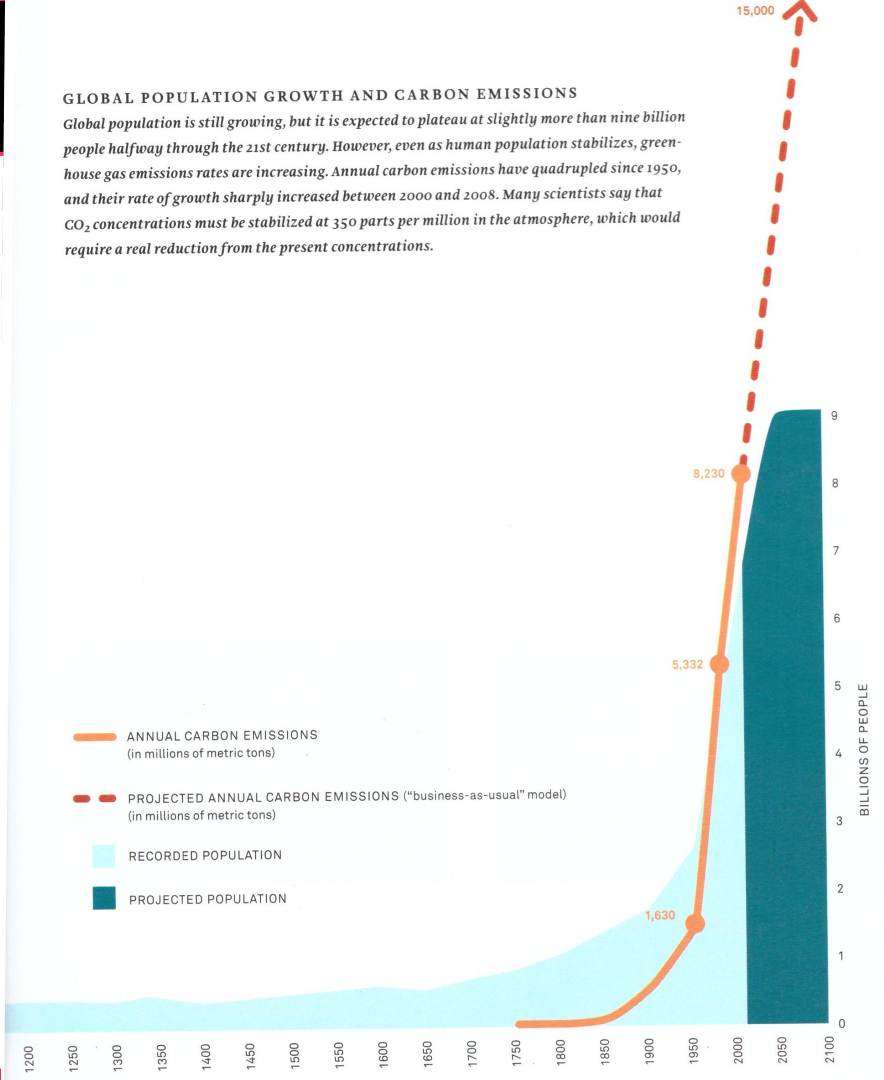
Possibilities of automation and complex problem solving using Machine Learning are very attractive

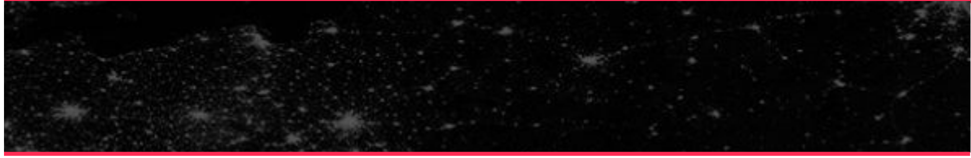
Exponential growth of technology





Vastly expanded human intelligence (predominantly nonbiological) spreads through the universe
Epoch 6 The Universe Wakes Up
Patterns of matter and energy in the universe become saturated with intelligent processes and knowledge
Technology masters the methods of biology (including human intelligence)
Epoch 5 Merger of Technology and Human Intelligence
The methods of biology (including human intelligence) are integrated into the (exponentially expanding) human technology base
Technology evolves
Epoch 4 Technology
Information in hardware and software designs
Brains evolve
Epoch 3 Brains
Information in neural patterns
DNA evolves
Epoch 2 Biology
Information in DNA
Epoch 1 Physics and Chemistry
Information in atomic structures
The Six Epochs of Evolution
Evolution works through indirection: it creates a capability and then uses that capability to evolve the next stage.

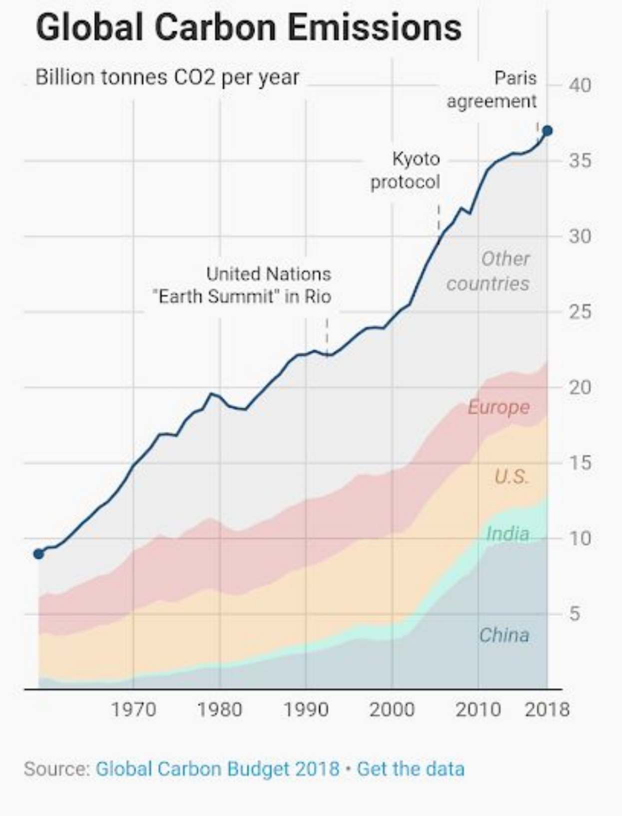


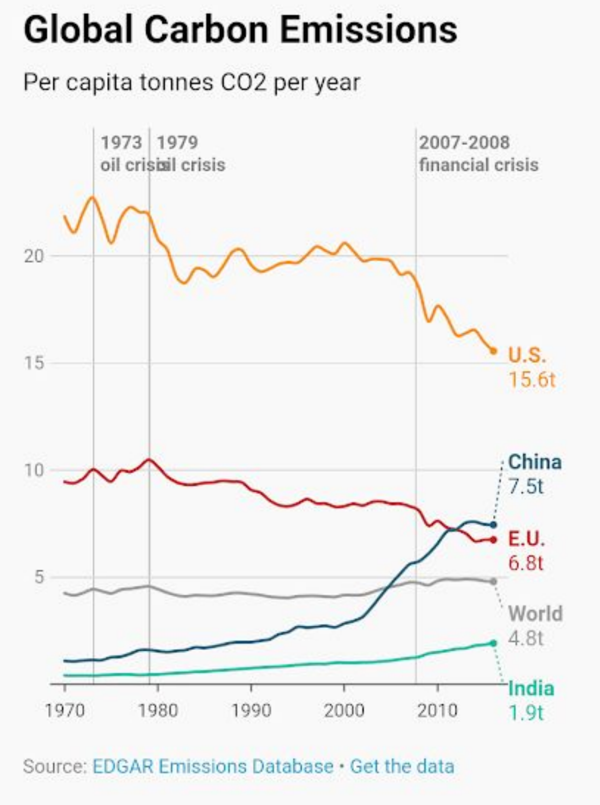


Not all human-related exponential plots are encouraging!
When you plot CO2 emissions for the last 150 years it looks much more steep than we are ready to accept.

Source: "Our Choice"

We are risking our children's lives



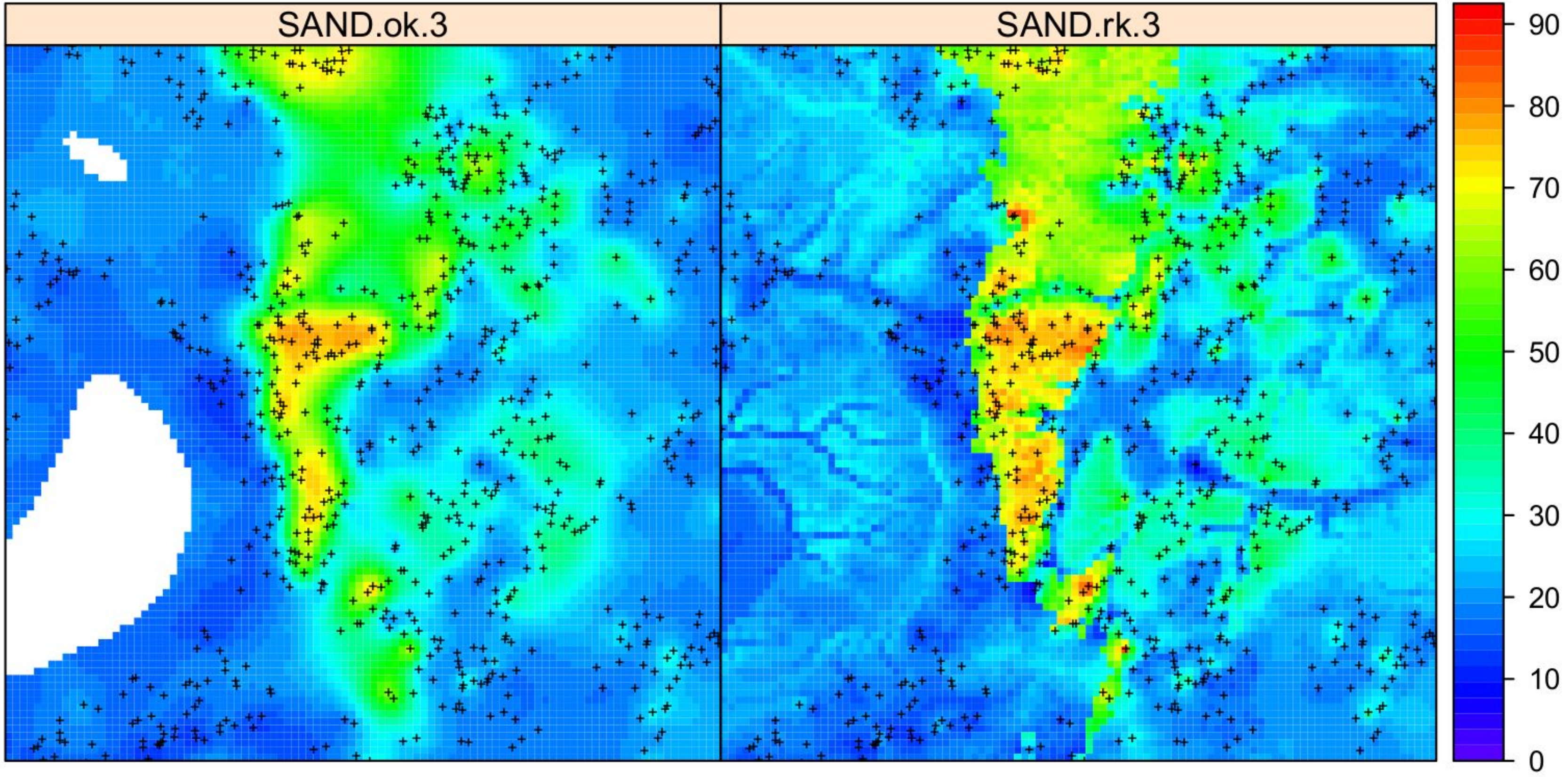


20 years after Kyoto Protocol, where does world stand on climate? NO IMPROVEMENT?!!

My evolution over the years

- ⇓ 1999–2003: Kriging (in 1999 I was very excited about it),
- ⇓ 2001–2007: Regression-kriging (main topic of my PhD thesis!),
- ⇓ 2014–...: Random forest (Africa 250m),
- ⇓ 2016–...: Random forest + XGboost + SVM (global 250m),
- ⇓ 2018–...: Random forest spatial (RFsp),
- ⇓ 2019–...: Ensemble Machine Learning spatial (EML),
- ⇓ 2020–...: Global spatiotemporal EML,

Ordinary Kriging vs RK (Year: 2013)



OpenDataScience

International Journal of Applied Earth Observation and Geoinformation

Volume 22, June 2013, Pages 127-138

Mapping efficiency and information content

T. Hengl [a], M. Nikolić [b], R.A. MacMillan [a]

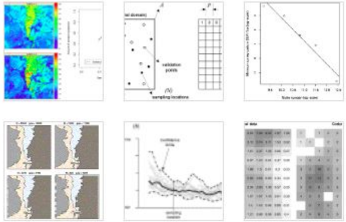
Show more

https://doi.org/10.1016/j.jag.2012.02.005

Abstract

This paper proposes two compound measures of mapping quality to support objective comparison of spatial prediction techniques for geostatistical mapping: (1) *mapping efficiency* – defined as the costs per area per amount of variation explained by the model, and (2) *information production efficiency* – defined as the cost per byte of effective information produced. These were inspired by concepts of complexity from mathematics and physics. Complexity i.e. the total effective information is defined as bytes remaining after compression and after rounding up the numbers using half the mapping accuracy (effective precision). It is postulated that the mapping efficiency, for an area of given size and limited budget, is basically a function of inspection intensity and mapping accuracy. Both measures are illustrated using the Meuse and Ebergötzen case studies (gstat, plotKML packages). The results demonstrate that, for mapping organic matter (Meuse data set), there is a gain in the mapping efficiency when using regression-kriging versus ordinary kriging: mapping efficiency is 7% better and the information production efficiency

Figures (9)



Show all figures

Tables (25)

PEER-REVIEWED Environmental Science section ›

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], Benedikt Gräler[5]

Published August 29, 2018

Note that a Preprint of this article also exists, first published March 14, 2018.

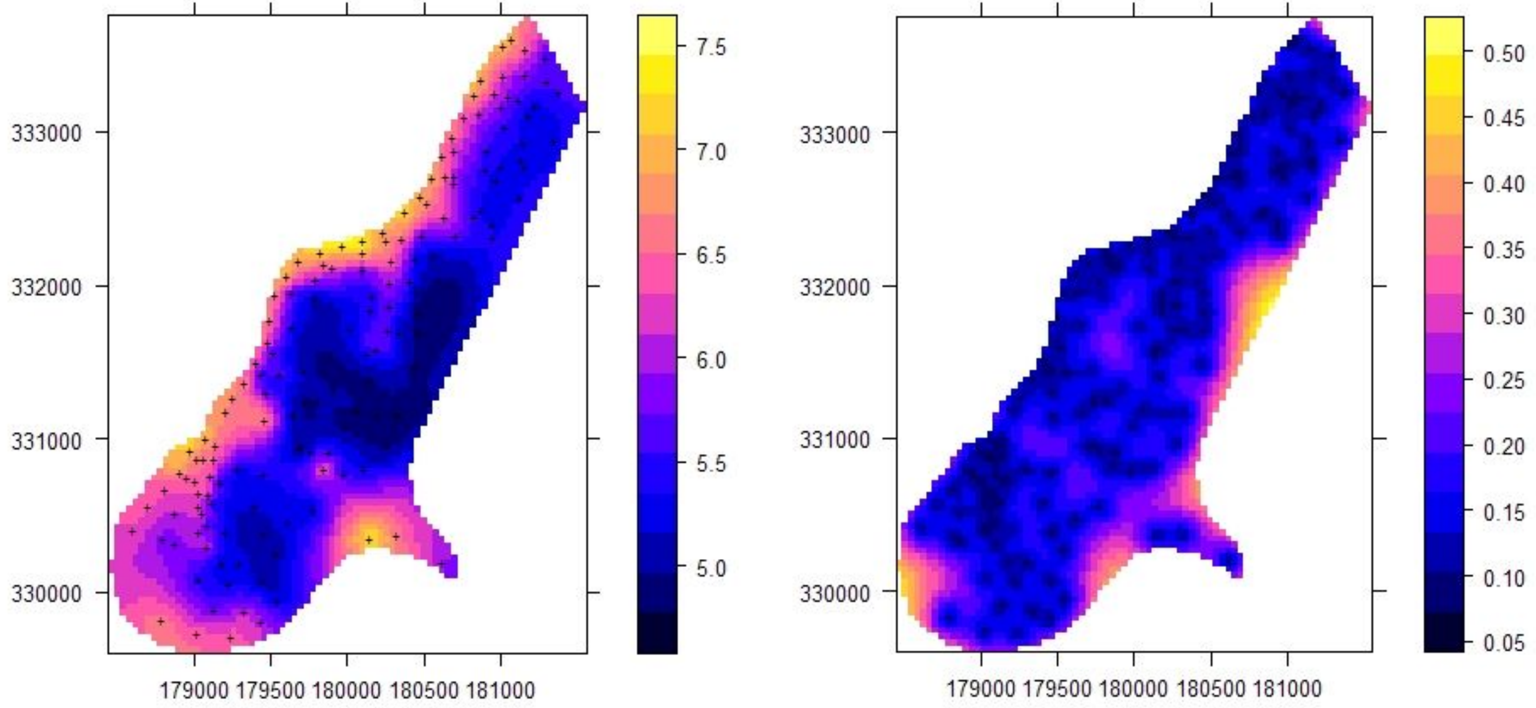
PubMed 30186691

› 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

“Kriging” is basically a synonym for Geostatistics



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)

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...

GLM vs RF

transform covariate

subset covariates

define link function

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

vs

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

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

Kriging using geoR package

```
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 modeling

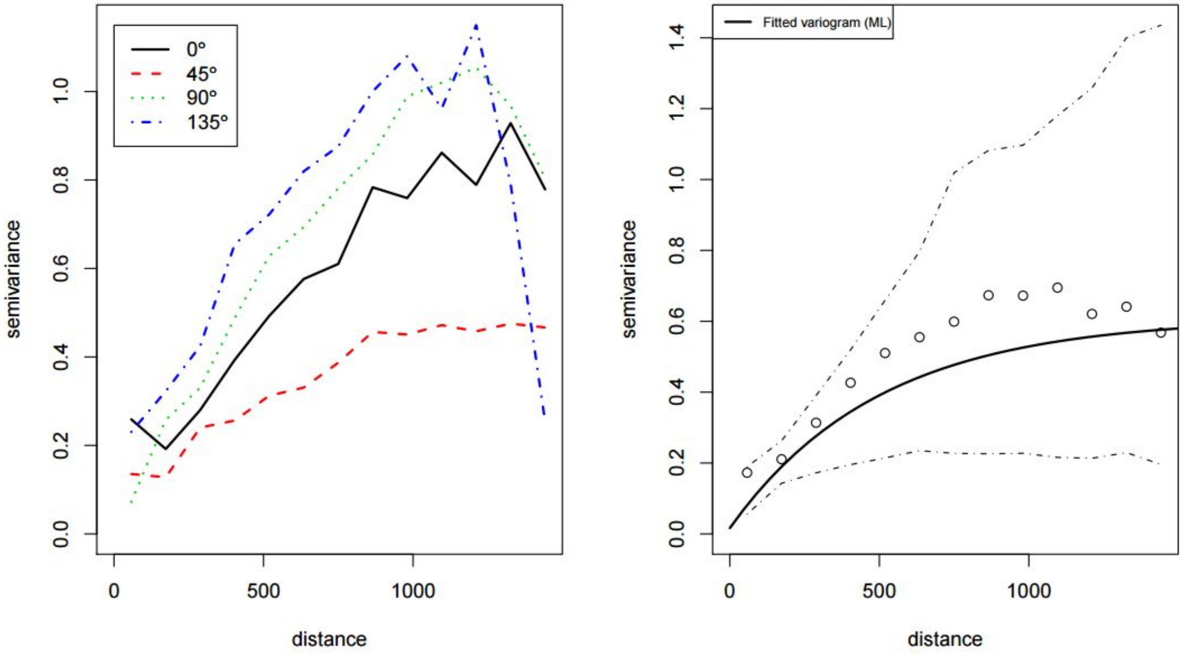


Fig. 5.15: Anisotropy (left) and variogram model fitted using the Maximum Likelihood (ML) method (right). The confidence bands (*envelopes*) show the variability of the sample variogram estimated using simulations from a given set of model parameters.

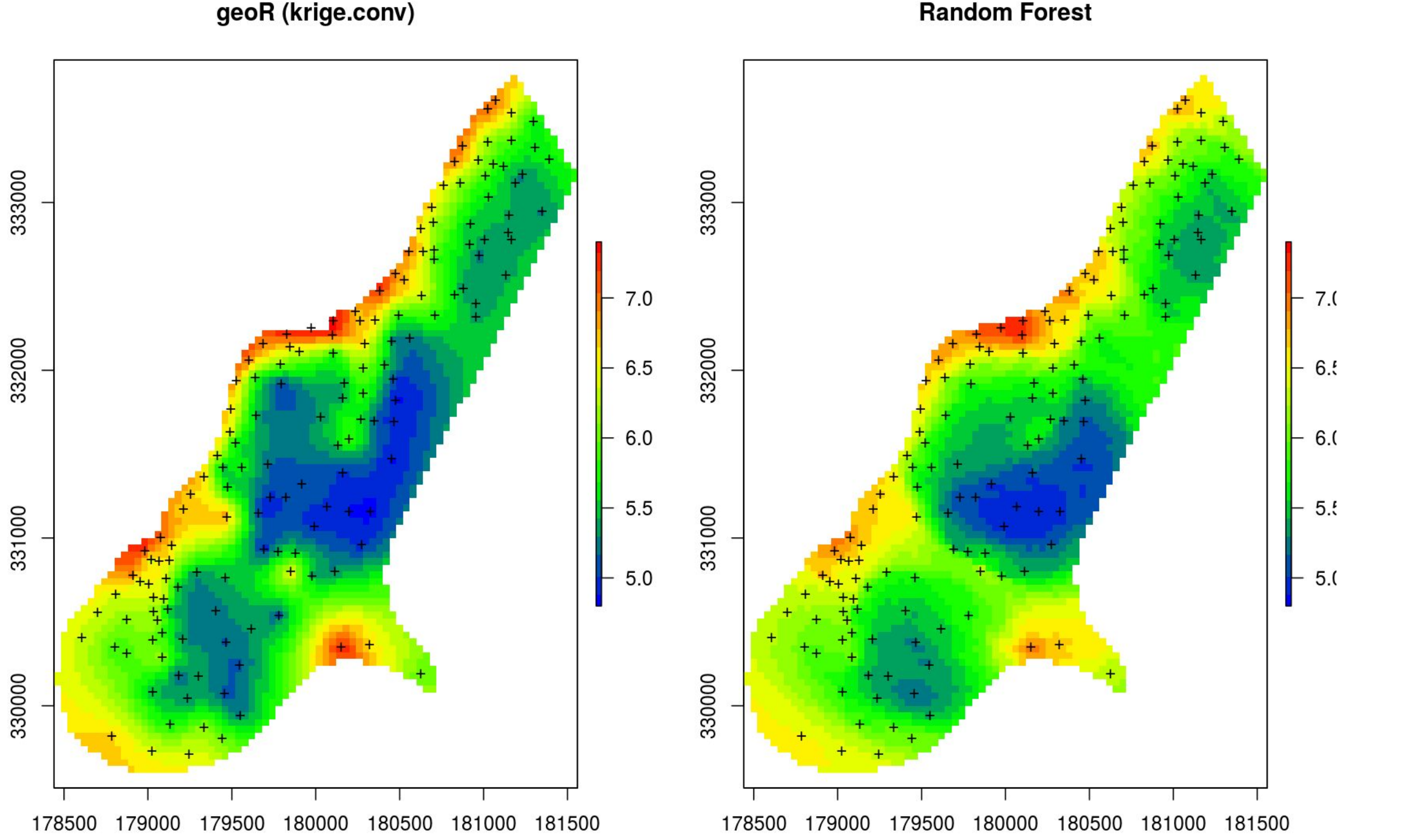
Ordinary Kriging vs RK (Year: 2013)

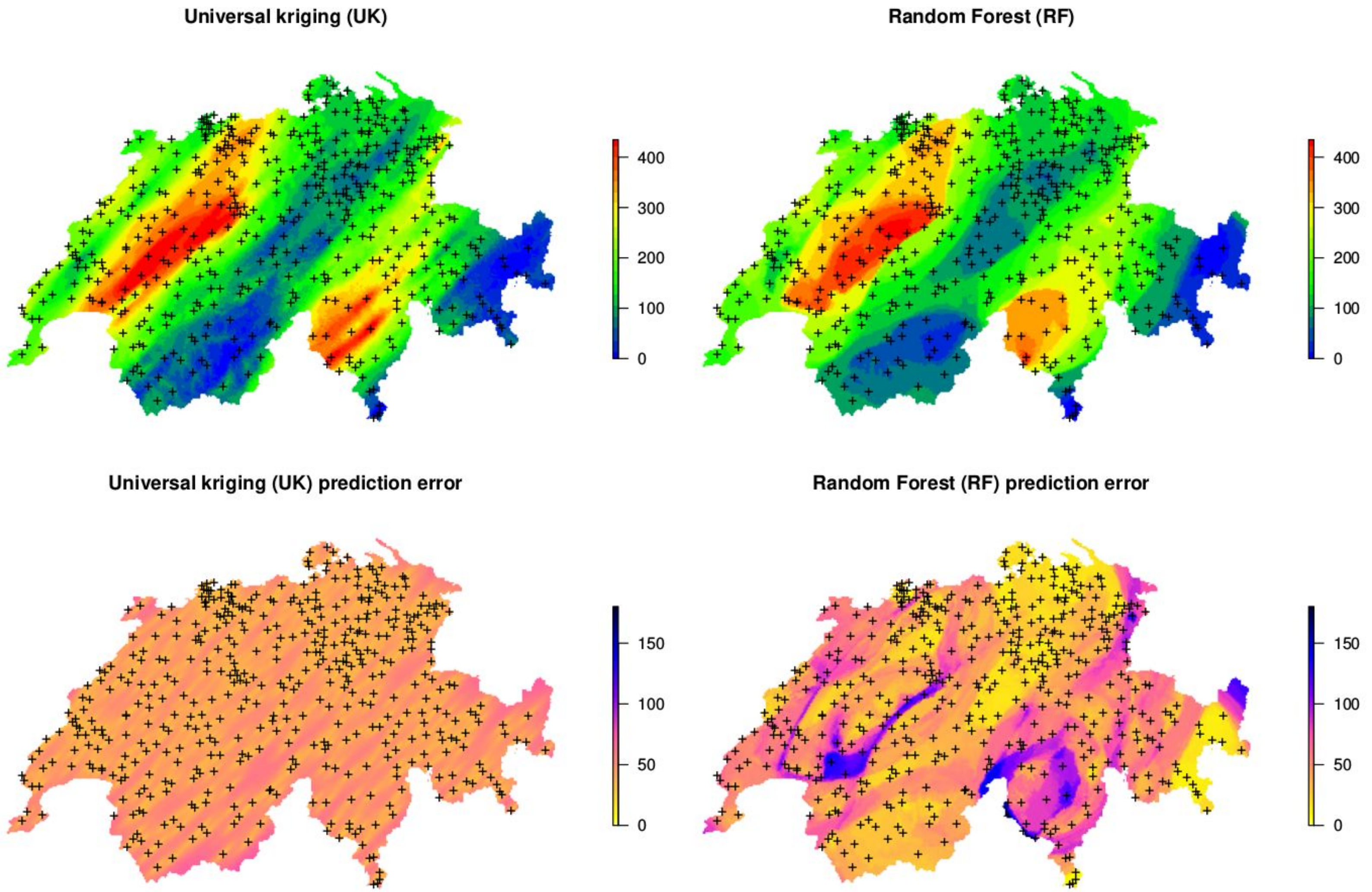
Imagine you can produce almost identical map without using / understanding any geostatistics

Correlate values of y 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))
ov.zinc <- over(meuse["zinc"], grid.dist0)
m.zinc <- ranger(fm0, cbind(meuse@data["zinc"],
ov.zinc))
zinc.rfd <- predict(m.zinc, grid.dist0@data)
```

Meuse data set




Universal kriging (UK)
Random Forest (RF)
400
300
200
100
0
Universal kriging (UK) prediction error
Random Forest (RF) prediction error
150
100
50
0

cience

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.

README.md

RFsp — Random Forest for spatial data (R tutorial)

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

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

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

4,263 views | Jan 25, 2019, 03:56am

DeepMind AI Beats Professional Human StarCraft II Players

Sam Shead Contributor
AI & Big Data
I cover artificial intelligence and Google DeepMind.

StarCraft II gameplay DEEPMIND

DeepMind has achieved yet another milestone in the gaming world.

The Google-owned artificial intelligence lab announced on Thursday that its new "AlphaStar" AI had beaten two of the world's best StarCraft II players.

The pros that AlphaStar beat are Dario Wunsch and Grzego rz Komincz — they're ranked 44th and 13th in the world respectively. DeepMind's AlphaStar beat them both 5-0 in two separate best-of-five matches.

The victories are being hailed as a major breakthrough by academics and AI industry watchers. While there have been several successes in video games such as Atari, Mario, Quake III Arena Capture the Flag,

Outline

Article | Published: 18 October 2017

Mastering the game of Go without human knowledge

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel & Demis Hassabis

Nature **550**, 354–359 (19 October 2017) | Download Citation

Editorial Summary

AlphaGo Zero goes solo

To beat world champions at the game of Go, the computer program AlphaGo has relied largely on supervised learning from millions of...
show more

Abstract

A long-standing goal of artificial intelligence is an algorithm that learns, *tabula rasa*, superhuman proficiency in challenging domains. Recently, AlphaGo became the first program to defeat a world champion in the game of Go. The tree search in AlphaGo evaluated positions and selected moves using deep neural networks. These neural networks were trained by supervised learning from human expert moves, and by reinforcement learning from self-play. Here we introduce an algorithm based solely on reinforcement learning, without human data, guidance or domain knowledge beyond game rules. AlphaGo becomes its own teacher: a neural network is trained to predict AlphaGo's own move selections and also the winner of AlphaGo's games. This neural network improves the strength of the tree search, resulting in higher quality move selection and

Associated Content

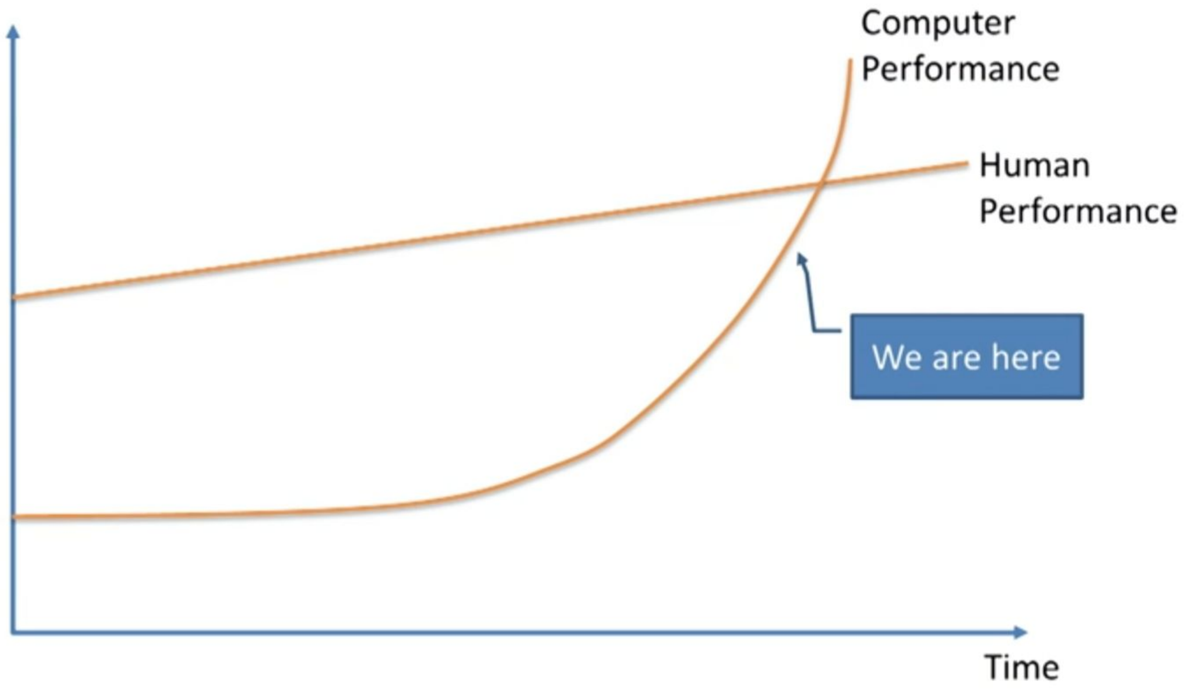
Nature | News & Views

Artificial intelligence: Learning to play Go from scratch

Satinder Singh, Andy Okun & Andrew Jackson

Sections | Figures | References

Independent Deepmind not that far away?



It does not go so well with ML always

Why ML does not always work for predictive mapping?

Top 5 problems

1. Extrapolation problems.
2. Bias in training data = for which you will be responsible!
3. Blunders / wrongly bind data can lead to serious artifacts = SPEND MUCH MORE TIME LOOKING CLOSELY AT YOUR TRAINING POINTS
4. Blindly applying ML to geographical (spatial) data problems leads to problems (Spatial CV) = WITH ML OVERFITTING HAPPENS TO EVERYONE!
5. Estimating prediction intervals can also get biased!

Outline

This is how AI bias really happens—and why it's so hard to fix

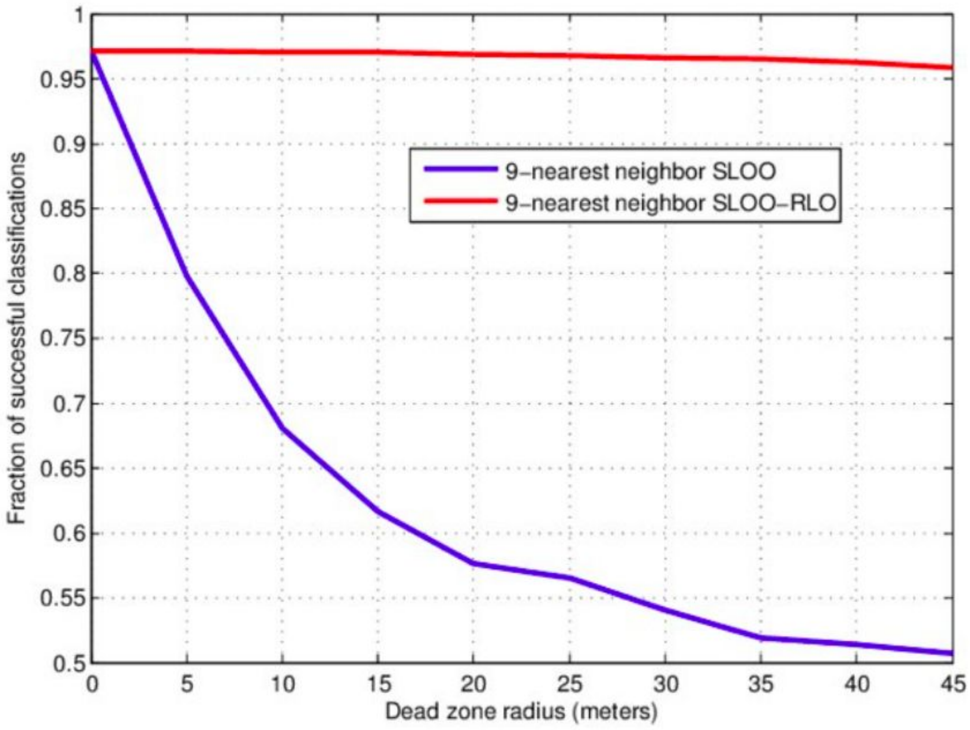
Bias can creep in at many stages of the deep-learning process, and the standard practices in computer science aren't designed to detect it.

by **Karen Hao** Feb 4, 2019

Over the past few months, we've documented how the vast majority of AI's applications today are based on the category of algorithms known as deep learning, and how deep-learning algorithms find patterns in data. We've also covered how these technologies affect people's lives: how they can perpetuate injustice in hiring, retail, and security and may already be doing so in the criminal legal system.

But it's not enough just to know that this bias exists. If we want to be able to fix it, we need to understand the mechanics of how it arises in the first place.

Cross-validation using blocking by distance



Figure

Caption

Figure 12.: The SLOO and SLOO-RLO results in the Pieksämäki analysis. The y-axis corresponds to the fraction of successful classifications and x-axis to the length of dead zone radius r δ .

RF doing extrapolation

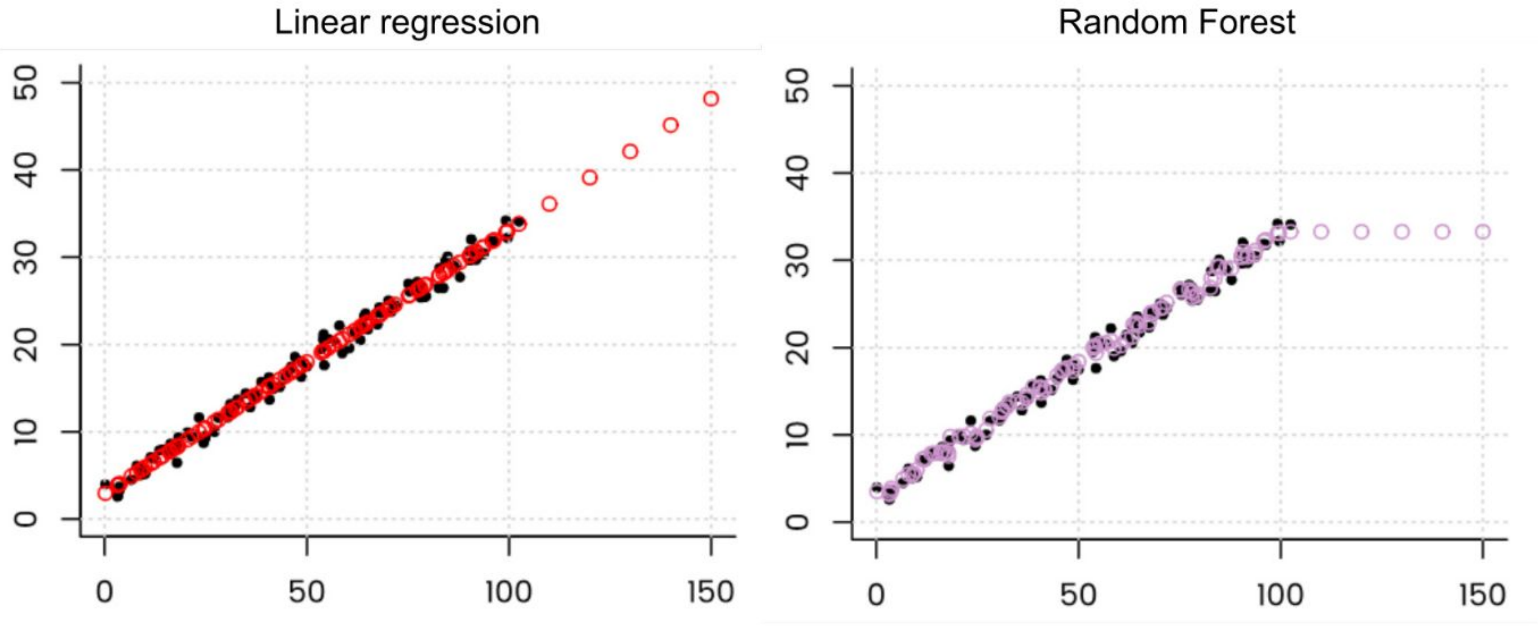
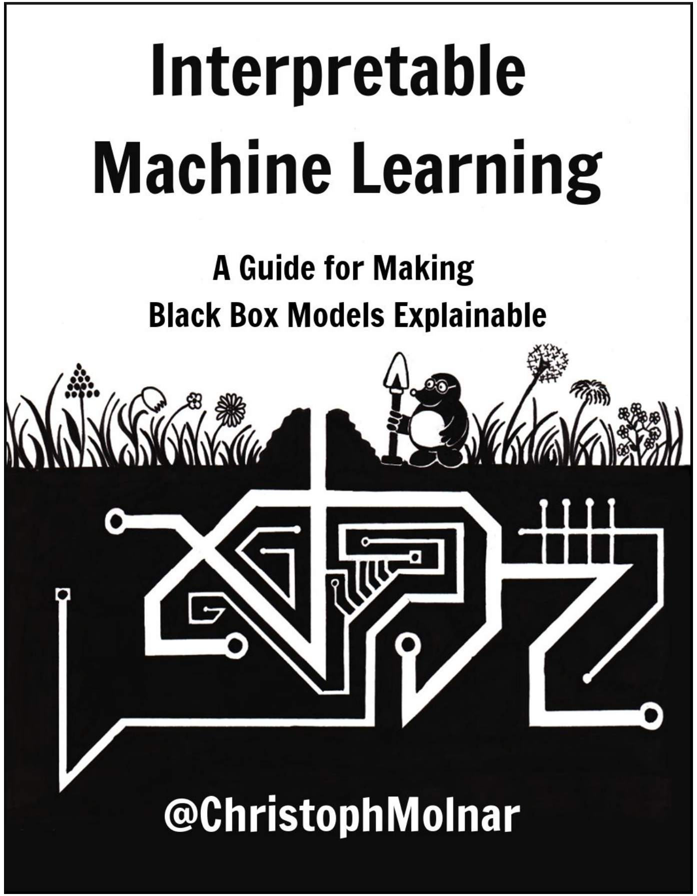
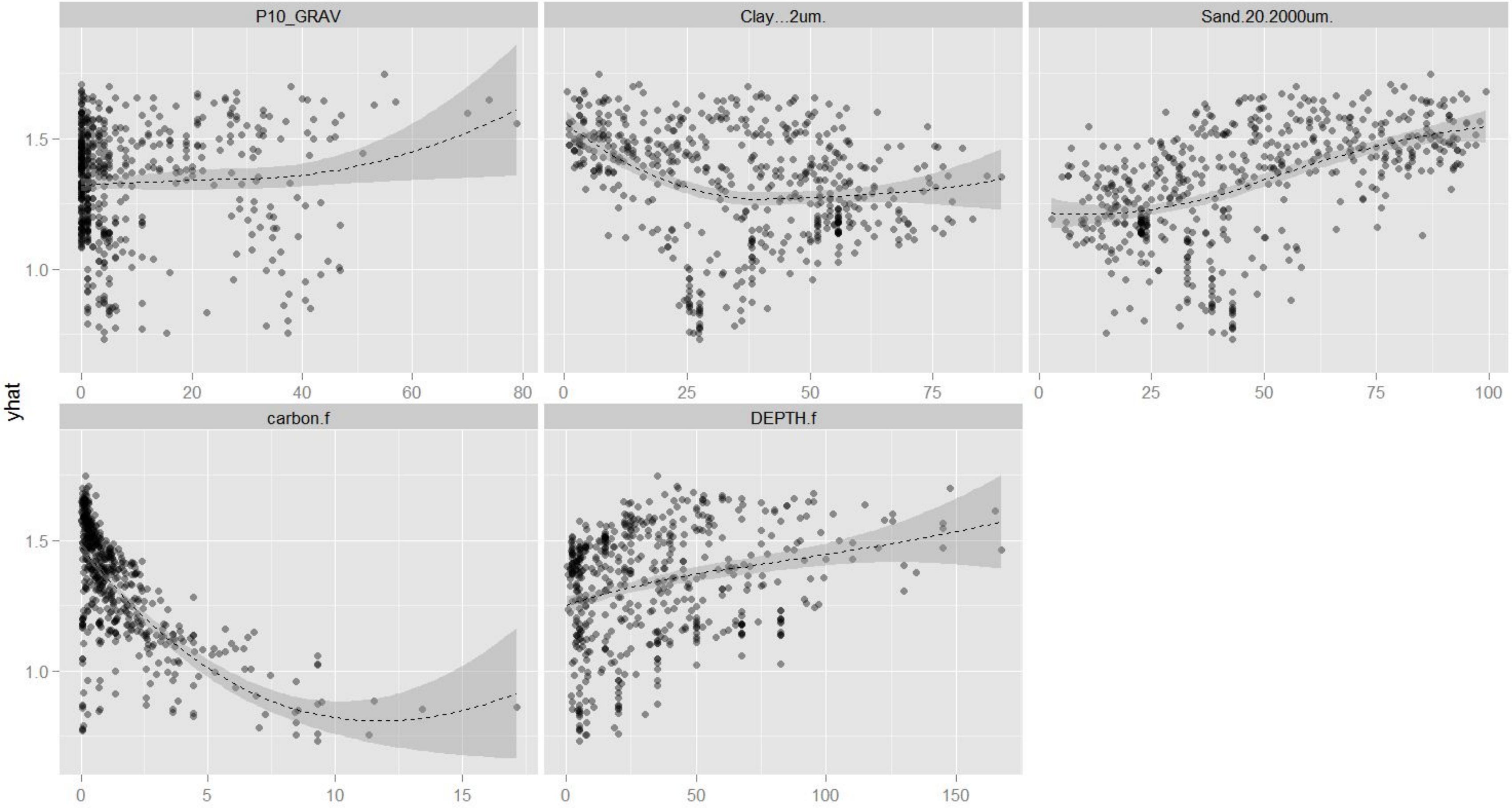


Figure 14. Illustration of the extrapolation problem of Random Forest based on the code examples from Peter Ellis (`http://freerangestats.info`). Even though Random Forest is more generic than linear regression and can be used also to fit complex non-linear problems, it can lead to completely nonsensical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

ML is NOT a black box system!





[RandomForestSRC package]

Summary

- Machine learning is rapidly developing and it is really amazing that computers can be “trained” to solve complex problems even beat humans in complex games;
- Tree-based ML’s seem to be suitable for spatial interpolation: random forest on geographical distances can make interpolations that are similar to using state-of-the-art model-based geostatistics;
- But ML if used blindly and without understanding the processes and potential extrapolation and bias problems can produce even worse results than some simple statistical method from 1960’s.