Imputing missing values in satellite data: From parametric to non-parametric approaches
Joint work

- Florian Gerber
- Emilio Porcu
- Francois Bachoc

and with contributions of several others
Outlook

- Motivation
- Parametric models and their issues
- A particular non-parametric approach
- Outcome of a biodiversity exercise

 ~> Questions and «Fast-Forward» appreciated!

Slides at http://www.math.uzh.ch/furrer/slides/
Global Change and Biodiversity

- University Research Priority Program, start in 2013
- ≈ 50 members in 2 Faculties and 5 Institutes/Departments
Global Change and Biodiversity

Past view

Environment

Biodiversity

Present view

Global Change
Species introduction, Land use, Biochemical cycles, Climate

Humans
Well-being, Welfare, Security, Freedom

Biodiversity
Genetic diversity, Diversity of species, Ecosystem and Landscape diversity

Ecosystems
Biological resources Cultural values Regulating services

www.gcb.uzh.ch
Linking MODIS with plot based data

Biodiversity promotes primary productivity and growing season lengthening at the landscape scale

Jacqueline Oehri, Bernhard Schmid, Gabriela Schaepman-Strub, and Pascal A. Niklaus

Department of Evolutionary Biology and Environmental Studies, University of Zurich, CH-8057 Zurich, Switzerland

Edited by Shahid Naeem, Columbia University, New York, NY, and accepted by Editorial Board Member Ruth S. DeFries August 4, 2017 (received for review March 13, 2017)

Experiments have shown positive biodiversity-ecosystem function (11). Organisms, in most cases plants (7). These settings markedly

“...we show that primary productivity, its temporal stability, and the decadal trend of a prolonged growing season strongly increase with biodiversity across heterogeneous landscapes, which is consistent over vast environmental, climatic, and altitudinal gradients. ...”
Linking MODIS with plot based data

Now analysis in the Arctic, not Switzerland

- Species abundance plot scale measurements from the International Tundra Experiment (ITEX)
- NDVI satellite images and ASTER elevation data

Source: F. Gerber
Arctic NDVI data

MODIS NDVI data (satellite product MOD13A1, \( \text{NDVI} = \frac{\text{NIR} - R}{\text{NIR} + R} \) )
Visual example

Day of the year

Source: Gerber et al (2018), TGRS
Visual example

Day of the year

Source: Gerber et al (2018), TGRS
Spatial statistics: prediction

Observations: \( y(s_1), \ldots, y(s_n) \)

Model:

\[ Y(s) = \text{signal} + \text{noise} \]

\[ Y(s) = \text{trend} + \text{stochastic part} + \text{noise} \]

\[ Y(s) = x^T(s)\beta + \alpha(s) + Z(s) + \varepsilon(s) \]
Spatial statistics: prediction

Predict the quantity of interest at an arbitrary location.

Why?

▶ Fill-in missing data
▶ Force data onto a regular grid
▶ Smooth out the measurement error

How?

▶ By eye
▶ Linear interpolation
▶ The correct way . . .
Spatial statistics: prediction

Describing the covariance structure

Covariance matrix $\Sigma$ contains elements $C(\text{dist}(s_i, s_j))$. 
Spatial statistics: prediction

Predict $Z(s_0)$ given $y(s_1), \ldots, y(s_n)$.

Minimize mean squared prediction error (over all linear unbiased predictors)

$\implies$ Best Linear Unbiased Predictor:

\[
\text{BLUP} = \text{Cov}[Z(s_{\text{predict}}), Y(s_{\text{obs}})] \ Var[Y(s_{\text{obs}})]^{-1} \text{obs}
\]

\[
\hat{Z}(s_0) = c^T \Sigma^{-1} y
\]

(one spatial process, no trend, known covariance structure; otherwise almost the same)
Outlook

- Motivation
- Parametric models and their issues
- A particular non-parametric approach
- Outcome of a biodiversity exercise
Issues of basic, classical kriging

\[ \text{Cov}(\text{pred}, \text{obs}) \cdot \text{Var}(\text{obs})^{-1} \cdot \text{obs} = c \Sigma^{-1} y \]

- “Simple” spatial interpolation . . .

  . . . on paper or in class!

- BUT:
  1. Complex mean structure
  2. Unknown parameters
  3. Large spatial fields
  4. Non-stationary covariances
  5. Space-time data on the sphere
Issues of basic, classical kriging

1. Complex mean structure
2. Unknown parameters
3. Large spatial fields
4. Non-stationary covariances
5. Space-time data on the sphere

- Parametric structure typically ok
- Non-parametric structure often creates “model clash”
Issues of basic, classical kriging

1. Complex mean structure
2. Unknown parameters
3. Large spatial fields
4. Non-stationary covariances
5. Space-time data on the sphere

▶ (method of moment estimation)
▶ Likelihood approaches

⇝ Cholesky factorizations
Issues of basic, classical kriging

1. Complex mean structure
2. Unknown parameters
3. Large spatial fields
4. Non-stationary covariances
5. Space-time data on the sphere

- Many R packages do perform kriging . . .
  . . . many black boxes . . .
  . . . to tailored situations


Computational limits are quickly attained!
Methods for large spatial datasets

▶ Sparse Covariance methods:
  — Covariance Tapering
  — Spatial Partitioning

▶ Sparse Precision methods:
  — Lattice Kriging
  — Multiresolution Approximations
  — Stochastic Partial Differential Equations
  — Periodic Embedding
  — Nearest Neighbor Processes

▶ Low rank approximation:
  — Fixed Rank Kriging
  — Predictive Processes

▶ Algorithmic approaches:
  — Gapfill
  — Local Approximate Gaussian Processes
  — Metakriging
Spatial modeling

Geostatistical model (GRF):

- Covariance matrix: $\Sigma$
- Lattice model (GMRF):
  - $\mathbb{E}(Z_i|z_{-i}) = \beta \sum_{j \text{ neighbor of } i} z_j$
  - $\text{Var}(Z_i|z_{-i}) = \tau^2$
- Gaussianity and regularity conditions:
  - $\Sigma = \tau^2(I - B)^{-1}$
Spatial modeling

Geostatistical model (GRF):

\[ \Sigma \]

\[ \Sigma_{\text{app}} \]

Lattice model (GMRF):

\[ \Sigma^{-1} \]

\[ \Sigma \]
Sparseness

Using sparse covariance functions for greater computational efficiency.

Sparseness is guaranteed when

- the covariance function has a compact support
- a compact support is (artificially) imposed \( \leadsto \) tapering
Spareness: prediction/estimation

▶ Univariate setting:

Proofs based on infill asymptotics and “misspecified” covariances
Conditions on the tail behaviour of the spectrum of the (tapered) covariance

Furrer, Genton, Nychka (2006) JCGS
Kaufman, Schervish, Nychka (2008) JMVA
Stein (2013) JCGS
Bevilacqua et al (2018?) AoS

▶ Multivariate setting:

Proofs based on domain increasing framework
Weak conditions on the taper

Furrer, Du, Bachoc (2016) JMVA
Software to exploit the sparse structure **spam64** for R:

- an R package for sparse matrix algebra
- storage economical and fast
- versatile, intuitive and simple

See Furrer et al. (2006) JCGS; Furrer, Sain (2010) JSS

- R objects have at most $2^{31}$ elements (almost)
- R does not ‘have’ 64-bit integers: stored as doubles
- 64-bit exploitation consists of type conversions between front-end R and pre-compiled code

Gerber, Mösinger, Furrer (2017) CaGeo
Arctic NDVI data

MODIS NDIV data (satellite product MOD13A1, $\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}}$)
Kriging is smoothing
Interpolation using gapfill

(a) Flow diagram of the gap filling method

- Observed data
- Position of the missing value

Extract subset

Update subset parameters

Subset component

(C1) and (C2) fulfilled?

no

yes

Rank images

Estimate quantile

Quantile regression

Fill value

(b) Example MODIS NDVI data

Day of the year

145 161 177 193

2004 2005 2006 2007
Interpolation using gapfill

(b) Example MODIS NDVI data

Day of the year

NDVI

0.2
0.4
0.6
0.8

145 161 177 193

2004 2005 2006 2007
**gapfill**: ranking of the images

![Graph showing NDVI values over time](image)

- **Method by example**
- Sort images by pixel-wise comparisons
- **NDVI**
- **Day of the year**
- **Year**
- **Low**
- **High**
**gapfill**: quantile regression
gapfill: prediction uncertainties

data and predictions

uncertainties
gapfill: location
### gapfill: comparison

<table>
<thead>
<tr>
<th></th>
<th>gapfill</th>
<th></th>
<th></th>
<th>Gapfill-Python</th>
<th></th>
<th>TIMESAT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#filled</td>
<td>RMSE</td>
<td>RMSE$_p$</td>
<td>RMSE$_T$</td>
<td>#filled</td>
<td>RMSE</td>
<td>#filled</td>
<td>RMSE</td>
</tr>
<tr>
<td>20%</td>
<td>92'822 (100%)</td>
<td>41.80</td>
<td>42.06</td>
<td>41.10</td>
<td>90’307 (97%)</td>
<td>45.00</td>
<td>59’948</td>
<td>83.43</td>
</tr>
<tr>
<td>30%</td>
<td>147’827 (100%)</td>
<td>42.54</td>
<td>42.39</td>
<td>37.09</td>
<td>146’686 (99%)</td>
<td>45.54</td>
<td>42’892</td>
<td>71.43</td>
</tr>
<tr>
<td>40%</td>
<td>192’456 (100%)</td>
<td>41.34</td>
<td>40.98</td>
<td>36.41</td>
<td>169’998 (88%)</td>
<td>42.49</td>
<td>31’279</td>
<td>71.93</td>
</tr>
<tr>
<td>50%</td>
<td>240’326 (100%)</td>
<td>59.58</td>
<td>44.94</td>
<td>37.24</td>
<td>134’540 (56%)</td>
<td>45.61</td>
<td>14’127</td>
<td>86.09</td>
</tr>
</tbody>
</table>

RMSE $\times 10^3$
(l) Uncertainty contribution from the indicated four steps of the gapfill procedure.

(m) Average width of the 90% prediction intervals (40% missing values).

(r) Average interval widths and coverage rate per day of the year.
## Summary

<table>
<thead>
<tr>
<th>Implementation:</th>
<th>spam64</th>
<th>gapfill</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intuition:</td>
<td>statistical</td>
<td>conceptual</td>
</tr>
<tr>
<td>Model:</td>
<td>frequentist based</td>
<td>distribution free</td>
</tr>
<tr>
<td>Uncertainties:</td>
<td>formal</td>
<td>resampling type</td>
</tr>
<tr>
<td>Practicality:</td>
<td>play ground</td>
<td>competitive</td>
</tr>
</tbody>
</table>
Outlook

- Motivation
- Parametric models and their issues
- A particular non-parametric approach
- Outcome of a biodiversity exercise
Biodiversity hypotheses

H1: Plant productivity (quantified through NDVI) is positively correlated with plot scale biodiversity

H2: Landscape variability (quantified through NDVI and slope) is positively correlated with plot scale biodiversity

H3: Slope induces a drainage effect and increases plot scale biodiversity
Data

- Species abundance plot scale measurements from the International Tundra Experiment (ITEX)
  - Shannon biodiversity index on site and plot scale
- Landsat NDVI satellite images and ASTER elevation data
  - Characterization of the landscape heterogeneity

Source: F. Gerber
## Data

- **Species abundance plot scale measurements from the International Tundra Experiment (ITEX)**

<table>
<thead>
<tr>
<th></th>
<th>abisko</th>
<th>alexfiord</th>
<th>anwr</th>
<th>atqasuk</th>
<th>barrow</th>
<th>bylot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td><img src="abisko_1980.png" alt="Graph" /></td>
<td><img src="alexfiord_1980.png" alt="Graph" /></td>
<td><img src="anwr_1980.png" alt="Graph" /></td>
<td><img src="atqasuk_1980.png" alt="Graph" /></td>
<td><img src="barrow_1980.png" alt="Graph" /></td>
<td><img src="bylot_1980.png" alt="Graph" /></td>
</tr>
<tr>
<td>1985</td>
<td><img src="abisko_1985.png" alt="Graph" /></td>
<td><img src="alexfiord_1985.png" alt="Graph" /></td>
<td><img src="anwr_1985.png" alt="Graph" /></td>
<td><img src="atqasuk_1985.png" alt="Graph" /></td>
<td><img src="barrow_1985.png" alt="Graph" /></td>
<td><img src="bylot_1985.png" alt="Graph" /></td>
</tr>
<tr>
<td>1990</td>
<td><img src="abisko_1990.png" alt="Graph" /></td>
<td><img src="alexfiord_1990.png" alt="Graph" /></td>
<td><img src="anwr_1990.png" alt="Graph" /></td>
<td><img src="atqasuk_1990.png" alt="Graph" /></td>
<td><img src="barrow_1990.png" alt="Graph" /></td>
<td><img src="bylot_1990.png" alt="Graph" /></td>
</tr>
<tr>
<td>1995</td>
<td><img src="abisko_1995.png" alt="Graph" /></td>
<td><img src="alexfiord_1995.png" alt="Graph" /></td>
<td><img src="anwr_1995.png" alt="Graph" /></td>
<td><img src="atqasuk_1995.png" alt="Graph" /></td>
<td><img src="barrow_1995.png" alt="Graph" /></td>
<td><img src="bylot_1995.png" alt="Graph" /></td>
</tr>
<tr>
<td>2000</td>
<td><img src="abisko_2000.png" alt="Graph" /></td>
<td><img src="alexfiord_2000.png" alt="Graph" /></td>
<td><img src="anwr_2000.png" alt="Graph" /></td>
<td><img src="atqasuk_2000.png" alt="Graph" /></td>
<td><img src="barrow_2000.png" alt="Graph" /></td>
<td><img src="bylot_2000.png" alt="Graph" /></td>
</tr>
<tr>
<td>2005</td>
<td><img src="abisko_2005.png" alt="Graph" /></td>
<td><img src="alexfiord_2005.png" alt="Graph" /></td>
<td><img src="anwr_2005.png" alt="Graph" /></td>
<td><img src="atqasuk_2005.png" alt="Graph" /></td>
<td><img src="barrow_2005.png" alt="Graph" /></td>
<td><img src="bylot_2005.png" alt="Graph" /></td>
</tr>
<tr>
<td>2010</td>
<td><img src="abisko_2010.png" alt="Graph" /></td>
<td><img src="alexfiord_2010.png" alt="Graph" /></td>
<td><img src="anwr_2010.png" alt="Graph" /></td>
<td><img src="atqasuk_2010.png" alt="Graph" /></td>
<td><img src="barrow_2010.png" alt="Graph" /></td>
<td><img src="bylot_2010.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

### % Cloud cover

<table>
<thead>
<tr>
<th>Month</th>
<th>0</th>
<th>25</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td><img src="January.png" alt="Graph" /></td>
<td><img src="January_25.png" alt="Graph" /></td>
<td><img src="January_50.png" alt="Graph" /></td>
<td><img src="January_75.png" alt="Graph" /></td>
<td><img src="January_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>February</td>
<td><img src="February.png" alt="Graph" /></td>
<td><img src="February_25.png" alt="Graph" /></td>
<td><img src="February_50.png" alt="Graph" /></td>
<td><img src="February_75.png" alt="Graph" /></td>
<td><img src="February_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>March</td>
<td><img src="March.png" alt="Graph" /></td>
<td><img src="March_25.png" alt="Graph" /></td>
<td><img src="March_50.png" alt="Graph" /></td>
<td><img src="March_75.png" alt="Graph" /></td>
<td><img src="March_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>April</td>
<td><img src="April.png" alt="Graph" /></td>
<td><img src="April_25.png" alt="Graph" /></td>
<td><img src="April_50.png" alt="Graph" /></td>
<td><img src="April_75.png" alt="Graph" /></td>
<td><img src="April_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>May</td>
<td><img src="May.png" alt="Graph" /></td>
<td><img src="May_25.png" alt="Graph" /></td>
<td><img src="May_50.png" alt="Graph" /></td>
<td><img src="May_75.png" alt="Graph" /></td>
<td><img src="May_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>June</td>
<td><img src="June.png" alt="Graph" /></td>
<td><img src="June_25.png" alt="Graph" /></td>
<td><img src="June_50.png" alt="Graph" /></td>
<td><img src="June_75.png" alt="Graph" /></td>
<td><img src="June_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>July</td>
<td><img src="July.png" alt="Graph" /></td>
<td><img src="July_25.png" alt="Graph" /></td>
<td><img src="July_50.png" alt="Graph" /></td>
<td><img src="July_75.png" alt="Graph" /></td>
<td><img src="July_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>August</td>
<td><img src="August.png" alt="Graph" /></td>
<td><img src="August_25.png" alt="Graph" /></td>
<td><img src="August_50.png" alt="Graph" /></td>
<td><img src="August_75.png" alt="Graph" /></td>
<td><img src="August_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>September</td>
<td><img src="September.png" alt="Graph" /></td>
<td><img src="September_25.png" alt="Graph" /></td>
<td><img src="September_50.png" alt="Graph" /></td>
<td><img src="September_75.png" alt="Graph" /></td>
<td><img src="September_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>October</td>
<td><img src="October.png" alt="Graph" /></td>
<td><img src="October_25.png" alt="Graph" /></td>
<td><img src="October_50.png" alt="Graph" /></td>
<td><img src="October_75.png" alt="Graph" /></td>
<td><img src="October_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>November</td>
<td><img src="November.png" alt="Graph" /></td>
<td><img src="November_25.png" alt="Graph" /></td>
<td><img src="November_50.png" alt="Graph" /></td>
<td><img src="November_75.png" alt="Graph" /></td>
<td><img src="November_100.png" alt="Graph" /></td>
</tr>
<tr>
<td>December</td>
<td><img src="December.png" alt="Graph" /></td>
<td><img src="December_25.png" alt="Graph" /></td>
<td><img src="December_50.png" alt="Graph" /></td>
<td><img src="December_75.png" alt="Graph" /></td>
<td><img src="December_100.png" alt="Graph" /></td>
</tr>
</tbody>
</table>

Source: F. Gerber
Results

- Data did not provide evidence for the hypothesis H1–H3.
- Statistical power could be improved by adding additional plot data.
- Limited amount of Landsat images makes it difficult to measure their seasonal and annual variability. This confounds the temporal aggregation.
Collaboration with:
- Florian Gerber
- Gabriela Schaepman-Strub
- Rogier de Jong
- Emilio Porcu
- Francois Bachoc
- Alfredo Alegria
- Kaspar Mösinger
- former & present ‘Applied Statistics’ team
  ... and many more

143282, 144973, 175529
Monte Verità Conference 2019

Global Change and Biodiversity: Integrating the impact of earth and world drivers across scales

30 June - 4 July 2019, Monte Verità

Source: www.gcb.uzh.ch/
References (some, alphabetical)


Furrer et al (2017) *spam64: 64-Bit Extension of the SPArse Matrix R Package 'spam'*. R package version 2.2-0

Gerber Moesinger Furrer (2017) Extending R Packages to Support 64-bit Compiled Code: An Illustration with spam64 and GIMMS NDVI3g Data *Comput Geosci* 104 109-119


Gerber Moesinger Furrer (2017) dotCall64: An Efficient Interface to Compiled C/C++ and Fortran Code Supporting Long Vectors *SoftwareX* 7 217-221


Complete list at: [www.math.uzh.ch/furrer/research/publications.shtml](http://www.math.uzh.ch/furrer/research/publications.shtml)