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

©Reinhard Furrer, I-Math/ICS, UZH

UPNA, Pamplona, 2018/09/27
Joint work

- Florian Gerber
- Emilio Porcu
- Francois Bachoc

and with contributions of several others
Outlook

- Gentle introduction to *Spatial Statistics*
- Parametric models and their issues
- Non-parametric approach, a particular example
- Outcome of a biodiversity exercise

〜 Questions and «Fast-Forward» appreciated!
Spatial statistics: prediction

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

First law of geography (Waldo Tobler):

Everything is related to everything else, but near things are more related than distant things.
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)

\[ \text{Best Linear Unbiased Predictor:} \]

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

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

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

Day of the year

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

Day of the year

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

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

Source: Gerber et al (2018), TGRS
Outlook

- Gentle introduction to *Spatial Statistics*
- Parametric models and their issues
- Non-parametric approach, a particular example
- Outcome of a biodiversity exercise
Issues of basic, classical kriging

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

- “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

See Heaton et al. arXiv:1710.05013 stat.ME.

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

Furrer
Heaton

Nychka
Katzfuss
Lindgren
Guinness
Datta

Zammit-Mangion
Finley

Gerber
Gramacy
Guhanifyogi
Spatial modeling

Geostatistical model (GRF):

Lattice model (GMRF):

E(Z_i | z_{-i}) = \beta \sum_{j \text{ neighbor of } i} z_j

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):

$\Sigma^{-1}$

$\Sigma$
Tapering: 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

![Covariance vs Distance, lag h graph]

- Green line: Matern $\nu = 1.5$
- Blue line: Wendland
- Red line: Matern $\times$ Wendland
Tapering: 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

Tapering obsolete, work directly with misspecified covariance
Bevilacqua et al (2018?) AoS

▶ Multivariate setting:

Proofs based on domain increasing framework
Weak conditions on the taper

Furrer, Du, Bachoc (2016) JMVA
Software

Software to exploit the sparse structure \texttt{spam64} for R:

- an R package for \textit{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
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

- Spectral representations/stochastic expansions
- Scale mixture representations
- Lagrangian framework

Porcu, Alegria, Furrer (2018) ISR
Outlook

- Gentle introduction to *Spatial Statistics*
- Parametric models and their issues
- Non-parametric approach, a particular example
- Outcome of a biodiversity exercise
Arctic NDVI data

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

<table>
<thead>
<tr>
<th>Year</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>161</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>177</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>193</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NDVI: 0 0.2 0.4 0.6 0.8 NA
Interpolation using gapfill

(b) Example MODIS NDVI data
**gapfill:** ranking of the images

Day of the year

<table>
<thead>
<tr>
<th>161</th>
<th>177</th>
<th>193</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Image 2004" /></td>
<td><img src="image" alt="Image 2005" /></td>
<td><img src="image" alt="Image 2006" /></td>
</tr>
<tr>
<td><img src="image" alt="Image 2007" /></td>
<td><img src="image" alt="Image 2008" /></td>
<td><img src="image" alt="Image 2009" /></td>
</tr>
</tbody>
</table>

NDVI

- Low
- High

$\begin{align*}
r = 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\end{align*}$
**gapfill**: quantile regression

![Flow diagram of the gap filling method](image1)

(a) Flow diagram of the gap filling method
(b) Example MODIS NDVI data

<table>
<thead>
<tr>
<th>Date</th>
<th>Score</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>193 doy 2004</td>
<td>0.65</td>
<td>8</td>
</tr>
<tr>
<td>177 doy 2006</td>
<td>0.71</td>
<td>9</td>
</tr>
<tr>
<td>177 doy 2005</td>
<td>0.77</td>
<td>10</td>
</tr>
<tr>
<td>193 doy 2006</td>
<td>0.88</td>
<td>11</td>
</tr>
<tr>
<td>193 doy 2005</td>
<td>0.91</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>q:</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA</td>
</tr>
<tr>
<td>0.64</td>
</tr>
<tr>
<td>NA</td>
</tr>
<tr>
<td>0.12</td>
</tr>
<tr>
<td>0.77</td>
</tr>
</tbody>
</table>

![Images from the subset ranked 8–12](image2)

(c) Images from the subset ranked 8–12
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></th>
<th>TIMESAT</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>
</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 (65%)</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 (29%)</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 (16%)</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 (6%)</td>
</tr>
</tbody>
</table>

RMSE $\times 10^3$
**gapfill: uncertainties**

(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

Implementation: spam64       gapfill
Intuition: statistical        conceptual
Model: frequentist based     distribution free
Uncertainties: formal        resampling type
Practicality: play ground    competitive
Outlook

- Gentle introduction to *Spatial Statistics*
- Parametric models and their issues
- Non-parametric approach, a particular example
- Outcome of a biodiversity exercise
Application in GCB: Playground

Scientifically:

- Complex interplay between climate and vegetation/ecosystems
- Reflectance measurements as proxy for “greenness”
Why do we care?

Scientifically:

- Complex interplay between climate and vegetation/ecosystems
- Reflectance measurements as proxy for “greenness”

Statistically:

- Large, spatio-temporal datasets with complex structures at low resolution
- …fusion 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)

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.

Source: F. Gerber
Collaboration with:
- Florian Gerber
- Emilio Porcu
- Francois Bachoc
- Alfredo Alegria
- Kaspar Mösinger
- former & present ‘Applied Statistics’ team
  ... and many more

URPP Global Change and Biodiversity

143282, 144973, 175529
References (some, alphabetical)


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


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