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

@ReinhardFurrer, I-Math/ICS, UZH

UPNA, Pamplona, 2018/09/27

NZZ.ch

University of Zurich[UZH]

Joint work

- Florian Gerber
- Emilio Porcu
- Francois Bachoc

and with contributions of several others

University of Zurich[UZH]

Outlook

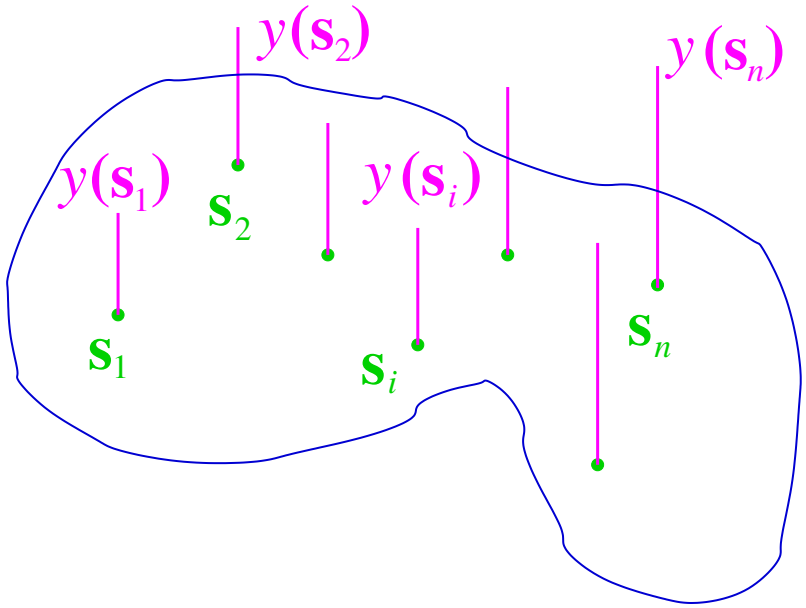
- Gentle introduction to *Spatial Statistics*
- Parametric models and their issues
- Non-parametric approach, a particular example
- Outcome of a biodiversity exercise

$\rightsquigarrow$ Questions and $\ll$Fast-Forward$\gg$ appreciated!

Spatial statistics: prediction

Observations: $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$

Source: wikipedia.org

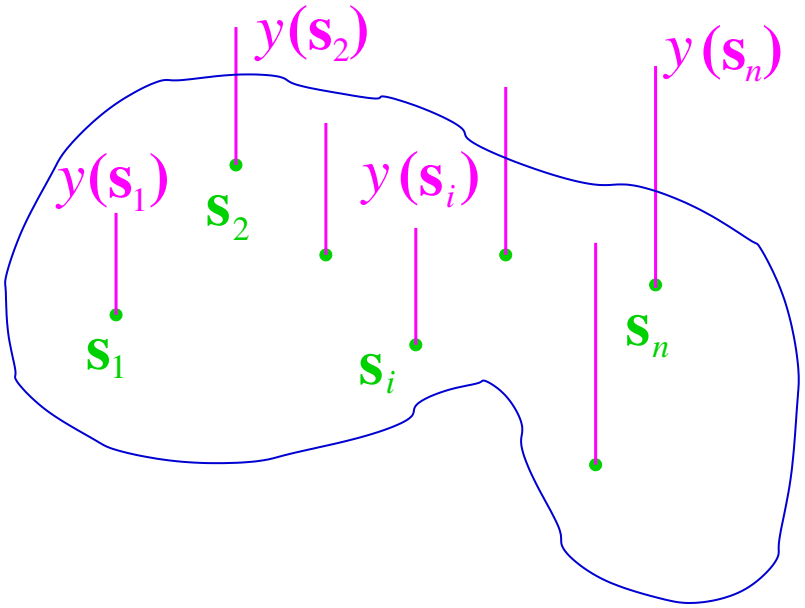


First law of geography (Waldo Tobler):

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

Spatial statistics: prediction

Observations: $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$



Model:

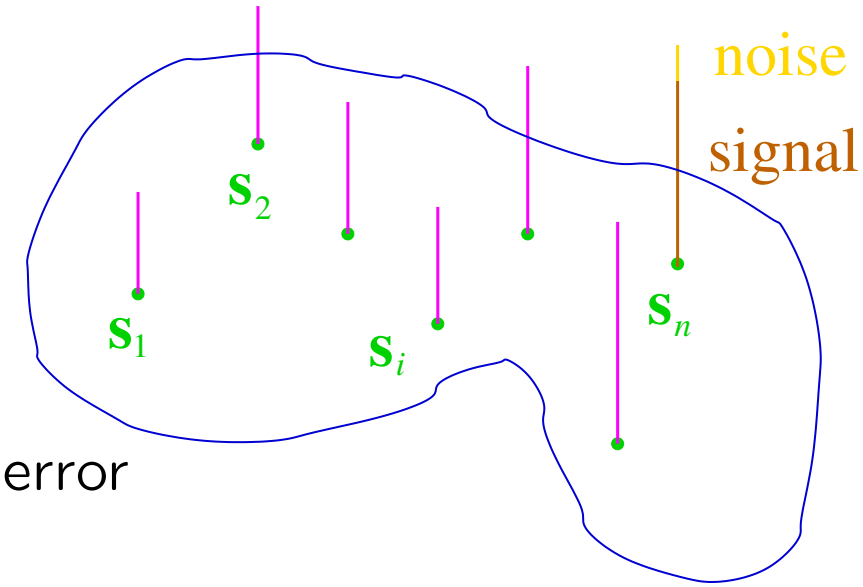
$$Y(\mathbf{s}) = \text{signal} + \text{noise}$$

$$Y(\mathbf{s}) = \text{trend} + \text{stochastic part} + \text{noise}$$

$$Y(\mathbf{s}) = \mathbf{x}^{\mathsf{T}}(\mathbf{s})\boldsymbol{\beta} + \alpha(\mathbf{s}) + Z(\mathbf{s}) + \varepsilon(\mathbf{s})$$

Spatial statistics: prediction

Predict the quantity of interest at an arbitrary location.



Why?

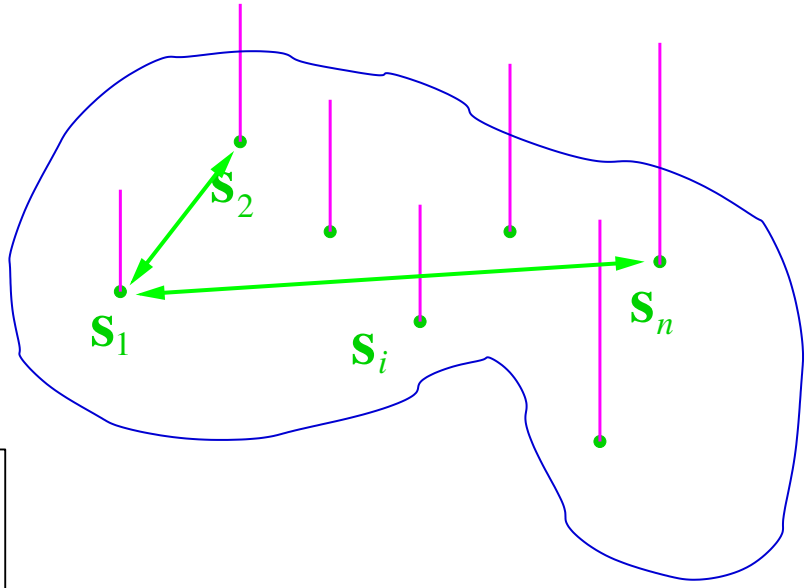
- Fill-in missing data
- Force data onto a regular grid
- Smooth out the measurement error

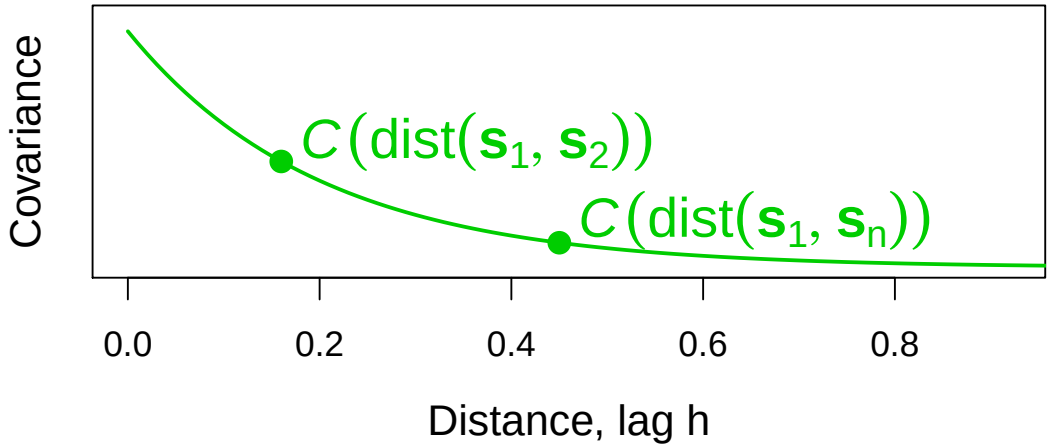
How?

- By eye
- Linear interpolation
- The correct way . . .

Spatial statistics: prediction

Describing the covariance structure

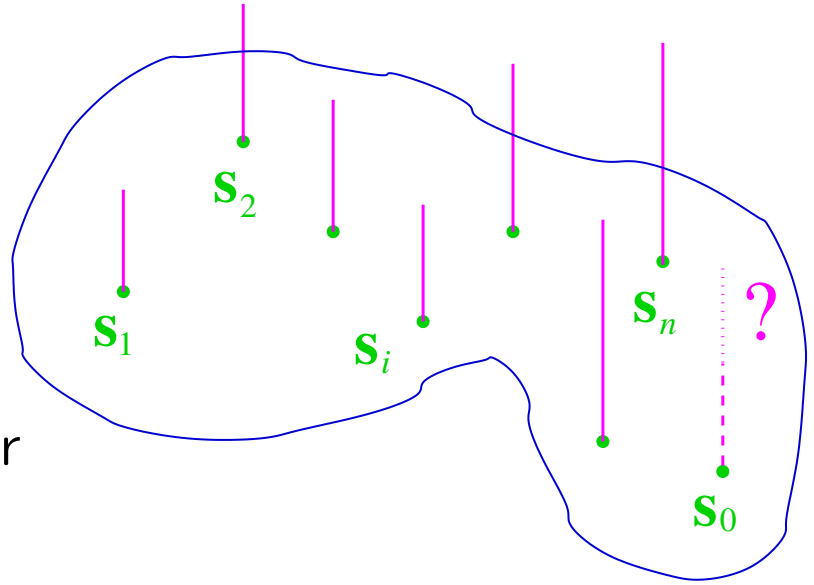




Covariance matrix Σ contains elements $C\Big(\text{dist}(\mathbf{s}_i, \mathbf{s}_j)\Big)$.

Spatial statistics: prediction

Predict $Z(\mathbf{s}_0)$ given $y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n)$.



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

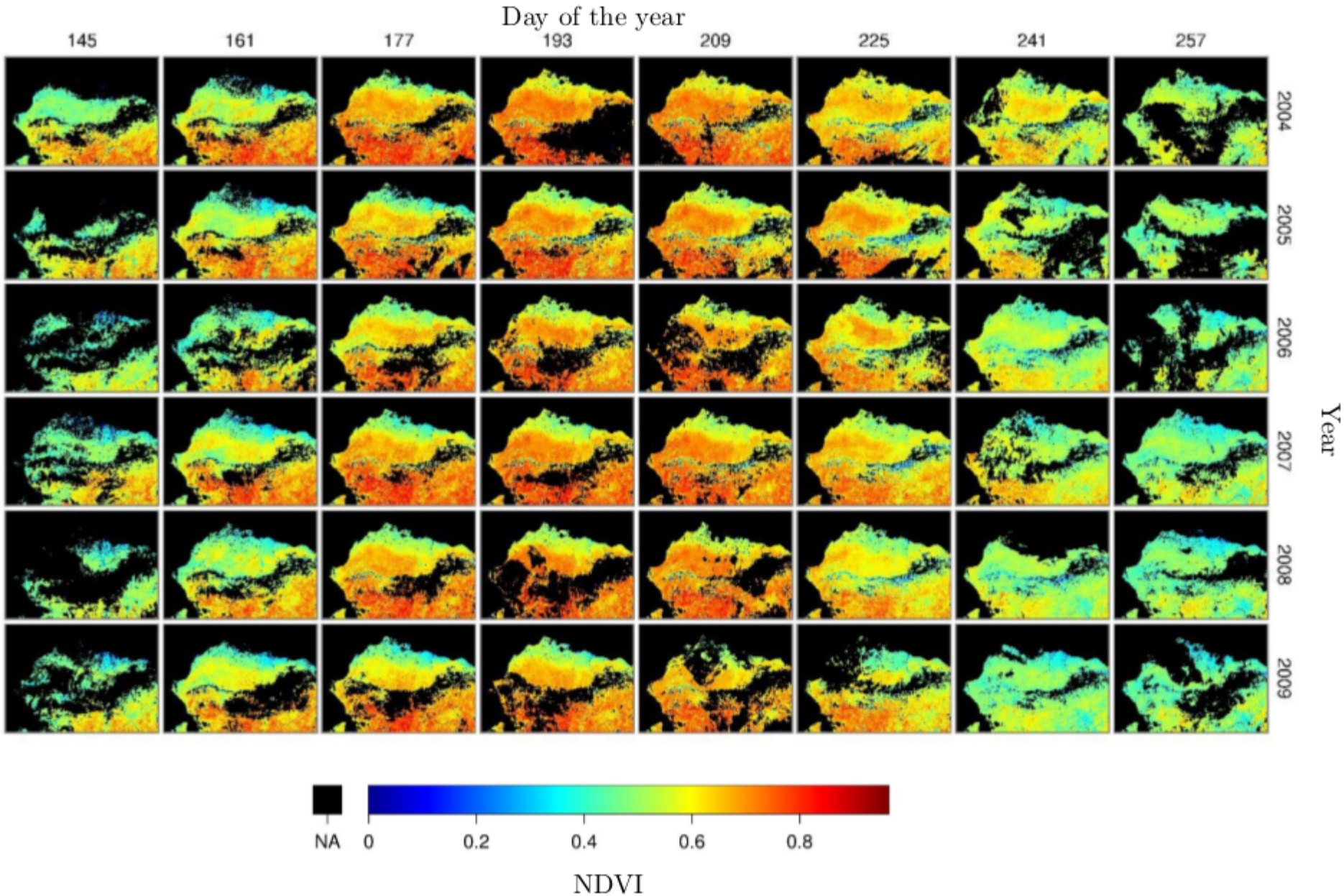
$\rightsquigarrow$ Best Linear Unbiased Predictor:

$$\text{BLUP} = \text{Cov}\Big[Z(\mathbf{s}_{\text{predict}}), Y(\mathbf{s}_{\text{obs}})\Big]\,\text{Var}\Big[Y(\mathbf{s}_{\text{obs}})\Big]^{-1}\text{obs}$$

$$\widehat{Z}(\mathbf{s}_0) = \mathbf{c}^{\mathsf{T}}\,\boldsymbol{\Sigma}^{-1}\,\mathbf{y}$$

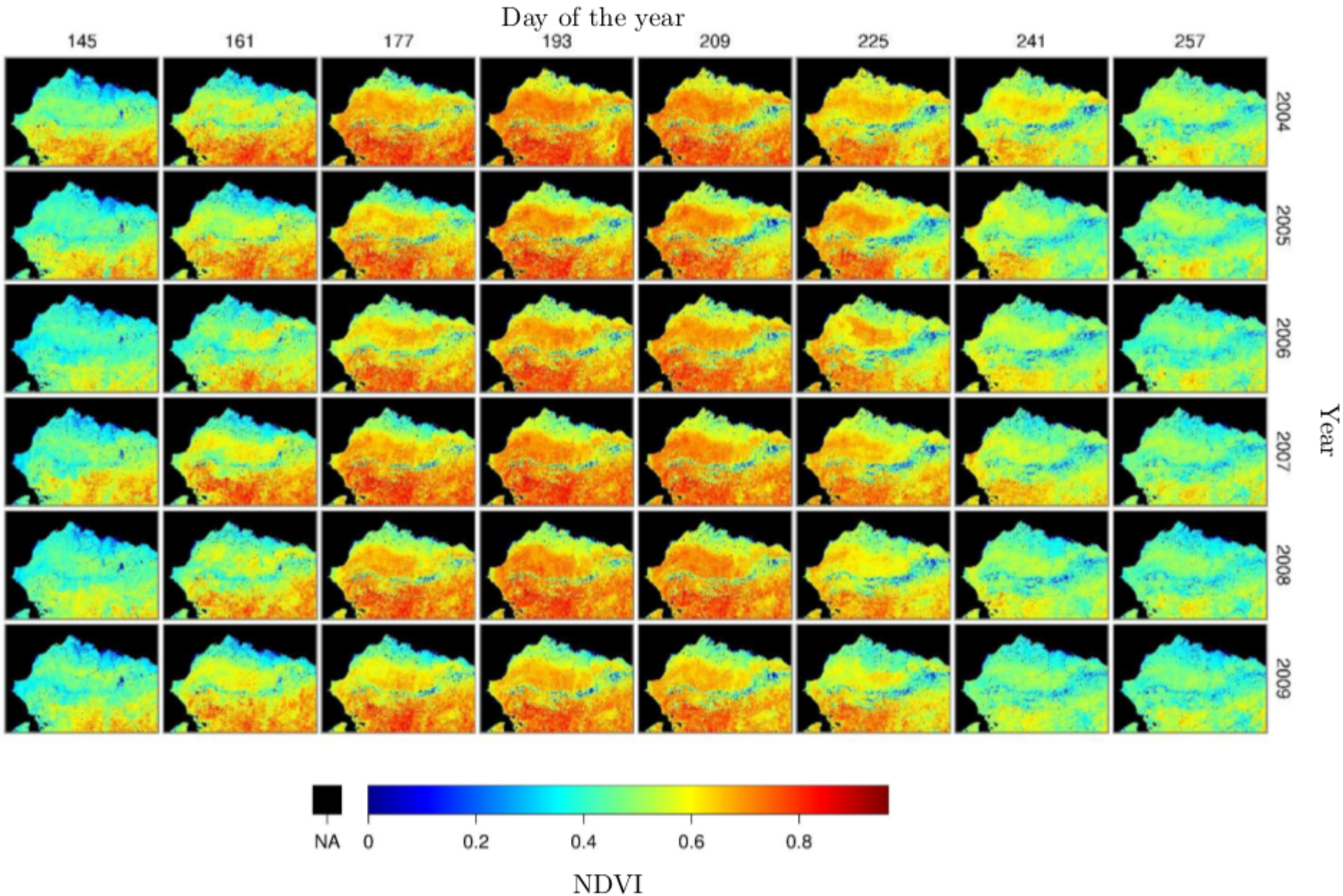
(one spatial process, no trend, known covariance structure;
otherwise *almost* the same)

Visual example



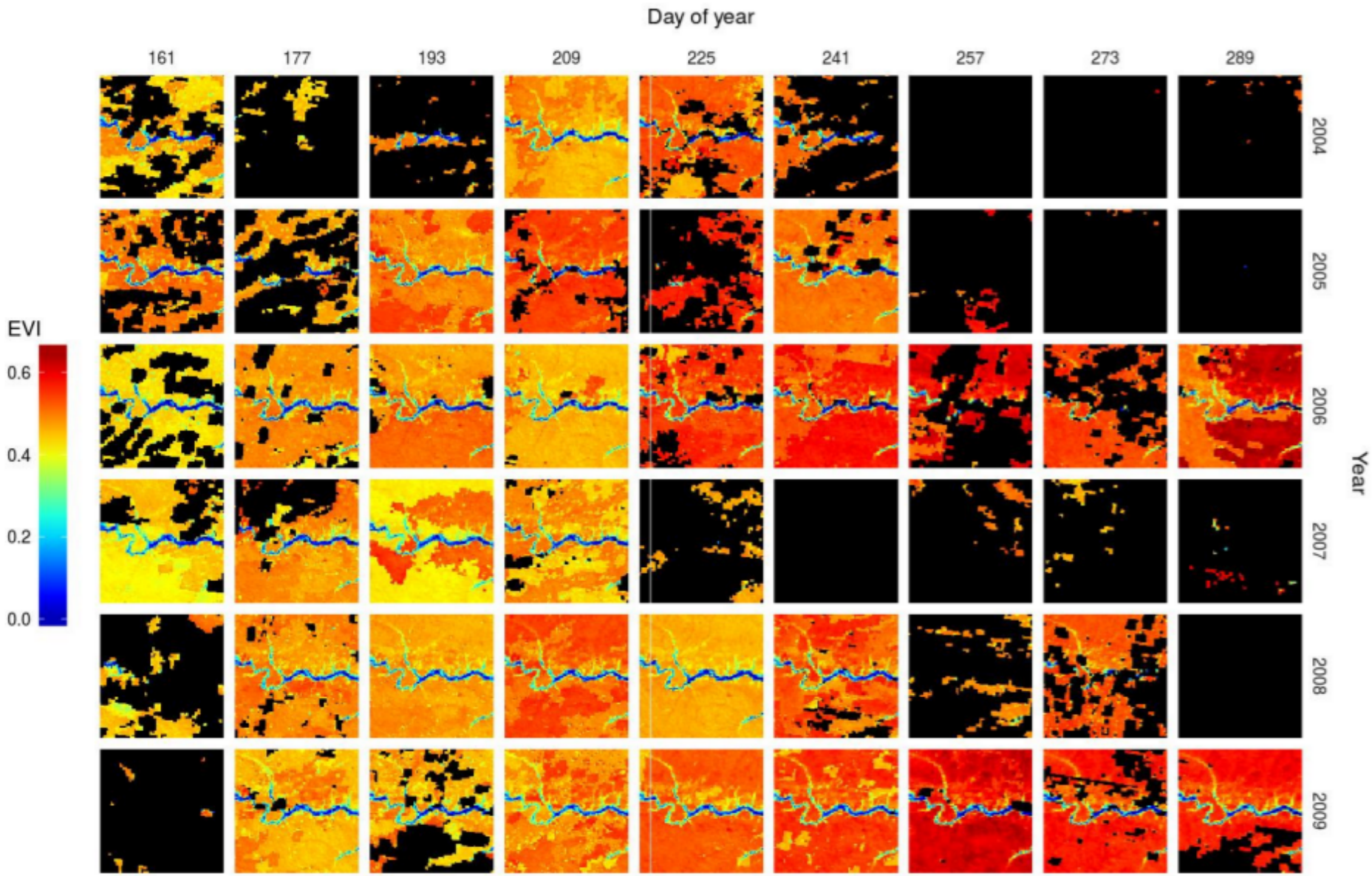
Gerber et al (2018), TGRS

Visual example



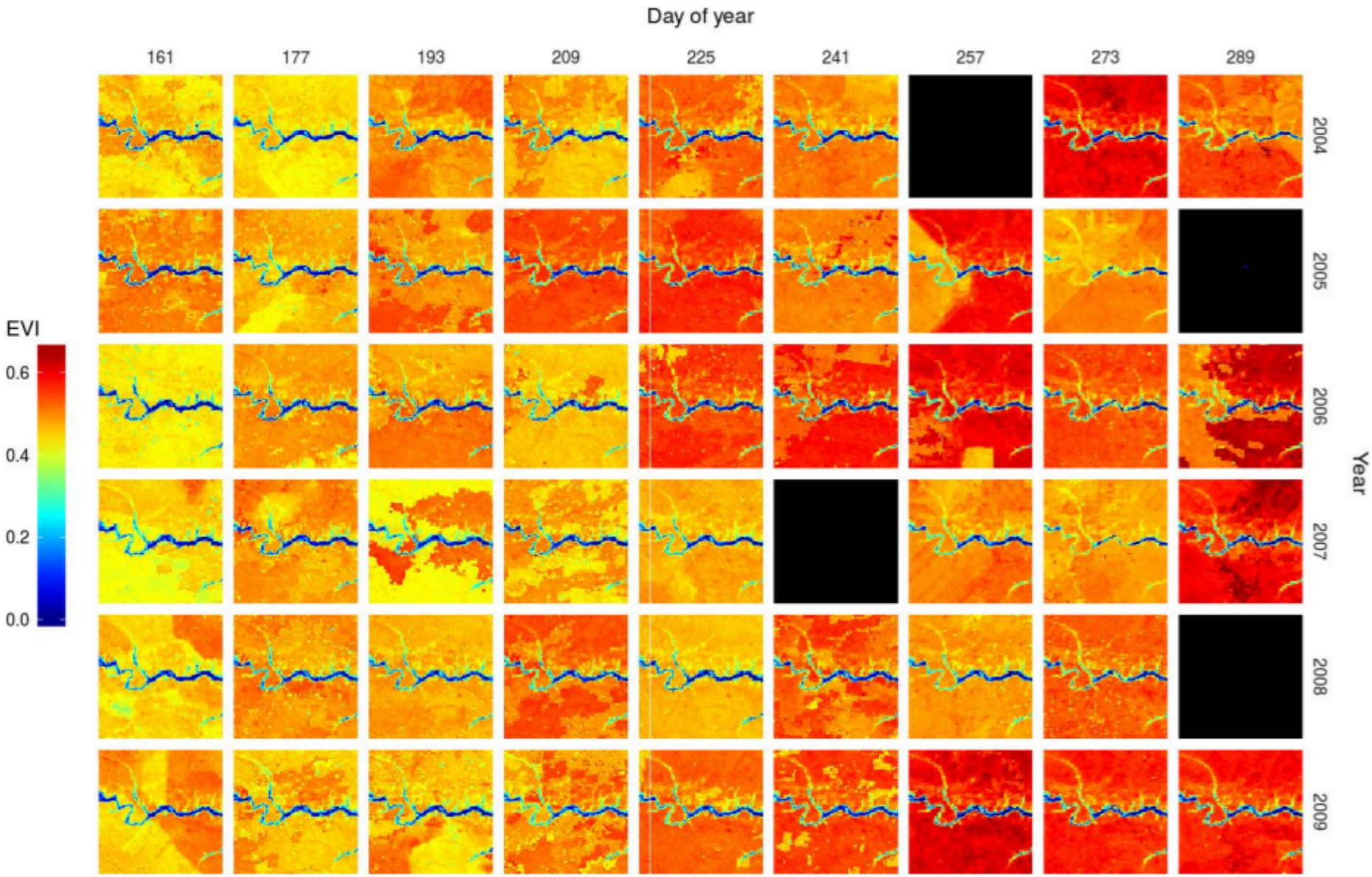
Gerber et al (2018), TGRS

Visual example



Gerber et al (2018), TGRS

Visual example



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

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

- "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

$\rightsquigarrow$ Cholesky factorization**s**

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.

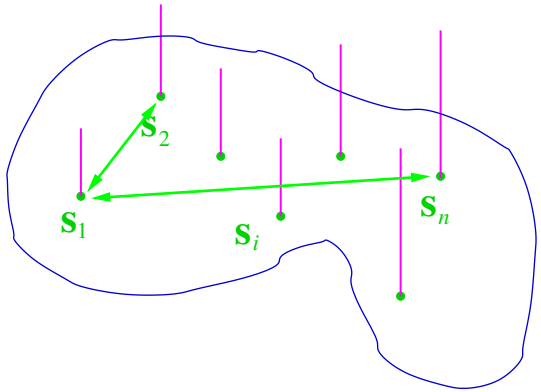
Computational limits are quickly attained!

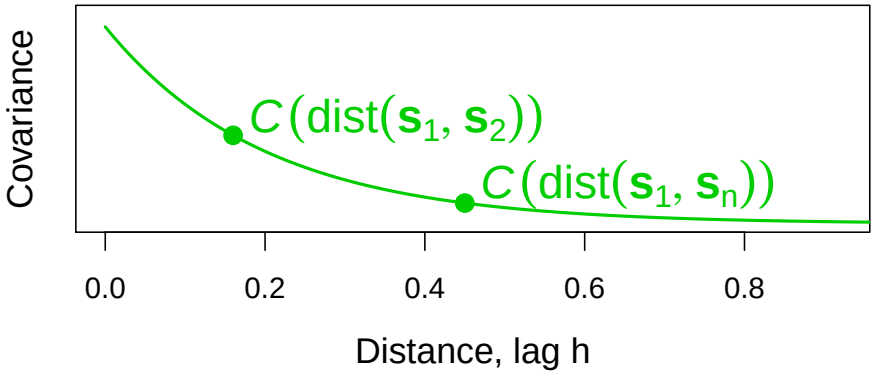
Methods for large spatial datasets

- Sparse Covariance methods:
  - Covariance Tapering — Furrer
  - Spatial Partitioning — Heaton
- Sparse Precision methods:
  - Lattice Kriging — Nychka
  - Multiresolution Approximations — Katzfuss
  - Stochastic Partial Differential Equations — Lindgren
  - Periodic Embedding — Guinness
  - Nearest Neighbor Processes — Datta
- Low rank approximation:
  - Fixed Rank Kriging — Zammit-Mangion
  - Predictive Processes — Finley
- Algorithmic approaches:
  - Gapfill — Gerber
  - Local Approximate Gaussian Processes — Gramacy
  - Metakriging — Guhaniyogi

Spatial modeling

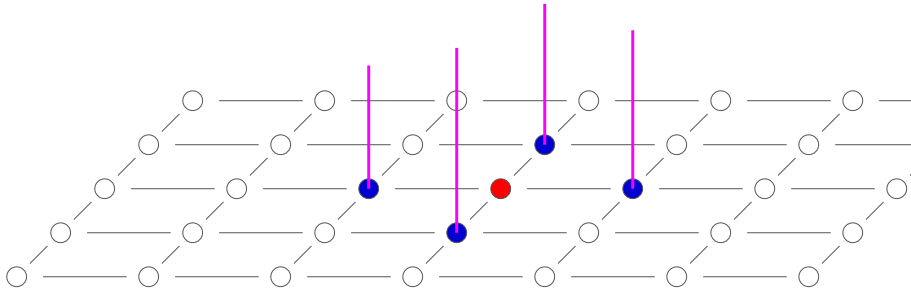
Geostatistical model (GRF):





Covariance matrix: $\boldsymbol{\Sigma}$

Lattice model (GMRF):



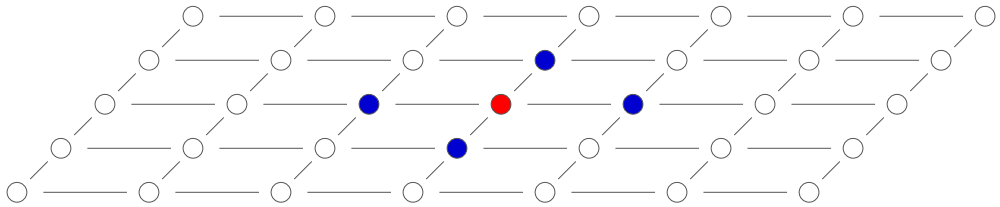
$$\mathrm{E}(Z_i|z_{-i}) = \beta \sum_{j \text{ neighbor of } i} z_j$$

$$\mathrm{Var}(Z_i|z_{-i}) = \tau^2$$

Gaussianity and regularity conditions:

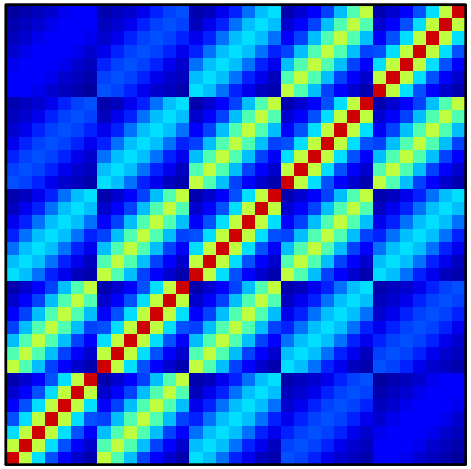
$$\boldsymbol{\Sigma} = \tau^2(\mathbf{I} - \mathbf{B})^{-1}$$

Spatial modeling

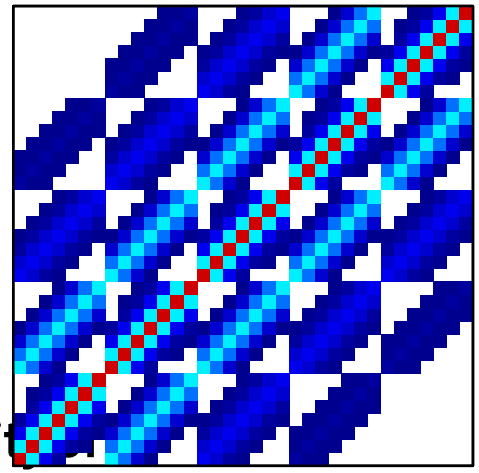


Geostatistical model (GRF):

Σ

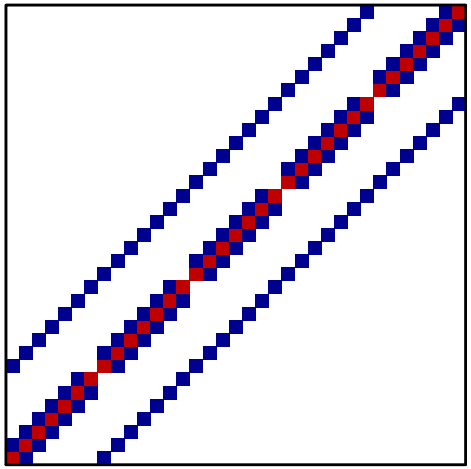


Σ_{app}

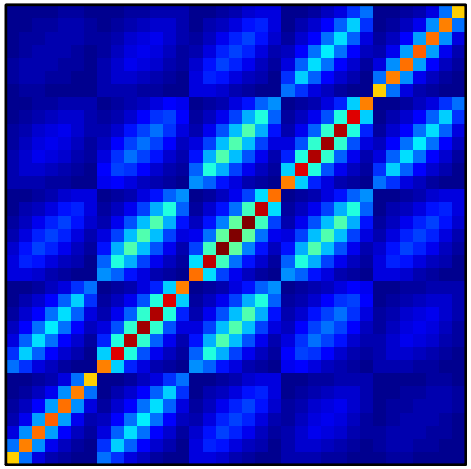


Lattice model (GMRF):

Σ^{-1}



Σ

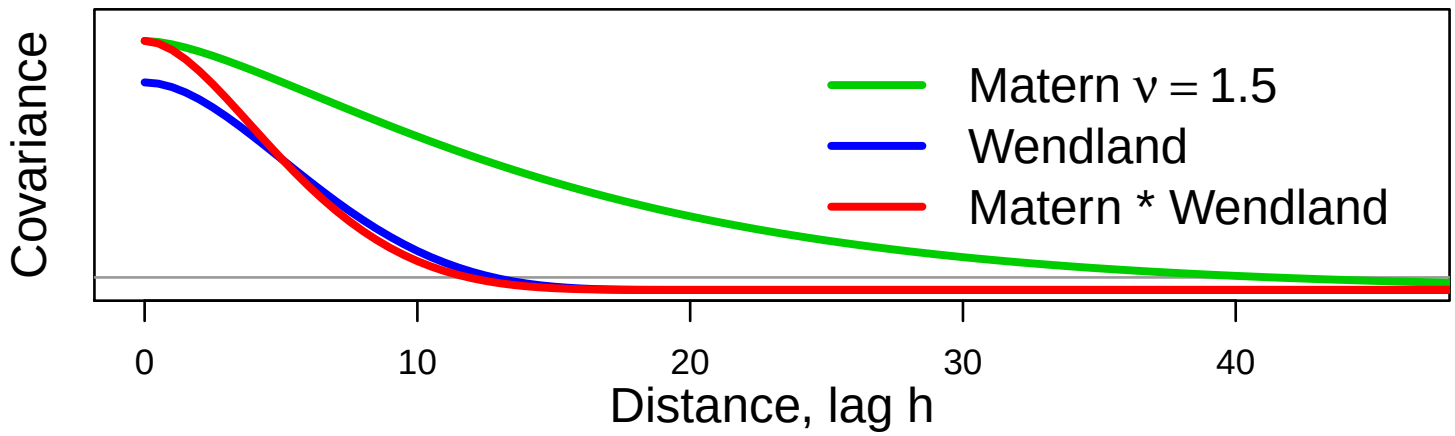


Tapering: sparseness

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 $\rightsquigarrow$ tapering



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 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
Gerber, Mösinger, Furrer (2018) Software X

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

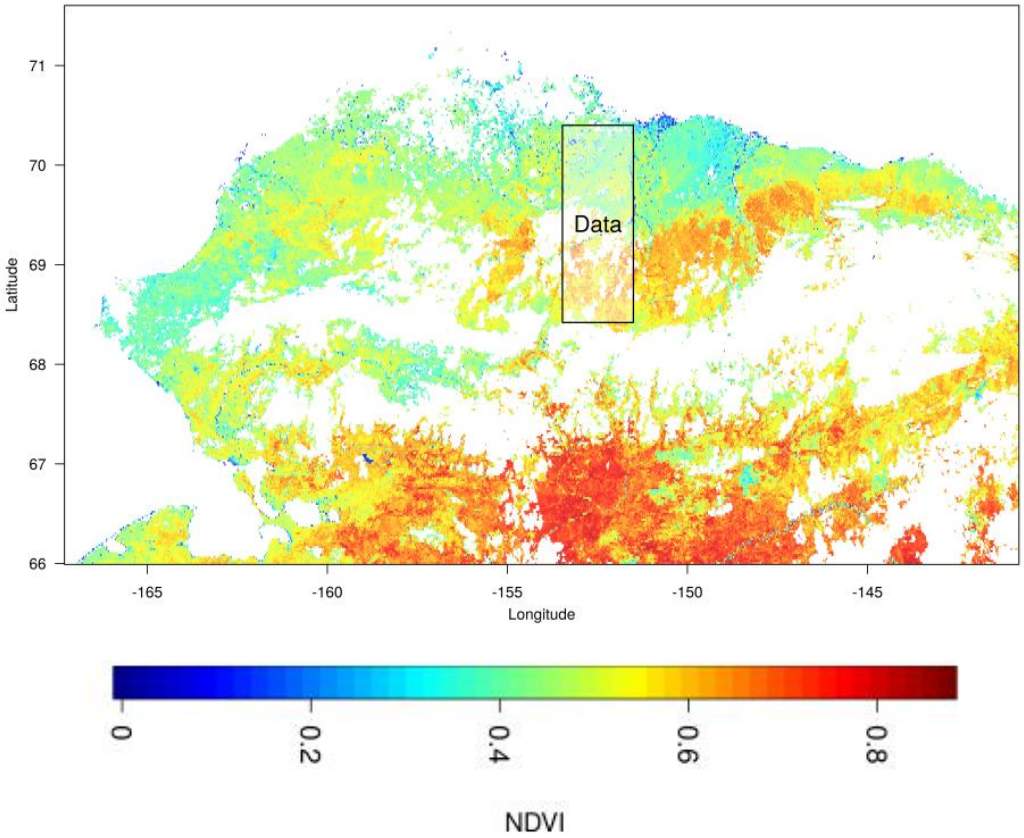
Jeong, Jun, Genton (2017) StatSci
Porcu, Alegria, Furrer (2018) ISR

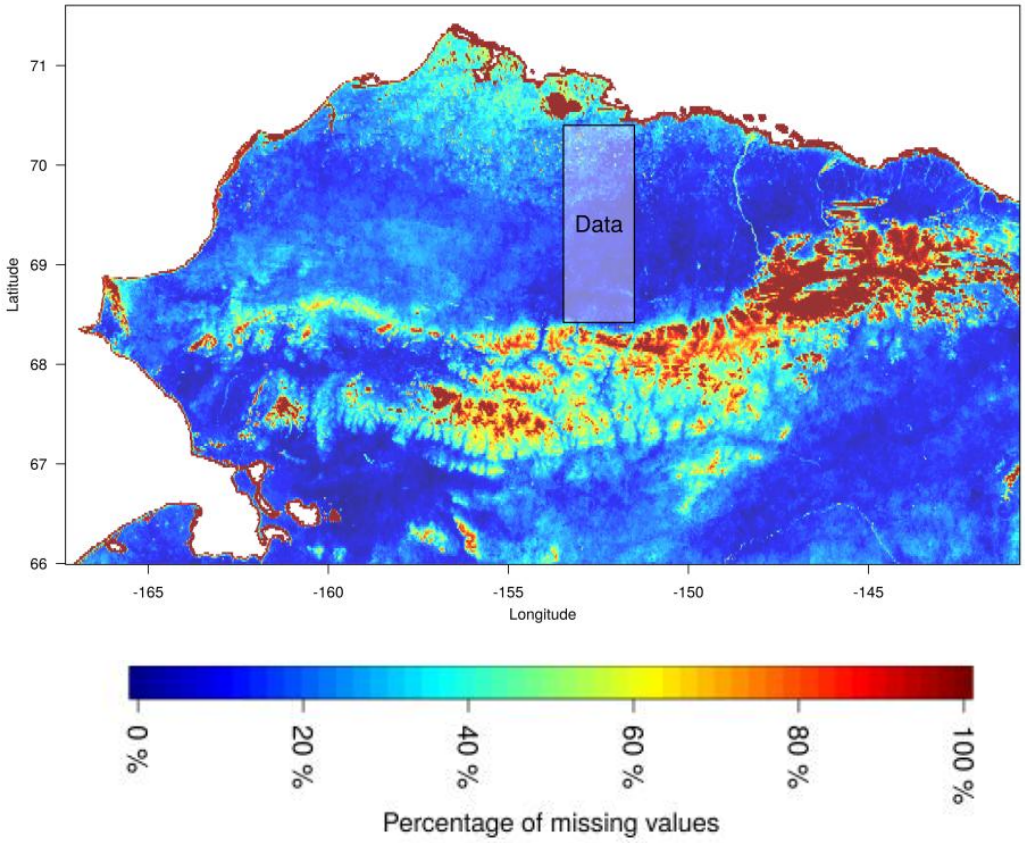
Outlook

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

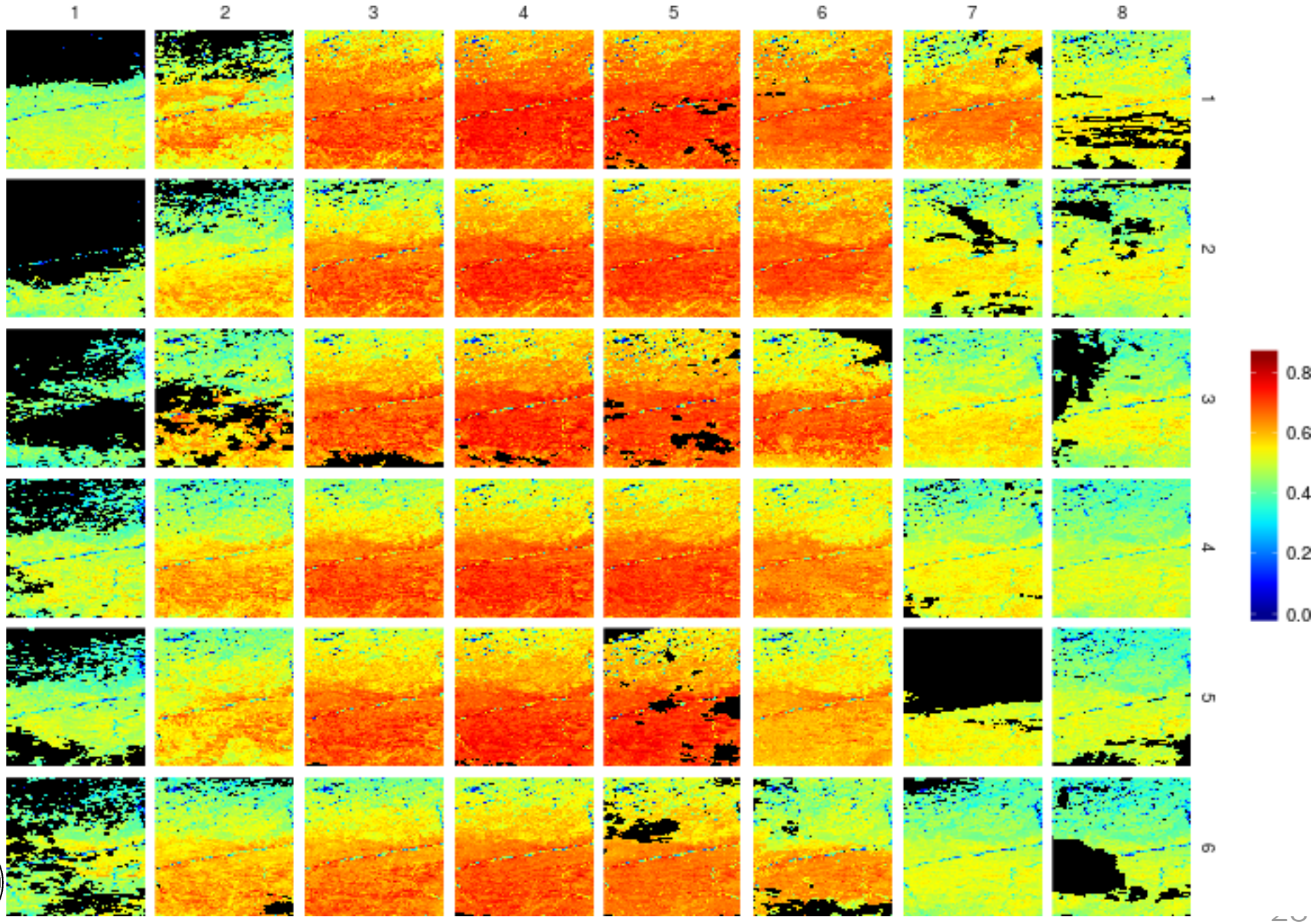
Arctic NDVI data

MODIS NDIV data (satellite product MOD13A1, $\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}}$)

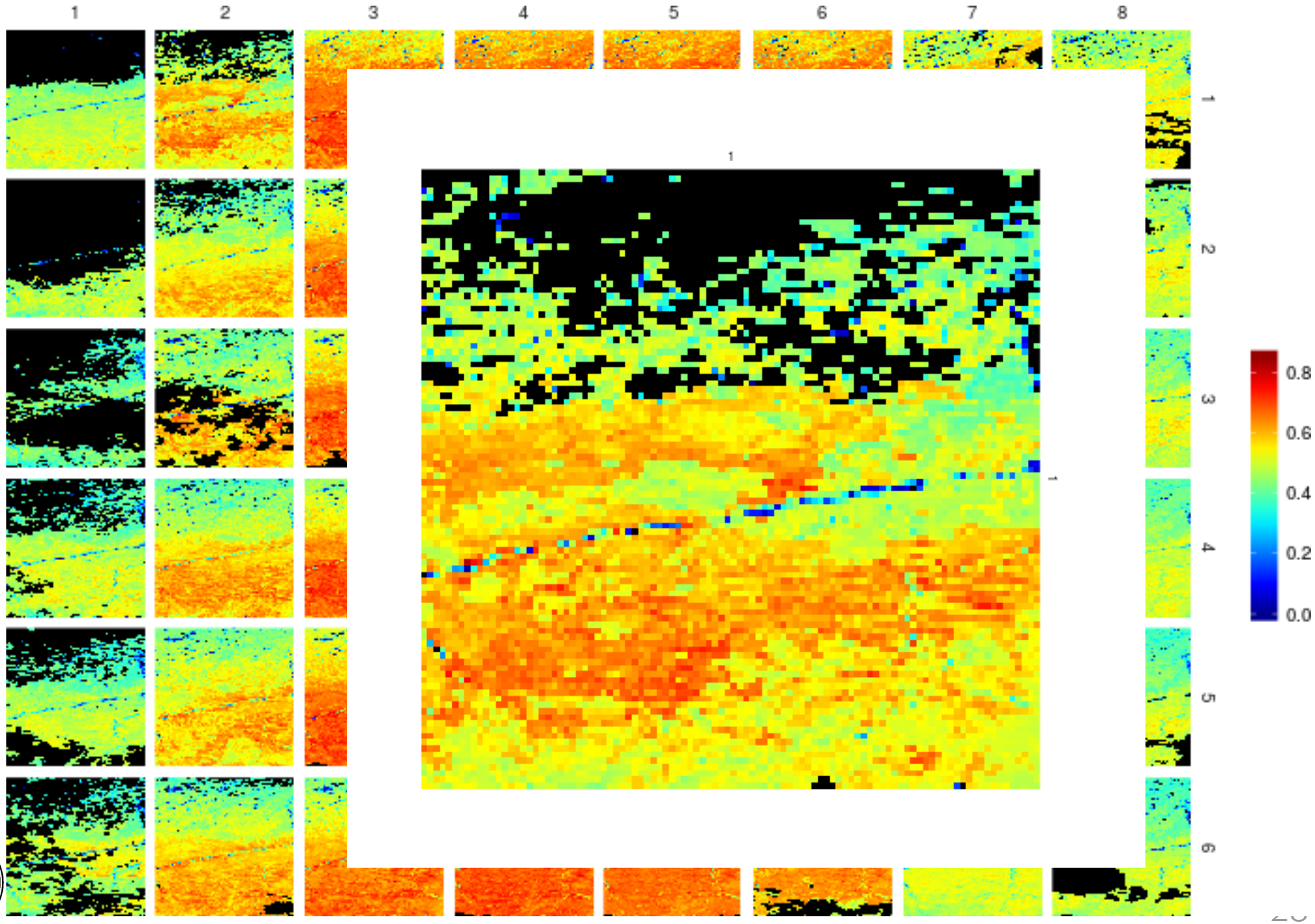




Kriging is smoothing

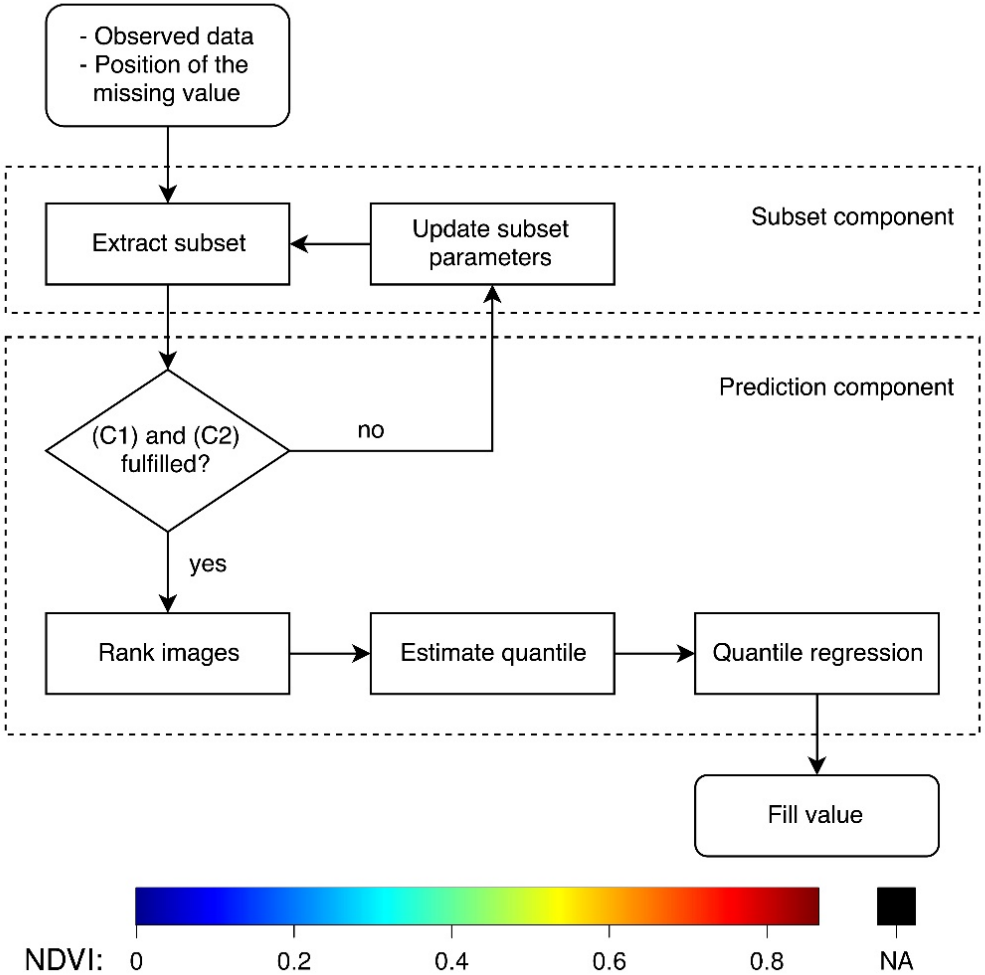


Kriging is smoothing

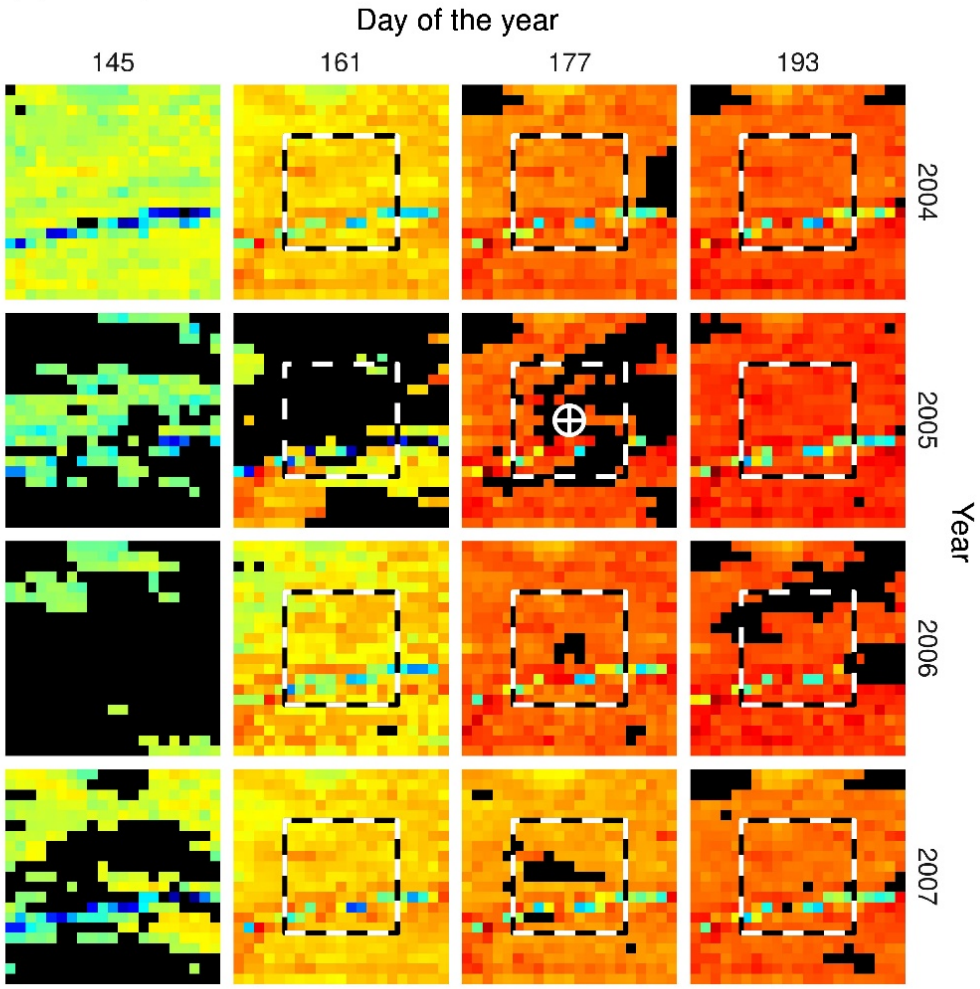


Interpolation using `gapfill`

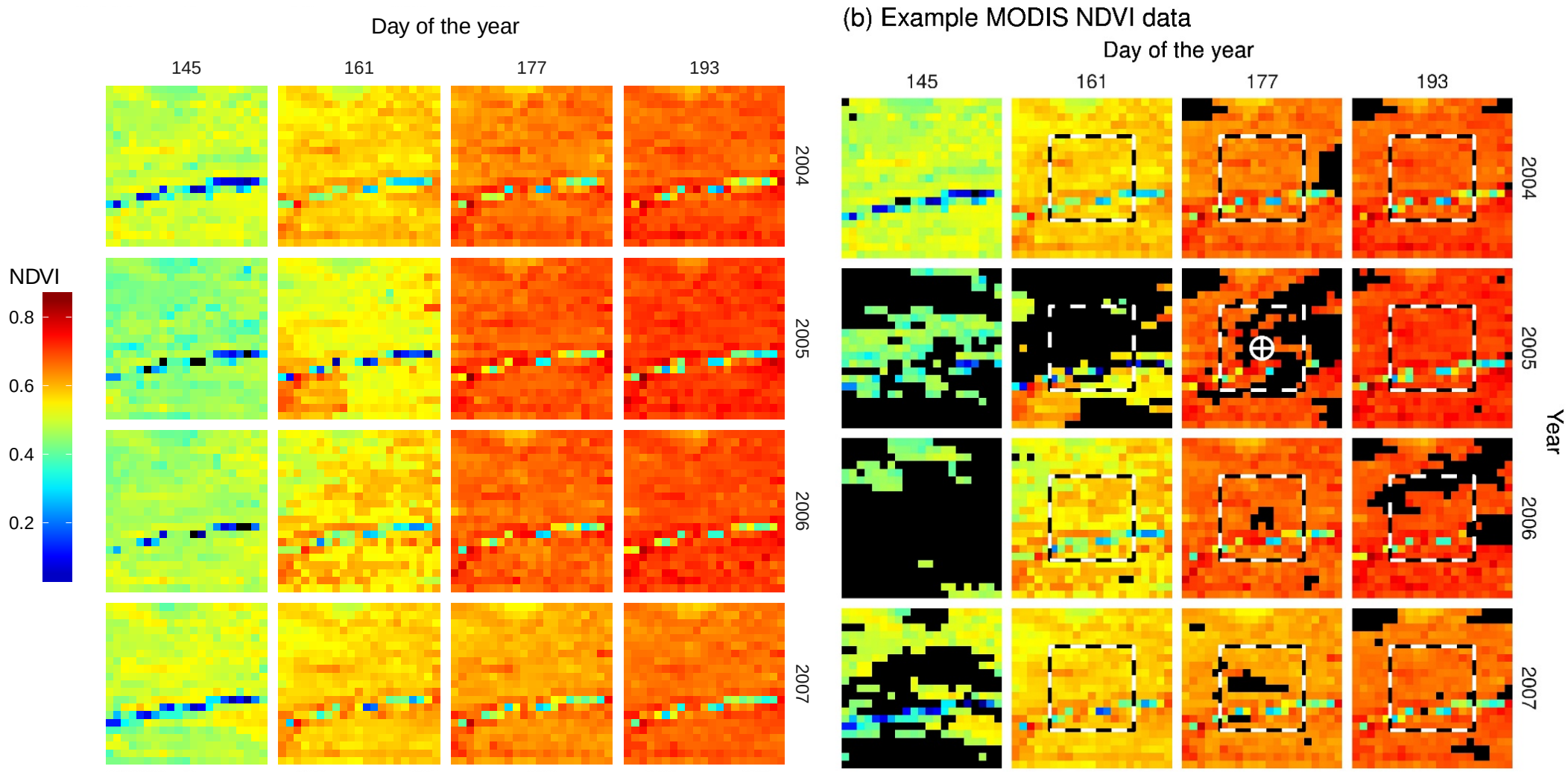
(a) Flow diagram of the gap filling method



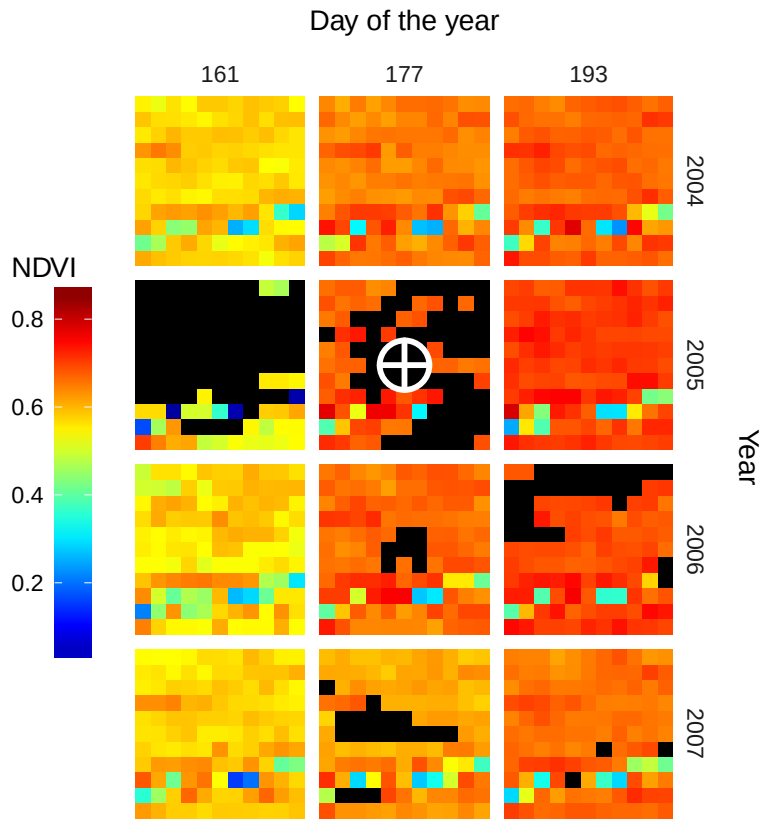
(b) Example MODIS NDVI data

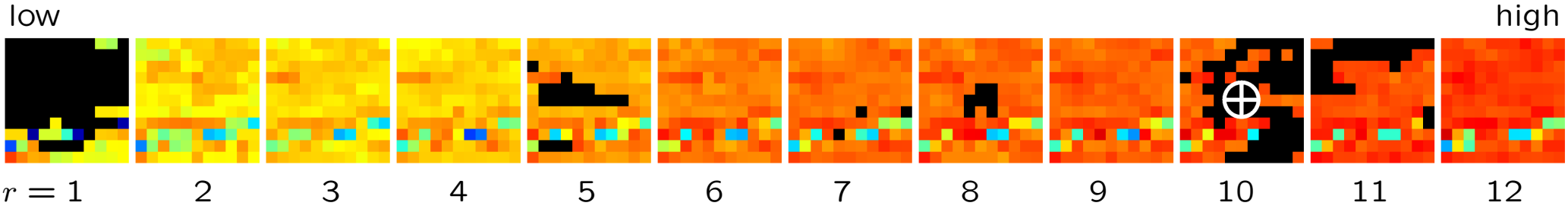


Interpolation using `gapfill`

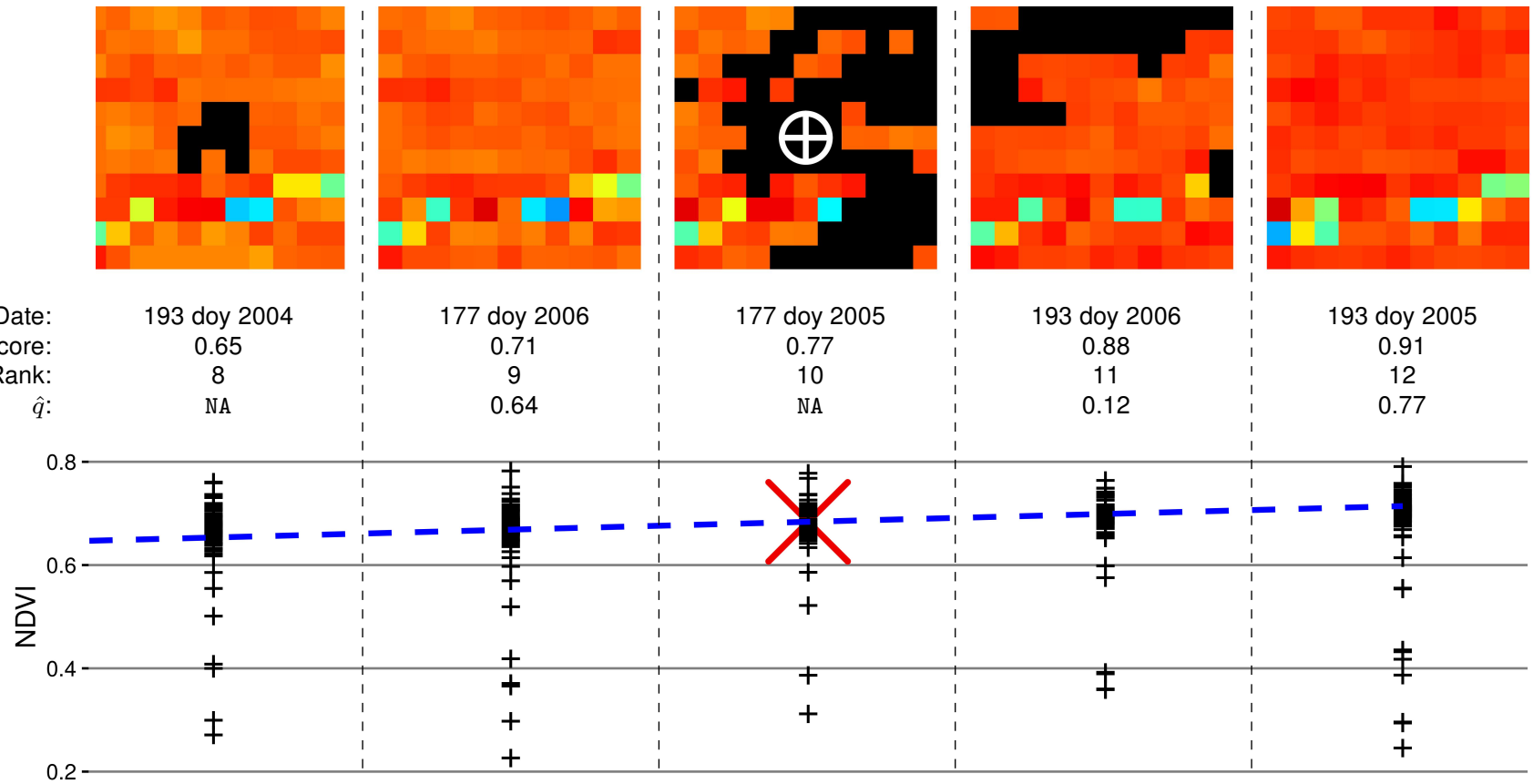


gapfill: ranking of the images

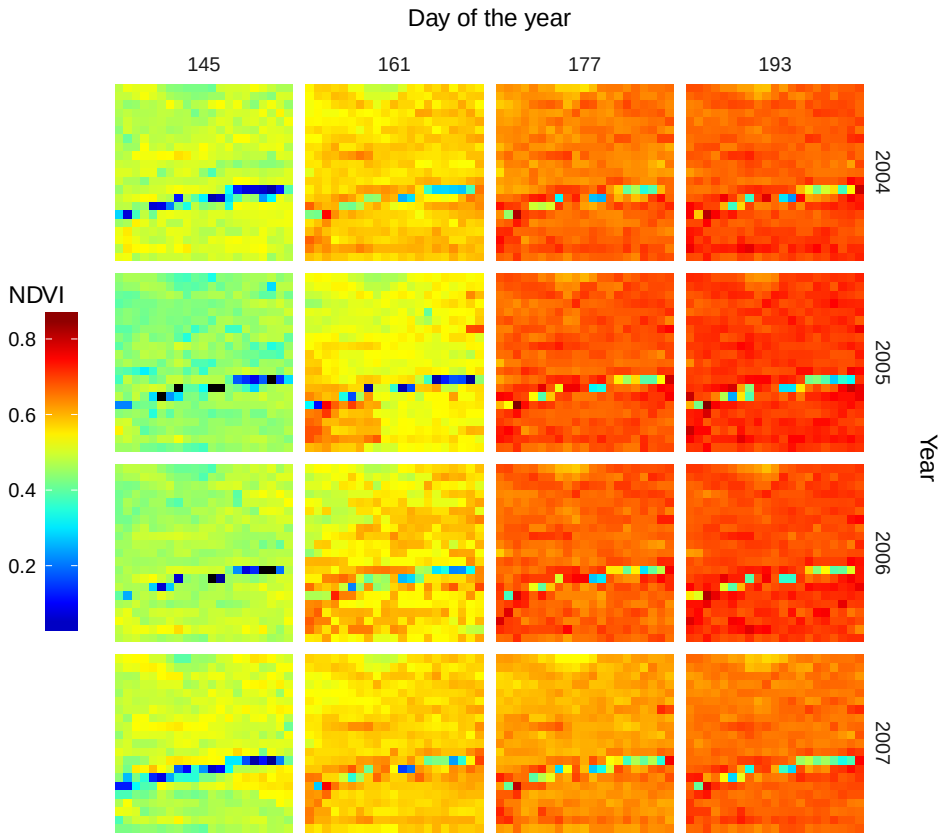




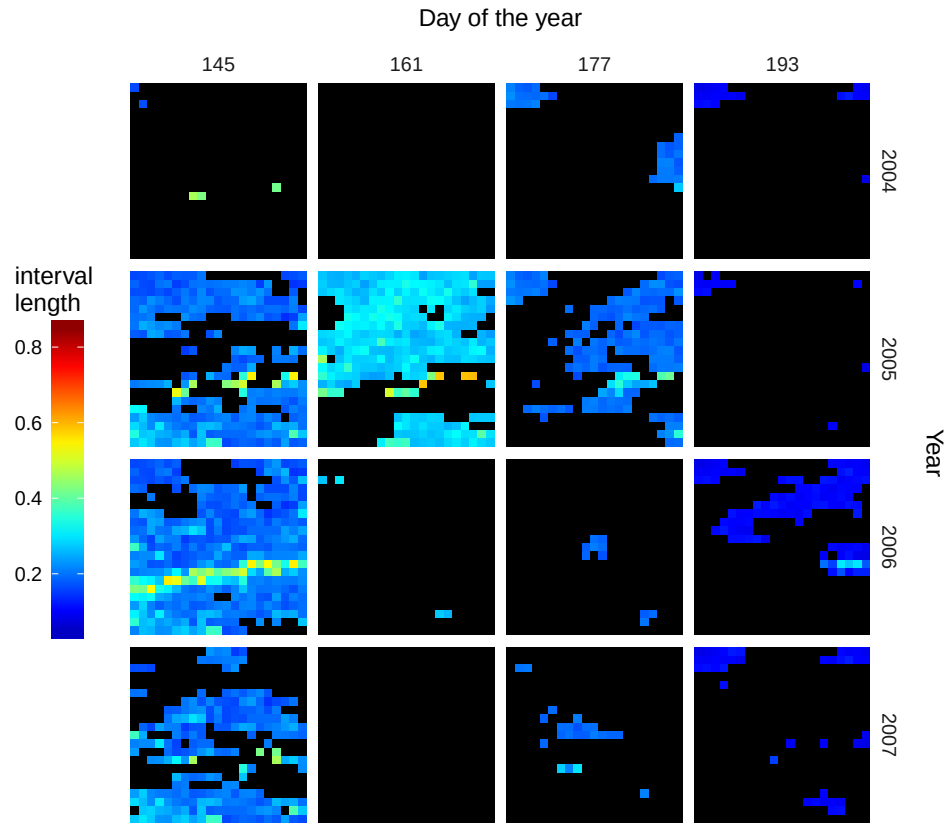
gapfill: quantile regression



gapfill: prediction uncertainties

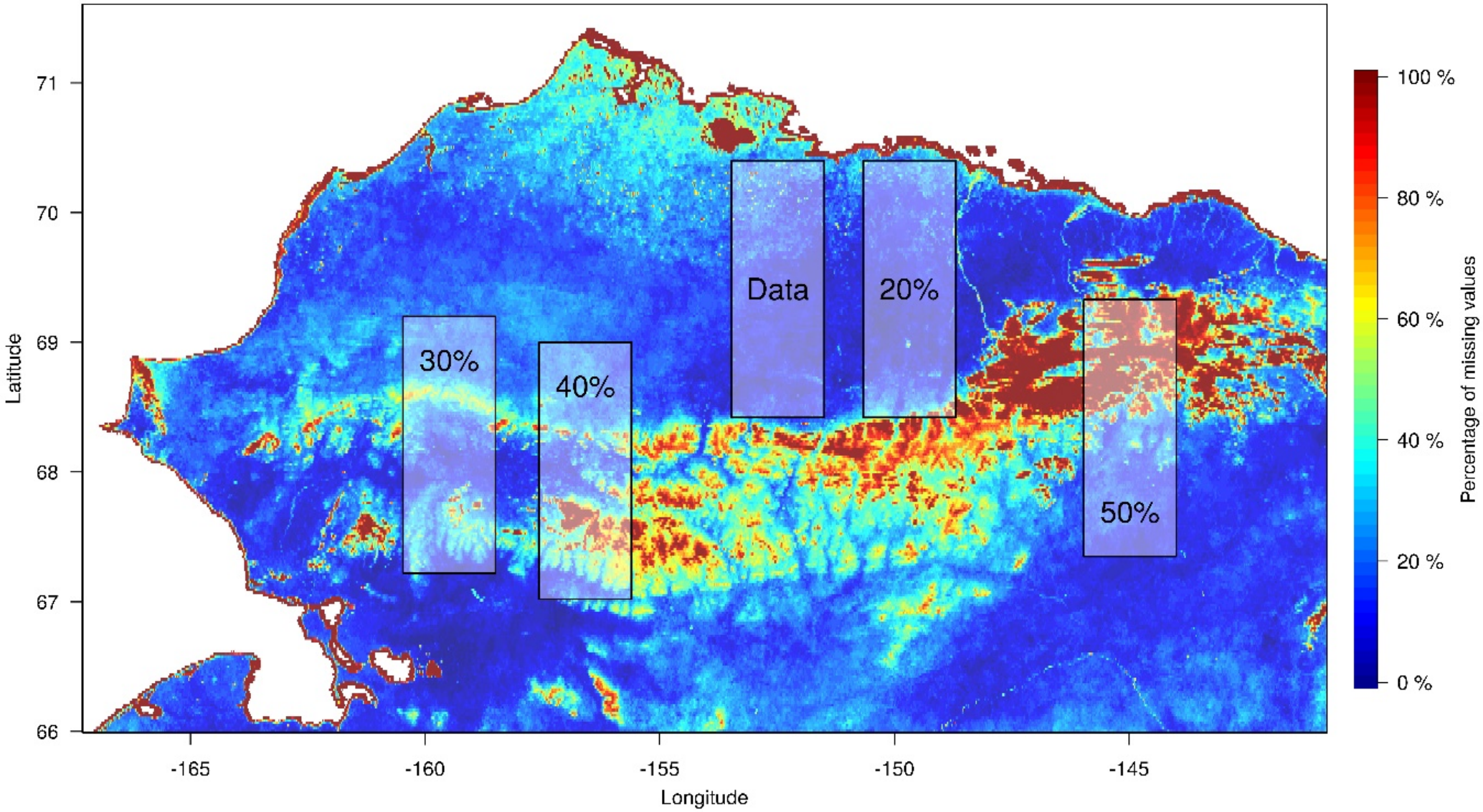


data and predictions



uncertainties

gapfill: location



University of Zurich[UZH]

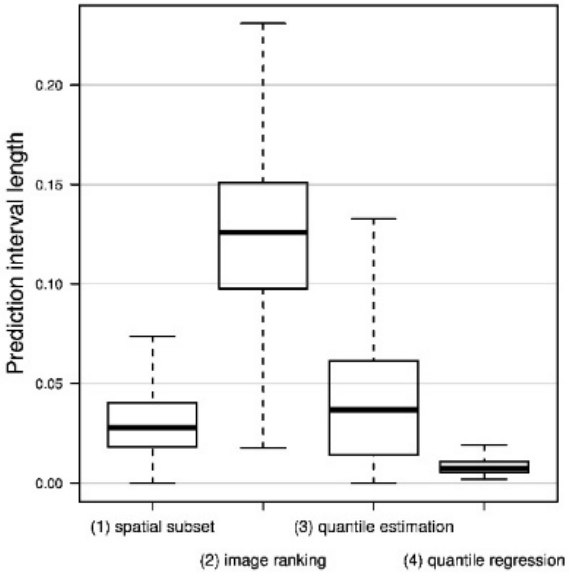
gapfill: comparison

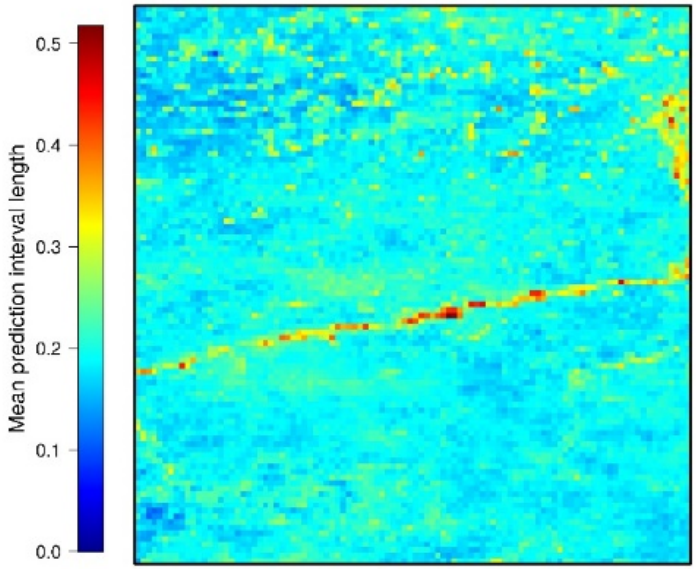
	gapfill				Gapfill-Python		TIMESAT	
	#filled	RMSE	$RMSE_P$	$RMSE_T$	#filled	RMSE	#filled	RMSE
20%	92'822 (100%)	41.80	42.06	41.10	90'307 (97%)	45.00	59'948 (65%)	83.43
30%	147'827 (100%)	42.54	42.39	37.09	146'686 (99%)	45.54	42'892 (29%)	71.43
40%	192'456 (100%)	41.34	40.98	36.41	169'998 (88%)	42.49	31'279 (16%)	71.93
50%	240'326 (100%)	59.58	44.94	37.24	134'540 (56%)	45.61	14'127 (6%)	86.09

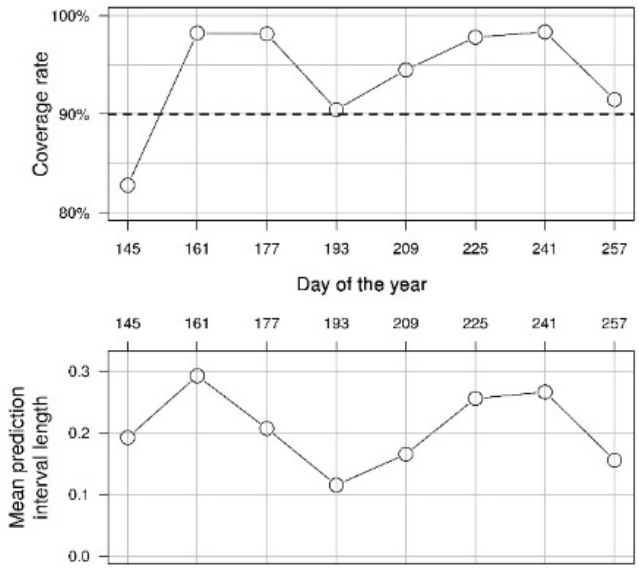
RMSE $\times 10^3$

University of Zurich[UZH]

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"

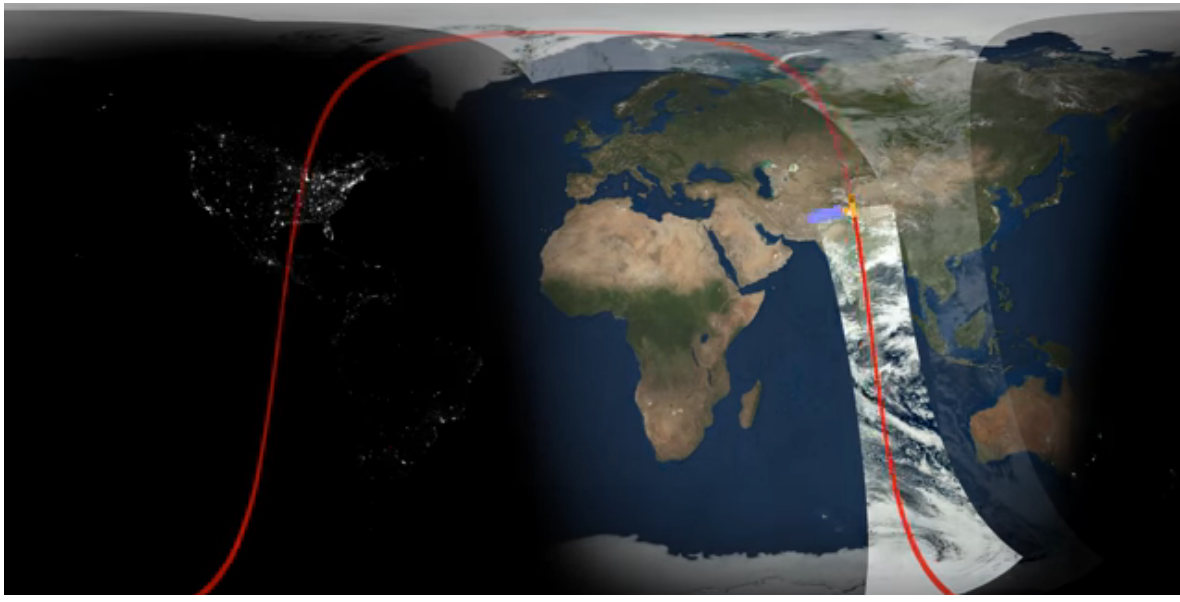
Why do we care?

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



University of Zurich[UZH]

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

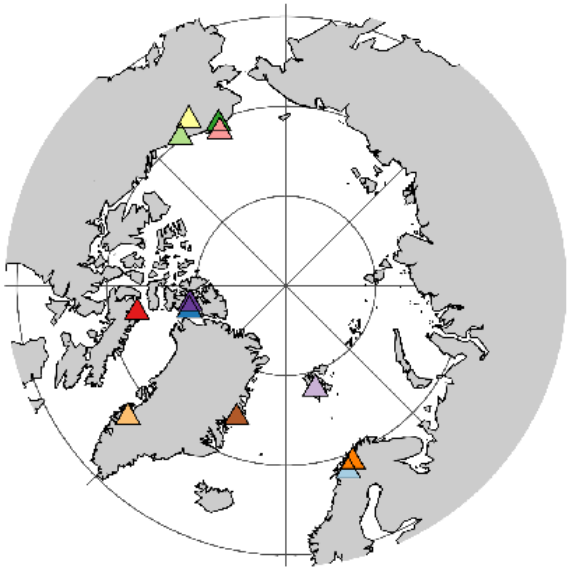
Data

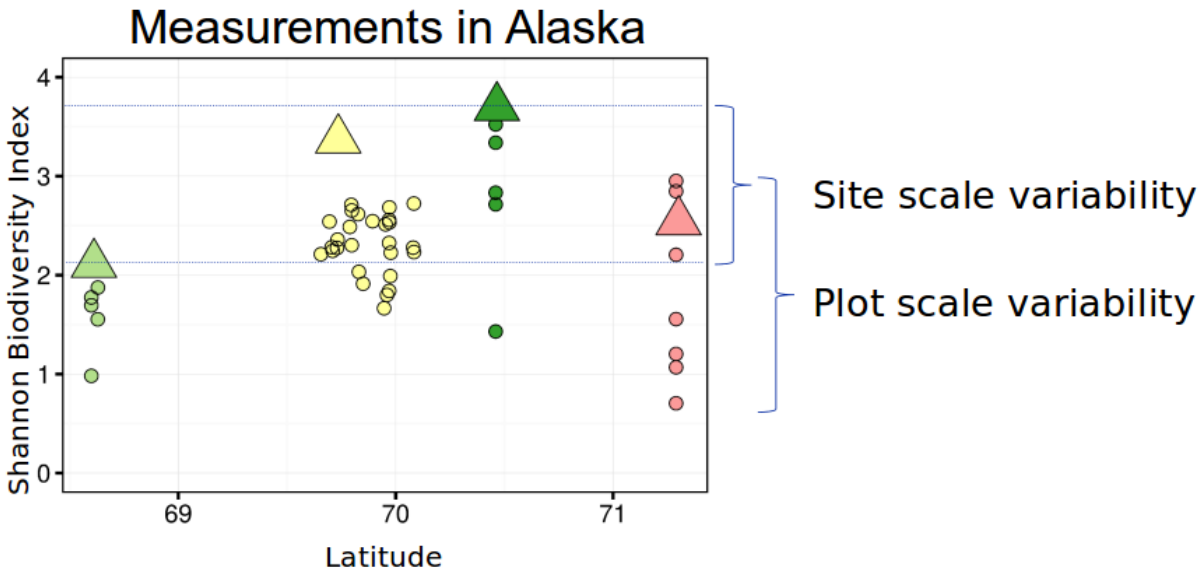
- Species abundance plot scale measurements from the International Tundra Experiment (ITEX)

  ⇝ Shannon biodiversity index on site and plot scale

- Landsat NDVI satellite images and ASTER elevation data

  ⇝ characterization of the landscape heterogeneity

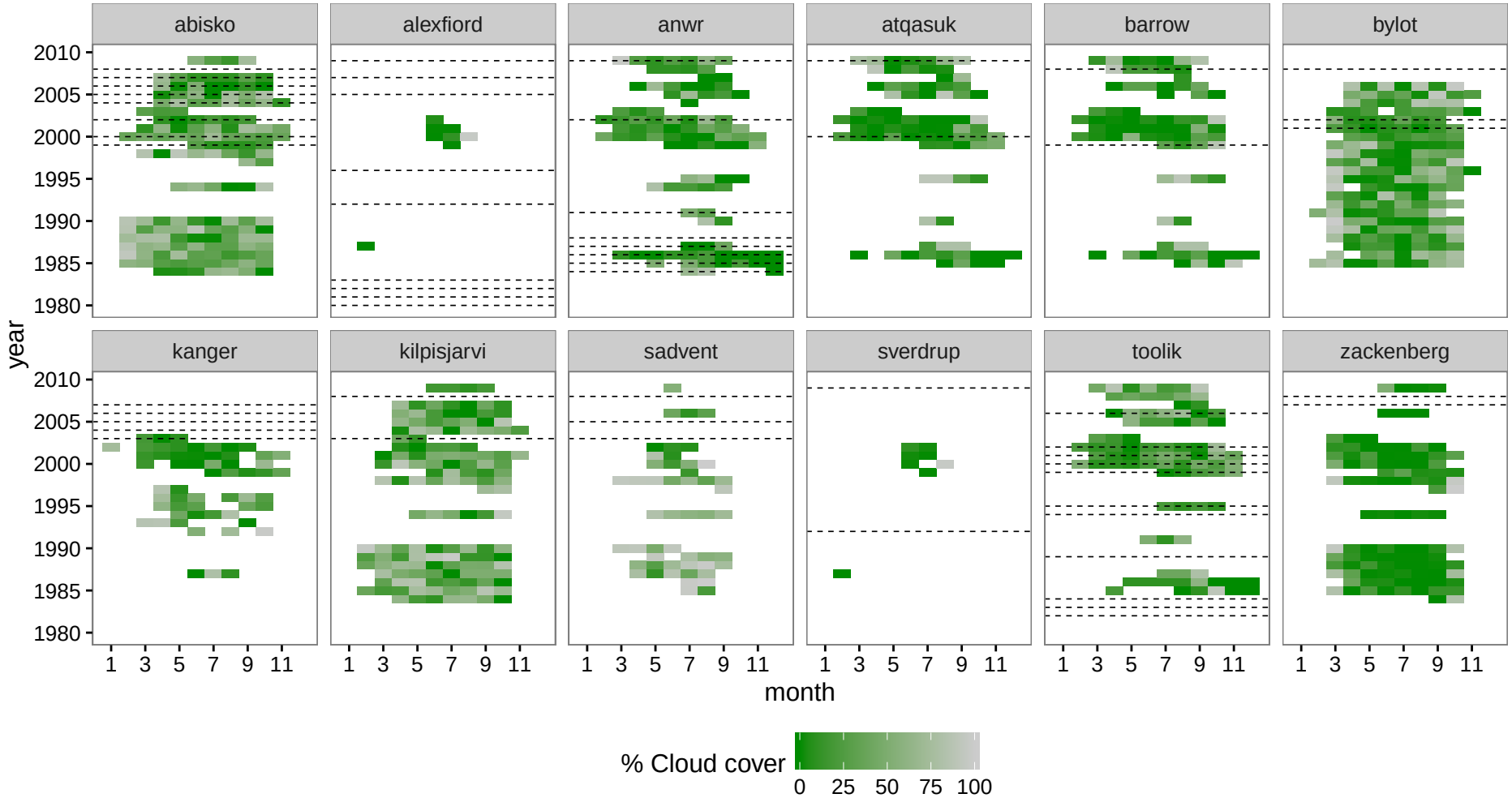




Source: E. Gerber

Data

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

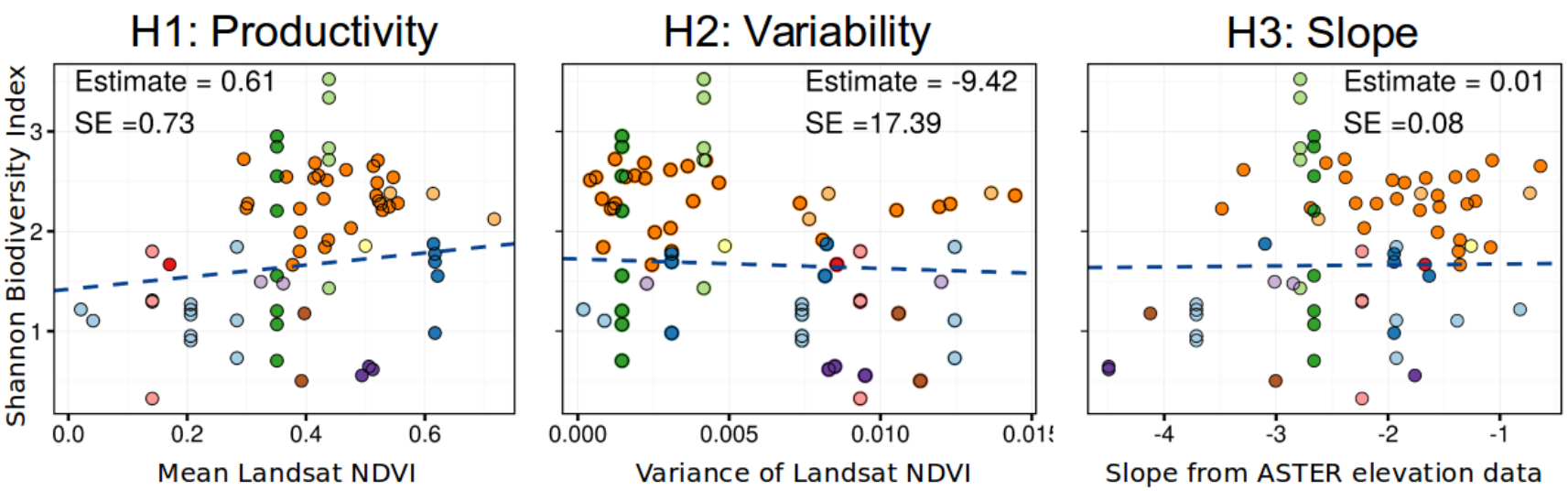


Source: E. Gerber

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: E. 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

FNSNF Swiss National Science Foundation 143282, 144973, 175529

References (some, alphabetical)

Furrer Sain (2010) spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields *JSS* **36** 1–25

Furrer et al (2017) *spam: Sparse Matrix algebra*. R package version 2.2-0

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 et al (2018) Predicting Missing Values in Spatio-Temporal Remote Sensing Data. *IEEE TGRS* 56(5) 2841-2853

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

Heaton et al (2017) A Case Study Competition among Methods for Analyzing Large Spatial Data *arXiv*:1710.05013

Porcu Alegria Furrer (2018) Modeling Temporally Evolving and Spatially Globally Dependent Data *International Statistical Review* 86 344–377

Complete list at: www.math.uzh.ch/furrer/research/publications.shtml