RECAP15
Re-thinking the Efficacy of International Climate Change Agreements Post COP15

Gefördert durch das Bundesministerium für Bildung und Forschung (BMBF) im Rahmen des Förderschwerpunkts „Ökonomie des Klimawandels“ unter FKZ 01LA1139A

www.europa-uni.de/recap15

Discussion Paper Series recap15

No 9 – August 2013

Efficient Approximation of the Spatial Covariance Function for Large Datasets - Analysis of Atmospheric CO_2 Concentrations

Patrick Gneuss, Wolfgang Schmid, Reimund Schwarze

Contact

Patrick Gneuss
European University Viadrina
PO Box 1786, 15207 Frankfurt (Oder), Germany
vetter@europa-uni.de

Efficient Approximation of the Spatial Covariance Function for Large Datasets - Analysis of Atmospheric CO_2 Concentrations

Patrick Gneuss
Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany
Wolfgang Schmid
Department of Statistics, European University Viadrina, Frankfurt (Oder), Germany
Reimund Schwarze
Department of Economics, European University Viadrina, Frankfurt (Oder), Germany

Abstract

Linear mixed effects models have been widely used in the spatial analysis of environmental processes. However, parameter estimation and spatial predictions involve the inversion and determinant of the $n \times n$ dimensional spatial covariance matrix of the data process, with n being the number of observations. Nowadays environmental variables are typically obtained through remote sensing and contain observations of the order of tens or hundreds of thousand on a single day, which quickly leads to bottlenecks in terms of computation speed and requirements in working memory. Therefore techniques for reducing the dimension of the problem are required. The present work analyzes approaches to approximate the spatial covariance function in a real dataset of remotely sensed carbon dioxide concentrations, obtained from the Atmospheric Infrared Sounder of NASA's "Aqua" satellite on the 1st of May 2009. In a cross-validation case study it is shown how fixed rank kriging, stationary covariance tapering and the full-scale approximation are able to notably speed up calculations. However the loss in predictive performance caused by the approximation strongly differs. The best results were obtained for the full-scale approximation, which was able to overcome the individual weaknesses of the fixed rank kriging and the covariance tapering.

Keywords: spatial covariance function, fixed rank kriging, covariance tapering, full-scale approximation, large spatial data sets, mid-tropospheric CO_2, remote sensing, efficient approximation.

1. Introduction

The monitoring of environmental processes has been revolutionized in the recent past through the upcoming of remotely sensed satellite measurements. The resulting spatial resolution is far superior compared to the traditional monitoring through networks of measurement stations. This of course constitutes a major improvement for scientific research, but also introduces the need for statistical models that can handle such large data sets, which often involve observations on the order of tens or hundreds of thousand per day. One such environmental data set of particular interest for the ongoing political discussion and the related negotiations on climate change and global warming is the remotely sensed measurement of carbon dioxide concentration in the mid-troposphere as measured by the Atmospheric Infrared Sounder (AIRS) of NASA's "Aqua" satellite. It has been contemplated that space-based observations of CO_2 could complement the weakly enforceable system of national reporting of sources of CO_2 emissions in a meaningful way (Mintzer, Leonard, and Valencia (2010, p. 28)). In fact, recent studies have reported a 'gap' in CO_2 reporting from China of 1.4 gigatonnes per year (Guan, Liu, Geng, Lindner, and Hubacek (2012)), which would amount to 5% of the global total. This gives rise for an objective assessment of CO_2 emissions based on measurements and a corresponding validation of national reporting standards. In that way statistical modeling of atmospheric CO_2 concentrations can serve as an important input to climate projection projects and for the estimation of CO_2 surface fluxes. Linear mixed effects models have been widely used in the spatial analysis of such environmental data sets. However, parameter estimation and spatial predictions involve the inversion and determinant of the $n \times n$ dimensional spatial covariance matrix of the data process, with n being the number of observations. As mentioned above, environmental variables as measured through remote sensing contain observations of the order of tens or hundreds of thousand on a single day, which quickly leads to bottlenecks in terms of computation speed and requirements in working memory.

1.1. Linear Mixed-Effects Models

Consider a real-valued spatial process $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$ defined on the domain of interest (e.g. the globe as in the CO_2 example). The process is observed at n locations and is a noisy version of the smooth process $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}$, which we are interested in making inference on. This defines the process $Z(\cdot)$ at location $\mathbf{s}$ as

$$Z(\mathbf{s}) = Y(\mathbf{s}) + \epsilon(\mathbf{s}) \tag{1}$$

where $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ is a spatial white-noise process with zero mean and $var(\epsilon(\mathbf{s})) = \sigma^2_\epsilon v(\mathbf{s})$. $\epsilon(\cdot)$ covers the *nugget effect*, or alternatively the measurement error of the instrument. The smooth process $Y(\cdot)$ contains two parts,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\alpha} + \nu(\mathbf{s}) \tag{2}$$

where the first one covers fixed-effects from a deterministic large-scale trend, modeled here as a linear function of p spatial covariates $\mathbf{x}(\cdot)$. The second term $\nu(\cdot)$ models small-scale spatial random variations through a zero-mean process with positive and finite variance and (generally non-stationary) covariance function

$$cov(\nu(\mathbf{u}), \nu(\mathbf{v})) \equiv C(\mathbf{u}, \mathbf{v}) \qquad \mathbf{u}, \mathbf{v} \in D \quad . \tag{3}$$

For the process $Z(\cdot)$ at the n observed locations $\mathbf{Z} = (Z(\mathbf{s}_1), \cdots, Z(\mathbf{s}_n))$ this becomes

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\nu} + \boldsymbol{\epsilon} \tag{4}$$

with $\mathbf{X}$ being the $n \times p$ matrix of covariate values at the observed data locations. Assuming $\boldsymbol{\epsilon}$ and $\boldsymbol{\nu}$ to be independent the resulting $n \times n$ covariance matrix of $\mathbf{Z}$ is

$$\boldsymbol{\Sigma} = var(\boldsymbol{\nu}) + var(\boldsymbol{\epsilon}) = \mathbf{C} + \sigma_\epsilon^2 \mathbf{V}_\epsilon \tag{5}$$

where $\mathbf{C}$ is the covariance matrix of $\boldsymbol{\nu}$ generated by the covariance function in (3) and $\mathbf{V}_\epsilon = diag\{v_\epsilon(\mathbf{s}_1), \cdots, v_\epsilon(\mathbf{s}_n)\}$. The model described in (1)-(5) is also called a *linear mixed-effects model*.

To obtain an optimal linear spatial prediction of the smooth process $Y(\cdot)$ at a specific location $\mathbf{s}_0$, *universal kriging* can be applied, as described for example in Cressie and Wikle (2011). Universal Kriging solves for the homogeneously linear combination of the data $\boldsymbol{\lambda}'\mathbf{Z}$, that minimizes the mean squared prediction error

$$MSPE(\boldsymbol{\lambda}) = E(Y(\mathbf{s_0}) - \boldsymbol{\lambda}'\mathbf{Z})^2.$$

In a purely gaussian setting this is also equivalent to deriving the posterior distribution $[Y(\mathbf{s_0})|\mathbf{Z}]$ and its first two moments $E(Y(\mathbf{s_0})|\mathbf{Z})$ and $var(Y(\mathbf{s_0})|\mathbf{Z})$. The resulting universal-kriging predictor and kriging variance are given in (6) and (7) (Cressie and Wikle (2011, p. 148))

$$\widehat{Y}(\mathbf{s_0}) = \mathbf{x}(\mathbf{s_0})'\widehat{\boldsymbol{\alpha}}_{gls} + \mathbf{c}_Y(\mathbf{s_0})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\alpha}}_{gls}) \tag{6}$$

$$\begin{aligned} \widehat{\sigma}^2(\mathbf{s_0}) &= Var(Y(\mathbf{s_0})) \\ &= \mathbf{c}_Y(\mathbf{s_0})'\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0}) \\ &\quad +(\mathbf{x}(\mathbf{s_0}) - \mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0}))'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}(\mathbf{s_0}) - \mathbf{X}\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0}))), \end{aligned} \tag{7}$$

where $\mathbf{c}_Y(\mathbf{s_0}) = cov(Y(\mathbf{s_0}), \mathbf{Z})$ describes the cross-covariance between $Y(\mathbf{s_0})$ and the observed data $\mathbf{Z}$, generated through the covariance function in (3), and $\widehat{\boldsymbol{\alpha}}_{gls} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$, is the generalized-least squares estimator of $\boldsymbol{\alpha}$. Computational problems of speed and storage may arise in the calculation of the inverse of the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$, which is needed for kriging predictions and variances in (6) and (7) and requires $O(n^3)$ computations. This becomes even more difficult in iterative Maximum-Likelihood parameter estimations, where $\boldsymbol{\Sigma}^{-1}$ has to be calculated in each iteration step. Another potential shortage can be identified for the case of a large number of prediction locations, for which the $m \times n$ cross-covariance matrix might become very large and needs huge amounts of storage in the current workspace. The following approaches have been recently developed to tackle the large-matrix-problem by applying a low-rank approximation of the spatial process $\nu(\cdot)$ (e.g. Cressie and Johannesson (2006), Shi and Cressie (2007), Cressie and Johannesson (2008) and Katzfuss and Cressie (2009)), by introducing sparseness to $\boldsymbol{\Sigma}$ (Furrer, Genton, and Nychka (2006)) and a combination of both approaches (Sang and Huang (2012)).

2. Approximating the Spatial Covariance Function

2.1. Fixed Rank Kriging

As a way of dealing with the inversion of the $n \times n$ covariance matrix in a large data setting, Cressie and Johannesson (2006, 2008) proposed to approximate the spatial process $\nu(\cdot)$ in

(2) by a vector $\boldsymbol{\eta}$ of r random effects with $r << n$ and a corresponding set of spatial basis functions $\mathbf{S}(\cdot)$. The model for $\nu(\cdot)$, which the authors call *spatial random-effects model*, is

$$\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}), \qquad ,\mathbf{s} \in D \tag{8}$$

and the corresponding smooth process $Y(\cdot)$ becomes

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\alpha} + \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}) \qquad ,\mathbf{s} \in D \tag{9}$$

which results in a *spatial mixed effects model*, where $\mathbf{S}(\mathbf{s}) \equiv (S_1(\mathbf{s}), \cdots, S_r(\mathbf{s}))'$ is the set of r basis functions evaluated at location $\mathbf{s} \in D$, and $\boldsymbol{\eta}$ is a r-dimensional zero-mean vector of random effects with $r \times r$ dimensional covariance matrix $var(\boldsymbol{\eta}) = \mathbf{K}$. The zero-mean micro-scale variation process $\xi(\cdot)$ with variance $\sigma_\xi^2 v_\xi(\cdot)$ accounts for the spatial variation not explained by the dimension reduced model. Assuming that the micro-scale variation $\xi(\cdot)$ is white-noise in space and that $\boldsymbol{\eta}$ and $\xi(\mathbf{s})$ are independent the covariance function of $\nu(\cdot)$ is consequently

$$C(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}), \qquad \mathbf{u}, \mathbf{v} \in D. \tag{10}$$

It is important to note, that no assumptions about stationarity or isotropy are made in the spatial random-effects model. The resulting theoretical covariance matrix $\boldsymbol{\Sigma}$ of the data process is

$$\boldsymbol{\Sigma} = \mathbf{S}\mathbf{K}\mathbf{S}' + \sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi, \tag{11}$$

with $\mathbf{S}$ being the $n \times r$ dimensional matrix of basis functions evaluated at each observed location $\mathbf{s}_1, \cdots, \mathbf{s}_n$ and $\mathbf{V}_\xi \equiv diag\{v_\xi(\mathbf{s}_1), \cdots, v_\xi(\mathbf{s}_n)\}$ covering the heterogeneity of the small-scale spatial variation. This representation of the covariance matrix allows for the application of the Sherman-Morrison-Woodsbury formula as in (12) (Henderson and Searle (1981, p. 53))

$$(\mathbf{A} + \mathbf{U}\mathbf{B}\mathbf{V})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{I} + \mathbf{B}\mathbf{V}\mathbf{A}^{-1}\mathbf{U})^{-1}\mathbf{B}\mathbf{V}\mathbf{A}^{-1} \tag{12}$$

and consequently the inverse of (11) can be written as

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} \quad = \quad & (\sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi)^{-1} \\ & -(\sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi)^{-1}\mathbf{S}\left\{\mathbf{K}^{-1} + \mathbf{S}'(\sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi)^{-1}\mathbf{S}\right\}^{-1}\mathbf{S}'(\sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi)^{-1}. \end{aligned} \tag{13}$$

Obviously using this representation only the inverse of the fixed-rank $r \times r$ matrix $\mathbf{K}$ and the diagonal $n \times n$ matrix $(\sigma_\epsilon^2\mathbf{V}_\epsilon + \sigma_\xi^2\mathbf{V}_\xi)$ describing the nugget effect are needed. In that way great savings in terms of storage and reductions of computing time can be achieved. As stated in Cressie and Johannesson (2008, p. 214) the computational cost reduces from $O(n^3)$ to $O(nr^2)$ and accordingly rises only linear with the size of the dataset. With the setting of (8)-(13) the corresponding kriging prediction and variance for the prediction location $\mathbf{s}_0 \in D$ are

$$\widehat{\mathbf{Y}}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)'\widehat{\boldsymbol{\alpha}}_{gls} + \mathbf{c}_Y(\mathbf{s}_0)'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\alpha}}_{gls}) \tag{14}$$

and

$$\begin{aligned} \widehat{\sigma}^2(\mathbf{s_0}) \quad = \quad & \mathbf{S}(\mathbf{s_0})'\mathbf{K}\mathbf{S}(\mathbf{s_0}) - \mathbf{c}_Y(\mathbf{s_0})'\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0}) \\ & +(\mathbf{x}(\mathbf{s_0}) - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0}))'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}(\mathbf{s_0}) - \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s_0})) \end{aligned} \tag{15}$$

with

$$\mathbf{c}_Y(\mathbf{s_0}) = \mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s_0}) + \sigma_\xi^2 v_\xi(\mathbf{s_0})I(\mathbf{s_0} \in \{\mathbf{s}_1, \cdots, \mathbf{s}_n\}), \tag{16}$$

and $\mathbf{S}(\mathbf{s_0})$ is the r dimensional vector of basis functions evaluated at the prediction location $\mathbf{s_0}$.

Basis Function Selection

Important for the spatial random-effects model in (8) is the specification of the basis functions. The application of basis functions to approximate non-stationary covariance functions has already been discussed in literature, e.g. in Nychka, Wikle, and Royle (2002). Popular classes of functions are smoothing spline basis functions (e.g. in Wahba (1990)), wavelet basis functions (e.g. in Vidakovic (1999)) and radial basis functions (e.g. in Hastie, Tibshirani, and Friedman (2003)). An overview of available classes is also given in Wikle (2010). Since the predictions in Section 3 are required for a sphere, namely the globe, the class of multi-resolutional local bisquare functions is used (as suggested in Cressie and Johannesson (2008))

$$S_{i,l}(\mathbf{s}) = \begin{cases} \left[1 - (\parallel \mathbf{s} - \mathbf{m}_{i,l} \parallel / r_l)^2\right]^2 & \text{if } \parallel \mathbf{s} - \mathbf{m}_{i,l} \parallel \leq r_l \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

where $\mathbf{m}_{i,l}$ is the center point of the ith basis function in resolution level l and r_l is defined through Cressie and Johannesson (2006) as

$$r_l = (1.5) \times (\text{shortest distance between the center points in resolution level } l).$$

By specifying multiple resolutions the covariance model is able to cover different scales of spatial variation. Along with the type of basis function the locations of the center points have to be specified. Ideally they should cover the entire domain and be equidistant. This can be achieved through the application of a multi-resolutional *Discrete Global Grid (DGG)* of hexagons (see Sahr, White, and Kimerling (2003)). In Figure 1 a DGG (ISEA3H[1]) was generated for the globe with 4 different resolutions, which is later used for the analysis of atmospheric carbon dioxide concentrations. Center points below 60 degrees South have been deleted, since no measurements of the satellite can be obtained from that area. The corresponding inter-cell distances, measured in great-arc distances, are 4430.85 km for resolution 1 (red dots) and 2558.15 km, 1476.95 km and 852.71 km for resolution 2 (blue dots), 3 (green dots) and 4 (black dots) respectively. Depending on how much resolutions are considered, there are 29, 116, 370 or 1127 basis functions to evaluate for each of the n observations, resulting in an increasing dimension of $\mathbf{K}$. Hence, there is a trade-off between the explained spatial variation and the computational advantages obtained through the basis function approximation.

Parameter Estimation

The Fixed Rank Kriging approach requires estimates for the parameters $\sigma_\epsilon^2, \sigma_\xi^2, \mathbf{K}$ and $\boldsymbol{\alpha}$. As already mentioned, a suitable estimator for $\boldsymbol{\alpha}$ is the generalized-least squares estimator $\widehat{\boldsymbol{\alpha}}_{gls} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Z}$. Since σ_ϵ^2 and σ_ξ^2 are not individually identifiable, σ_ϵ^2 is assumed to be known and can be estimated through the semi-variogram at spatial lags close to zero using robust variogram estimates (see Cressie and Hawkins (1980)). The intercept of a fitted line using Weighted Least Squares (see Cressie (1985)) represents the estimate for σ_ϵ^2. The parameters $\mathbf{K}$ and σ_ξ^2 of the spatial random-effects model can be estimated either through a

[1] Characteristics of different DGGs can be found on URL:http://webpages.sou.edu/ sahrk/dgg/isea/tables/tables.html

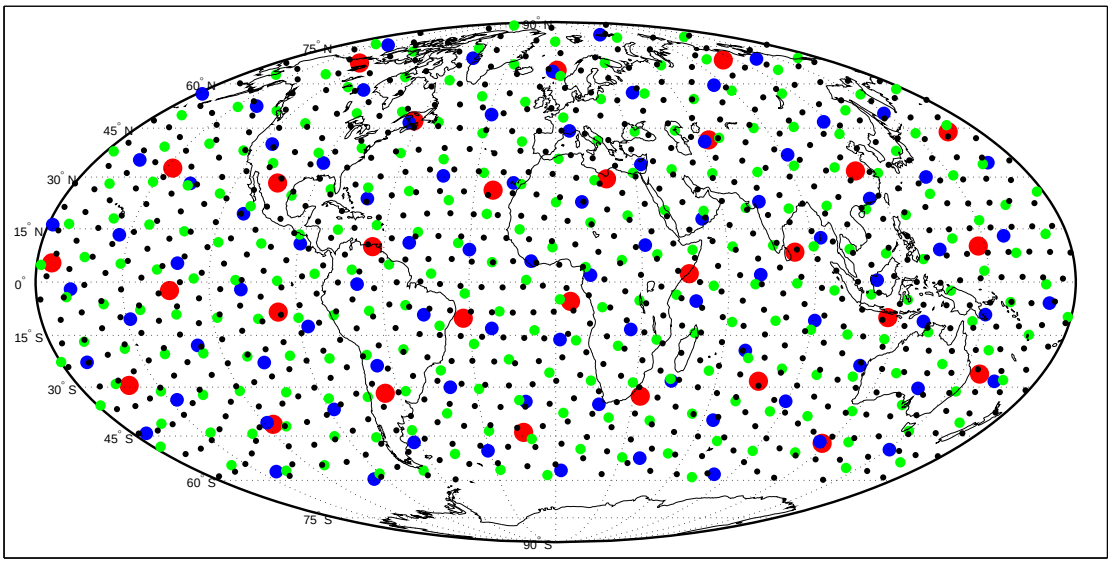


Figure 1: Basis function center points for 4 different resolutions of a Discrete Global Grid

Binned Method-of-Moments (BMoM) estimation procedure (as described in Cressie and Johannesson (2008)) or by Maximum Likelihood Estimation using an Expectation-Maximization (EM) algorithm (see Katzfuss and Cressie (2009)). In the BMoM approach the parameters $\mathbf{K}$ and σ_ξ^2 are estimated through minimizing a weighted Frobenius Norm between the theoretical covariance matrix $\mathbf{\Sigma}$ and an empirical counterpart obtained by binning the data. However, as the authors in Katzfuss and Cressie (2009) state, BMoM estimation is inferior in providing accurate estimates of prediction uncertainty, is much more complicate to apply and requires many subjective decisions. Therefore our focus is on the EM algorithm and the BMoM procedure will not be described in detail. The interested reader is referred to Cressie and Johannesson (2008).

Maximum Likelihood Estimation via EM-Algorithm

First, some distributional assumptions for the data have to be made in order to apply Maximum Likelihood estimation. For simplicity $\mathbf{Z}$ represents the vector of detrended data in this case and it is assumed that $\mathbf{Z}$ follows a multivariate normal distribution

$$\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{SKS}' + \sigma_\epsilon^2 \mathbf{V}_\epsilon + \sigma_\xi^2 \mathbf{V}_\xi).$$

Together with σ_ϵ^2, $\mathbf{V}_\epsilon$ and $\mathbf{V}_\xi$ assumed known, the Log-Likelihood becomes

$$\ell(\mathbf{K}, \sigma_\xi^2; \mathbf{Z}) \equiv \log f(\mathbf{Z}; \mathbf{K}, \sigma_\xi^2) = -\frac{1}{2} \log \det \mathbf{\Sigma} - \frac{1}{2} tr(\mathbf{\Sigma}^{-1} \mathbf{Z}\mathbf{Z}'). \tag{18}$$

However, as Katzfuss and Cressie (2009, p. 3381) state, finding estimates for $\mathbf{K}$ and σ_ξ^2 that maximize the likelihood equations is complicated. For that reason an EM algorithm (Dempster, Laird, and Rubin (1977)) is suggested. Instead of maximizing the likelihood of the observed data, it is assumed that knowing the distribution of some unobserved random variables, in this case $\boldsymbol{\eta} \sim N_r(\mathbf{0}, \mathbf{K})$ and $\boldsymbol{\xi} \sim N_n(\mathbf{0}, \sigma_\xi^2 \mathbf{V}_\xi)$ independently distributed, would

result in the joint distribution function of both, the observed and the missing data, which in turn can be maximized much easier. The algorithm consists of two steps. The first one is the Expectation step, in which the conditional expectation of the complete-data likelihood at a certain value of the parameter vector $\boldsymbol{\theta}^{[t]}$ at the t-th iteration, given the observed data, has to be calculated

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{[t]}) = E_{\boldsymbol{\theta}^{[t]}}\{\log f(\boldsymbol{\eta}, \boldsymbol{\xi}; \boldsymbol{\theta}) | \mathbf{Z}\}. \tag{19}$$

In the Maximization step the parameters are updated, so that (19) is maximized, which results in the updating scheme

$$\mathbf{K}^{[t+1]} = \mathbf{K}^{[t]} + \mathbf{K}^{[t]} \left(\mathbf{S}' {\boldsymbol{\Sigma}^{[t]}}^{-1} \left[\mathbf{Z}\mathbf{Z}' {\boldsymbol{\Sigma}^{[t]}}^{-1} - \mathbf{I}_n \right] \mathbf{S} \right) \mathbf{K}^{[t]} \tag{20}$$

$$\sigma_\xi^{2^{[t+1]}} = \sigma_\xi^{2^{[t]}} + \sigma_\xi^{2^{[t]}} tr \left(\frac{1}{n} {\boldsymbol{\Sigma}^{[t]}}^{-1} \left[\mathbf{Z}\mathbf{Z}' {\boldsymbol{\Sigma}^{[t]}}^{-1} - \mathbf{I}_n \right] \mathbf{V}_\xi \right) \sigma_\xi^{2^{[t]}}. \tag{21}$$

For a thorough derivation of the Expectation and Maximization steps the reader is referred to Katzfuss and Cressie (2009). Both steps are repeated until a convergence criterion based either on the change in the maximized likelihood or on the change in the parameter values is fulfilled. The estimates found are solutions to the likelihood equations. However, the user has to make subjective decisions concerning the convergence criterion and the starting values for the iteration procedure, which might influence both the efficiency and the accuracy of the algorithm. Furthermore the algorithm might lead to a local maximum depending on the choice of the initial values.

The outlined fixed rank kriging approach is very suitable for dealing with datasets of massive size. Due to the low dimensional spatial random effects vector $\boldsymbol{\eta}$ inverting the data covariance matrix $\boldsymbol{\Sigma}$ requires operations that rise only linear with the size of the dataset. In addition if the number of random effects r is sufficiently small no assumptions on the form of $var(\boldsymbol{\eta}) = \mathbf{K}$ are necessary and the restricting assumptions of stationarity and/or isotropy can be avoided. Fixed rank kriging is able to cover spatial variations on larger scales with a small number of basis functions. However in order to capture many scales of spatial variation of the phenomenon finer resolutions of basis functions, and consequently a larger r are needed. This reduces the computational efficiency and introduces the need of parametric covariance functions for $\mathbf{K}$ and simplifying assumptions. Obviously the choice of the basis functions is critical for the fixed rank kriging approach. Through choosing the number, the location and the type of the basis functions, the user is left with many subjective decisions, which possibly affect the outcome.

2.2. Covariance Tapering

Another way of efficiently dealing with $\boldsymbol{\Sigma}^{-1}$ is to introduce sparseness. With $\boldsymbol{\Sigma}$ and $\mathbf{c}_Y(\mathbf{s}_0)$ being sparse, significant computational savings in calculating the kriging predictions and variances in (6) and (7) can be achieved. The operation $\mathbf{u} = \boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\alpha}}_{gls})$ in (6) or $\mathbf{u} = \boldsymbol{\Sigma}^{-1}\mathbf{c}_Y(\mathbf{s}_0)$ in (7) can be solved efficiently through sparse matrix techniques based on the Cholesky factorization $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$. Then solving the triangular systems $\mathbf{A}\mathbf{w} = \mathbf{Z} - \mathbf{X}\widehat{\boldsymbol{\alpha}}_{gls}$ or $\mathbf{A}\mathbf{w} = \mathbf{c}_Y(\mathbf{s}_0)$ respectively and $\mathbf{A}'\mathbf{u} = \mathbf{w}$ yields the desired quantity. The computational complexity of the preceding operations is of order $O(nk^2)$, where k denotes the average number of non-zero elements in each row of $\boldsymbol{\Sigma}$, and consequently rises only linear with the size of the dataset.

Based on Furrer *et al.* (2006) sparseness can be introduced by setting covariances to zero for

observations more than a specific distance apart. The intuition behind this, is that observations far from the prediction location are not expected to have a large influence on the prediction and can therefore be neglected. Another argument for restricting to a local neighborhood is that even if the process inhibits long-range spatial dependence, the conditional correlation is expected to be very small after observing a closely located neighbor, since most of the information was already covered in the correlation with the neighboring observation. Let $C_{\boldsymbol{\theta}}(h)$ be a second order stationary and isotropic covariance function, with $h = ||\mathbf{s}_i - \mathbf{s}_j||$ and parameter vector $\boldsymbol{\theta}$. Using a taper function $T(h, \gamma)$, which is an isotropic and second order stationary covariance function with compact support, being equal to zero for $h \geq \gamma$, the tapered covariance function is the Schur product of $C_{\boldsymbol{\theta}}(\cdot)$ and $T(\cdot)$

$$C_{tap}(h, \gamma) = C_{\boldsymbol{\theta}}(h) \circ T(h, \gamma). \tag{22}$$

The tapered covariance function will also be a valid covariance function, since the Schur product of two positive definite matrices is again positive definite according to Horn and Johnson (1994, Theorem 5.2.1). An overview and some suggestions on choosing the type of taper function can be found in Furrer *et al.* (2006), including the spherical covariance function and functions from the Wendland family

$$T_{spherical}(h, \gamma) = \left(1 - \frac{h}{\gamma}\right)_+^2 \left(1 + \frac{h}{2\gamma}\right), \quad h > 0, \tag{23}$$

$$T_{wendland,1}(h, \gamma) = \left(1 - \frac{h}{\gamma}\right)_+^4 \left(1 + 4\frac{h}{\gamma}\right), \quad h > 0, \tag{24}$$

$$T_{wendland,2}(h, \gamma) = \left(1 - \frac{h}{\gamma}\right)_+^6 \left(1 + 6\frac{h}{\gamma} + \frac{35h^2}{2\gamma^2}\right), \quad h > 0. \tag{25}$$

Furrer *et al.* (2006) also investigated the asymptotic behavior of the kriging estimators with the tapered covariance function and proved that under certain conditions the estimator is asymptotically equivalent to the one obtained by using the original covariance function.

Parameter Estimation

The concept of tapering the covariance function not only improves the computational efficiency in kriging applications, in fact it can also be used in maximum likelihood estimation procedures, as it is demonstrated in Kaufman, Schervish, and Nychka (2008). In assuming multivariate normality a simple approximation of the log-likelihood function is obtained through replacing the model covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ by its tapered counterpart $\boldsymbol{\Sigma}_{tap} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \circ \mathbf{T}(\gamma)$, where $\mathbf{T}(\gamma)_{i,j} = T(||\mathbf{s}_i - \mathbf{s}_j||, \gamma)$. However this may lead to biased estimates in practice for small values of γ, as Kaufman *et al.* (2008, p. 1546) points out. For that reason the authors suggest to apply a two-taper approximation, where both the model and the sample covariance matrix are tapered. In this case $\mathbf{Z}$ represents the vector of detrended data and the one- and two-taper likelihood functions become

$$\ell_{1,taper}(\boldsymbol{\theta}) = -\frac{1}{2}\log\det(\boldsymbol{\Sigma}_{tap}) - \frac{1}{2}\mathbf{Z}'\boldsymbol{\Sigma}_{tap}^{-1}\mathbf{Z} \tag{26}$$

$$\ell_{2,taper}(\boldsymbol{\theta}) = -\frac{1}{2}\log\det(\boldsymbol{\Sigma}_{tap}) - \frac{1}{2}\mathbf{Z}'\left(\boldsymbol{\Sigma}_{tap}^{-1} \circ \mathbf{T}(\gamma)\right)\mathbf{Z}. \tag{27}$$

Maximizing $\ell_{2,taper}(\boldsymbol{\theta})$ leads to unbiased estimators, but at the cost of an increased computational complexity, as the two-taper approximation involves calculating the full inverse $\boldsymbol{\Sigma}_{tap}^{-1}$,

whereas the simple approach (one-taper approximation) only requires solving the sparse system of equations $\boldsymbol{\Sigma}_{tap}^{-1}\mathbf{Z}$. For the functional form of $C_Y(h, \boldsymbol{\theta})$ used to generate $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ there are many choices in the literature (see e.g. Cressie and Wikle (2011) for an overview) including the popular class of Matérn covariance functions (Matérn (1986)) defined as

$$C_Y(h, \sigma^2, \rho, \nu) = \frac{\sigma^2(h/\rho)^\nu}{\Gamma(\nu)2^{\nu-1}} K_\nu(h/\rho), \qquad h \geq 0, \sigma^2, \rho, \nu > 0, \tag{28}$$

with K_ν being the modified Bessel function of order ν (see Abramowitz and Stegun (1964)), σ^2 is the sill of the semi-variogram, ρ is a range parameter and ν controls for the smoothness of the process.
Despite Maximum Likelihood Estimation there is also the possibility for Variogram-model fitting using the empirical semi-variogram. Assuming stationarity and isotropy of the semi-variogram and the covariance function of the process Y, their relationship can be established through

$$\gamma_Y(||h||, \boldsymbol{\theta}) = C_Y(0, \boldsymbol{\theta}) - C_Y(||h||, \boldsymbol{\theta}). \tag{29}$$

An empirical estimate of $\gamma_Y(\cdot)$ (see Cressie and Wikle (2011, p. 131)) can be computed through

$$\begin{aligned} \widehat{\gamma}_Y(h) &= \frac{1}{2}\left(\widehat{\gamma}_Z(h) - \sigma_\epsilon^2\right) \\ &= \frac{1}{2}\left(ave\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : ||\mathbf{s}_i - \mathbf{s}_j|| \in T(h); i, j = 1, \cdots, n\} - \sigma_\epsilon^2\right), \end{aligned} \tag{30}$$

where $T(h)$ is a small tolerance region around h, resulting in a binning of the data. Using Least-Squares variogram fitting approaches a theoretical covariance function (e.g. (28)) can be fitted to the empirical semi-variogram by minimizing a weighted or non-weighted loss function (Cressie (1993)), given as

$$loss(\boldsymbol{\theta})_{OLS} = \sum_k \left(\widehat{\gamma}_k - \gamma_k(\boldsymbol{\theta})\right)^2 \tag{31}$$

$$loss(\boldsymbol{\theta})_{WLS} = \sum_k n_k \left(\frac{\widehat{\gamma}_k - \gamma_k(\boldsymbol{\theta})}{\gamma_k(\boldsymbol{\theta})}\right)^2, \tag{32}$$

where $\widehat{\gamma}_k$ denotes the value of the empirical semi-variogram for the kth bin, $\gamma_k(\boldsymbol{\theta})$ is the value of the theoretical semi-variogram in the kth bin and n_k is the number of pairs in the kth bin. Through tapering the covariance function many zeroes are introduced to $\boldsymbol{\Sigma}$ and the required operations rise only linear with the size of the dataset through the application of sparse matrix techniques. Disregarding covariances for locations with large distances only leads to a slight loss in predictive performance, since much of the information in the distant location is already covered in the correlation with closely located observations. In that way covariance tapering is able to efficiently capture spatial dependence at small spatial scales. However, depending on the choice of the taper range γ dependences at large scales are ignored and this might lead to a decline in predictive performance in regions with few data points. Another problem arises when using stationary covariance tapers on a non-stationary process. In this case small taper ranges are recommended in order to keep the bias small.

2.3. Full-Scale Approximation

Both approaches, the Fixed Rank Kriging and the Covariance Tapering, can be combined in a way, so that their advantages are fully exploited, as in Sang and Huang (2012). The former is able to efficiently capture the large-scale spatial dependence, since for that purpose only a small number of basis functions is needed (i.e. a small choice of r), however in order to describe local behaviour the dimension of $\mathbf{S}$ and $\mathbf{K}$ has to be increased accordingly. In contrast the Covariance Tapering is efficient for the spatial dependence at small spatial scales (i.e. a small choice of γ), whereas larger scales require larger taper ranges. In combining both approaches even small choices of r and γ are sufficient for providing a good approximation of the spatial dependence at the full scale. Consider the spatial random effects model in (8)

$$\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\boldsymbol{\eta} + \xi(\mathbf{s}), \qquad ,\mathbf{s} \in D,$$

where the residual process $\xi(\cdot)$ from the Fixed Rank Kriging is no longer assumed to be independent in space, but instead has the following covariance function approximated through Covariance Tapering

$$C_\xi(\mathbf{u},\mathbf{v}) = \left\{C_Y(\mathbf{u},\mathbf{v}) - \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v})\right\} \circ T(\mathbf{u},\mathbf{v},\gamma), \qquad \mathbf{u},\mathbf{v} \in D. \tag{33}$$

Consequently the covariance matrix $\boldsymbol{\Sigma}$ becomes

$$\boldsymbol{\Sigma} = \mathbf{S}\mathbf{K}\mathbf{S}' + \mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon, \tag{34}$$

where $\mathbf{C}_\xi$ denotes the sparse covariance matrix of the residual process $\xi(\cdot)$ generated through (33). The approximate log-likelihood function, ignoring the constant term and assuming normality on $\mathbf{Z}$ representing the vector of the detrended data, is

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log\det\left\{\mathbf{S}\mathbf{K}\mathbf{S}' + \mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon\right\} - \frac{1}{2}\mathbf{Z}'\left\{\mathbf{S}\mathbf{K}\mathbf{S}' + \mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon\right\}^{-1}\mathbf{Z}. \tag{35}$$

In evaluating the log-likelihood function, the inverse and the determinant of the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ are needed. The form of (34) allows for efficiently inverting $\boldsymbol{\Sigma}$ using the Sherman-Morrison-Woodsbury formula.

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{C}_\xi+\sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1} - (\mathbf{C}_\xi+\sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}\mathbf{S}\left\{\mathbf{K}^{-1} + \mathbf{S}'(\mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}\mathbf{S}\right\}^{-1}\mathbf{S}'(\mathbf{C}_\xi+\sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1} \tag{36}$$

The determinant of (34) can be computed through

$$\det(\boldsymbol{\Sigma}) = \det(\mathbf{K}^{-1} + \mathbf{S}'(\mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon)^{-1}\mathbf{S})\det(\mathbf{K}^{-1})^{-1}\det(\mathbf{C}_\xi + \sigma_\epsilon^2\mathbf{V}_\epsilon) \tag{37}$$

In computing (36) and (37) only inverses and determinants of sparse $n \times n$ and of $r \times r$ matrices are needed, which have a computational complexity of $O(nr^2 + nk^2)$. Alternatively the model can be fitted in a two-step procedure, so that the fixed rank kriging is estimated first through the efficient EM-Algorithm outlined in Section 2.2 and covariance tapering is applied to the residual process afterwards by applying variogram-model fitting as in Section 2.3. This has the advantage of avoiding the computational demanding maximization of the full likelihood. With this setting the full-scale approximation is able to combine the capabilities of the fixed rank kriging and the covariance tapering and to overcome their individual weaknesses. Whereas the r dimensional spatial random effect $\boldsymbol{\eta}$ captures large scale spatial dependence,

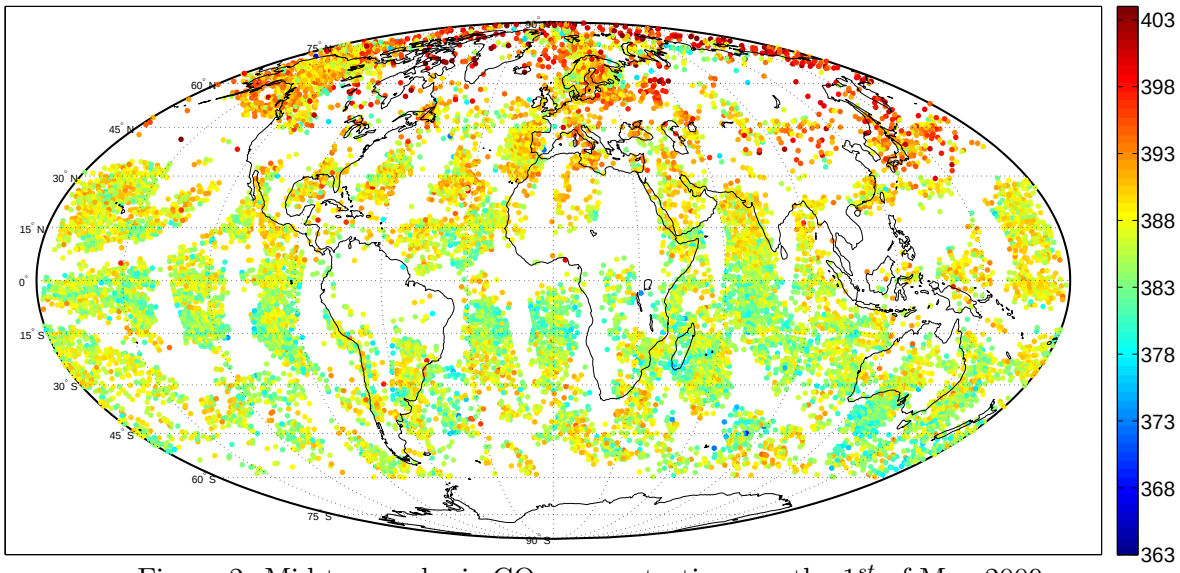


Figure 2: Mid-tropospheric CO_2 concentrations on the 1^{st} of May 2009

the residual process $\xi(\cdot)$ with tapered covariance function C_ξ efficiently describes local behavior. Importantly non-stationary and anisotropic behavior can be captured by the fixed rank part. However, an open problem remains the choice of the approximation parameters γ and r that lead to the most efficient outcome.

3. Efficiency Evaluation - Analysis of Atmospheric CO_2 Concentrations

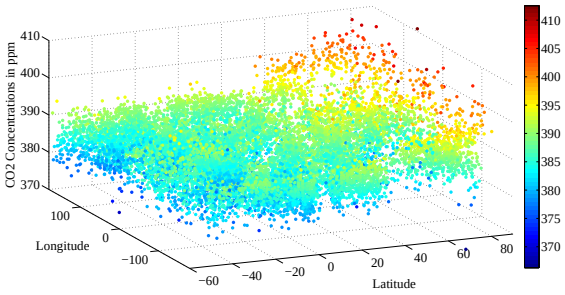
3.1. Data Description

The spatial dataset used for the comparative study consists of 12842 measurements of mid-tropospheric CO_2 concentrations obtained from the Atmospheric InfraRed Sounder (AIRS) on board NASA's Aqua satellite. The unit of measurement is *ppm* corresponding to 10^{-6} and denotes the number of CO_2 molecules in one million parts of air. This Level-2 product (AIRX2STC)[2] contains observations at 90×90 km nominal horizontal resolution at nadir, measured between -180° and 180° longitude and -60° and 90° latitude. However, in order to avoid change-of-support issues, it is assumed that measurements are at point support. The Aqua satellite reaches global coverage twice a day. Since this study considers spatial-only processes, the data are treated as if they were taken at one specific time point, neglecting the time discrepancy between measurements. The dataset consists of observations taken on the 1^{st} of May in 2009 and it already reveals some characteristic patterns of the natural carbon dioxide process. Particularly the data inhibits higher CO_2 concentrations and volatility in the northern hemisphere corresponding to a seasonal pattern caused by the growth stage of plants occurring in springtime.

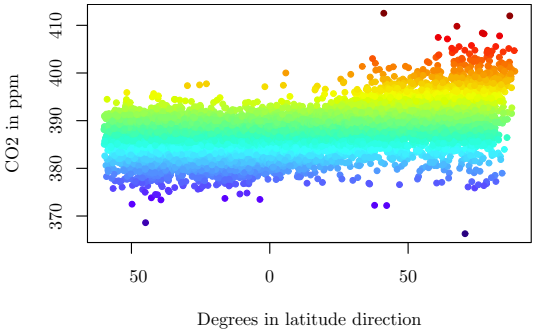
[2]Level 2 and 3 products are freely downloadable at http://disc.sci.gsfc.nasa.gov/AIRS/data-holdings

3.2. Preliminary Steps

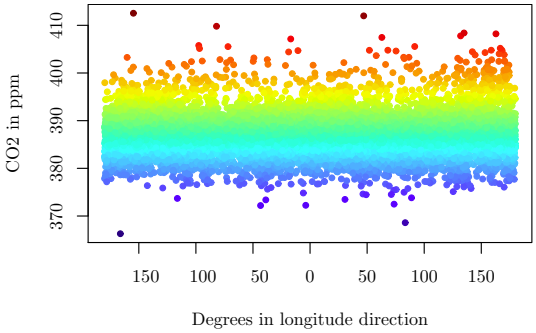
In order to evaluate the efficiency of the different approaches in approximating the spatial covariance function, it has to be ensured that the respective model assumptions are fulfilled. In Figures 3a, 3b and 3c the data is plotted against the degrees in longitude and/or latitude direction. Obviously the CO_2 process evolves differently in space depending on the orientation. Whereas a trend pattern can hardly be identified in the East-West direction, CO_2 concentrations tend to rise with decreasing distance to the north pole. Figure 3b also indicates that this latitude trend is of a non-linear kind. For that reason a non-linear regression was performed using polynomials up to order 3 for the latitude direction as covariates and a linear trend is assumed for the longitude direction. The subsequent analysis will be based on the detrended data. In addition the process variance appears to be higher in the northern hemisphere, as can be seen in Figure 3b. In Figures 3d and 3e empirical directional variograms of the detrended data with orientation 0° (North), 45° (North-East), 90° (East) and 135° (South-East) for the northern and southern hemisphere are shown. The empirical directional variograms were generated by using a tolerance angle of 22.5°. Importantly, since the spatial process evolves over the globe (for simplicity it is assumed that the earth is a perfect sphere with radius $R = 6371km$) great-circle distances have to be used. As can be seen in Figure 3e the empirical directional variograms for the data in the southern hemisphere are quite similar, indicating that deviations from isotropy can be regarded as small. However due to the seasonal effect in the northern hemisphere caused by the growth of plants in springtime the volatility is much higher. This can also be seen in the variograms in Figure 3d, where the sill is much higher in the East-West direction. In effect, the variogram becomes increasingly anisotropic with increasing degrees in latitude direction, leading to a non-stationary behavior of the process. Another indicator for non-stationarity is apparent in Figure 3f showing the estimated variance of the measurement error process depending on the degrees in latitude direction. Estimates for the nugget variance were computed for several subregions along the latitude direction using the approach mentioned in Section 2.2. Obviously measurements of the satellite become increasingly noisy the closer they are located to the poles. To account for this heterogeneity of the nugget variance a nonlinear regression was performed using polynomials up to order 4. This was used to provide values for $\mathbf{V}_\epsilon$ in (5). As has been shown, the spatial process is characterized by a non-stationary dependence structure and a variogram that varies with the orientation, depending on the degrees in latitude direction. Consequently the stationary covariance tapering can be regarded as inappropriate and adjustments to non-stationary and anisotropic covariance functions are needed for spatial predictions on a global scale. In contrast, the fixed rank kriging approximation is able to work without these assumptions and is therefore suitable for this problem. Likewise the full-scale approximation can handle non-stationary and anisotropic processes through its fixed rank part. However, to ensure comparability of the outlined approaches, a subset of the data consisting of 5073 measurements between $-20°$ and 20° latitude is considered first. For this region around the equator deviations from stationarity and isotropy can be neglected, as can be seen in Figure 4. The empirical directional variograms of the subset of the data (Figure 4a) indicate a comparable spatial dependence structure for all directions and as Figure 4b suggests, this property holds irrespectively of the location. The empirical omni-directional variograms were calculated at four equally spaced reference regions within the subset of the data. In addition, the variance of the measurement error of the instrument can be considered as constant in the subset of the data as can be seen in Figure 3f. In that way the subset serves as a sta-



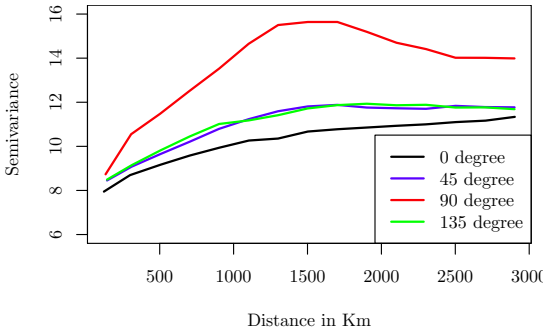
(a) 3D Scatter-Plot of the Data vs. Longitude and Latitude



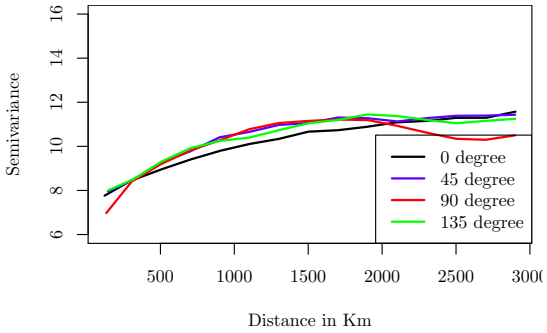
(b) Scatter-Plot of the Data vs. Latitude



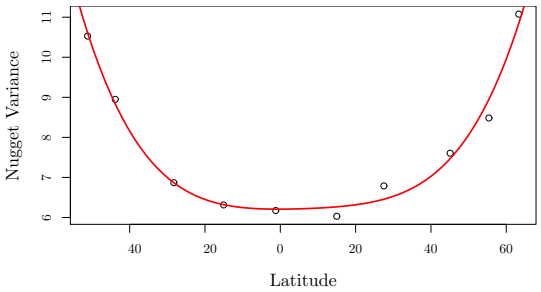
(c) Scatter-Plot of the Data vs. Longitude



(d) Empirical Directional Variograms (North)

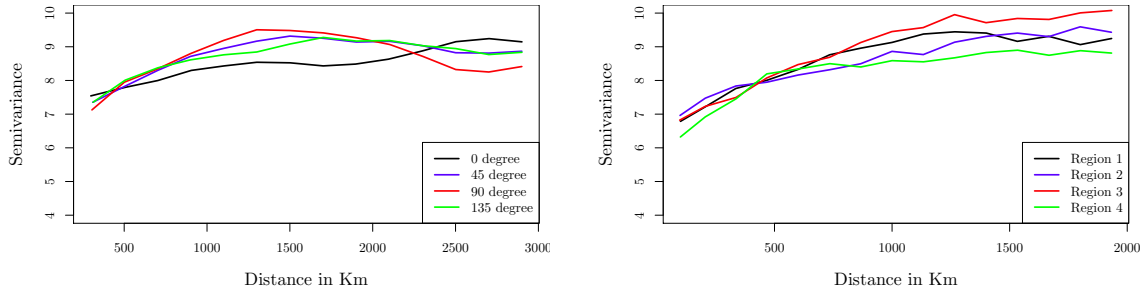


(e) Empirical Directional Variograms (South)



(f) Estimated Nugget Variance vs. Latitude

Figure 3: Exploratory Data Analysis



(a) Directional variogram of the subset of the data (b) Omni-directional variogram for different regions

Figure 4: Spatial dependence structure in the subset of the data

tionary scenario for evaluating the efficiency of the approximation approaches, which will be compared to the case, when these assumptions are not fulfilled.

3.3. Comparative Study

The focus of the study is to compare the efficiency of the approaches outlined in Section 2 in approximating the spatial covariance function by relating their predictive performance with the corresponding demand in computational resources. Both quantities are directly affected by the choice of the number of basis functions r and/or the taper range γ. Increasing values result in a higher approximation quality but at the same time in higher computing times and storage requirements. Consequently it is of interest, which approach is able to solve this trade-off best. For the choice of the basis functions either 1, 2 or 3 resolutions from the DGG in Figure 1 are selected, resulting in 10, 42 or 132 basis functions for the subset and in 29, 166 or 370 basis functions for the complete dataset, respectively. The taper range γ is varied between 50km and 1500km. This results in 3 Fixed Rank Kriging models, 10 Covariance Tapering models and in another 30 Full-scale approximations covering each parameter combination. The predictive performance is evaluated by a series of cross-validation experiments. For each model a 10-fold cross-validation is performed, where the dataset has been divided randomly into 10 subsamples. In each round one subsample is retained as a validation set for testing purposes and the remaining subsamples are used to fit the model. This procedure is repeated 10 times, so that each observation was part of the validation set once. The predictions of the validation set can then be compared to the original data to construct out-of-sample performance measures, whereas the MSPE will be used in this study

$$MSPE\left(\widehat{Y}(\mathbf{s}_0)\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\widehat{Y}(\mathbf{s}_i) - Y(\mathbf{s}_i)\right)^2 \qquad ,\mathbf{s}_0 = (\mathbf{s}_1, \cdots, \mathbf{s}_m)\,. \tag{38}$$

However, the MSPE has to be adjusted, since predictions are based on the smooth process $Y(\cdot)$ but only the noisy process $Z(\cdot)$ is observed and consequently the squared residuals would be affected by the measurement error variance. Recall that $\mathbf{Z} = \mathbf{Y} + \boldsymbol{\epsilon}$ and that $var(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 \mathbf{V}_\epsilon$ was assumed known and fitted through a polynomial function of order 4, so that the correct representation of the MSPE in the presence of measurement error can be obtained by subtracting the location specific nugget variance from the squared residual (see

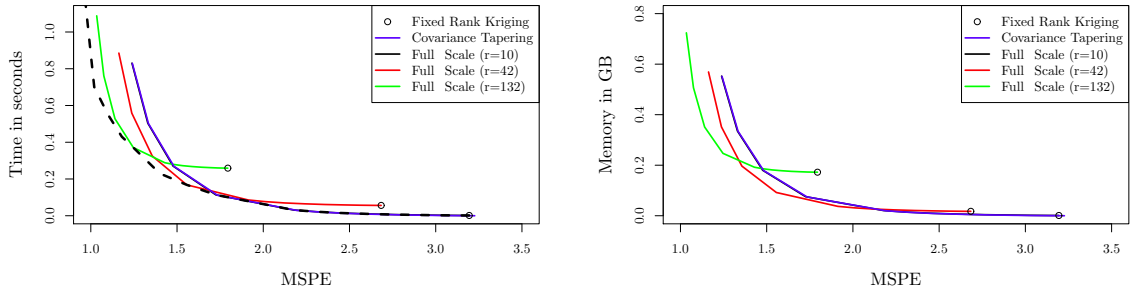
Cressie (1993, p. 128))

$$MSPE\left(\widehat{Y}(\mathbf{s}_0)\right) = \frac{1}{m}\sum_{i=1}^{m}\left\{\left(\widehat{Y}(\mathbf{s}_i) - Z(\mathbf{s}_i)\right)^2 - \sigma_\epsilon^2 v_\epsilon(\mathbf{s}_i)\right\} \qquad ,\mathbf{s}_0 = (\mathbf{s}_1, \cdots, \mathbf{s}_m)\,. \tag{39}$$

Besides the predictive performance, it is also of interest how much computational resources have been used by the models. In particular the computing time needed for calculating the important quantities in kriging predictions and in likelihood maximizations, which are the solution of the system of linear equations $\mathbf{\Sigma}^{-1}\mathbf{Z}$ and the determinant of $\mathbf{\Sigma}$, is monitored. Furthermore, the maximum amount of working memory used in these calculations is recorded, disregarding all preliminary calculations. However, it has to be noted that depending on how much prediction locations $\mathbf{s_0}$ are considered, the operation $\mathbf{c}_Y(\mathbf{s_0})\mathbf{\Sigma}^{-1}$ might also need significant amounts of working memory, especially for smooth prediction surfaces.

3.4. Subset Results - Stationary scenario

For the subset of the data, which serves as a stationary scenario, the trade-off between predictive performance and demand in computational resources is visualized in Figure 5. In Figure 5a the MSPE is plotted against the time in seconds needed to calculate the important quantities $\mathbf{\Sigma}^{-1}\mathbf{Z}$ and $\det\mathbf{\Sigma}$ and in Figure 5b the maximum amount of working memory in GB is shown. In comparing the fixed rank kriging (black dots) with the covariance tapering (blue line) it can be seen that the latter approach is more efficient in approximating the spatial covariance function, since for every level of the MSPE less or equal time and memory is needed. However it has to be noted, that this result strongly depends on the range of spatial dependence relative to the total extent of the spatial domain and the spatial distribution of the data locations. As denoted earlier, covariance tapering has advantages in describing local and fixed rank kriging in large-scale dependencies. Accordingly having a process with a small range of spatial dependence in relative terms will result in efficiency advantages for the covariance tapering. In contrast a high proportion of clustered data decreases the sparsity of $\mathbf{\Sigma}$ and increases the demand in computational resources for the covariance tapering. To overcome the individual weaknesses and to exploit the advantages of the fixed rank kriging and the covariance tapering their combination in a full-scale approximation leads to further efficiency gains, as can be seen in Figures 5a and 5b. For lower approximation qualities the full-scale approximation (black line) is slightly more efficient as the covariance tapering (blue line), however in order to achieve lower values of the MSPE it is worth including higher resolutions of basis functions in the full-scale approximation (red and green line) to further reduce the computational complexity. The complete summary of results for the 43 different models is shown in Table 3 in the Appendix, whereas a characteristic snapshot is shown in Table 1. To evaluate the overall quality of the approximations, the results for the full model without any approximation, i.e. the model with an untapered Matérn covariance function, as a baseline are shown. The full model achieved a MSPE of 0.975 and the approximation that came closest to that level is the full-scale approximation with 132 basis functions and a taper range of 1500km. As can be seen, almost the same predictive performance can be accomplished, but about 20 times faster and with only around 30% of the maximum working memory required. Table 1 also compares the efficiency of the approximations for a fixed level of the MSPE of about 1.75. Clearly the full-scale approximation outperforms the other approaches in terms of speed and storage, whereas the advantage over the covariance tapering is rather small for

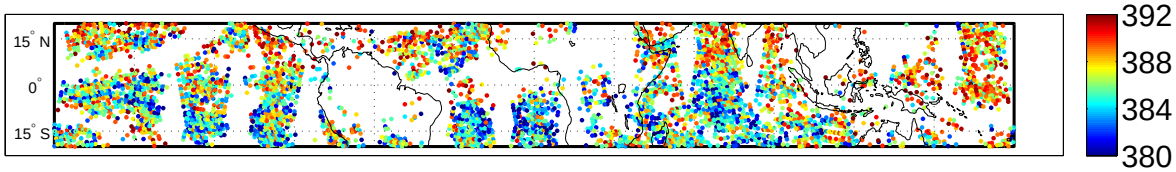


(a) MSPE vs. Computing Time (Subset) (b) MSPE vs. Maximum Memory Usage (Subset)

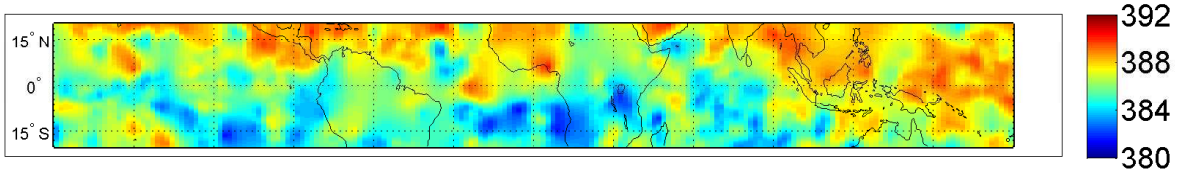
Figure 5: Efficiency Evaluation of the Covariance Approximization Approaches (Subset)

	Full Model	Full-scale Approximation r=132,γ=1500	Full-scale Approximation r=42,γ=625	Fixed Rank Kriging r=132	Fixed Rank Kriging r=42	Covariance Tapering γ=750	Covariance Tapering γ=625
MSPE	0.975	1.034	1.734	1.795	2.684	1.734	1.963
Time $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	11.086	0.76104	0.06663	0.18632	0.01967	0.06561	0.04159
Time det $\boldsymbol{\Sigma}$	12.726	0.32702	0.03035	0.07271	0.00656	0.04635	0.02916
Memory $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	2.464	0.72400	0.06453	0.17236	0.01745	0.07450	0.04708
Memory det $\boldsymbol{\Sigma}$	2.365	0.23800	0.02121	0.05666	0.00574	0.02449	0.01548

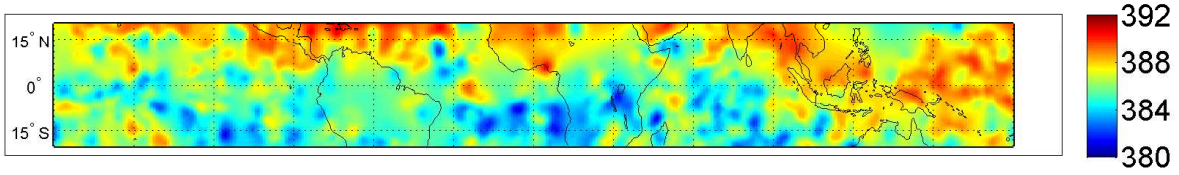
Table 1: Efficiency Evaluation - Subset



(a) The Data Subset

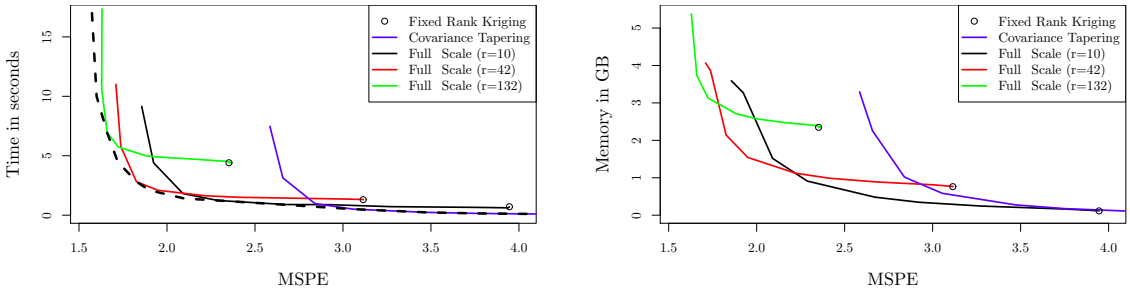


(b) Prediction surface using the full model



(c) Prediction surface using the full-scale approximation

Figure 6: Prediction Surfaces of the Subset



(a) MSPE vs. Computing Time

(b) MSPE vs. Maximum Memory Usage

Figure 7: Efficiency Evaluation of the Covariance Approximization Approaches

this level of the MSPE, but becomes much larger for better approximation qualities. Finally Figure 6 shows kriging surfaces of the full model (Figure 6b) and the full-scale approximated model with $r = 132$ and $\gamma = 1500$km (Figure 6c), whereas 250000 prediction locations were used to produce the latter and, due to memory restrictions, only 40000 pixels can be produced for the full model. As the similarity of both plots indicate, the quality of the approximation is very good and comes together with remarkably high computational savings.

3.5. Global Dataset Results - Non-stationary Scenario

As outlined in Section 3.2 the global dataset is characterized by a non-stationary and anisotropic dependence structure, which affects the efficiency of the stationary covariance tapering. Compared to the stationary case, the efficiency curves (blue lines) are shifted to the right in Figures 7a and 7b. The stationary covariance tapering is not able to provide high quality approximations of the non-stationary covariance function, because the increasing process variance in the northern hemisphere is not captured. Consequently estimated prediction errors will not yield reliable estimates of the prediction uncertainty. In contrast the Fixed Rank Kriging (black dots) is able to account for the spatially varying dependence structure and yields lower values of the MSPE. Nevertheless the stationary covariance tapering is still more efficient for low approximation qualities. The directional semi-variograms in Figure 3d already revealed the anisotropic and non-stationary character of the dependence structure by the increased process variance in the northern hemisphere. However, at small spatial scales the spatial dependence patterns are very similar, despite the differing nugget variances. Consequently a stationary covariance tapering is still capable of providing good approximations of the covariance function in a local neighborhood although a non-stationary behavior is apparent at larger spatial scales. In that way, a full-scale approximation is again able to further increase efficiency and to supply high quality approximations of the spatial covariance function at all spatial scales, as indicated by the efficiency curves (black, red and green lines) in Figure 7. A complete summary of the results of all models can be found in Table 4 in the Appendix, whereas a characteristic snapshot is shown in Table 2. As in the subset, the results of the full model with an untapered stationary Matérn covariance function are provided for comparative purpose. However, analogously to the stationary covariance tapering, it does not show a high predictive performance and results in a MSPE of 2.556 accompanied by high computation times and a huge amount of 14 GB of working memory. Using the Full-scale approximation with $r = 29$ and $\gamma = 625$ a comparable value can be obtained about 335 times faster and with only about

	Full Model	Full-scale Approximation r=29,γ=625	r=29,γ=750	Fixed Rank Kriging r=370	r=29	Covariance Tapering γ=1500	γ=750
MSPE	2.556	2.480	2.288	2.352	3.946	2.658	3.055
Time $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	155.6	0.620	0.736	2.911	0.396	1.881	0.317
Time det $\boldsymbol{\Sigma}$	186.0	0.399	0.453	1.499	0.264	1.259	0.194
Memory $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	14.18	0.696	0.911	2.347	0.115	2.254	0.584
Memory det $\boldsymbol{\Sigma}$	13.79	0.210	0.272	1.058	0.046	0.815	0.230

Table 2: Efficiency Evaluation - Global Dataset

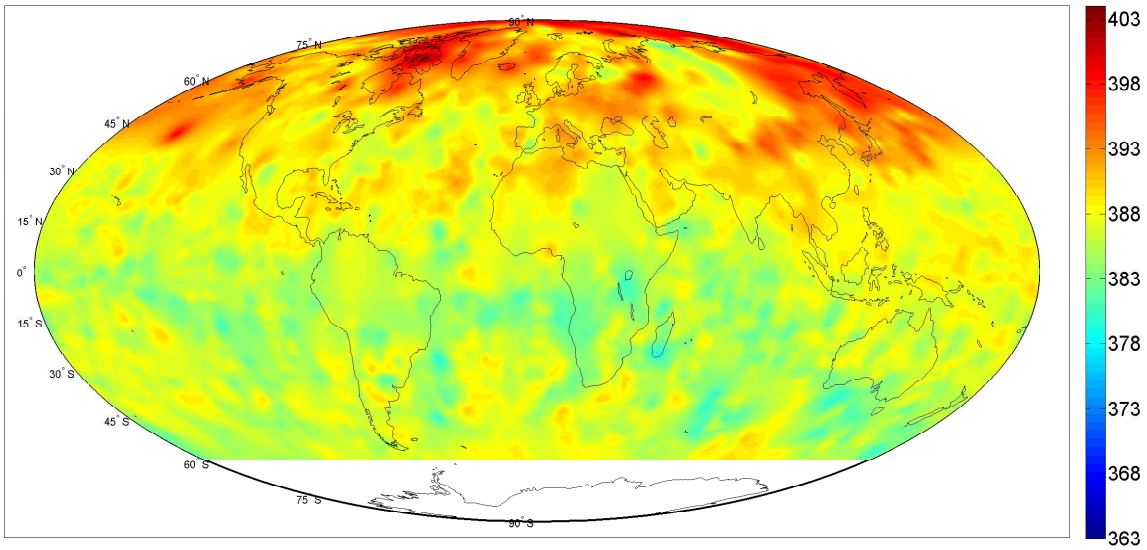


Figure 8: Mid-tropospheric CO_2 concentrations on the 1^{st} of May 2009

5% of the memory required at maximum. For comparing the efficiency of the approximation approaches a fixed level of about 2.3 of the MSPE is considered. Again the full-scale approximation was superior in terms of efficiency compared to both single approaches and the lead is even more pronounced in the non-stationary scenario than for the subset. Finally the full-scale approximation can be used to compute a high quality prediction surface for the process of atmospheric CO_2 concentrations over the globe, as it is shown in Figure 8, where a kriging surface containing 250000 prediction locations for the full-scale approximation with $r = 370$ and $\gamma = 1500$km was produced.

3.6. Choice of the approximation parameters

The trade-off between predictive performance and computational complexity can be directly controlled through the choice of the approximation parameters r and γ in the full-scale approximation. In Figures 5a and 7a this trade-off was illustrated for a fixed number of basis functions r through the black, red and green lines. However these Figures can also give an idea on how the overall efficiency curve of the full-scale approximation would look like, as the enveloping black doted lines sketch. These curves depict the optimal combinations of r and γ yielding the lowest achievable computational complexity at all levels of the MSPE. Interestingly the tangency points of the black, red and green lines on the hypothetical black

doted line correspond to a certain taper range γ^* between 750km and 1000km. Consequently Full-scale approximations with a taper range higher/lower than γ^* are always dominated through models with more/less basis functions. Moreover γ^* appears to coincide with the estimated effective range of the fitted Matérn covariance function. Intuitively this also makes sense, since this is exactly the scale of spatial dependence where the covariance tapering has advantages over the fixed rank kriging.

4. Conclusions

This paper investigated approaches to approximate the spatial covariance function and analyzed the trade-off between the loss in information due to the approximation and the reductions in computational complexity. Based on a remotely sensed data set of carbon dioxide concentrations in the mid-troposphere an efficiency evaluation was conducted, monitoring the predictive performance through the MSPE and the computational complexity through computation speed and storage requirements. All outlined approaches, namely fixed rank kriging (Cressie and Johannesson (2006, 2008)), covariance tapering (Furrer *et al.* (2006)) and the full-scale approximation (Sang and Huang (2012)) were able to notably speed up the calculations of the important quantities in maximum likelihood estimation and in kriging predictions, which are the determinant of the covariance matrix of the observed data $\mathbf{\Sigma}$ and the solution of the system of linear equations $\mathbf{\Sigma}^{-1}\mathbf{Z}$. The required computations rise only linear with the size of the data set, instead of cubic. However, depending on the degree of the approximation, controlled by parameters r as the number of random effects in the fixed rank kriging approach and γ as the taper range in the covariance tapering approach, the loss in predictive performance differs substantially. In a subset of the data, where the process can be regarded as stationary, it was shown that covariance tapering outperformed the fixed rank kriging. However through combining both approaches in a full-scale approximation even more efficient approximations can be generated. The individual weaknesses, namely the inefficiency of the fixed rank kriging to describe local spatial dependence and of the covariance tapering to cover large-scale spatial dependence, can be overcome. In the full data set, involving a strong non-stationary behavior in the latitude direction, the advantage of the fixed rank kriging is apparent, since no assumptions on stationarity and/or isotropy have to be made and therefore the increased process variance in the northern hemisphere can be easily captured. This feature also translates into the full-scale approximation. Interestingly the analysis gives an idea on how to choose the approximation parameters r and γ optimally. For each level of the MSPE the most efficient combination of parameters involves a certain taper range γ^*, which coincides with the effective range of the fitted Matérn covariance function for the CO_2 example. However a thorough investigation of the optimal choice of the approximation parameters, i.e. model selection, is left open for future research.

Table 3: Efficiency Evaluation - Subset

Model Type	Parameters r	Parameters γ	MSPE	Time in sec $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Time in sec $\det \boldsymbol{\Sigma}$	Memory in GB $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Memory in GB $\det \boldsymbol{\Sigma}$
Full Model			0.975	11.0866	12.7267	2.4645	2.3657
Fixed Rank Kriging	10	0	3.194	0.00111	0.00037	0.00099	0.00033
	42	0	2.684	0.01967	0.00656	0.01745	0.00574
	132	0	1.795	0.18632	0.07271	0.17236	0.05666
Covariance Tapering	0	50	3.216	0.00001	0.00001	0.00001	0.00001
	0	100	3.186	0.00005	0.00003	0.00005	0.00002
	0	200	3.055	0.00067	0.00041	0.00072	0.00024
	0	300	2.783	0.00304	0.00185	0.00325	0.00107
	0	400	2.466	0.00826	0.00526	0.00900	0.00296
	0	500	2.192	0.01758	0.01197	0.01966	0.00646
	0	625	1.963	0.04159	0.02916	0.04708	0.01548
	0	750	1.734	0.06561	0.04635	0.07450	0.02449
	0	1000	1.482	0.15570	0.11324	0.17895	0.05883
	0	1500	1.241	0.47393	0.35510	0.55164	0.18134
Full-scale Approximation (r=10)	10	50	3.184	0.00113	0.00038	0.00100	0.00033
	10	100	3.155	0.00115	0.00041	0.00104	0.00034
	10	200	3.026	0.00182	0.00074	0.00171	0.00056
	10	300	2.759	0.00438	0.00200	0.00424	0.00140
	10	400	2.447	0.01292	0.00209	0.00999	0.00328
	10	500	2.177	0.02048	0.01055	0.02065	0.00679
	10	625	1.951	0.04693	0.02531	0.04807	0.01580
	10	750	1.726	0.07337	0.04008	0.07549	0.02481
	10	1000	1.477	0.16990	0.10052	0.17994	0.05915
	10	1500	1.239	0.51520	0.31532	0.55263	0.18167
Full-scale Approximation (r=42)	42	50	2.676	0.01968	0.00656	0.01746	0.00574
	42	100	2.654	0.01958	0.00673	0.01750	0.00575
	42	200	2.558	0.02024	0.00706	0.01817	0.00597
	42	300	2.358	0.02266	0.00845	0.02070	0.00681
	42	400	2.121	0.02801	0.01173	0.02645	0.00869
	42	500	1.912	0.03929	0.01648	0.03711	0.01220
	42	625	1.734	0.06663	0.03035	0.06453	0.02121
	42	750	1.556	0.09396	0.04422	0.09195	0.03023
	42	1000	1.356	0.19425	0.10091	0.19640	0.06456
	42	1500	1.163	0.55367	0.30158	0.56909	0.18708
Full-scale Approximation (r=132)	132	50	1.791	0.18665	0.07240	0.17237	0.05666
	132	100	1.780	0.18651	0.07260	0.17241	0.05668
	132	200	1.736	0.18797	0.07214	0.17308	0.05690
	132	300	1.644	0.19186	0.07206	0.17561	0.05773
	132	400	1.531	0.19894	0.07361	0.18136	0.05962
	132	500	1.429	0.21101	0.07757	0.19202	0.06312
	132	625	1.338	0.23900	0.09079	0.21944	0.07214
	132	750	1.247	0.26698	0.10401	0.24686	0.08115
	132	1000	1.140	0.37446	0.15351	0.35131	0.11549
	132	1500	1.034	0.76104	0.32702	0.72400	0.23800

Table 4: Efficiency Evaluation - Global Dataset

Model Type	Parameters r	Parameters γ	MSPE	Time in sec $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Time in sec $\det \boldsymbol{\Sigma}$	Memory in GB $\boldsymbol{\Sigma}^{-1}\widetilde{\mathbf{Z}}$	Memory in GB $\det \boldsymbol{\Sigma}$
Full Model			2.556	155.6	186.0	14.18	13.79
Fixed Rank Kriging	29	0	3.946	0.396	0.264	0.115	0.046
	166	0	3.115	0.620	0.385	0.762	0.341
	370	0	2.352	2.911	1.499	2.347	1.058
Covariance Tapering	0	50	5.069	0.008	0.005	0.009	0.003
	0	100	5.034	0.011	0.007	0.013	0.005
	0	200	4.726	0.030	0.019	0.045	0.017
	0	300	4.149	0.059	0.035	0.102	0.039
	0	400	3.752	0.098	0.055	0.175	0.068
	0	500	3.474	0.153	0.085	0.274	0.106
	0	625	3.264	0.235	0.140	0.429	0.168
	0	750	3.055	0.317	0.194	0.584	0.230
	0	1000	2.840	0.600	0.380	1.017	0.392
	0	1500	2.658	1.881	1.259	2.254	0.815
Full-scale Approximation (r=29)	29	50	3.946	0.396	0.264	0.115	0.046
	29	100	3.940	0.303	0.266	0.120	0.048
	29	200	3.752	0.330	0.278	0.165	0.060
	29	300	3.277	0.373	0.294	0.244	0.082
	29	400	2.925	0.519	0.315	0.345	0.111
	29	500	2.672	0.503	0.345	0.482	0.148
	29	625	2.480	0.620	0.399	0.696	0.210
	29	750	2.288	0.736	0.453	0.911	0.272
	29	1000	2.091	1.130	0.640	1.517	0.435
	29	1500	1.924	2.839	1.518	3.273	0.858
Full-scale Approximation (r=166)	166	50	3.115	0.620	0.385	0.762	0.341
	166	100	3.116	0.632	0.388	0.767	0.343
	166	200	3.006	0.659	0.399	0.811	0.356
	166	300	2.677	0.722	0.419	0.889	0.377
	166	400	2.411	0.769	0.436	0.988	0.406
	166	500	2.220	0.864	0.465	1.123	0.444
	166	625	2.085	1.043	0.520	1.334	0.506
	166	750	1.949	1.221	0.574	1.545	0.568
	166	1000	1.827	1.754	0.759	2.141	0.730
	166	1500	1.738	3.838	1.639	3.867	1.152
Full-scale Approximation (r=370)	370	50	2.352	2.911	1.499	2.347	1.058
	370	100	2.368	2.928	1.496	2.352	1.060
	370	200	2.337	3.008	1.515	2.396	1.073
	370	300	2.160	3.157	1.552	2.474	1.094
	370	400	2.002	3.276	1.566	2.574	1.123
	370	500	1.885	3.401	1.577	2.708	1.161
	370	625	1.804	3.726	1.632	2.919	1.222
	370	750	1.724	4.050	1.686	3.129	1.284
	370	1000	1.660	5.047	1.884	3.734	1.447
	370	1500	1.629	7.899	2.707	5.372	1.869

Acknowledgements

This work was supported by the German Ministry of Education and Research (BMBF) under its funding program 'Economics of Climate Change' [grant number 01LA1139A]

References

Abramowitz M, Stegun IA (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York.

Cressie N (1985). "Fitting variogram models by weighted least squares." *Journal of the International Association for Mathematical Geology*, **17**(5), 563–586.

Cressie N (1993). *Statistics for Spatial Data*. Revised edition edition. Wiley-Interscience.

Cressie N, Hawkins D (1980). "Robust estimation of the variogram: I." *Mathematical Geology*, **12**(2), 115–125.

Cressie N, Johannesson G (2006). "Spatial prediction of massive datasets." In *Proceedings of the Australian Academy of Science Elizabeth and Frederick White Conference*, volume 1247, pp. 1–11.

Cressie N, Johannesson G (2008). "Fixed rank kriging for very large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 209–226.

Cressie N, Wikle C (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Wiley.

Dempster A, Laird N, Rubin D (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39**(1), 1–38.

Furrer R, Genton MG, Nychka D (2006). "Covariance tapering for interpolation of large spatial datasets." *Journal of Computational and Graphical Statistics*, **15**(3), 502–523.

Guan D, Liu Z, Geng Y, Lindner S, Hubacek K (2012). "The gigatonne gap in China's carbon dioxide inventories." *Nature Climate Change*, **2**, 672–675.

Hastie T, Tibshirani R, Friedman J (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Henderson H, Searle S (1981). "On deriving the inverse of a sum of matrices." *Siam Review*, **23**(1), 53–60.

Horn RA, Johnson CR (1994). *Topics in Matrix Analysis*. Cambridge University Press.

Katzfuss M, Cressie N (2009). "Maximum likelihood estimation of covariance parameters in the spatial-random-effects model." In *Proceedings of the Joint Statistical Meetings*, pp. 3378–3390. American Statistical Association.

Kaufman CG, Schervish MJ, Nychka DW (2008). "Covariance tapering for likelihood-based estimation in large spatial data sets." *Journal of the American Statistical Association*, **103**(484), 1545–1555.

Matérn B (1986). *Spatial Variation*. Lecture Notes in Statistics. Springer-Verlag.

Mintzer I, Leonard J, Valencia I (2010). "Counting the Gigatonnes: Building trust in greenhouse gas inventories from the United States and China." *World Wildlife Federation*.

Nychka D, Wikle C, Royle JA (2002). "Multiresolution models for nonstationary spatial covariance functions." *Statistical Modelling*, **2**(4), 315–331.

Sahr K, White D, Kimerling AJ (2003). "Geodesic discrete global grid systems." *Cartography and Geographic Information Science*, **30**(2), 121–134.

Sang H, Huang J (2012). "A full scale approximation of covariance functions for large spatial data sets." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**(1), 111–132.

Shi T, Cressie N (2007). "Global statistical analysis of MISR aerosol data: a massive data product from NASA's Terra satellite." *Environmetrics*, **18**(7), 665–680.

Vidakovic B (1999). *Statistical Modeling by Wavelets*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Wahba G (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics.

Wikle C (2010). "Low-rank representations for spatial srocesses." In *Handbook of Spatial Statistics*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, pp. 107–118. CRC Press.

Affiliation:

Patrick Gneuss
Department of Statistics
Europa-Universität Viadrina Frankfurt (Oder), Germany
Postal Code 15230
E-mail: vetter@europa-uni.de