Environmental monitoring is a task that requires to surrogate system-wide information with limited sensor readings. Under the proximity principle, an environmental monitoring system can be based on the virtual sensing logic and then rely on distance-based prediction methods, such as $k$-nearest-neighbors, inverse distance weighted regression and spatio-temporal kriging. The last one is cumbersome with large datasets, but we show that a suitable separability assumption reduces its computational cost to an extent broader than considered insofar. Only spatial interpolation needs to be performed in a centralized way, while forecasting can be delegated to each sensor. This simplification is mostly related to the fact that two separate models are involved, one in time and one in the space domain. Any of the two models can be replaced without re-estimating the other under a composite likelihood approach. Moreover, the use of convenient spatial and temporal models eases up computation. We show that this perspective on kriging allows to perform virtual sensing even in the case of tall datasets.

Keywords Spatio-temporal kriging · Distance-based prediction · Separability · Isotropy · Indoor environments · Composite likelihood · Distributed calculus

1 Introduction

Environmental monitoring systems rely on virtual sensing logic to predict relevant variables of their target environment. While the information on the whole system is of interest, this is typically based on sensor readings, which are limited in both space and time, so it is necessary to surrogate them, based on some suitable statistical method [1]. Variables of interest may include, for instance, room temperature [2], energy consumption [3] and air quality [4]. We consider the case of enclosed environments [5][2][4], as contrasted to other applications that are aimed at larger environments like ecosystems [6][7]. Moreover, our focus is on applications that need real-time control [8].

The motivating example for this work comes from a virtual sensing project at Silicon Austria Labs GmbH, a European research center for electronic based systems [9]. The data relate to an office room in Villach, Austria, that has been monitored for 19 weeks between October 2019 and March 2020. The temperature (in °C) is reported by twelve sensors every 10 seconds, along with other physical measurements like pressure (in Pa) and light (in lx). The room is 127 m$^2$ large and it is structured as reported in Figure 1. The picture also shows the locations of sensors and windows, along with the cardinal points.

The twelve sensors are all Raspberry Pi Zero boards. Their measurements are broadcast over a wireless network to a database server, which is a Raspberry Pi 3 instead. Raspberry Pis are popular and affordable single-board computers that are widely used in home automation, smart systems [10]. This example has some key aspects, including data referenced both temporally and spatially, high-frequency measurements, resulting in multivariate times series with as many as $10^6$ observations for each sensor. The server has a limited yet non-negligible computational power, which is also important, as it allows to process data locally if this task is planned carefully, by taking into consideration the limitations of the monitoring system.
Separable spatio-temporal kriging for fast virtual sensing

As common in modern data analysis, there are at least two main and opposite approaches to deal with sensor data. These two opposites are represented by interpretable models and black-box algorithms, respectively. The former include specifications based on actual knowledge about physical aspects of the system [8], often hard to formulate; the latter include neural networks and other machine learning techniques that achieve remarkable performance levels and are readily available in general software. Other authors addressed the same datasets of our present analysis [9], but they used techniques such as XGBoost regression [11] and LSTM recurrent neural networks [12]. These methods do not provide interpolation by design, but they can do so only after some suitable engineering, thus some generalization issues emerge.

Here we advocate for an approach lying between the two extremes, which is statistically sound without compromising prediction performance. We are interested in simple models and distributed computing, thus on scalable methods that leverage on the proximity principle. It is then natural to resort to distance-based prediction methods [13], which include for instance inverse distance weighted regression (IDWR), k-nearest-neighbors (k-NN) [14] and spatio-temporal kriging [15]. These methods are somewhat related to pure spatial data analysis [16]. Our focus is on kriging, in particular. This approach relies on a correlation model, which depends on distances between measurements in time and space. A crucial assumption for distributed calculus is spatio-temporal separability, which implies two separate models for spatial and temporal correlations [7]. This assumption is hardly suitable for large environments, where some locations can anticipate events that will occur somewhere else. In smaller environments, separability provides instead a useful approximation that catches up with more complicated, non-separable models [17].

While involving a simple model, kriging is cursed by the enormous cost of computation, mostly due to the inversion of large matrices. Some approximations have been devised to make kriging tractable like, for instance, covariance tapering [18]. A composite likelihood approach can be used to estimate a separable model, which allows to separate the estimation of the spatial model from the estimation of the temporal model. Also, some models in the time domain can spare the cumbersome matrix inversions and thus simplify both estimation and prediction. For instance, auto-regressive models can be estimated just by minimizing the conditional sum of squares, and they come with compact forecasting rules [19]. As to spatio-temporal predictions, we show that these can be seen as a spatial interpolation of temporal forecasts under separability, which allows to leverage on specific advantages of time series and spatial models. All these possibilities seem somewhat overlooked in modeling-aware literature, despite the attention received by separability itself.

The plan of the paper is as follows. In Section 2, we recall basic kriging formulation, while emphasizing some correlation structures of practical importance. In Section 3, we detail our inferential and predictive framework, focusing on distributed computing. Section 4 illustrates the application of the proposed methodology to the motivating example, whereas Section 5 is devoted to some possible twists and extensions. Finally, Section 6 presents some concluding remarks.

2 Model specification

We deal with the case of data that are both spatially and temporally referenced, so we use a space index \( s \) and a time index \( t \). The space index \( s \) takes its value in \( S = \{1, \ldots, S\} \) and is a pointer to one out of \( S \) locations in space, while the time index \( t \) takes its value in \( T = \{1, \ldots, T\} \) and is a pointer to one out of \( T \) time frames. A joint index \( st \) denotes

```latex
\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Room with windows, cardinal points and sensor locations, modified from the original map [9].}
\end{figure}
```
Separable spatio-temporal kriging for fast virtual sensing

location $s$ at time $t$. Since dealing with a constant sampling rate, we consider a discrete time system with equispaced time frames. In the long run, it holds $T \gg S$.

Let $d_{s,s'}$ be a spatial distance, defined for all pairs of locations $s, s' \in S$, thus endowed with non-negativity, symmetry and triangle inequality. Distance is ordinarily evaluated along straight lines in the absence of physical obstacles; otherwise, the length of the shortest path is considered. We choose the Euclidean distance for this purpose. Temporal distance is defined analogously \([20]\) as

$$d_{t,t'} = |t - t'| .$$

Here, $d_{s,s'}$ and $d_{t,t'}$ are the generic elements of the $S \times S$ spatial distance matrix $d_S$ and of the $T \times T$ temporal distance matrix $d_T$, respectively.

The data $y$ are structured as follows:

$$y = \begin{bmatrix} y_{11} & \cdots & y_{s1} \\ \vdots & \ddots & \vdots \\ y_{1T} & \cdots & y_{ST} \end{bmatrix},$$

so the data related to the location $s$ are all stored in the same column, while those related to the time frame $t$ are all stored in the same row. As new data are observed, they are appended to $y$ as a new row. The data $y$ are modeled by the random matrix $Y$ and the mean matrix $\mu$ with the same number of rows and columns of $y$.

Let $\text{vec}$ be a unary operator defined for matrices that stacks their columns into a single vector \([21]\). $Y$ is assumed to be a multivariate normal with scale parameter $\sigma$ and correlation matrix $R$, in the sense that $R$ is the correlation matrix of $\text{vec}(Y)$. More formally, we assume that $Y$ has the following density function.

$$f(y; \mu, \sigma, R) = \frac{1}{\sqrt{\det(2\pi \sigma^2 R)}} \exp \left\{ -\frac{1}{2} \text{vec} \left( \frac{y - \mu}{\sigma} \right)^\top R^{-1} \text{vec} \left( \frac{y - \mu}{\sigma} \right) \right\} ,$$

We assume that $\mu_{st}$ is a smooth function of $t$ and shared across locations, so it makes sense to estimate it with an asymmetric moving average $m_t$, defined below, which pools data across locations from $w$ time frames preceding $t$. Namely,

$$\hat{\mu}_{st} = m_t, \quad \text{with} \quad m_t = \frac{1}{S \cdot w} \sum_{s=1}^{S} \sum_{i=1}^{w} Y_{s(t-i)} .$$

We set $w$ equal to the number of observations per sensor in the 24 hours. Thus, the latest estimate available for $\mu$ can also serve as an estimate for future trend $\mu'$, assuming stability in the short term. Such an assumption can be credible in cases where the univariate time series may agree on a single latent factor ruling all of them.

The parameter $\sigma > 0$ contributes only in making predictions probabilistically calibrated \([22]\), because it is just a scale parameter, like the error standard deviation in classical linear regression, thus involved in prediction variance but not in mean predictions; see Appendix A. A simple estimator of $\sigma$ is the following, based on the assumption of constant variance through time and space.

$$\hat{\sigma} = \sqrt{\frac{1}{ST} \sum_{s=1}^{S} \sum_{t=1}^{T} (Y_{st} - \hat{\mu}_{st})^2} .$$

We resort to the classical kriging approach, which belongs to frequentist statistics, but this methodology also has a Bayesian counterpart involving a prior distribution on parameters. \([23]\) Ch. 5–7] As per kriging approach, we assume that correlations between components of $Y$ are stationary and thus depend on their distances in space and time. The covariance between any two components of $Y$, say, $Y_{st}$ and $Y_{s't'}$, is modeled as

$$\text{cov}(Y_{st}, Y_{s't'}) = \sigma^2 \cdot \text{cor}_S(d_{s,s'}) \cdot \text{cor}_T(d_{t,t'}) ,$$

where $\text{cor}_S(\cdot)$ is the spatial auto-correlation function (ACF) \([24]\), while $\text{cor}_T(\cdot)$ is the temporal ACF \([19]\). The product between spatial and temporal correlations is implied by the separability assumption. ACFs depending on distances and not directions are implied by an isotropy assumption. Both separability and isotropy can simplify modeling and computing \([6, 7, 25, 2, 1]\).

For the sake of illustration, the spatial ACF can be, for instance, one of the following \([20, 16, 13]\):

- Matérn ACF \([20]\)

$$\text{cor}_S(d) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} (d/\lambda)^\alpha K_\alpha(d/\lambda)^\alpha, \quad \lambda, \alpha > 0 ,$$

with $\Gamma(\cdot)$ the gamma function and $K_\alpha(\cdot)$ the modified Bessel function of the second kind;
Separable spatio-temporal kriging for fast virtual sensing

- power exponential ACF \[ \text{cor}_S(d) = \exp\{-d/\lambda\beta\}, \quad \lambda, \beta > 0. \]

Here, \( \lambda \) can be regarded as a range parameter. We refer to \( \alpha \) and \( \beta \) as smoothness parameters. Both Matérn and power exponential ACFs include two notable sub-cases:

- when \( \alpha = 1/2 \) and \( \beta = 1 \), the exponential ACF is implied \[6];
- when \( \alpha \to \infty \) and \( \beta = 2 \), the Gaussian ACF is implied \(1, 27\), which is also known as squared-exponential ACF \(28, 4, 3\) and involved in the double-stable model \(24\).

The temporal ACF in discrete time can be, for instance, one of the following \[19\]:

- ACF of a stationary auto-regressive model of order 1
  \[ \text{cor}_T(d) = \phi^{\lfloor d \rfloor}, \quad |\phi| < 1; \]
- ACF of an invertible moving-average model of order 1
  \[ \text{cor}_T(d) = \begin{cases} 1 & d = 0 \\ \alpha & d = \pm 1 \\ 0 & \text{otherwise} \end{cases}, \quad |\alpha| < 1/2. \]

More complicated ACFs are possible when looking at more flexible time series models, such as the multiplicative seasonal AR models that are used in the empirical application presented later. Some approaches assume weak or no correlation structure, like empirical kriging \[15\]. These approaches are necessarily less scalable but may still work for suitably targeted tasks.

Let \( \text{cor}_S(\cdot) \) and \( \text{cor}_T(\cdot) \) be vectorized function, that is, they transform matrices in an entry-wise fashion. Then, \( R_S = \text{cor}_S(d_S) \) will be a spatial correlation matrix and \( R_T = \text{cor}_T(d_T) \) a temporal correlation matrix. We call

\[ R = R_S \otimes R_T, \]

the spatio-temporal correlation matrix. Here, \( \otimes \) denotes the Kronecker product.

Both temporal and spatial ACFs can be modified in order to account for noisy data by including a so-called nugget effect \[26, 1\]. This means that the spatial ACF \( \text{cor}_S(d_S) \) and the temporal ACF \( \text{cor}_T(d_T) \) are multiplied by \( \beta_S \) and \( \beta_T \) when \( d_S = 0 \) and \( d_T = 0 \), respectively, the parameters \( \beta_S \) and \( \beta_T \) taking values in the interval \([0, 1]\) \[15\]. We refer to \( 1 - \beta_S \) and \( 1 - \beta_T \) as the nugget parameters, in space and time domains, respectively. Some authors apply the nugget directly to the spatio-temporal covariance function, but this will break up separability \[29, 25\]. The latter way of modeling is more natural in the case of additive covariance models \[30\].

3 Inferential aspects

This section presents two strategies that allow to perform estimation and prediction under a separable model, with a low computational cost. In particular, we base estimation on a novel composite likelihood approach. Then, leveraging on the peculiar expression of the kriging mean formula, we show how to compute predictions efficiently under separability.

3.1 Estimation

The distribution of \( Y \) in Equation \[1\] is assumed to be indexed by a parameter vector \( \theta \) via \( \mu, \sigma, R \). Let \( \theta \) be partitioned as \( \theta = (\mu^\top, \sigma, \psi^\top)^\top \), with \( \psi \) the correlation parameters ruling \( \hat{R} \). Moreover, \( \psi \) can be partitioned as \( \psi = (\psi_S^\top, \psi_T^\top)^\top \), where \( \psi_S \) and \( \psi_T \) are the spatial and temporal correlation parameters, respectively. In particular, \( R_S \) depends only on \( \psi_S \), while \( R_T \) depends only on \( \psi_T \). As a starting point, we consider the likelihood function \( \mathcal{L}(\theta; y) \), defined as

\[ \mathcal{L}(\theta; y) = f(y; \mu, \sigma, R). \]

Let the sample correlation matrix be defined as rank-1 matrix

\[ \hat{M} = \frac{1}{\sigma^2} \text{vec}(y - \hat{\mu})\text{vec}(y - \hat{\mu})^\top, \]

where \( \mu \) and \( \sigma \) have been replaced with some estimates, denoted by \( \hat{\mu} \) and \( \hat{\sigma} \), respectively. When replacing \( \mu \) and \( \sigma \) with estimates, the likelihood function turns into a so-called pseudo likelihood \( \mathcal{L}^p(\psi; y) \), which allows to make inference on \( \psi \) alone \[31, 26\], defined as

\[ \mathcal{L}^p(\psi; y) = |R|^{-1/2} \exp \left\{ -\frac{1}{2} \text{trace} \left( R^{-1}\hat{M} \right) \right\}. \]


Separable spatio-temporal kriging for fast virtual sensing

Kriging is often cumbersome due to the inversion of $R$, and actually the pseudo likelihood in Equation (4) is intractable with high-dimensional data. Separability reduces the dimensionality of the problem, as it holds

$$R^{-1} = (R_S \otimes R_T)^{-1} = R_S^{-1} \otimes R_T^{-1},$$

so two smaller inverse matrices must be computed instead of a single and larger one. However, as $T \gg S$, inverting $R_T$ alone can also be difficult.

Our proposal is two-fold. First, a marginal composite likelihood approach can be used, exploiting separability more in depth so that the tasks of estimating $\psi_S$ and $\psi_T$ can even be tackled separately. Second, a suitable time-series model can help handling the temporal correlation implicitly, in the sense that $R_T$ needs not be evaluated at all, and make it possible to address tall data and high sampling rates.

Composite likelihoods are known in spatial statistics mainly as tools that simplify estimation and inference, like the pairwise composite likelihood, and they can also be used in model selection. Composite likelihoods allow to make inference on under-specified models but, even in the case of fully specified models, like in kriging, a suitable composite likelihood can reduce the computational cost of estimation. The estimator based on the full model can be computed in some cases, but it would be cumbersome with the dataset under investigation. We remark the tractability of our estimator by carrying out inference based on bootstrap. Cheaper computation comes at a price, since the estimator is naturally sub-optimal with respect to the maximum likelihood estimator. Nonetheless, the loss of efficiency might be not significant when dealing with high-frequency data.

With composite likelihoods, as with any so-called pseudo likelihood, estimation variance cannot be assessed as with classical likelihood functions, that is, based on the Hessian matrix. Parametric bootstrap can be used in the case of kriging because the model is fully specified, and one can simulate datasets based on it. It is straightforward to simulate datasets under separability and some temporal models. We detail a bootstrap strategy in Appendix B.

Within a single time frame $t \in T$, it holds $R = R_S$, as per Equation (3). Let the spatial composite likelihood be defined as

$$L_S(\theta; y) = \prod_{t \in T} f(y_t; \mu, \sigma, R_S),$$

with $y_t$ and $\mu$, the data and mean vector at time frame $t$. The expression is the same of a “small blocks” marginal composite likelihood. This composite likelihood, when replacing $\mu_t$ and $\sigma$ with estimates, turns into a pseudo likelihood for the spatial correlation parameters $\psi_S$, which is defined as

$$L^p(\psi_S; y) = |R_S|^{-T/2} \exp \left\{ -\frac{T}{2} \text{trace} \left( R_S^{-1} \hat{M}_S \right) \right\}.$$  

(5)

Here, $\hat{M}_S$ is the sample correlation matrix between the univariate time series at distinct locations, defined as

$$\hat{M}_S = \frac{1}{T \sigma^2} (Y - \hat{\mu})^\top (Y - \hat{\mu}).$$  

(6)

The true correlations are better reflected by $\hat{M}_S$ if it is standardized to have a unit-valued diagonal, at least during the estimation step.

The estimator $\hat{\psi}_S$ for the spatial correlation parameters $\psi_S$ can be defined as the maximizer of the spatial pseudo likelihood $L^p(\psi_S; y)$. This estimator is consistent, assuming $S$ is bounded and $T$ increases. If $R_S$ is a smooth function of $\psi_S$, the estimator can be seen as the solution of a system of estimating equations, the equation for the generic parameter $\alpha \in \psi_S$ being defined as follows:

$$\alpha : \text{trace} \left\{ \left( R_S - \hat{M}_S \right) \frac{\partial R_S^{-1}}{\partial \alpha} \right\} = 0.$$  

The above equation looks like a weighted average of the equations from the over-determined system $R_S - \hat{M}_S = 0$.

It is worth noticing the close resemblance between the pseudo likelihoods in Equations (5) and (4), but the correlation matrices involved are strikingly different, as $R_S$ is much smaller than $R$ in practical applications.

The temporal correlation parameter $\psi_T$ could be estimated analogously to $\psi_S$, by defining a temporal composite likelihood $L_T$ as

$$L_T(\theta; y) = \prod_{s \in S} f(y_s; \mu_s, \sigma^2 R_T),$$  

(7)

with $y_s$ and $\mu_s$ the univariate time series and the mean vector of location $s$. The strategy is again to define a marginal composite likelihood. A temporal pseudo likelihood is obtained by replacing $\mu$ and $\sigma$ with estimates; this operation
Separable spatio-temporal kriging for fast virtual sensing

will completely disentangle the estimators of spatial and temporal parameters from each other. As contrasted to $R_S$, $R_T$ will be high-dimensional, so it is even more pressing the need to use a convenient correlation structure in the time domain. In particular, we resort to an auto-regressive model (AR) with multiple overlapping seasonal lags. For its definition, the classical lag operator $B$ can be introduced, such that

$$BY_{st} = Y_{s(t-1)} ,$$

and $B^\Delta Y_{st} = Y_{s(t-\Delta)}$ more in general. Then, letting $\epsilon_s$ be a white noise time series process, the multiplicative seasonal auto-regressive model is defined as follows [19].

$$\prod_{k=1}^K (1 - \phi_k B^{\Delta_k}) \cdot (Y_s - \mu_s) = \epsilon_s , \tag{8}$$

where $\Delta_1, \ldots, \Delta_k, \ldots, \Delta_K$ are the structural lags, and $\phi_1, \ldots, \phi_k, \ldots, \phi_K \in ] -1, +1 [$ the model coefficients.

Equation (8) reduces the temporal correlation parameter to $\psi_T = (\phi_1, \ldots, \phi_K)^T$. AR modeling also makes it easy to approximately maximize the likelihood by minimizing the conditional sum of squares [19], which can be attained via the coordinate descent method [50]. In this sense, we define

$$V_s^{-h} = \prod_{k \neq h} (1 - \phi_k B^{\Delta_k}) \cdot (Y_s - \mu_s) ,$$

and, by Equation (3), it holds

$$(1 - \phi_h B^{\Delta_h})V_s^{-h} = \epsilon_s ,$$

so, the transformed time series $V_s^{-h}$ satisfies

$$\phi_h = \text{cor}(V_s^{-h}, B^{\Delta_h}V_s^{-h}) .$$

This relation motivates an iterative procedure that loops over the correlation parameters, and repeats until convergence. For each $\phi_h$, one evaluates the process $V_s^{-h}$ and then updates $\phi_h$ as the ACF of $V_s^{-h}$ at lag $\Delta_h$.

An online learning approach may be considered as an alternative. For instance, estimates can be updated continually via batch learning [14] or exponentially-weighted moving means and covariances [37, 38], which require additional tuning.

### 3.2 Prediction

Separability assumptions allow to simplify also prediction, as we show in this section. To this end, we must expand our notation. Then, let $Y'$ be from unobserved locations or times and with the mean matrix $\mu'$, while previously introduced matrices $Y$ and $\mu$ are now related to the available data $y$. The correlation matrix of $\text{vec}(Y')$ is $R'$, whereas $\hat{R}$ is the correlation matrix of $\text{vec}(Y)$. The cross-correlation matrix between $\text{vec}(Y')$ and $\text{vec}(Y)$ is generally not square and it is denoted by $\rho$, which is based on a suitable joint distance matrix. Like $\hat{R}$ in Equation (3), also $R'$ and $\rho$ can be decomposed via Kronecker product as, respectively,

$$R' = R_S' \otimes R_T', \quad \rho = \rho_S \otimes \rho_T . \tag{9}$$

As typical in kriging, we treat $Y$ and $Y'$ as jointly normally distributed.

Kriging revolves around the conditional distribution of $Y'$ given $Y = y$ [11], though it was originally motivated as the linear unbiased prediction that is optimal with respect to the squared prediction error criterion [39]. Conditional to $Y = y$, the distribution of $Y'$ is multivariate normal. The conditional mean matrix of $Y'$ and the conditional variance-covariance matrix of $\text{vec}(Y')$, denoted by $\hat{y}'$, respectively, satisfy the following conditions.

$$\text{vec}(\hat{y}') = \text{vec}(\mu') + \rho R^{-1} \text{vec}(y - \mu) , \quad R'_{\text{cond}} = R' - \rho R^{-1} \rho^T . \tag{10}$$

Let $\beta_S$ and $\beta_T$ be spatial and temporal regression coefficients, defined as

$$\beta_S = \rho_S R_S^{-1} , \quad \beta_T = \rho_T R_T^{-1} , \tag{11}$$

then the kriging mean formula in Equation (10) can be written as

$$\hat{y}' - \mu' = \hat{\varepsilon} \beta_S^T , \tag{12}$$

where $\hat{\varepsilon}$ is a matrix of centered temporal forecasts only, defined as

$$\hat{\varepsilon} = \beta_T (y - \mu) . \tag{13}$$
We prove this fact in Appendix A.1. The above result allows for at least two interesting uses.

First, our result allows for distributed calculus in kriging. Indeed, spatio-temporal prediction under separability can be carried out in two steps. The first step, in Equation (13), is temporal forecasting only within sensor locations. The second step, in Equation (12), is the spatial interpolation of such forecasts for the needed locations. Then, spatio-temporal predictions can be seen as spatial interpolation of temporal forecasts. Geometrically, instead of moving along straight lines between spatio-temporal points, we cross one domain at a time. So, a separability assumption allows to separate domains also in prediction. Distributed calculus can be used for evaluating Equation (13), as each univariate time series is transformed separately by the matrix product. If the locations of interest are fewer than sensors, it will be more convenient to anticipate the interpolation step and thus to forecast only afterwards.

Correlation models affect prediction only through $\beta_S$ and $\beta_T$. So, any specification of a time model or a space model can be employed if it implies tractable prediction. This view motivates, for instance, the use of general interpolators, or state space models [40], or integrated auto-regressive time series models [19], that allow for simple prediction since $\beta_T$ in Equation (11) is sparse and explicit. The rather general auto-regressive moving-average model (ARMA) has already been considered by [41] though the MA component makes the model harder to estimate.

In applications, one may consider adding covariates to the analysis. Assume that there are $p$ variables indexed by $j = 1, \ldots, p$, so $Y_{jst}$ denotes the $j$-th variable at spatio-temporal coordinates $st$. In this cases, the marginal means $\mu_{jst}$ will likely depend on $j$. The correlation structure can be simplified according to a fully-factored model [7], such that

$$\text{cov}(Y_{jst}, Y_{j's't'}) = \gamma_{j,j'} \cdot \text{cov}(s,s') \cdot \text{cov}(d,d').$$

Here, the new quantity $\gamma_{j,j'}$ represents the cross-sectional covariance between the $j$-th and $j'$-th variables at the same time and location. Thus, under a fully-factored model, our kriging mean formula scales easily, as a set of regression coefficients $\beta_C$ is defined besides $\beta_S$ and $\beta_T$, implied by the newly added domain of covariates.

We provided a simple expression for mean predictions, along with some optimization strategies. For the sake of completeness, we now illustrate how to compute prediction variances. In our view, the model can be used to design a prediction rule, resulting in mean prediction, while prediction variance is a performance measure, which can for instance be evaluated on a test set. This approach may be favored when the focus is especially on prediction.

Let $V$ be a matrix with the same size as $Y'$ with the entrywise variances of $Y'$ conditional to $Y = y$. Then, $\text{vec}(V) = \sigma^2 \text{diag}(R'_{\text{cond}})$, so it holds

$$V = \sigma^2 \left\{ \text{diag}(R'_{T}) \text{diag}(R'_{S})^\top - \text{diag}(R'_{T} - R'_{T,\text{cond}}) \text{diag}(R'_{S} - R'_{S,\text{cond}})^\top \right\},$$

where $\text{diag}$ is the operator that returns the diagonal of a square matrix as a column vector, and $R'_{S,\text{cond}}$ and $R'_{T,\text{cond}}$ are defined as follows.

$$R'_{S,\text{cond}} = R'_{S} - \rho_S S^{-1} S^\top, \quad R'_{T,\text{cond}} = R'_{T} - \rho_T T^{-1} T^\top.$$  

We prove this result in Appendix A.2. Equation (14) has an advantage over (10), as the expression for $R'_{T,\text{cond}}$ is simple to retrieve under some time series models. For instance, with stationary AR processes, in one-step-forward forecasting, $R'_{T,\text{cond}}$ is the ratio between the variance of the innovation term and the marginal variance of the process.

4 Empirical application

Now we illustrate our proposed methodology on the SAL dataset presented in the introduction. We ran all our analyses in the statistical computing environment R [42]. We used the package gstat [43] to compute a spatio-temporal variogram and then to perform variogram fitting. The first operation was the most compute-intensive and memory-consuming, and it was done with the aid of a virtual machine on Microsoft Azure [1]. All the other analyses were instead carried out using typical laptop computers.

4.1 Explorative analysis

The SAL dataset collects temperature sensor readings from an office room and has been briefly described in the introductory section. The goal is to develop a spatio-temporal prediction rule for temperature based on these data. Some candidate spatial models are estimated on a training set, and the best one is selected based on a test set. The full dataset comprises 19 weeks of data and it is partitioned accordingly, with the leading eight weeks of data for training and the trailing 11 weeks as the test set. This choice is challenging for our method, as it is more exposed to shifts in the

---

1 Microsoft Azure is a cloud computing service, see [www.azure.microsoft.com](http://www.azure.microsoft.com)
Figure 2: Univariate time series of temperature per sensor, training set (October through December). Peaks and troughs are thresholded and highlighted with dots.
Figure 3: Sample quantiles of temperature, aggregation according to the weekday, training set (October through December).
We then carried out maximum composite likelihood estimation of the same spatial and temporal models as with VF. We compare our proposed composite likelihood estimation approach with least squares variogram fitting (VF), which is a standard in the estimation of spatio-temporal correlation structures [45]. VF requires to compute the empirical spatio-temporal variogram. The temporal covariance is feasibly tapered by considering only data pairs with time lag less than some custom threshold $T_{\text{max}}$. A variogram model is then fitted to the empirical variogram in a non-linear least squares fashion. The empirical variogram is computed with cost $O(S^2 \cdot T \cdot T_{\text{max}})$. With data of this size, VF seems very demanding for the implementation provided by the R package gstat.

In computing the empirical variogram, we used a virtual machine with four cores to emulate the computational power of the server. However, after running some pilot tests, we deemed it necessary to configure the virtual machine with 16 gigabytes of RAM, which would be overwhelming for typical Raspberry Pi boards. The variogram was computed using the eight weeks of training data, which took about 40 hours to complete with cores working in parallel. It is fair to notice that the comparison with gstat is somewhat uneven, because that library is general purpose and not optimized for separable models.

After calculating the empirical variogram, the subsequent step in VF is model fitting. The model was chosen to be separable, as discussed insofar. We consider four candidate models in the space domain, namely, exponential, Gaussian, power exponential and Matérn ACFs. As to the time domain, gstat supports the exponential ACF, which implies an AR(1) model with structural lag equal to the sampling interval, that is, $\Delta_1 = 10$ seconds in Equation (8). Fitting the variogram models took just a few seconds, as contrasted to computing the empirical variogram, but only boundary estimates were obtained for either the temporal correlation parameters or the spatial ones.

Unfortunately, these boundary results make comparisons between VF and our estimation approach not very informative. We then carried out maximum composite likelihood estimation of the same spatial and temporal models as with VF. The estimation of spatial parameters is based on the sample spatial correlation matrix, whose computational cost is $O(S^2 \cdot T)$. Temporal parameters are estimated by iteratively transforming data and the number of iterations is generally stochastic, but fitting an AR(1) is simple because it involves evaluating the empirical ACF only once and at a specific lag. After computing the sample correlation matrix, the spatial composite likelihoods are maximized for each of the four candidate spatial models. As a remark, the temporal model was estimated once, independently of the four candidate spatial models, as they are not involved in the temporal composite likelihood.

The estimate based on the composite likelihood for the AR coefficient is also boundary, i.e. close to unit. This result reflects the high sampling rate and the likely strong auto-correlation in the process. We then use longer structural lags, so that the model can be used even at lower sampling rates. In particular, we consider a custom AR model with lags $\Delta_1 = 10$ minutes, $\Delta_2 = 1$ day and $\Delta_3 = 1$ week in Equation (9). This model allows to assess multiplicative daily and weekly seasonal effects. It would be challenging to estimate with VF, as it relies on the ACF, which is hard to
Table 1: Point estimates and standard errors for spatial and temporal correlation parameters.

| spatial model | parameter | est.  | std. err. (×10^3) |
|---------------|-----------|-------|-------------------|
| exponential   | \( \phi_1 \) | 0.977 | 0.206             |
|               | \( \phi_2 \) | 0.078 | 0.950             |
|               | \( \phi_3 \) | 0.047 | 0.990             |
|               | nugget    | 0.191 | 2.307             |
|               | range     | 24.692| 446.329           |
| Gaussian      | \( \phi_1 \) | 0.977 | 0.203             |
|               | \( \phi_2 \) | 0.078 | 0.936             |
|               | \( \phi_3 \) | 0.047 | 0.975             |
|               | nugget    | 0.250 | 2.470             |
|               | range     | 15.011| 140.078           |
| Matérn       | \( \phi_1 \) | 0.977 | 0.206             |
|               | \( \phi_2 \) | 0.078 | 0.950             |
|               | \( \phi_3 \) | 0.047 | 0.991             |
|               | nugget    | 0.187 | 4.805             |
|               | range     | 25.993| 1547.047          |
|               | smoothness| 0.479 | 22.544            |
| power exponential | \( \phi_1 \) | 0.977 | 0.205             |
|               | \( \phi_2 \) | 0.078 | 0.946             |
|               | \( \phi_3 \) | 0.047 | 0.986             |
|               | nugget    | 0.217 | 3.057             |
|               | range     | 19.883| 424.822           |
|               | smoothness| 1.312 | 28.654            |

formulate with such a model. After estimating the custom AR and the four spatial models, we carry out inference based on parametric bootstrap by simulating and analyzing 1000 datasets for each joint spatio-temporal model. Some details on this procedure are given in Appendix B. Performing the bootstrap took about 3 hours on a single laptop computer with our composite likelihood approach, despite the large sample size, while it would have been out of reach with VF. The model fit is summarized in Table 1. The implied spatial ACFs are reported in Figure 4 along with the empirical correlogram.

4.3 Sensor network optimization

After estimating the temporal and spatial models, a practical concern is the selection of few operation sensors from the initial set, as twelve of them are too many for a room that is 127m^2 large. Under the proximity principle, some sensors could be dropped, and their location could be just virtually sensed since their data can be surrogated with predictions from the remaining sensors.

Different network configurations can be evaluated and compared according to a metric, which should reflect priorities and objectives of stakeholders. The 95th percentile of absolute prediction errors on all but active sensors can be used for an approximate minimax decision. For comparison, we illustrate a sensor selection based on this criterion alongside one that uses a more classical mean absolute prediction error. The performance of spatial models and sensor configurations is evaluated and compared on the test set. For each sensor configuration, we interpolate data from selected locations to the unselected ones within time frames, as implied by separability. Prediction errors are then summarized according to the metric. We perform selection in a forward fashion, by starting with the best performer alone and then adding the sensor that led to the best improvement at each step. Adding sensors can worsen the performance because we are evaluating models on the test set. In Figure 5, a summary of the selection process is reported. The sensor added at each step appears within a box and is numbered as in Figure 1. An alternative prediction is given by the simple mean, which assumes that a single latent temperature is ruling the whole room. The selection took no more than 10 minutes in total, so it would be easy to perform it multiple times ad interim to check on the quality of predictions.

We show only the power exponential ACF. The result based on the Matérn function is very similar, and the exponential and Gaussian ones are slightly outperformed. The percentile performance seems in line with k-NN and IDW benchmarks. Based on performances in Figure 5, the power exponential ACF may be preferred over the mean prediction because this choice seems more robust with respect to the metric. Moreover, the mean prediction yields some narrowly spaced sensor configurations under both metrics.
Figure 4: Empirical auto-correlation (dots) and estimated ACFs for two spatial models (lines).

Figure 5: Forward selection of sensors, distinct per metric, prediction error versus number of selected sensors. The sensor added at each step appears within a box. Run on test set (December through March).
4.4 Behavior of the monitoring system

By using different sensor configurations, we show few examples of actual situations along with the behavior of a possible monitoring system. The environment in focus has some peculiarities that somehow affect the performance of the system. As an illustration, we combine interpolation and forecasting by predicting temperatures 10 minutes forward for the whole floor plan, as in Figure 6.

We anticipated that there would be differences between regular and anomalous sensors. Concerns relate to sensors facing direct sunlight around noon (numbered 5, 10) or close to other sources of anomalies (7, 12). If these were used to make predictions, the results would hardly reflect the normalcy ruling the interior of the room, see Figure 6a. On the contrary, a system based on more regular sensors only, like in Figure 6b, would yield more constant predictions that are likely exchangeable with mean prediction. The sensor selection illustrated in the previous section is aimed at selecting a solution between these two extremes.

The two metrics considered suggest similar solutions. Figure 6c reports the prediction provided under the power exponential ACF by the four best sensors according to the percentile metric. Figure 6d shows the selection based on the mean absolute error instead. Both configurations attempt to replicate the north-south gradient, which seems to require choosing between sensors 5 or 10. However, neither of these choices can surrogate sensors 11 and 3. Selecting sensors 5 and 10 would fail the whole interior of the room and, by converse, the sensors in the interior of the room cannot surrogate 5 and 10.

5 Some extensions

In the previous section, we presented an application of our compute-efficient approach to spatio-temporal kriging. Here we describe two possible extensions that look legitimate under our prediction-oriented view of kriging. For instance, it is possible to include some spatial interpolators or temporal forecasting methods that do not necessarily underlie any stationary ACF. It is also possible to use distinct temporal models for each sensor location to provide more specialized forecasting. Both these extensions share at least the advantage of further distributing calculus across the sensor network and simplify server-side computations.

5.1 Non-stationary modeling

The kriging approach relies on stationary ACF models, which offer a wide variety of possibilities, but some problems may be addressed only with non-stationary models. For instance, integrated AR models need no trend formulation. Another alternative is \( k \)-NN, which returns the sample average of response values from the \( k \) sensors closest to the desired location.

This could be useful in cases like ours, where distinct sensors might have different equilibria. We used a moving average to accommodate for non-stationarity in the mean, which conciliates with stationarity in ACF, but one can use integrated AR and \( k \)-NN as an alternative. With high-frequency data, long-term stationarity may coexist with short-term non-stationarity, the latter resulting from locally linear trends, so it might be useful to consider a non-stationary model that copes with both aspects.

We considered an integrated AR model in our analysis but without obtaining any significant improvement. Indeed, the chosen AR model was already able to cope with non-stationarity due to both a moving average trend and a near-unit first AR coefficient, which implied integration \textit{de facto}.

5.2 Sensor-specific temporal correlation parameters

A limitation of separable ACFs is that they imply all locations having the same marginal dynamics. As an extension, the temporal correlation parameters can be sensor-specific: each sensor can estimate and update a distinct temporal model that is valid at least for its location and approximately also for a neighborhood.

Distinct temporal models can each be based on a different composite likelihood and use different portions of data, so they will not affect each other. In mean prediction, as per Equation (13), \( \hat{z} \) can be replaced with a matrix, where each column is made up of the temporal forecasts based on a model with limited scope that works for just one location or a neighborhood. When interpolating these forecasts spatially, via Equation (12), more weight is given to forecasts close to the needed locations. This implies that all temporal models are involved but to a varied extent, depending on the distance.

This extension with distinct temporal models per location adds flexibility to monitoring in at least two ways.
Figure 6: Examples of prediction via four sensors, based on power exponential model. Forecasts for February 17, 12:00, based on data from 10 minutes before. Circles are colored based on observed temperatures, the floor based on predicted temperatures, selected sensors only are numbered as in Figure 1.

(a) Anomalous sensors.
(b) Regular sensors.
(c) Selection via 95th percentile absolute error.
(d) Selection via mean absolute error.

- It adds flexibility in network management. Each sensor has to estimate and update its own temporal model, so this has not to be handled by the server, which thus must be in charge only with the spatial interpolation task.

- Statistical modeling becomes more flexible too. Prior to this, a single overall temporal model is formulated that has to fit all locations forcefully. Distinct temporal models may now be used to address subsets of locations, so they can cope with more local and specific dynamics.

Anomalous sensors may be more effectively dealt with by allowing them to make predictions based on a more specific model targeted to them only. The office room in our example is too small to allow for a variety of temporal models. Larger environments will likely be more heterogeneous and will thus need many local models to provide better forecasts. Indeed, since spatio-temporal prediction is made up of both interpolation and forecasting, the quality of the latter is a necessary ingredient to joint predictive performance.
6 Discussion

We have proposed a separable kriging approach that allows to analyze large datasets by exploiting some overlooked aspects of separability. Even high-frequency data can be processed in a reasonable amount of time using a maximum composite likelihood estimator and optimized calculus in prediction. Separability allows to distribute calculus across the sensor network by delegating as many operations as possible to the components that gather the relevant data.

To our knowledge, the use of marginal composite likelihood is novel to sensor data analysis. Its most appealing aspect is that the spatial and temporal models under separability can be estimated in parallel without affecting each other. The spatial model must be estimated in a centralized way, but the temporal model may be addressed in a decentralized way by allowing sensors to estimate a temporal model valid for their location or neighborhood, as described in Section 5.2. This idea relates to stratified variograms [47,16], but it has even more in common with the estimation of a single variogram with data pairs sharing some identical conditions [48].

The predictive part of our approach was already common in climate and weather sciences, though in a modeling-unaware fashion, and confined mostly to spatial interpolation [49]. Instead, we provide a formal motivation for this way of computing predictions based on separability. We found a related simplification in jointly spatio-temporal prediction, which we guess can be easily extended to separability in more than two domains via Tucker products instead of Kronecker ones. For instance, covariates could be included in kriging via fully-factored modeling [7], but feature engineering seemed necessary in our case [9], which is not typical in kriging. For the sake of completeness, alternative modeling strategies include additive covariances [50], process convolution [51] and linear mixed models [30], for which simplifications might be different where possible.

In developing our proposal, we require data to be gridded, which means that all sensors provide simultaneous readings. However, this requirement can be weakened, since data can be at least projected onto a grid [52].

Kriging computation is hugely simplified by assuming separability and by choosing a suitable spatial or (especially) temporal model. Both estimation and prediction can bypass the evaluation and inversion of large correlation matrices by treating the data as univariate time series or cross-sections and thus splitting a generally complicated calculation into simpler operations. For instance, AR models may have an intractable ACF, but they can be estimated easily by minimizing a conditional sum of squares, and their forecasts based on Equation (8) are simple to calculate as well. Our approach allows to blend together and leverage on known strengths of time series analysis and spatial statistics, without the need to outline a joint spatio-temporal framework from scratch.

7 Acknowledgements

This project was performed within the COMET Centre ASSIC Austrian Smart Systems Integration Research Center, which is funded by BMK, BMDW, and the Austrian Provinces of Carinthia and Styria, within the framework of COMET – Competence Centers for Excellent Technologies. The COMET programme is run by FFG. The research of Michele Lambardi di San Miniato was supported by the European Social Fund (Investimenti in favore della crescita e dell’occupazione, Programma Operativo del Friuli Venezia Giulia 2014/2020) - Programma specifico 89/2019 - Sostegno alla realizzazione di dottorati e assegni di ricerca, operazione PS 89/2019 ASSEGNI DI RICERCA - UNIUD (FP1956292002, canale di finanziamento 1420_SRDAR8919).

No conflict of interest has been detected.

The dataset used in the application example is publicly available on GitHub, as submitted by the authors of the first paper addressing it [9], at

\[https://github.com/dslab-uniud/virtual-sensing\].

References

[1] Robert B. Gramacy. *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman Hall/CRC, Boca Raton, Florida, 2020.

[2] Linh Nguyen, Guoqiang Hu, and Costas J. Spanos. Spatio-temporal environmental monitoring for smart buildings. In 2017 13th IEEE International Conference on Control & Automation (ICCA). IEEE, July 2017.

[3] Joseph Carpenter, Keith A. Woodbury, and Zheng O'Neill. Using change-point and Gaussian process models to create baseline energy models in industrial facilities: A comparison. *Applied Energy*, 213:415–425, March 2018.
Separable spatio-temporal kriging for fast virtual sensing

[4] Hongbin Liu, Chong Yang, Mingzhi Huang, Dongsheang Wang, and ChangKyoo Yoo. Modeling of subway indoor air quality using Gaussian process regression. *Journal of Hazardous Materials*, 359:266–273, October 2018.

[5] Haorong Li, Dainhong Yu, and James E. Braun. A review of virtual sensing technology and application in building systems. *HVAC&R Research*, 17(5):619–645, 2011.

[6] Ignacio Rodríguez-Iturbe and José M. Mejía. The design of rainfall networks in time and space. *Water Resources Research*, 10(4):713–728, aug 1974.

[7] Kanti V Mardia and Colin R Goodall. Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*, 6(76):347–385, 1993.

[8] Bing Dong and Khee Poh Lam. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. *Building Simulation*, 7(1):89–106, September 2013.

[9] Andrea Brunello, Andrea Urgolo, Federico Pittino, András Montvay, and Angelo Montanari. Virtual sensing and sensors selection for efficient temperature monitoring in indoor environments. *Sensors*, 21(8):2728, apr 2021.

[10] Sheikh Ferdoush and Xinrong Li. Wireless sensor network system design using Raspberry Pi and Arduino for environmental monitoring applications. *Procedia Computer Science*, 34:103–110, 2014.

[11] Tianqi Chen and Carlos Guestrin. XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, aug 2016.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, nov 1997.

[13] Esi Oktavia, Widyawan, and I Wayan Mustika. Inverse distance weighting and kriging spatial interpolation for data center thermal monitoring. In *2016 1st International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISSEE)*. IEEE, August 2016.

[14] A. Azzalini, B. Scarpa, and G. Walton. *Data Analysis and Data Mining: An Introduction*. Oxford University Press, USA, 2012.

[15] Aloysius W. Aryaputra, Dazhi Yang, Lu Zhao, and Wilfred M. Walsh. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. *Solar Energy*, 122:1266–1278, dec 2015.

[16] Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer New York, 2013.

[17] Marc G. Genton. Separable approximations of space-time covariance matrices. *Environmetrics*, 18(7):681–695, 2007.

[18] Cari G. Kaufman, Mark J. Schervish, and Douglas W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, December 2008.

[19] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: Forecasting and control*. John Wiley & Sons, 2015.

[20] Giorgio Franceschetti and Daniele Riccio. Surface classical models. In *Scattering, Natural Surfaces, and Fractals*, pages 21–59. Elsevier, 2007.

[21] Robert E. Hartwig. Ax - xb = c, resultants and generalized inverses. *SIAM Journal on Applied Mathematics*, 28(1):154–183, January 1975.

[22] Veronica J. Berrocal, Adrian E. Raftery, and Tilmann Gneiting. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135(4):1386–1402, apr 2007.

[23] Peter J. Diggle and Paulo J. Ribeiro. *Model-based Geostatistics*. Springer New York, 2007.

[24] Noel Cressie. *Statistics for Spatial Data*. John Wiley & Sons, Inc., sep 1993.

[25] Bohai Zhang, Huiyan Sang, and Jianhua Z. Huang. Full-scale approximations of spatio-temporal covariance models for large datasets. *Statistica Sinica*, 25(1):99–114, 2015.

[26] Carlo Gaetan and Xavier Guyon. *Spatial Statistics and Modeling*. Springer New York, 2010.

[27] Marc A. Maes, Karl Breitung, and Markus R. Dann. At issue: The Gaussian autocorrelation function. In *Lecture Notes in Civil Engineering*, pages 191–203. Springer International Publishing, 2021.

[28] Cyrill Stachniss, Christian Plagemann, and Achim J. Lilienthal. Learning gas distribution models using sparse Gaussian process mixtures. *Autonomous Robots*, 26(2-3):187–202, March 2009.
Separable spatio-temporal kriging for fast virtual sensing

[29] Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, sep 2008.

[30] Michael Dumelle, Jay M. Ver Hoef, Claudio Fuentes, and Alix Gitelman. A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product–sum covariance functions. *Spatial Statistics*, 43:100510, jun 2021.

[31] Gail Gong and Francisco J. Samaniego. Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, 9(4), jul 1981.

[32] Petruța C. Caragea and Richard L. Smith. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7):1417–1440, aug 2007.

[33] Cristiano Varin, Nancy Reid, and David Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

[34] Petruța C. Caragea and Richard L. Smith. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *Journal of Multivariate Analysis*, 98(7):1417–1440, aug 2007.

[35] Cristiano Varin and Paolo Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, September 2005.

[36] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

[37] Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, March 2015.

[38] J. Stuart Hunter. The exponentially weighted moving average. *Journal of Quality Technology*, 18(4):203–210, October 1986.

[39] Longcheen Huwang, Arthur B. Yeh, and Chien-Wei Wu. Monitoring multivariate process variability for individual observations. *Journal of Quality Technology*, 39(3):258–278, jul 2007.

[40] Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, apr 1990.

[41] Dominique Courault and Pascal Monestiez. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast france. *International Journal of Climatology*, 19(4):365–378, mar 1999.

[42] Pascal Monestiez, Dominique Courault, Denis Allard, and Françoise Ruget. Spatial interpolation of air temperature using environmental context: Application to a crop model. *Environmental and Ecological Statistics*, 8(4):297–309, 2001.

[43] Noel Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586, jul 1985.

[44] Dominique Courault and Pascal Monestiez. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast france. *International Journal of Climatology*, 19(4):365–378, mar 1999.

[45] Pascal Monestiez, Dominique Courault, Denis Allard, and Françoise Ruget. Spatial interpolation of air temperature using environmental context: Application to a crop model. *Environmental and Ecological Statistics*, 8(4):297–309, 2001.

[46] Robert B. Gramacy, Jarad Niemi, and Robin M. Weiss. Massively parallel approximate gaussian process regression. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):564–584, jan 2014.

[47] Dominique Courault and Pascal Monestiez. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast france. *International Journal of Climatology*, 19(4):365–378, mar 1999.

[48] Pascal Monestiez, Dominique Courault, Denis Allard, and Françoise Ruget. Spatial interpolation of air temperature using environmental context: Application to a crop model. *Environmental and Ecological Statistics*, 8(4):297–309, 2001.

[49] MR Holdaway. Spatial modeling and interpolation of monthly temperature using kriging. *Climate Research*, 6:215–225, 1996.

[50] Pulong Ma, Bledar A. Konomi, and Emily L. Kang. An additive approximate Gaussian process model for large spatio-temporal data. *Environmetrics*, 30(8), apr 2019.

[51] Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer London, 2002.

[52] Christopher J. Paciorek. Computational techniques for spatial logistic regression with large data sets. *Computational Statistics & Data Analysis*, 51(8):3631–3653, may 2007.
A Proofs

A.1 Kriging mean formula

Predictions can be computed in a vectorized form, as follows, after Equation \( 10 \).

\[
\text{vec}(\mathbf{y}' - \mu') = \mathbf{R}^{-1}\text{vec}(\mathbf{y} - \mu).
\]

By using definitions in Equations \( 3 \) and \( 9 \), it follows that

\[
\text{vec}(\mathbf{y}' - \mu') = (\mathbf{p}_{\mathbf{S}} \otimes \mathbf{p}_{\mathbf{T}})(\mathbf{R}_{\mathbf{S}} \otimes \mathbf{R}_{\mathbf{T}})^{-1}\text{vec}(\mathbf{y} - \mu).
\]

Next, we use the inversion behavior of the Kronecker product.

\[
\text{vec}(\mathbf{y}' - \mu') = (\mathbf{p}_{\mathbf{S}} \otimes \mathbf{p}_{\mathbf{T}})(\mathbf{R}_{\mathbf{S}}^{-1} \otimes \mathbf{R}_{\mathbf{T}}^{-1})\text{vec}(\mathbf{y} - \mu).
\]

Then, it comes in handy to use the mixed-product property of the Kronecker product.

\[
\text{vec}(\mathbf{y}' - \mu') = \{(\mathbf{p}_{\mathbf{S}}\mathbf{R}_{\mathbf{S}}^{-1}) \otimes (\mathbf{p}_{\mathbf{T}}\mathbf{R}_{\mathbf{T}}^{-1})\}\text{vec}(\mathbf{y} - \mu).
\]

At this point, the regression coefficients of Equation \( 11 \) can be recognized.

\[
\text{vec}(\mathbf{y}' - \mu') = (\beta_{\mathbf{S}} \otimes \beta_{\mathbf{T}})\text{vec}(\mathbf{y} - \mu).
\]

Lastly, we use Roth’s column lemma \( [21] \).

\[
\text{vec}(\mathbf{y}' - \mu') = \text{vec}\{\beta_{\mathbf{T}}(\mathbf{y} - \mu)\mathbf{\beta}_{\mathbf{S}}^\top\}.
\]

Then, the vec operator can be dropped and the matrix \( \mathbf{y}' \) is obtained.

A.2 Kriging variance formula

Using Equation \( 10 \) as a starting point, it holds

\[
\text{diag}(\mathbf{R}'_{\text{cond}}) = \text{diag}(\mathbf{R}') - \text{diag}(\rho\mathbf{R}^{-1}\rho^\top).
\]

Similarly to the proof in Appendix \( A.1 \), we exploit again the inversion behavior and the mixed-product property of the Kronecker product. We also use Equations \( 4 \) and \( 9 \) to obtain

\[
\text{diag}(\mathbf{R}'_{\text{cond}}) = \text{diag}(\mathbf{R}_{\mathbf{S}}^\top \otimes \mathbf{R}_{\mathbf{T}}^\top) - \text{diag}\{(\mathbf{R}_{\mathbf{S}}^\top \mathbf{R}_{\mathbf{S},\text{cond}}) \otimes (\mathbf{R}_{\mathbf{T}}^\top \mathbf{R}_{\mathbf{T},\text{cond}})\}.
\]

After Equation \( 15 \), it follows that

\[
\text{diag}(\mathbf{R}'_{\text{cond}}) = \text{diag}(\mathbf{R}_{\mathbf{S}}^\top \otimes \mathbf{R}_{\mathbf{T}}^\top) - \text{diag}\{(\mathbf{R}_{\mathbf{S}}^\top \mathbf{R}_{\mathbf{S},\text{cond}}) \otimes (\mathbf{R}_{\mathbf{T}}^\top \mathbf{R}_{\mathbf{T},\text{cond}})\}.
\]

One can use the self-evident property \( \text{diag}(\mathbf{A} \otimes \mathbf{B}) = \text{diag}(\mathbf{A}) \otimes \text{diag}(\mathbf{B}) \) for \( \mathbf{A} \) and \( \mathbf{B} \) square matrices, which implies

\[
\text{diag}(\mathbf{R}'_{\text{cond}}) = \text{diag}(\mathbf{R}_{\mathbf{S}}^\top) \otimes \text{diag}(\mathbf{R}_{\mathbf{T}}^\top) - \text{diag}(\mathbf{R}_{\mathbf{S}}^\top \mathbf{R}_{\mathbf{S},\text{cond}}) \otimes \text{diag}(\mathbf{R}_{\mathbf{T}}^\top \mathbf{R}_{\mathbf{T},\text{cond}}).
\]

Now, the components of \( \text{diag}(\mathbf{R}'_{\text{cond}}) \) can be partitioned into vectors with the same length as \( \text{diag}(\mathbf{R}_{\mathbf{T}}') \). Such vectors can be the columns of the matrix \( \mathbf{V} \), which is thus defined as in our claims.

B Bootstrap

Parametric bootstrap \( [35] \) under separable kriging is particularly convenient because the implied model is easy to simulate, and its parameters are simple to estimate with the approach proposed in this paper.

The temporal and spatial model together identify the full model, under which one can simulate artificial datasets. In particular, separability allows to simulate a dataset \( \mathbf{Y} \) as

\[
\mathbf{Y} \sim \mu + \mathbf{\sigma} \cdot \mathbf{R}_{\mathbf{T}}^{1/2} \cdot \epsilon \cdot \mathbf{R}_{\mathbf{S}}^{1/2},
\]

where \( \epsilon \) is a Gaussian white noise structured into a \( T \times S \) matrix, and \( \mathbf{R}_{\mathbf{T}}^{1/2} \) and \( \mathbf{R}_{\mathbf{S}}^{1/2} \) are the matrix square roots of the matrices \( \mathbf{R}_{\mathbf{T}} \) and \( \mathbf{R}_{\mathbf{S}} \), respectively.

Assuming \( T \gg S, \mathbf{R}_{\mathbf{S}}^{1/2} \) may be tractable, while \( \mathbf{R}_{\mathbf{T}}^{1/2} \) will hardly be so. The operator \( \mathbf{R}_{\mathbf{T}}^{1/2} \) just makes \( \mathbf{R}_{\mathbf{T}}^{1/2} \cdot \epsilon \) a matrix with independent columns that share the same correlation structure \( \mathbf{R}_{\mathbf{T}} \). So, as an alternative to directly evaluating \( \mathbf{R}_{\mathbf{T}}^{1/2} \), one can generate each column of \( \mathbf{R}_{\mathbf{T}}^{1/2} \cdot \epsilon \) according to the temporal model. AR(p) processes can be simulated efficiently according to the factorized MA(\( \infty \)) form \( [19] \). The first observations should be initialized according to the stationary distribution of the process, but with complicated models one may instead provide an arbitrary initialization and then simulate additional observations as a burn-in. We adopted this latter strategy. Actually, in simulating eight weeks of data, we needed to simulate 32 more leading weeks of data as a burn-in.