Received: 29 September 2021 | Revised: 27 April 2022 | Accepted: 20 May 2022

DOI: 10.1002/asmb.2697

RESEARCH ARTICLE

WILEY

Separable spatio-temporal kriging for fast virtual sensing

Michele Lambardi di San Miniato | **Ruggero Bellio** | **Luca Grassetti** | **Paolo Vidoni**

Department of Economics and Statistics, University of Udine, Udine, Italy

Correspondence
Michele Lambardi di San Miniato, Department of Economics and Statistics, University of Udine, via Tomadini, 30/a, 33100 Udine, Italy.
Email: michele.lambardidisanminiato@gmail.com

Funding information
Competence Centres for Excellent Technologies Programme (COMET); European Social Fund; Università degli Studi di Udine, Grant/Award Number: FP1956292002

Abstract

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, foremostly 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 typically. Only spatial interpolation needs to be performed in a centralized way, while forecasting can be delegated to each sensor. This simplification is 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 (CL) approach. Moreover, the use of convenient spatial and temporal models eases up computation, not only in the CL approach, but also in maximum likelihood estimation. We show that this perspective on kriging allows to perform virtual sensing even in the case of tall datasets.

KEYWORDS

composite likelihood, distance-based prediction, distributed calculus, indoor environments, separability, isotropy, spatio-temporal kriging

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,[2,4,5] 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 $^{\circ}C$) is reported by 12 sensors every 10 s, 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 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

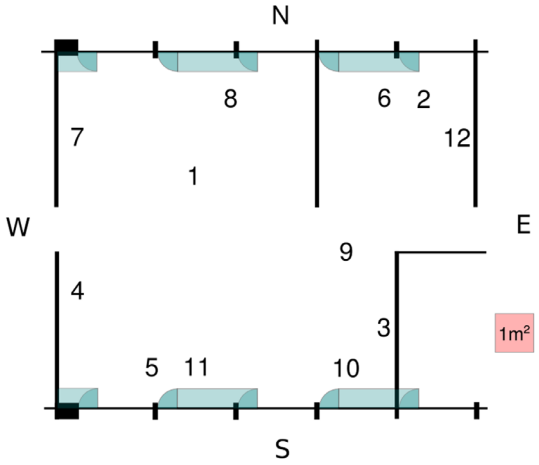

FIGURE 1 Room with windows, cardinal points, and sensor locations, modified from the original map[9]

widely used in home automation and 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, considering the limitations of the monitoring system.

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 spatial 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 the predictive performance. We are interested in simple models and distributed computing, thus on scalable methods that leverage 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] The focus of the present article is on the kriging method. 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 may perform similarly to more complicated, nonseparable models.[17]

While involving a simple model, kriging is cursed by the enormous cost of computation, mostly due to the inversion of large correlation matrices. Some approximations have been devised to make kriging tractable like, for instance, covariance tapering.[18] A composite likelihood (CL) 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, autoregressive (AR) 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 specific advantages of time series and spatial models. These possibilities seem somewhat overlooked in the literature, in spite of the attention received by separability itself.

All in all, the main contributions of this article can be summarized as follows. First, a CL-based fitting procedure is proposed that is both memory- and compute-efficient. An affordable implementation of the maximum likelihood (ML) estimator is also presented, which is slower to compute but serves as the natural benchmark for our proposal. Second, this article provides significant simplifications of computational order in both estimation and prediction related to Gaussian processes (GPs), in the case of separable covariance structures, making it possible to carry out scalable sensor data analysis. As a result, it expands the range of models that can be addressed in spatio-temporal kriging, as we point out that the

allegedly required covariance matrices need not to be evaluated directly. Thus, even rich seasonal models can be addressed without directly evaluating the autocorrelation function.

The plan of this article is as follows. In Section 2, we review the literature in brief as it concerns GPs and their applications to large datasets. In Section 3, we recall basic kriging formulation, while emphasizing some correlation structures of practical importance. In Section 4, we detail our inferential and predictive framework, focusing on distributed computing. Section 5 illustrates the application of the proposed methodology to the motivating example, whereas Section 6 is devoted to some possible twists and extensions. Finally, Section 7 presents some concluding remarks.

2 | LITERATURE REVIEW

We deal with the task of surrogate modeling, estimation and prediction.[1] Consider a function or a process defined over a domain. This can represent the temperature at given times and locations,[9] groundwater levels over a region,[20] material properties of different portions of a solid,[21] wind speed at turbine locations,[22] turbulence around airships,[23] or robots' reliability over time,[24] to name a few examples. The process is feasibly observed only at selected points in the domain. For instance, room temperature may be known only at sensor locations and at discrete times. In many of these examples, inputs and outputs of some production systems can be thought of, respectively, as domain points and process outcomes. The unavailable information, that is, the process at unobserved locations and times, has thus to be surrogated. Optimization tasks may involve some interpolation steps,[23,25] for which a proximity principle is implicitly trusted. From a statistical viewpoint, a stochastic process is assumed to rule the target of knowledge, and suitable regularity assumptions will be made: these reflect the heuristic that similarity comes with proximity and independence comes with distance. For instance, atmospheric events and air conditions often exhibit such regularity.[26] From an operational standpoint, a distance measure must be defined, which must be meaningful within the specific geometry of the domain on focus.[27] For example, a domain can even be made up of functions: even in such a case, a meaningful distance measure has been proposed recently.[28,29]

Gaussianity, coupled with a suitable correlation structure, is a common assumption in statistical modeling that is central to kriging, as it translates straightforwardly into first- and second- order conditions.[30,31] Moreover, GPs arise naturally in the case of spatially and temporally referenced data. The term kriging in broad generality refers to the task of making predictions based on conditional normal distributions, where the covariance structure depends on a suitably defined distance measure.[32] By extension, the model underlying kriging has been called the kriging model.[33] Kriging has become an umbrella name to include many variants, like ordinary, simple, and universal kriging, which differ in the specification of the mean. Moreover, co-kriging involves regression on covariates that are known beforehand; on the contrary, semantic kriging involves covariates that are not known in advance and thus need to be predicted in turn.[34] In machine learning, kriging often uses rich and nonlinear trend formulations,[35] see, for instance, the polynomial chaos approach.[36] Intrinsic kriging revolves around the process of differences, which is approximately de-trended if the mean varies slowly with distance.[37] The main point of all these variants of kriging is in the way of formulating or estimating the mean of the process.

The estimation of kriging models often involves two subsequent steps: the first one is devoted to estimating the mean, which may depend on covariates, and the second one is devoted to estimating the correlation structure parameters, based on the residual or de-trended process that results from the first step. This procedure is known as residual kriging.[38] When the mean model is complicated, Kaufman et al.[39] advocate for separating the estimation of trend and covariance structures. The residual process is indeed hard to model in terms of correlations.[40] Some kriging applications may actually fail due to little spatial regularities left in the process after accounting for covariates.[41] By converse, missing covariates can be thought of as a source of residual correlation, which leaves room for spatio-temporal modeling.[42] Similar issues may arise when attempting to estimate both spatial and temporal correlations. One may thus generalize Kaufman et al.'s approach to separate the estimation of dependence for each domain, being it either space, time or covariates, thus borrowing predictive power from all the domains.[43]

The scalability of GPs is concerning in the case of large datasets. Accessible reviews are available on this topic, focusing on big data applications.[44] All scalable GPs generally involve approximating the information that could be extracted from a full dataset, provided a sufficient computational power. Data subsetting, partitioning,[45] or subsampling[46] involve retaining only a fraction of training data points: moving kriging,[47] as a local estimation method, belongs to this category. Related but distinct are local approximations such as Markov assumptions,[48] which are based on selecting a smaller training dataset that is relevant for a specific prediction task. Covariance tapering[18] involves bounded support kernels[49] that

induce sparsity in large correlation matrices, which are then more manageable both memory- and compute- wise. Other algebraic approximations allow for parsimonious, low-rank representations for large matrices: for instance, separability assumptions involve writing large matrices as a Kronecker product of multiple smaller matrices, which are easier to invert.[7]

Computational concerns make it difficult to deploy an otherwise interesting, functional variant of kriging called GP regression (GPR).[50] However, even plain GPs can be troubled within the same setting. The aforementioned approximations can be useful, but they require tuning of some sort. For instance, in Markov modeling, the extent of Markov neighborhoods is a tuning aspect of the problem.

The focus of this article is on separability assumptions, which may fit into the category of parsimonious, low-rank approximations for spatio-temporal correlation matrices.[17] Separability assumptions arise naturally in multivariate normal modeling. In the literature, one may find this model stated in terms of a normally distributed random vector with covariance matrices in Kronecker form, which come in handy even in problems with fewer data than the one we present here.[42] An iterative procedure attributed to Arendt et al.[51] allows to estimate the correlation structures in time and space domains. This technique consists in estimating the spatial correlation of data that have been temporally decorrelated, then estimating the temporal correlation of data that have been (by converse) spatially decorrelated; the two steps are iterated until convergence. What we highlight in our contribution is that the practitioner may find it useful to adopt some suitable time series models[19] for which the matrix transformations need not be evaluated explicitly, as it may be cumbersome. The advantages of separability have not been exploited in full yet. A likelihood function for separable kriging models in many domains has already been considered and maximized with iterative procedures, though without leveraging convenient correlation models.[52]

Separable models have already been addressed using CLs, which are nongenuine likelihoods based on working independence assumptions.[53] Pairwise CLs help avoid matrix inversion.[54] Basically, a CL allows to make inference even based on under-specified models, on trusted assumptions, thus making the analysis more robust[55] to model misspecification. Furthermore, under a separable spatio-temporal model, temporal and spatial parameters can be estimated separately for added robustness with respect to each other's misspecification.[56] The composite estimator provides stronger consistency guarantees in this sense.[57] The potential of composite estimation has also been investigated in terms of the other kinds of robustness that are common in statistics, with respect to outliers and data contamination.[58] Given these interesting results, the present article aims at presenting a composite estimation approach to the reader, as it promises to make kriging applications more computationally sustainable and, hopefully, robust.

3 | 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 $\mathcal{S} = \{1, \dots, S\}$ and is a pointer to one out of S locations in space, while the time index t takes its value in $\mathcal{T} = \{1, \dots, T\}$ and is a pointer to one out of T time frames. A joint index ts denotes 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 $s, s' \in \mathcal{S}$, 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, namely,

$$d_{s,s'} = \sqrt{(\mathrm{x}_s - \mathrm{x}_{s'})^2 + (\mathrm{y}_s - \mathrm{y}_{s'})^2}\,, \tag{1}$$

where $(\mathrm{x}_s, \mathrm{y}_s)$ and $(\mathrm{x}_{s'}, \mathrm{y}_{s'})$ are the two-dimensional Cartesian coordinates of $s, s' \in \mathcal{S}$, respectively. Temporal distance is defined analogously[59] for all pairs of $t, t' \in \mathcal{T}$ as

$$d_{t,t'} = |t - t'|\,, \tag{2}$$

with $|\cdot|$ denoting the absolute value. Here, $d_{s,s'}$ is the generic (s, s')th entry of the $S \times S$ spatial distance matrix d_S and $d_{t,t'}$ is the generic (t, t')th entry of the $T \times T$ temporal distance matrix d_T, respectively.

The observed data y are structured as follows:

$$y = \begin{bmatrix} y_{11} & \cdots & y_{1S} \\ \vdots & y_{ts} & \vdots \\ y_{T1} & \cdots & y_{TS} \end{bmatrix}, \tag{3}$$

so the data related to location s are all stored in the same sth column, while those related to the time frame t are all stored in the same tth 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 μ with the same number of rows and columns of y.

Let vec be a unary operator defined for matrices that stacks their columns into a single vector.[60] Y is assumed to be a multivariate normal with scale parameter $\sigma > 0$ and correlation matrix R, in the sense that R is the correlation matrix of $\mathrm{vec}(Y)$. More formally, we assume that Y has the following density function.

$$f(y; \mu, \sigma, R) = \frac{1}{\sqrt{|2\pi\sigma^2 R|}} \cdot \exp\left\{ -\frac{1}{2} \mathrm{vec}\left(\frac{y-\mu}{\sigma}\right)^{\top} R^{-1} \ \mathrm{vec}\left(\frac{y-\mu}{\sigma}\right) \right\}, \tag{4}$$

where $|\cdot|$ is the determinant of a square matrix. In kriging-related applications, R is a positive definite square matrix that depends on spatial and temporal distances through spatial and temporal autocorrelation functions discussed further in this section. The definition in use throughout this article will be given later in Equation (11).

We assume that the expected value of Y at time t and location s, denoted by μ_{ts}, is a smooth function of t and is shared across locations, so it makes sense to estimate it with the asymmetric moving average m_t defined below, which pools data from across all the observed locations, limited to w time frames preceding t. The trend μ_{ts} is thus estimated via $\hat{\mu}_{ts}$, defined as

$$\hat{\mu}_{ts} = m_t, \quad \text{for all } s, \text{with} \quad m_t = \frac{1}{S \cdot w} \sum_{s=1}^{S} \sum_{i=1}^{w} Y_{(t-i)s} \ . \tag{5}$$

Here, $Y_{(t-i)s}$ is the process at time $t - i$ and location s. We set w equal to the number of observations per sensor in 24 h to remove the observed daily seasonality. The latest estimate available for μ_{ts} can also serve as an estimate for future trend $\mu_{t's}$, $t' > t$, assuming stability in the short term. Such an assumption can be credible when the univariate time series may agree on a single latent factor ruling all of them.

The parameter $\sigma > 0$ contributes only to making predictions probabilistically calibrated,[61] 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.

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.[62] 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_{ts} and $Y_{t's'}$, is modeled as

$$\mathrm{cov}(Y_{ts}, Y_{t's'}) = \sigma^2 \cdot \mathrm{cor}_S(d_{s,s'}) \cdot \mathrm{cor}_T(d_{t,t'}) \ , \tag{6}$$

where $\mathrm{cor}_S(\cdot)$ is the spatial autocorrelation function (ACF),[63] while $\mathrm{cor}_T(\cdot)$ is the temporal ACF;[19] few examples of definition for $\mathrm{cor}_S(\cdot)$ and $\mathrm{cor}_T(\cdot)$ are provided in the next paragraph. 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.[1,2,6,7,64]

For the sake of illustration, we recall few possible definitions for the spatial ACF $\mathrm{cor}_S(\cdot)$ and the temporal ACF $\mathrm{cor}_T(\cdot)$. The spatial ACF can be, for instance, one of the following:[13,16,59]

- Matérn ACF[65]

$$\mathrm{cor}_S(d) = \frac{2^{1-\alpha}}{\Gamma(\alpha)} (d/\lambda)^{\alpha} \mathrm{K}_{\alpha}(d/\lambda) \ , \tag{7}$$

with $\Gamma(\cdot)$ the gamma function and $K_\alpha(\cdot)$ the modified Bessel function of the second kind; $\alpha > 0$ is a smoothing parameter, $\lambda > 0$ is a range parameter.

- Power exponential ACF[1]

$$\mathrm{cor}_S(d) = \exp\{-(d/\lambda)^\beta\} \,. \tag{8}$$

Here, $\beta > 0$ is a smoothing parameter, $\lambda > 0$ is a range parameter.

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,[1,66] which is also known as squared-exponential ACF[3,4,67] and involved in the double-stable model.[2]

The temporal ACF $\mathrm{cor}_T(\cdot)$ in discrete time can be, for instance, one of the following:[19]

- ACF of a stationary AR of order 1

$$\mathrm{cor}_T(d) = \phi^{|d|} \,, \tag{9}$$

where $\phi \in]-1, +1[$ is the correlation parameter.

- ACF of an invertible moving-average model of order 1

$$\mathrm{cor}_T(d) = \begin{cases} 1, & d = 0, \\ \eta, & d = \pm 1, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\eta \in]-1/2, +1/2[$ is the correlation parameter.

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

$$R = R_S \otimes R_T \,, \tag{11}$$

the spatio-temporal correlation matrix, this expression denoting also a Kronecker correlation structure due to the Kronecker product $\otimes$.

Both temporal and spatial ACFs can be modified to account for noisy data by including a so-called nugget effect.[1,65] This means that the spatial ACF $\mathrm{cor}_S(d)$ and the temporal ACF $\mathrm{cor}_T(d)$ are multiplied by β_S and β_T when $d > 0$, respectively, the parameters β_S and β_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.[64,68] The latter way of modeling is more natural in the case of additive covariance models.[69]

4 | 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 CL approach. Then, leveraging

the peculiar expression of the kriging mean formula, we show how to compute predictions efficiently under separability.

4.1 | Estimation

The distribution of Y in Equation (4) is assumed to be indexed by μ, σ, R. As per residual kriging, μ is replaced with its estimate $\hat{\mu}$ via Equation (5). The parameters left to be estimated are $\theta = (\sigma, \psi^\top)^\top$, with ψ the correlation parameters ruling R. Moreover, ψ can be partitioned as $\psi = (\psi_S^\top, \psi_T^\top)^\top$, where ψ_S and ψ_T are the spatial and temporal correlation parameters, respectively. In particular, R_S depends only on ψ_S, while R_T depends only on ψ_T. As a starting point, we consider the pseudo likelihood function $\mathcal{L}(\theta)$, defined as

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; \tilde{y}) = f(\tilde{y}; 0, \sigma, R) \,, \tag{12}$$

with $\tilde{y} = y - \hat{\mu}$ the de-trended data. The maximizer of $\mathcal{L}(\theta)$ with respect to θ is referred to as the ML estimate. Under the separability assumption in Equation (11), it is possible to evaluate ML efficiently as illustrated in Appendix C. The ML approach is the natural benchmark for any other estimator, due to its asymptotic efficiency. Nevertheless, robustness concerns and computational issues may arise in practical data analysis, for which the following CL estimation approach can be more suited.

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

$$\hat{M} = \frac{1}{\hat{\sigma}^2} \operatorname{vec}(y - \hat{\mu}) \operatorname{vec}(y - \hat{\mu})^\top \,, \tag{13}$$

where $\hat{\sigma}$ is a working estimate of σ, defined as

$$\hat{\sigma} = \sqrt{\frac{1}{TS} \sum_{s=1}^{S} \sum_{t=1}^{T} (y_{ts} - \hat{\mu}_{ts})^2} \,. \tag{14}$$

The likelihood function is thus replaced with a so-called pseudo likelihood $\mathcal{L}^p(\psi)$, defined as follows, that allows to make inference on ψ alone.[65,70]

$$\mathcal{L}^p(\psi) = |R|^{-1/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left(R^{-1} \hat{M}\right) \right\} \,. \tag{15}$$

Here, $\operatorname{tr}(\cdot)$ is the trace operator.

Kriging is often cumbersome due to the inversion of R, and actually the pseudo likelihood in Equation (15) is intractable with high-dimensional data. Separability reduces the dimensionality of the problem, as it holds

$$R^{-1} = (R_S \otimes R_T)^{-1} = R_S^{-1} \otimes R_T^{-1} \,, \tag{16}$$

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 CL approach[56] can be used, exploiting separability more in depth so that the tasks of estimating ψ_S and ψ_T can even be addressed separately. Second, a suitable time series model can help handle the temporal correlation implicitly, as R_T must not be evaluated at all, and make it possible to address tall data and high sampling rates.

CLs are known in spatial statistics mainly as tools that simplify estimation and inference, like the pairwise likelihood,[53] and they can also be used in model selection.[71] CLs allow to make inference on under-specified models but, even in the case of fully specified models, like in kriging, a suitable CL can reduce the computational cost of estimation. The estimator based on the full model can be computed in some cases, but it would be generally cumbersome with large datasets, without the convenient temporal models discussed here. 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 ML estimator. Nonetheless, the loss of efficiency might be not significant when dealing with high-frequency data.

With CLs, as with any so-called pseudo likelihood, estimation variance cannot be assessed as with classical likelihood functions based on the Hessian matrix. Parametric bootstrap can be used[72] 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 $t \in \mathcal{T}$, it holds $R = R_S$, as per Equation (11). Let the spatial pseudo CL be as follows,

$$\mathcal{L}_S(\psi_S) = \prod_{t \in \mathcal{T}} f(\tilde{y}_t; 0, \hat{\sigma}, R_S) = |R_S|^{-T/2} \exp\left\{ -\frac{T}{2} \mathrm{tr}\left(R_S^{-1} \hat{M}_S\right) \right\} , \tag{17}$$

with $\tilde{y}_t$ the de-trended data vector at time frame t. This allows to make inference on the spatial correlation parameters ψ_S alone. The expression is the same of a "small blocks" marginal CL.[56] Here, $\hat{M}_S$ is the sample correlation matrix between the univariate time series at distinct locations, defined as

$$\hat{M}_S = \frac{1}{T\hat{\sigma}^2} (y - \hat{\mu})^\top (y - \hat{\mu}) . \tag{18}$$

The estimator $\hat{\psi}_S$ for the spatial correlation parameters ψ_S is defined as the maximizer of (17). If R_S is a smooth function of ψ_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:

$$\alpha \ : \ \mathrm{tr}\left\{ \left(R_S - \hat{M}_S\right) \frac{\partial R_S^{-1}}{\partial \alpha} \right\} = 0 . \tag{19}$$

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 (15) and (17), but the correlation matrices involved are strikingly different in their sizes, as R_S is much smaller than R in practical applications.

The temporal correlation parameter ψ_T can be estimated analogously to ψ_S, by defining a temporal pseudo CL $\mathcal{L}_T$ as

$$\mathcal{L}_T(\psi_T) = \prod_{s \in \mathcal{S}} f(\tilde{y}_s; 0, \hat{\sigma}, R_T) = |R_T|^{-S/2} \exp\left\{ -\frac{S}{2} \mathrm{tr}\left(R_T^{-1} \hat{M}_T\right) \right\} , \tag{20}$$

with $\tilde{y}_s$ the de-trended univariate time series at location s. Here, $\hat{M}_T$ is the sample temporal correlation matrix, defined as

$$\hat{M}_T = \frac{1}{S\hat{\sigma}^2} (y - \hat{\mu})(y - \hat{\mu})^\top . \tag{21}$$

This operation 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, to simplify the estimation of ψ_T otherwise based on directly maximizing (20). In particular, we resort to an AR model with multiple overlapping seasonal lags. For its definition, the classical lag operator B can be introduced, such that

$$BY_{ts} = Y_{(t-1)s} , \tag{22}$$

and $B^\Delta Y_{ts} = Y_{(t-\Delta)s}$ more in general, where $Y_{(t-\Delta)s}$ is the process Y at time $t - \Delta$ and location s. Then, letting ε_s be a white noise time series process, the multiplicative seasonal AR model is defined as follows.[19]

$$\left[\prod_{k=1}^{K} (1 - \phi_k B^{\Delta_k}) \right] \cdot (Y_s - \mu_s) = \varepsilon_s , \tag{23}$$

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 (23) reduces the temporal correlation parameter to $\psi_T = (\phi_1, \dots, \phi_K)^\top$. 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.[73] In this sense, we define

$$V_s^{-h} = \left[\prod_{k \neq h}(1 - \phi_k B^{\Delta_k})\right] \cdot (Y_s - \mu_s) \,, \tag{24}$$

and, by Equation (23), it holds

$$(1 - \phi_h B^{\Delta_h}) V_s^{-h} = \varepsilon_s \,, \tag{25}$$

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

$$\phi_h = \mathrm{cor}(V_s^{-h}, B^{\Delta_h} V_s^{-h}) \,. \tag{26}$$

This relation motivates an iterative procedure that loops over the correlation parameters, and repeats until convergence. For each ϕ_h, one evaluates the current estimate of the process V_s^{-h} and then updates ϕ_h as the sample ACF of V_s^{-h} at lag Δ_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,[74,75] which require additional tuning.

4.2 | Prediction

Separability assumptions also simplify 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 μ', while previously introduced matrices Y and μ are now related to the available data y. The correlation matrix of $\mathrm{vec}(Y')$ is R', whereas R is the correlation matrix of $\mathrm{vec}(Y)$. The cross-correlation matrix between $\mathrm{vec}(Y')$ and $\mathrm{vec}(Y)$ is generally not square, and it is denoted by ρ, which is based on a suitable joint distance matrix. Like R in Equation (11), also R' and ρ are defined in terms of Kronecker products as, respectively,

$$R' = R'_S \otimes R'_T \,, \quad \rho = \rho_S \otimes \rho_T \,, \tag{27}$$

where R'_S and R'_T are the spatial and temporal correlation matrices of Y', analogously to R_S and R_T, while ρ_S and ρ_T are the spatial and temporal cross-correlation matrices between Y' and Y. These in turn depend on suitable spatial and temporal distance matrices. As typical in kriging, we treat Y and Y' as jointly normally distributed.

Kriging revolves around the conditional distribution of Y' given $Y = y$,[1] though it was originally motivated as the linear unbiased prediction that is optimal with respect to the squared prediction error criterion.[76] Conditional to $Y = y$, the distribution of Y' is multivariate normal. The conditional mean matrix of Y', denoted by $\hat{y}'$, and the conditional variance-covariance matrix of $\mathrm{vec}(Y')$, denoted by R'_{cond}, satisfy the following conditions.

$$\mathrm{vec}(\hat{y}') = \mathrm{vec}(\mu') + \rho R^{-1} \mathrm{vec}(y - \mu) \,, \quad R'_{\mathrm{cond}} = R' - \rho R^{-1} \rho^\top \,. \tag{28}$$

Let β_S and β_T be spatial and temporal regression coefficients, defined as

$$\beta_S = \rho_S R_S^{-1} \,, \quad \beta_T = \rho_T R_T^{-1} \,, \tag{29}$$

then the kriging mean formula in Equation (28) can be written as

$$\hat{y}' - \mu' = \hat{z} \beta_S^\top \,, \tag{30}$$

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

$$\hat{z} = \beta_T(y - \mu) \ . \tag{31}$$

We prove this fact in Appendix A.1. The proof also covers existing applications of multivariate normal models.[77]

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 (31), is temporal forecasting only within sensor locations. The second step, in Equation (30), is the spatial interpolation of such forecasts for the needed locations. Then, spatio-temporal predictions can be seen as spatial interpolation of temporal forecasts. So, a separability assumption allows to separate domains also in prediction. Distributed calculus can be used for evaluating Equation (31), as each univariate time series is transformed separately by the matrix product.

This insight into the meaning of matrix quantities cannot be found in other works,[77] as they lack both the interpretation of β_S and β_T. Moreover, an automatic application of their findings would be too compute-intensive with large datasets, because it would require evaluating β_T explicitly.

Correlation models affect prediction only through β_S and β_T. So, any specification of time or space models can be employed if it implies tractable prediction. This view motivates, for instance, the use of general interpolators, or state-space models,[78] or integrated AR time series models,[19] that allow for simple prediction since β_T in Equation (29) is sparse and explicit. The rather general auto-regressive moving-average model (ARMA) has already been considered by some authors[79] 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_{jts} denotes the jth variable at spatio-temporal coordinates ts. In this case, the marginal means μ_{jts} will likely depend on j. The correlation structure can be simplified according to a fully factored model,[7] such that

$$\operatorname{cov}(Y_{jts}, Y_{j't's'}) = \gamma_{j,j'} \cdot \operatorname{cor}_S(d_{s,s'}) \cdot \operatorname{cor}_T(d_{t,t'}) \ . \tag{32}$$

Here, the new quantity $\gamma_{j,j'}$ represents the cross-sectional covariance between the jth and j'th variables at the same time and location. Thus, under a fully factored model, our kriging mean formula scales easily, as an additional set of regression coefficients β_C is defined besides β_S and β_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 $Y = y$. Let $\operatorname{diag}(M)$ be the operator that returns the diagonal entries of a square matrix M as a column vector. Then, $\operatorname{vec}(V) = \sigma^2 \operatorname{diag}(R'_{\text{cond}})$, so it holds

$$V = \sigma^2 \left\{ \operatorname{diag}(R'_T) \operatorname{diag}(R'_S)^\top - \operatorname{diag}(R'_T - R'_{T,\text{cond}}) \operatorname{diag}(R'_S - R'_{S,\text{cond}})^\top \right\} \ , \tag{33}$$

where $R'_{S,\text{cond}}$ and $R'_{T,\text{cond}}$ are defined as follows.

$$R'_{S,\text{cond}} = R'_S - \rho_S R_S^{-1} \rho_S^\top \ , \quad R'_{T,\text{cond}} = R'_T - \rho_T R_T^{-1} \rho_T^\top \ . \tag{34}$$

We prove this result in Appendix A.2. Equation (33) has an advantage over direct evaluation of (28), 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.

5 | EMPIRICAL APPLICATION

Now we illustrate our proposed methodology on the SAL dataset presented in the introduction. The dataset 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. We performed all our analyses in the statistical computing environment R.[80] Both the ML and CL estimators were given a custom implementation that minimizes the conditional sum of squares in the estimation of the temporal model and maximizes a spatial pseudo likelihood for the estimation of the spatial model. All the analyses were carried out using typical laptop computers.

5.1 | Explorative analysis

The SAL dataset consists of 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 8 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 regime, but a longer training phase might not be reasonable for some applications, where a monitoring system shall be calibrated in short amounts of time.

Data were missing less than 1% of the times, resulting from miscommunication faults unrelated to the data and thus statistically random. Kriging can handle missings at random, but we used a simpler imputation method called *last observation carried forward* (LOCF), which uses the last valid reading to impute missings.[81] In fact, we require gridded data and the LOCF approach solves the problem rapidly and efficiently, so we can focus on other aspects of the problem.

Figure 2 presents the training set. The sensors are numbered from 1 to 12 as in Figure 1. Troughs are concentrated in the mornings, as windows are opened, and cold air flows in the room during routine cleaning. Peaks are concentrated around noon, as direct sunlight overheats the sensors facing south, the ones numbered 5, 10, and 11. Weekly trends are highlighted in Figure 3, with temperature median and other percentiles reported throughout weekdays. Troughs seem to occur mostly on Mondays and Fridays, so on the first and last workdays in the week. Peaks instead concentrate on Fridays and weekends.

Similar conditions between subsequent days motivate using time series models with seasonal components, the period being one day long. Similar events occurring on the same weekday motivate considering one more seasonal component, whose period should be 1 week long.

One may consider adding covariates into the model, in particular the physical variables mentioned in the introduction. Some preliminary analyses show that their explanatory power is limited, since spatial and temporal dependencies seem able to explain most of the variability of temperature data. As a consequence, they will not be considered further.

5.2 | Model estimates

We compared our proposed CL estimator with the ML approach. The model chosen was separable, as discussed in previous sections. We considered four candidate models in the space domain, namely, exponential, Gaussian, power exponential, and Matérn ACFs. In the time domain, we considered a few candidate models, all multiplicative AR as in Equation (23), with a short term lag Δ_1 and two seasonal lags, $\Delta_2 = 1$ day and $\Delta_3 = 1$ week. We probed some alternative values for the short term lag Δ_1, ranging in 10, 20, 30, 40 min, 1, 2, 3, 4, 6, 12 h. The longer lags were of interest because they implied higher estimates of ϕ_2 and ϕ_3, so that one could leverage seasonal dynamics and spatial regularities. All the candidate spatio-temporal models were then compared on the test set in terms of their performance in the spatial interpolation of each sensor based on the other ones, see Figure 5 in the following paragraphs.

The estimation of all the spatio-temporal models via ML took no more than 10 min. The estimation of spatial models via CL was essentially instantaneous by comparison, as the small size of the sample spatial correlation matrix made the computation fast. The estimation of temporal models via CL takes a few seconds, namely about one-fourth of the time required for ML. The four spatial ACFs estimated with either ML or CL appears to be very similar. In Figure 4, this similarity is illustrated based on just the CL estimate.

Figure 5 summarizes the predictive performance of the estimated spatio-temporal models on the test set. The performance is assessed in terms of spatial interpolation of each sensor, surrogated via all the others in turn. Three performance metrics are used, namely, the mean absolute prediction error (L1), the root of mean square error (L2), and the 95th percentile of the absolute prediction error (P95). In comparing ML and CL methods, these are seen to provide model estimates that turn out to be rather interchangeable for predictive tasks. Sensor 10 is uniformly poorly surrogated by the other sensors, which marks its outlying nature in the example. This sensor is better accommodated by the CL estimate, in line with its higher degree of robustness. CL seems otherwise a good approximation to ML in general.

Lastly, Table 1 summarizes the CL estimates and standard errors for the four spatio-temporal models we are going to consider in the next section with $\Delta_1 = 10$ min. Point estimates of the temporal parameters are shared across the four spatio-temporal models. Their standard errors are based on parametric bootstrap, as outlined in Appendix B. The

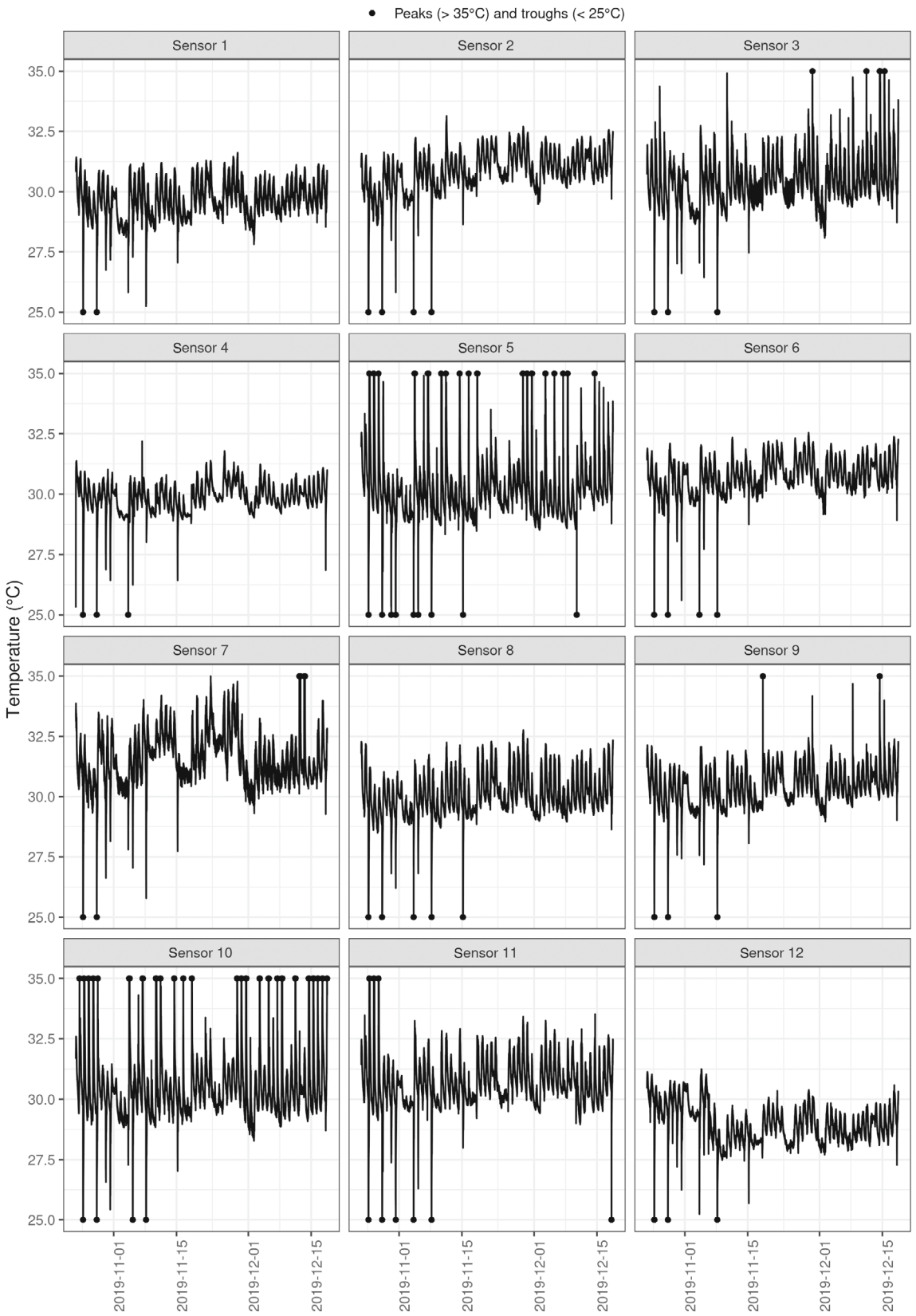

FIGURE 2 Univariate time series of temperature per sensor, training set (October through December). Peaks and troughs are thresholded and highlighted with dots.

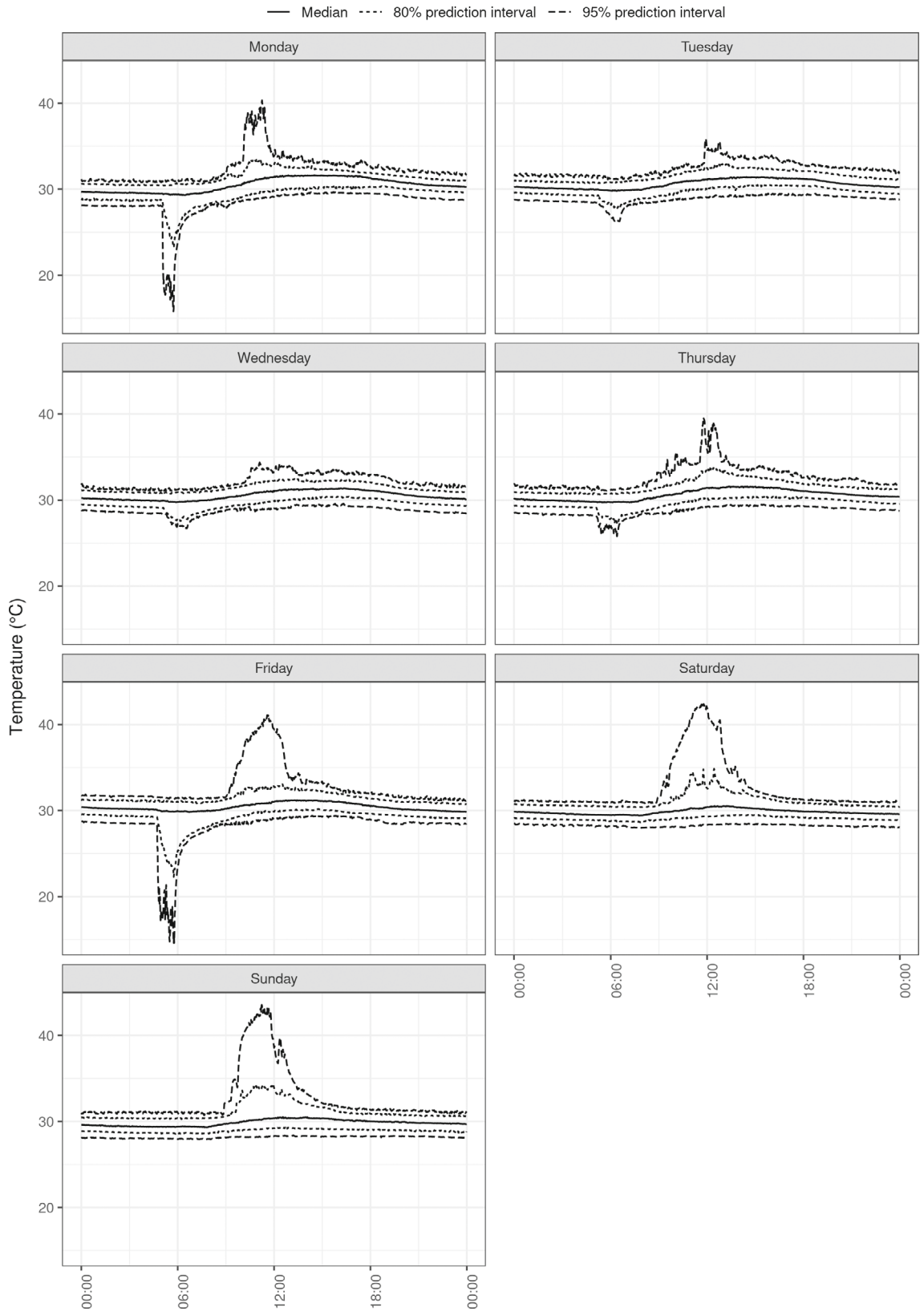

FIGURE 3 Sample quantiles of temperature, aggregation according to the weekday, training set (October through December)

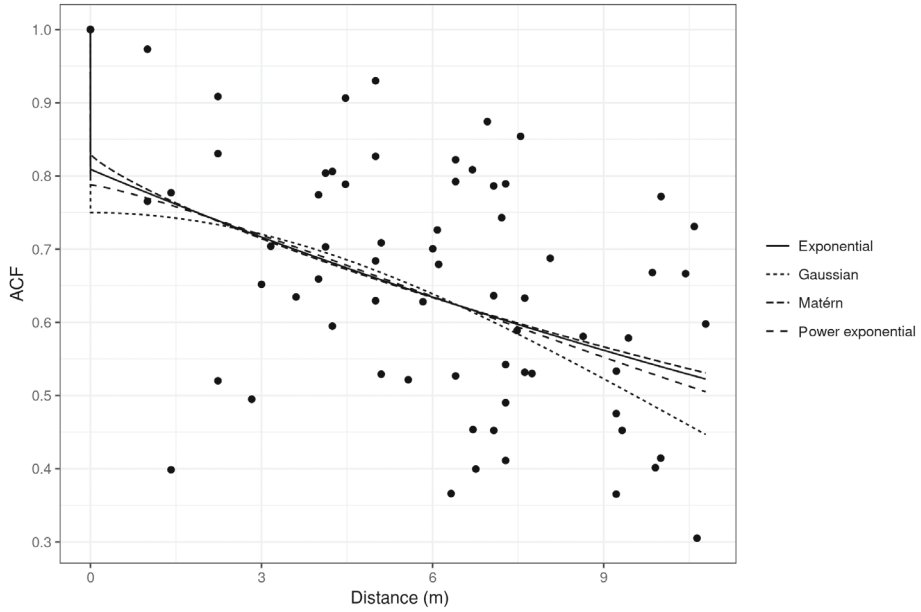

FIGURE 4 Spatial ACFs estimated via CL (lines) with sample correlations (dots). See Table 1 for the estimated parameters

estimation variance does, instead, depend on the joint spatio-temporal model, so it is different across them. For bootstrap, we carried out inference based on parametric bootstrap by simulating and analyzing 1000 datasets for each joint spatio-temporal model. Performing the bootstrap took about 3 h on a single laptop computer based on the CL approach. Standard errors, all multiplied by 10^3 in the table, are all rather small, as expected based on the large sample size.

5.3 | Sensor network optimization

In the previous subsection, CL turned out to be capable of surrogating the efficiency of ML, with the help of a large training dataset. Now a practical concern is the selection of a few operation sensors from the initial set, as twelve of them are too many for a room that is 127 m^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*[82] with predictions from the remaining sensors.

Different network configurations can be evaluated and compared according to a metric, which should reflect the priorities and objectives of stakeholders. The 95th percentile of absolute prediction errors on all but active sensors[9] can be used for an approximate minimax decision. For comparison, we illustrate a sensor selection based on this criterion alongside one that uses the 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 metrics. 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 6, 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 min 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.[9] Based on performances in Figure 6, 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.

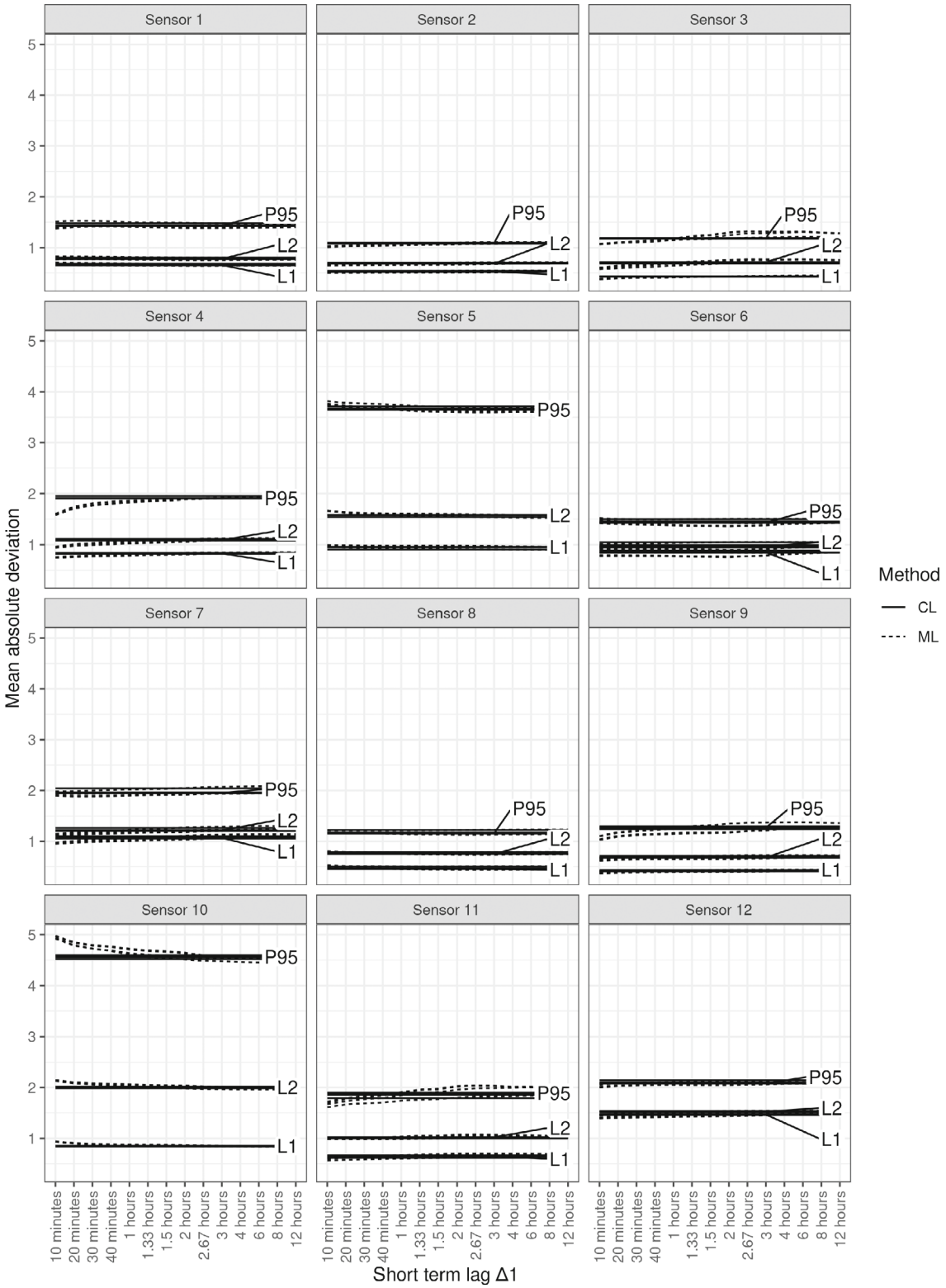

FIGURE 5 Performance in spatial interpolation of each sensor based on the other ones, for ML and CL (linetype) with different metrics (labels, *y*-axis) depending on the short term lag Δ_1 (*x*-axis). Multiple lines refer to distinct spatial models, essentially overlapped. Metrics in use are: mean absolute prediction error (L1), root mean square error (L2), and 95th percentile of the absolute prediction error (P95).

TABLE 1 Point estimates and standard errors for spatial and temporal correlation parameters, case $\Delta_1 = 10$ min

Spatial model	Parameter	Est.	Std. err. ($\times 10^3$)
Exponential	ϕ_1	0.977	0.206
	ϕ_2	0.078	0.950
	ϕ_3	0.047	0.990
	Nugget	0.191	2.307
	Range	24.692	446.329
Gaussian	ϕ_1	0.977	0.203
	ϕ_2	0.078	0.936
	ϕ_3	0.047	0.975
	Nugget	0.250	2.470
	Range	15.011	140.078
Matérn	ϕ_1	0.977	0.206
	ϕ_2	0.078	0.950
	ϕ_3	0.047	0.991
	Nugget	0.187	4.805
	Range	25.993	1547.047
	Smoothness	0.479	22.544
Power exponential	ϕ_1	0.977	0.205
	ϕ_2	0.078	0.946
	ϕ_3	0.047	0.986
	Nugget	0.217	3.057
	Range	19.883	424.822
	Smoothness	1.312	28.654

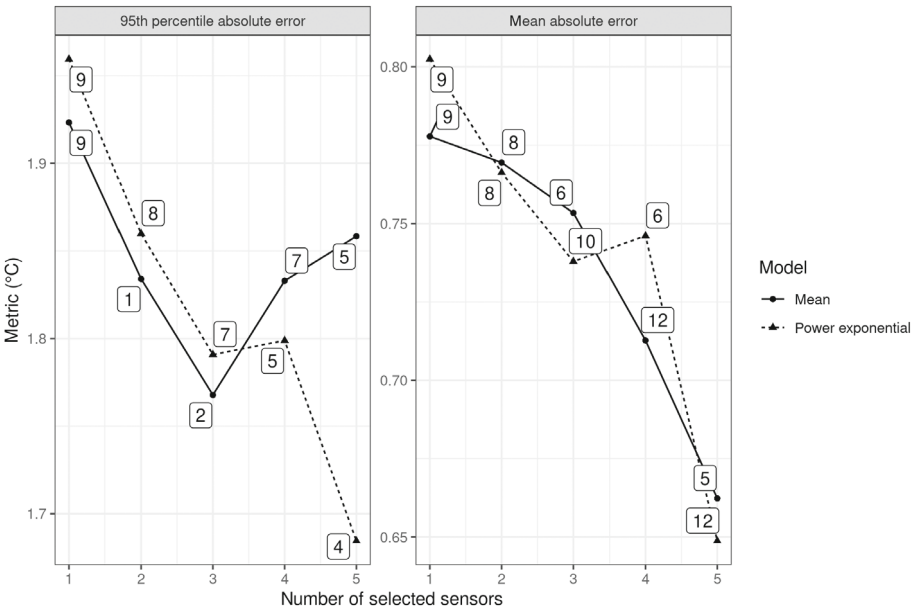

FIGURE 6 Forward selection of sensors, distinct per metric, prediction error versus number of selected sensors. The sensor added at each step appears within a box. Run on test set (December through March)

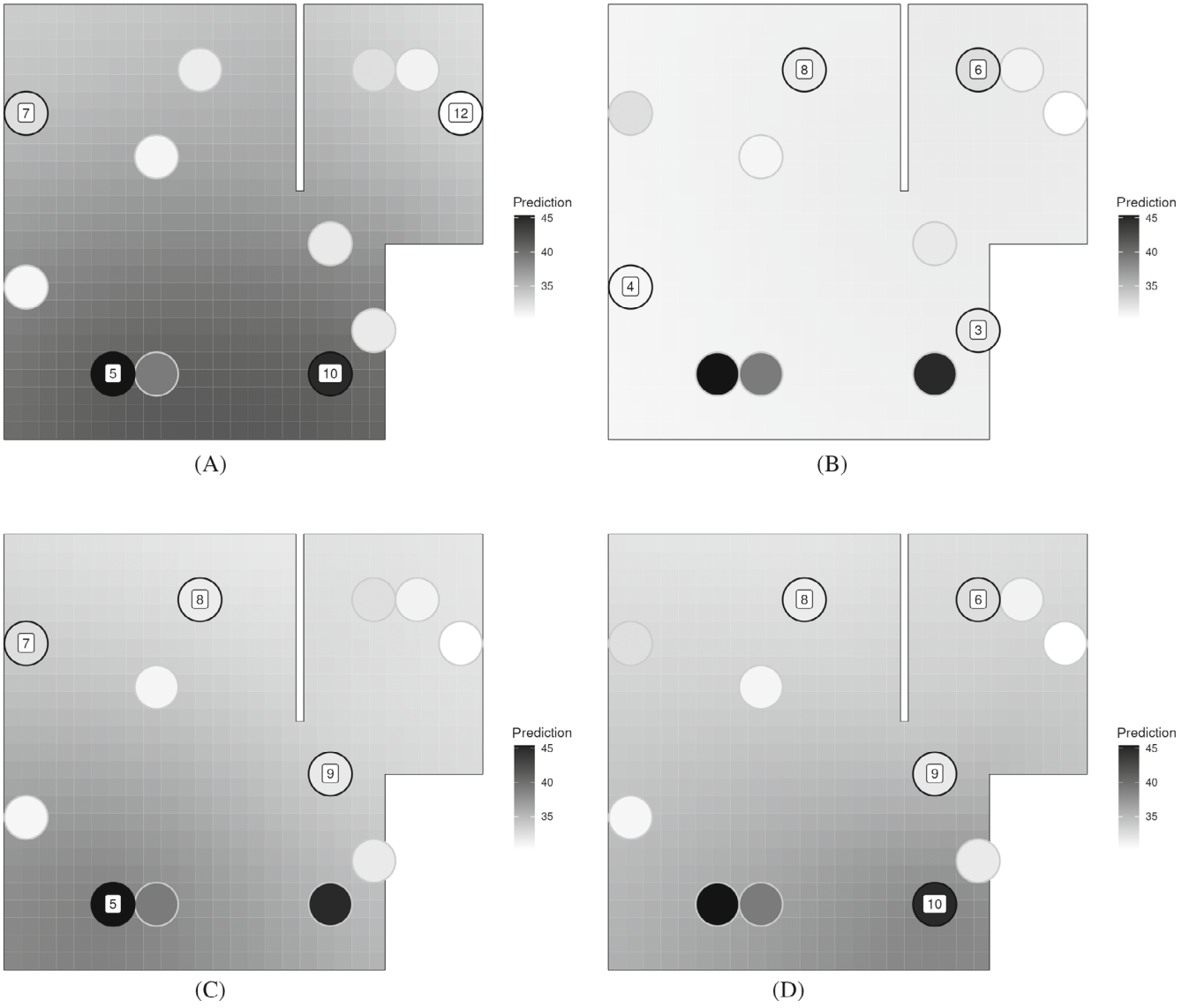

FIGURE 7 Examples of prediction via four sensors, based on the power exponential model. Forecasts for February 17, 12:00, based on data from 10 min before. Circles are colored based on *observed* temperatures, the floor based on *predicted* temperatures, selected sensors only are numbered as in Figure 1. (A) Anomalous sensors, (B) regular sensors, (C) selection via 95th percentile absolute error, and (D) selection via mean absolute error

5.4 | Behavior of the monitoring system

Using different sensor configurations, we show a few examples of actual situations and 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 min forward for the whole floor plan, as in Figure 7.

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 1A. On the contrary, a system based on more regular sensors only, like in Figure 1B, 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 1C reports the prediction provided under the power exponential ACF by the four best sensors according to the percentile metric. Figure 1D shows the selection based on the

mean absolute error instead. Both configurations attempt to replicate the north-south gradient, which requires to choose 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.

6 | 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 simplifying server-side computations.

6.1 | Nonstationary modeling

The kriging approach relies on stationary ACF models, which offer a wide variety of possibilities, but some problems may be addressed only with nonstationary 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 solution could be useful in cases like ours, where distinct sensors might have different equilibria. We used a moving average to accommodate nonstationarity 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 nonstationarity, resulting from locally linear trends, so it might be useful to consider a nonstationary 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 nonstationarity due to both a moving average trend and a near-unit first AR coefficient, which implied integration *de facto*.

6.2 | Sensor-specific temporal correlation parameters

A limitation of separable ACFs is that they imply the same marginal dynamics for all locations. 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 CL and use different portions of data, so they will not affect each other. In mean prediction, as per Equation (31), $\hat{z}$ 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 (30), 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.

- It adds flexibility to 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 of 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 that 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.

7 | 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 CL estimator and optimized calculus in prediction. The assumption of separability allows to distribute calculus across the sensor network by delegating as many operations as possible to the components that gather the relevant data. Separability is a strong assumption, though, which can be trusted at least in settings close to ours with sensor data from indoor environments. Probably, when considering less controlled environments and even weather data, this assumption is restrictive and unrealistic, and other nonseparable models may be considered instead. Future research may investigate the viability of CL approaches also in such settings.

To our knowledge, the use of marginal CL 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 6.2. This idea relates to stratified variograms,[16,83] but it has even more in common with the estimation of a single variogram with data pairs sharing some identical conditions.[84] The CL approach is thus computationally convenient and potentially more robust to model misspecification, as a wrong temporal model does not affect the estimation of the spatial model, and vice versa. There is, necessarily a loss in efficiency, which may be more noticeable in small samples, but should be less relevant to big data applications. In the case of challenging estimation problems, CL estimates are also readily available and may help initialize other iterative estimation procedures, if a more efficient estimate is of interest.

The predictive part of our approach was already common in climate and weather sciences but mostly confined to spatial interpolation.[85] 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,[86] process convolution,[87] and linear mixed models,[69] 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.[88]

Kriging computation is hugely simplified by assuming separability and 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 (23) are simple to calculate as well. Our approach allows to blend together and leverage known strengths of time series analysis and spatial statistics without the need to outline a joint spatio-temporal framework from scratch.

ACKNOWLEDGMENTS

This work was supported by the Competence Centre ASSIC – Austrian Smart Systems Integration Research Center, co-funded by the Austrian Federal Ministry for Transport Innovation and Technology (BMVIT), the Austrian Federal Ministry for Education, Science and Research (BMWFW), and the Austrian Federal Provinces of Carinthia and Styria within the Competence Centres for Excellent Technologies Programme (COMET). 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). Open Access Funding provided by Universita degli Studi di Udine within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Github at https://github.com/dslab-uniud/Virtual-sensing.

ORCID

Michele Lambardi di San Miniato https://orcid.org/0000-0003-2423-4250
Ruggero Bellio https://orcid.org/0000-0002-7633-087X
Luca Grassetti https://orcid.org/0000-0003-1997-8001
Paolo Vidoni https://orcid.org/0000-0003-0063-011X

REFERENCES

1. Gramacy RB. *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. 1st ed.; Chapman and Hall/CRC; 2020.
2. Nguyen L, Hu G, Spanos CJ. Spatio-temporal environmental monitoring for smart buildings. Proceedings of the 13th IEEE International Conference on Control & Automation; 2017; IEEE, Ohrid, North Macedonia.
3. Carpenter J, Woodbury KA, O'Neill Z. Using change-point and Gaussian process models to create baseline energy models in industrial facilities: a comparison. *Appl Energy*. 2018;213:415-425. doi:10.1016/j.apenergy.2018.01.043
4. Liu H, Yang C, Huang M, Wang D, Yoo C. Modeling of subway indoor air quality using Gaussian process regression. *J Hazard Mater*. 2018;359:266-273. doi:10.1016/j.jhazmat.2018.07.034
5. Li H, Yu D, Braun JE. A review of virtual sensing technology and application in building systems. *HVAC&R Res*. 2011;17(5):619-645.
6. Rodríguez-Iturbe I, Mejía JM. The design of rainfall networks in time and space. *Water Resour Res*. 1974;10(4):713-728. doi:10.1029/wr010i004p00713
7. Mardia KV, Goodall CR. Spatial-temporal analysis of multivariate environmental monitoring data. In: Patil GP, Rao CR, eds. *Multivariate Environmental Statistics*; Elsevier; 1993:347-385.
8. Dong B, Lam KP. A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting. *Build Simul*. 2013;7(1):89-106. doi:10.1007/s12273-013-0142-7
9. Brunello A, Urgolo A, Pittino F, Montvay A, Montanari A. Virtual sensing and sensors selection for efficient temperature monitoring in indoor environments. *Sensors*. 2021;21(8):2728. doi:10.3390/s21082728
10. Ferdoush S, Li X. Wireless sensor network system design using Raspberry Pi and Arduino for environmental monitoring applications. *Proc Comput Sci*. 2014;34:103-110. doi:10.1016/j.procs.2014.07.059
11. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2016:785-794; New York.
12. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735
13. Oktavia E, Widyawan MIW. Inverse distance weighting and kriging spatial interpolation for data center thermal monitoring. Proceedings of the 1st International Conference on Information Technology, Information Systems and Electrical Engineering; 2016; IEEE, Yogyakarta, Indonesia.
14. Azzalini A, Scarpa B. *Data Analysis and Data Mining: An Introduction*. Illustrated ed. Oxford University Press; 2012.
15. Aryaputera AW, Yang D, Zhao L, Walsh WM. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. *Sol Energy*. 2015;122:1266-1278. doi:10.1016/j.solener.2015.10.023
16. Bivand RS, Pebesma E, Gómez-Rubio V. *Applied Spatial Data Analysis with R*. Springer; 2013.
17. Genton MG. Separable approximations of space-time covariance matrices. *Environmetrics*. 2007;18(7):681-695. doi:10.1002/env.854
18. Furrer R, Genton MG, Nychka D. Covariance tapering for interpolation of large spatial datasets. *J Comput Graph Stat*. 2006;15(3):502-523. doi:10.1198/106186006x132178
19. Box GEP, Jenkins GM, Reinsel GC, Ljung GM. *Time Series Analysis: Forecasting and Control*. 5th ed. Wiley; 2015.
20. Ruybal CJ, Hogue TS, McCray JE. Evaluation of groundwater levels in the Arapahoe aquifer using spatiotemporal regression kriging. *Water Resour Res*. 2019;55(4):2820-2837. doi:10.1029/2018wr023437
21. Thai CH, Tran TD, Phung-Van P. A size-dependent moving kriging meshfree model for deformation and free vibration analysis of functionally graded carbon nanotube-reinforced composite nanoplates. *Eng Anal Bound Elem*. 2020;115:52-63. doi:10.1016/j.enganabound.2020.02.008
22. Slot RMM, Sørensen JD, Sudret B, Svenningsen L, Thøgersen ML. Surrogate model uncertainty in wind turbine reliability assessment. *Renew Energy*. 2020;151:1150-1162. doi:10.1016/j.renene.2019.11.101
23. García-Gutiérrez A, Gonzalo J, Domínguez D, López D. Stochastic optimization of high-altitude airship envelopes based on kriging method. *Aerosp Sci Technol*. 2022;120:107251. doi:10.1016/j.ast.2021.107251
24. Qian HM, Li YF, Huang HZ. Time-variant reliability analysis for industrial robot RV reducer under multiple failure modes using kriging model. *Reliab Eng Syst Saf*. 2020;199:106936. doi:10.1016/j.ress.2020.106936
25. Luo Y, Xing J, Kang Z. Topology optimization using material-field series expansion and kriging-based algorithm: an effective non-gradient method. *Comput Methods Appl Mech Eng*. 2020;364:112966. doi:10.1016/j.cma.2020.112966

26. Yang Y, Christakos G, Yang X, He J. Spatiotemporal characterization and mapping of PM2.5 concentrations in southern Jiangsu province. *China Environ Pollut*. 2018;234:794-803. doi:10.1016/j.envpol.2017.11.077
27. Mohammadi V, Dehghan M, Khodadadian A, Wick T. Numerical investigation on the transport equation in spherical coordinates via generalized moving least squares and moving kriging least squares approximations. *Eng Comput*. 2019;37(2):1231-1249. doi:10.1007/s00366-019-00881-3
28. Menafoglio A, Gaetani G, Secchi P. Random domain decompositions for object-oriented kriging over complex domains. *Stoch Env Res Risk A*. 2018;32(12):3421-3437. doi:10.1007/s00477-018-1596-z
29. Chen J, Mak S, Joseph VR, Zhang C. Function-on-Function kriging, with applications to three-dimensional printing of aortic tissues. *Technometrics*. 2020;63(3):384-395. doi:10.1080/00401706.2020.1801255
30. Stein ML. Space-time covariance functions. *J Am Stat Assoc*. 2005;100(469):310-321. doi:10.1198/016214504000000854
31. Politis DN. *Model-Free Prediction and Regression*. 1st ed. Springer; 2016.
32. Chilès JP, Desassis N. Fifty years of kriging. In: Sagar BSD, Cheng Q, Agterberg F, eds. *Handbook of Mathematical Geosciences*. Springer; 2018:589, 612.
33. You MY. Multi-objective optimal design of permanent magnet synchronous motor for electric vehicle based on deep learning. *Appl Sci*. 2020;10(2):482. doi:10.3390/app10020482
34. Bhattacharjee S, Mitra P, Ghosh SK. Spatial interpolation to predict missing attributes in GIS using semantic kriging. *IEEE Trans Geosci Remote Sens*. 2014;52(8):4771-4780. doi:10.1109/tgrs.2013.2284489
35. de Medeiros ES, de Lima RR, de Olinda RA, Dantas LG, de Santos CAC. Space-time kriging of precipitation: modeling the large-scale variation with model GAMLSS. *Watermark*. 2019;11(11):2368. doi:10.3390/w11112368
36. Totis G, Sortino M. Polynomial chaos-kriging approaches for an efficient probabilistic chatter prediction in milling. *Int J Mach Tools Manuf*. 2020;157:103610. doi:10.1016/j.ijmachtools.2020.103610
37. Delfiner P, Chilès JP. *Geostatistics: Modeling Spatial Uncertainty*. 2nd ed. Wiley; 2012.
38. Prudhomme C, Reed DW. Mapping extreme rainfall in a mountainous region using geostatistical techniques: a case study in Scotland. *Int J Climatol*. 1999;19(12):1337-1356. doi:10.1002/(sici)1097-0088(199910)19:12<1337::aid-joc421>3.0.co;2-g
39. Kaufman CG, Bingham D, Habib S, Heitmann K, Frieman JA. Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology. *Ann Appl Stat*. 2011;5(4):2470-2492. doi:10.1214/11-aoas489
40. Griffith DA. Hidden negative spatial autocorrelation. *J Geogr Syst*. 2006;8(4):335-355. doi:10.1007/s10109-006-0034-9
41. dos Reis AA, Carvalho MC, de Mello JM, Gomide LR, ACF F, FWA J. Spatial prediction of basal area and volume in Eucalyptus stands using Landsat TM data: an assessment of prediction methods. *N Z J For Sci*. 2018;48(1). doi:10.1186/s40490-017-0108-0
42. Angelini ME, Heuvelink GBM. Including spatial correlation in structural equation modelling of soil properties. *Spatial Statistics*. 2018;25:35-51. doi:10.1016/j.spasta.2018.04.003
43. McLean MI, Evers L, Bowman AW, Bonte M, Jones WR. Statistical modelling of groundwater contamination monitoring data: a comparison of spatial and spatiotemporal methods. *Sci Total Environ*. 2019;652:1339-1346. doi:10.1016/j.scitotenv.2018.10.231
44. Liu H, Ong YS, Shen X, Cai J. When Gaussian process meets big data: a review of scalable GPs. *IEEE Trans Neural Netw Learn Syst*. 2020;31(11):4405-4423. doi:10.1109/tnnls.2019.2957109
45. Guhaniyogi R, Banerjee S. Meta-kriging: scalable Bayesian modeling and inference for massive spatial datasets. *Technometrics*. 2018;60(4):430-444. doi:10.1080/00401706.2018.1437474
46. Opitz T, Bonneu F, Gabriel E. Point-process based Bayesian modeling of space-time structures of forest fire occurrences in Mediterranean France. *Spatial Stat*. 2020;40:100429. doi:10.1016/j.spasta.2020.100429
47. Gu L. Moving kriging interpolation and element-free Galerkin method. *Int J Numer Methods Eng*. 2002;56(1):1-11. doi:10.1002/nme.553
48. Hartman L, Hössjer O. Fast kriging of large data sets with Gaussian Markov random fields. *Comput Stat Data Anal*. 2008;52(5):2331-2349. doi:10.1016/j.csda.2007.09.018
49. Hristopulos DT, Agou VD. Stochastic local interaction model with sparse precision matrix for space-time interpolation. *Spatial Stat*. 2020;40:100403. doi:10.1016/j.spasta.2019.100403
50. Strandberg J, de Luna SS, Mateu J. Prediction of spatial functional random processes: comparing functional and spatio-temporal kriging approaches. *Stoch Env Res Risk A*. 2019;33(10):1699-1719. doi:10.1007/s00477-019-01705-y
51. Arendt PD, Apley DW, Chen W, Lamb D, Gorsich D. Improving identifiability in model calibration using multiple responses. *J Mech Des*. 2012;134(10):100909-1-100909-9. doi:10.1115/1.4007573
52. Welch WJ, Buck RJ, Sacks J, Wynn HP, Mitchell TJ, Morris MD. Screening, predicting, and computer experiments. *Technometrics*. 1992;34(1):15-25. doi:10.1080/00401706.1992.10485229
53. Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Stat Sin*. 2011;21:5-42.
54. Bevilacqua M, Gaetan C, Mateu J, Porcu E. Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach. *J Am Stat Assoc*. 2012;107(497):268-280. doi:10.1080/01621459.2011.646928
55. Maronna RA, Martin RD, Yohai VJ. *Robust Statistics: Theory and Methods*. Wiley; 2006.
56. Caragea PC, Smith RL. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J Multivar Anal*. 2007;98(7):1417-1440. doi:10.1016/j.jmva.2006.08.010
57. Xu X, Reid N. On the robustness of maximum composite likelihood estimate. *J Stat Plan Inference*. 2011;141(9):3047-3054. doi:10.1016/j.jspi.2011.03.026
58. Agostinelli C, Yohai VJ. Composite robust estimators for linear mixed models. *J Am Stat Assoc*. 2016;111(516):1764-1774. doi:10.1080/01621459.2015.1115358

59. Franceschetti G, Riccio D. Ch. 6 Surface classical models. *Scattering, Natural Surfaces, and Fractals*. 1st ed. Elsevier; 2007:21-59.
60. Hartwig RE. AX - XB = C, resultants and generalized inverses. *SIAM J Appl Math*. 1975;28(1):154-183. doi:10.1137/0128014
61. Berrocal VJ, Raftery AE, Gneiting T. Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon Weather Rev*. 2007;135(4):1386-1402. doi:10.1175/mwr3341.1
62. Diggle PJ, Ribeiro PJ. *Model-based Geostatistics*. 1st ed. Springer; 2007.
63. Cressie N. *Statistics for Spatial Data*. Revised ed. Wiley; 1993.
64. Zhang B, Sang H, Huang JZ. Full-scale approximations of spatio-temporal covariance models for large datasets. *Stat Sin*. 2015;25(1):99-114. doi:10.5705/ss.2013.260w
65. Gaetan C, Guyon X. *Spatial Statistics and Modeling*. 1st ed. Springer; 2010.
66. Maes MA, Breitung K, Dann MR. At issue: the Gaussian autocorrelation function. Proceedings of the 18th International Probabilistic Workshop; 2021:191-203; Springer, Cham.
67. Stachniss C, Plagemann C, Lilienthal AJ. Learning gas distribution models using sparse Gaussian process mixtures. *Auton Robot*. 2009;26(2-3):187-202. doi:10.1007/s10514-009-9111-5
68. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. *J Royal Stat Soc Ser B (Stat Methodol)*. 2008;70(4):825-848. doi:10.1111/j.1467-9868.2008.00663.x
69. Dumelle M, Hoef JMV, Fuentes C, Gitelman A. A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product-sum covariance functions. *Spatial Stat*. 2021;43:100510. doi:10.1016/j.spasta.2021.100510
70. Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: theory and applications. *Ann Stat*. 1981;9(4):861-869. doi:10.1214/aos/1176345526
71. Varin C, Vidoni P. A note on composite likelihood inference and model selection. *Biometrika*. 2005;92(3):519-528. doi:10.1093/biomet/92.3.519
72. Davison AC, Hinkley DV. *Bootstrap Methods and Their Application*. Illustrated ed. Cambridge University Press; 1997.
73. Wright SJ. Coordinate descent algorithms. *Math Program*. 2015;151(1):3-34. doi:10.1007/s10107-015-0892-3
74. Hunter JS. The exponentially weighted moving average. *J Qual Technol*. 1986;18(4):203-210. doi:10.1080/00224065.1986.11979014
75. Huwang L, Yeh AB, Wu CW. Monitoring multivariate process variability for individual observations. *J Qual Technol*. 2007;39(3):258-278. doi:10.1080/00224065.2007.11917692
76. Cressie N. The origins of kriging. *Math Geol*. 1990;22(3):239-252. doi:10.1007/bf00889887
77. Qian HM, Huang T, Huang HZ. A single-loop strategy for time-variant system reliability analysis under multiple failure modes. *Mech Syst Signal Process*. 2021;148:107159. doi:10.1016/j.ymssp.2020.107159
78. Hyndman RJ, Koehler AB, Snyder RD, Grose S. A state space framework for automatic forecasting using exponential smoothing methods. *Int J Forecast*. 2002;18(3):439-454. doi:10.1016/s0169-2070(01)00110-8
79. Ma C. Semiparametric spatio-temporal covariance models with the ARMA temporal margin. *Ann Inst Stat Math*. 2005;57(2):221-233. doi:10.1007/bf02507023
80. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021.
81. Zhou H, Yu KM, Lee MG, Han CC. The application of last observation carried forward method for missing data estimation in the context of industrial wireless sensor networks. Proceedings of the 7th IEEE Asia-Pacific Conference on Antennas and Propagation; 2018; IEEE, Auckland, New Zealand.
82. Gramacy RB, Niemi J, Weiss RM. Massively parallel approximate Gaussian process regression. *SIAM/ASA J Uncertainty Quantif*. 2014;2(1):564-584. doi:10.1137/130941912
83. Courault D, Monestiez P. Spatial interpolation of air temperature according to atmospheric circulation patterns in southeast France. *Int J Climatol*. 1999;19(4):365-378. doi:10.1002/(sici)1097-0088(19990330)19:4<365::aid-joc369>3.0.co;2-e
84. Monestiez P, Courault D, Allard D, Ruget F. Spatial interpolation of air temperature using environmental context: application to a crop model. *Environ Ecol Stat*. 2001;8(4):297-309. doi:10.1023/a:1012726317935
85. Holdaway M. Spatial modeling and interpolation of monthly temperature using kriging. *Clim Res*. 1996;6:215-225. doi:10.3354/cr006215
86. Ma P, Konomi BA, Kang EL. An additive approximate Gaussian process model for large spatio-temporal data. *Environmetrics*. 2019;30(8):e2569. doi:10.1002/env.2569
87. Higdon D. Space and space-time modeling using process convolutions. In: Anderson CW, Barnett V, Chatwin PC, El-Shaarawi AH, eds. *Quantitative Methods for Current Environmental Issues*. Springer; 2002:37-56.
88. Paciorek CJ. Computational techniques for spatial logistic regression with large data sets. *Comput Stat Data Anal*. 2007;51(8):3631-3653. doi:10.1016/j.csda.2006.11.008
89. Steeb WH. *Problems and Solutions in Introductory and Advanced Matrix Calculus*. 1st ed. World Scientific; 2006.

How to cite this article: Lambardi di San Miniato M, Bellio R, Grassetti L, Vidoni P. Separable spatio-temporal kriging for fast virtual sensing. *Appl Stochastic Models Bus Ind*. 2022;1-24. doi: 10.1002/asmb.2697

APPENDIX A. PROOFS

A.1 Kriging mean formula

Predictions can be computed in a vectorized form, as follows, after Equation (28).

$$\mathrm{vec}(\hat{y}' - \mu') = \rho R^{-1} \mathrm{vec}(y - \mu) \ . \tag{A1}$$

By using definitions in Equations (11) and (27), it follows that

$$\mathrm{vec}(\hat{y}' - \mu') = (\rho_S \otimes \rho_T)(R_S \otimes R_T)^{-1} \mathrm{vec}(y - \mu) \ . \tag{A2}$$

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

$$\mathrm{vec}(\hat{y}' - \mu') = (\rho_S \otimes \rho_T)(R_S^{-1} \otimes R_T^{-1}) \mathrm{vec}(y - \mu) \ . \tag{A3}$$

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

$$\mathrm{vec}(\hat{y}' - \mu') = \left\{ (\rho_S R_S^{-1}) \otimes (\rho_T R_T^{-1}) \right\} \mathrm{vec}(y - \mu) \ . \tag{A4}$$

At this point, the regression coefficients of Equation (29) can be recognized.

$$\mathrm{vec}(\hat{y}' - \mu') = (\beta_S \otimes \beta_T) \mathrm{vec}(y - \mu) \ . \tag{A5}$$

Lastly, we use Roth's column lemma.[60] One can find it uncredited in recent algebra handbooks, see Steeb's book, Chapter 9, Problem 22.[89]

$$\mathrm{vec}(\hat{y}' - \mu') = \mathrm{vec} \left\{ \beta_T (y - \mu) \beta_S^\top \right\} \ . \tag{A6}$$

Then, the vec operator can be dropped and the matrix $\hat{y}'$ is obtained.

A.2 Kriging variance formula

Using Equation (28) as a starting point, it holds

$$\mathrm{diag}(R'_{\mathrm{cond}}) = \mathrm{diag}(R') - \mathrm{diag}(\rho R^{-1} \rho^\top) \ . \tag{A7}$$

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 (11) and (27) to obtain

$$\mathrm{diag}(R'_{\mathrm{cond}}) = \mathrm{diag}(R'_S \otimes R'_T) - \mathrm{diag} \left\{ (\rho_S R_S^{-1} \rho_S^\top) \otimes (\rho_T R_T^{-1} \rho_T^\top) \right\} \ . \tag{A8}$$

After Equation (34), it follows that

$$\mathrm{diag}(R'_{\mathrm{cond}}) = \mathrm{diag}(R'_S \otimes R'_T) - \mathrm{diag} \left\{ (R'_S - R'_{S,\mathrm{cond}}) \otimes (R'_T - R'_{T,\mathrm{cond}}) \right\} \ . \tag{A9}$$

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

$$\mathrm{diag}(R'_{\mathrm{cond}}) = \mathrm{diag}(R'_S) \otimes \mathrm{diag}(R'_T) - \mathrm{diag}(R'_S - R'_{S,\mathrm{cond}}) \otimes \mathrm{diag}(R'_T - R'_{T,\mathrm{cond}}) \ . \tag{A10}$$

Now, the components of $\mathrm{diag}(R'_{\mathrm{cond}})$ can be partitioned into vectors with the same length as $\mathrm{diag}(R'_T)$. Such vectors can be the columns of the matrix V, which is thus defined as in our claims.

APPENDIX B. BOOTSTRAP

Parametric bootstrap[72] 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 article.

The temporal and spatial models together identify the full model, under which one can simulate artificial datasets. In particular, separability allows to simulate a dataset Y as

$$Y \sim \mu + \sigma \cdot R_T^{1/2} \cdot \varepsilon \cdot R_S^{1/2}\ , \tag{B1}$$

where ε is a Gaussian white noise structured into a $T \times S$ matrix, and $R_T^{1/2}$ and $R_S^{1/2}$ are the matrix square roots of the matrices R_T and R_S, respectively.

Assuming $T \gg S$, $R_S^{1/2}$ may be tractable, while $R_T^{1/2}$ will hardly be so. The operator $R_T^{1/2}$ just makes $R_T^{1/2} \cdot \varepsilon$ a matrix with independent columns that share the same correlation structure R_T. So, as an alternative to directly evaluating $R_T^{1/2}$, one can generate each column of $\sigma \cdot R_T^{1/2} \cdot \varepsilon$ according to the temporal model. AR(p) processes can be simulated efficiently according to the factorized MA(∞) 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 8 weeks of data, we needed to simulate 32 more leading weeks of data as a burn-in.

APPENDIX C. MAXIMUM LIKELIHOOD ESTIMATION

The pseudo ML estimator $\hat{\theta}$ for θ is defined as a maximizer of (12). An iterative optimization procedure can be devised to compute $\hat{\theta}$, by leveraging the convenient separability assumption in Equation (16). Operationally, the components of the parameter vector θ can be partitioned into groups of parameters, which can be updated one at a time until convergence. In particular, we define three groups of parameters, which coincide with the scale parameter σ, the spatial and temporal correlation parameters ψ_S and ψ_T.

In Equation (B1), $R_S^{1/2}$ and $R_T^{1/2}$ were matrix square roots of R_S and R_T, respectively. Now, let $R_S^{-1/2}$ and $R_T^{-1/2}$ be their inverses. These matrices can be feasibly estimated, as $\hat{R}_S^{-1/2}$ and $\hat{R}_T^{-1/2}$, respectively, based on available estimates of ψ_S and ψ_T. An approximately efficient estimator can be computed iteratively by

- updating $\hat{\sigma}^2$ as the mean of squared entries of the matrix $\hat{R}_T^{-1/2}(y - \hat{\mu})\hat{R}_S^{-1/2}$, based on current estimates for ψ_S and ψ_T;
- updating $\hat{\psi}_S$ as the maximizer of

$$\Omega(\psi_S) = -\frac{T}{2}\left\{\log|R_S| + \mathrm{tr}(R_S^{-1}\hat{M}_S)\right\}, \tag{C1}$$

  with respect to ψ_S, where $\hat{M}_S = z_S^\top z_S / T$ and $z_S = \hat{R}_T^{-1/2}(y - \hat{\mu})/\hat{\sigma}$, based on current estimates for ψ_T and σ; notice the resemblance with Equations (17) and (18); this operation is equivalent to estimating ψ_S based on T independent cross-sections;
- updating $\hat{\psi}_T$ as the maximizer of

$$\Omega(\psi_T) = -\frac{S}{2}\left\{\log|R_T| + \mathrm{tr}(R_T^{-1}\hat{M}_T)\right\}\ , \tag{C2}$$

  with respect to ψ_T, where $\hat{M}_T = z_T z_T^\top / S$ and $z_T = (y - \hat{\mu})\hat{R}_S^{-1/2}/\hat{\sigma}$, based on current estimates for ψ_S and σ; notice the resemblance with Equations (20) and (21); this operation is equivalent to estimating ψ_T based on S independent time series.

In Appendix B, we stress that some convenient spatial or temporal models allow for sparse formulations of matrices $R_S^{-1/2}$ and $R_T^{-1/2}$. In evaluating z_S, there is thus no need to allocate in memory any large matrix like $R_T^{-1/2}$.