International Journal of Geographical Information Science
Vol. 00, No. 00, Month 2017, 1–20

Research Paper

Leveraging Correlation across Space and Time to Interpolate Geophysical Data via CoKriging

Sonja Pravilovic[a] and Annalisa Appice[bc*] and Donato Malerba[bc]

[a]*Faculty of Information Technology, Mediterranean University, Vaka Djurovica - 81000 Podgorica - Montenegro*
[b]*Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, via Orabona, 4 - 70125 Bari - Italy*
[c]*CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Bari - Italy*
(*Received 00 Month 200x; final version received 00 Month 200x*)

Managing geophysical data generated by emerging spatio-temporal data sources (e.g. geosensor networks) presents a growing challenge to GIS science. The presence of correlation (i.e. spatial correlation across several geosensor sites and time correlation within each site) poses difficulties with respect to traditional spatial data analysis. This paper describes a novel spatio-temporal analytical scheme that allows us to yield a characterization of correlation in geophysical data along the spatial and temporal dimensions. We resort to a multivariate statistical model, namely CoKriging, in order to derive accurate spatio-temporal interpolation models. These predict unknown data by utilizing not only their own geosensor values at the same time, but also information from near past data. We use a window-based computation methodology that leverages the power of temporal correlation in a spatial modeling phase. This is done by also fitting the computed interpolation model to data which may change over time. In an assessment, using various geophysical data sets, we show that the presented algorithm is often able to deal with both spatial and temporal correlations. This helps to gain accuracy during the interpolation phase, compared to spatial and spatio-temporal competitors. Experimental results validate the use of the window-based methodology in the computation of time-evolving, spatio-temporal interpolation models of a geosensor network. In particular, we evaluate the efficacy of the interpolation phase by using established machine learning metrics (i.e. root mean squared error, Akaike information criterion and computation time).

Keywords: Spatio-temporal data, CoKriging, Multi-variate analysis, Interpolation

*Corresponding author. Email: annalisa.appice@uniba.it

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

1. Introduction

Natural processes and physical fields (e.g. solar radiation and wind speed) are being increasingly observed across space and over time. The ubiquity of this kind of spatio-temporal data, namely geophysical data, has motivated us to investigate and develop appropriate models to analyze and interpolate them. Interpolation, i.e. the prediction of missing data in each location across space or time, can be useful in supplementing, smoothing and standardizing observational data. It allows us to transform irregular point or line data into a raster representation, or to perform resampling between different raster resolutions (Mitas and Mitasova 1999).

The geophysical interpolation context we consider in this study is georeferenced and timestamped data which are routinely sampled for a physical numeric field by geosensors, installed at fixed-to-ground sites of a network. The spatial location of the geosensors is known, distinct and invariant. The acquisition activity is synchronized on the geosensors of the network. The time points of the acquisition activity are equally spaced in time. This scenario recurs in various geophysical applications (e.g. solar radiation measurements in photovoltaic plants, wind and weather data sensors and ambient pollution monitoring sensors (Appice *et al.* 2014a)).

At present, several studies (Guttorp and Schmidt 2013, Appice *et al.* 2014b, Pravilovic *et al.* 2017) have already assessed that the temporal information within a site and the spatial information across distinct sites are both informative in a geophysical scenario. However, the analysis of both spatial and temporal correlations is more complicated than modeling purely spatial or purely temporal correlations. On the other hand, the classes of spatio-temporal dependence structures differ from each other in the way in which space and time are coupled. At one extreme, space and time are considered to be independent, giving rise to the separable covariance model that allows us to represent the spatio-temporal correlation function as the product of a spatial and temporal term. Otherwise, at the cost of a heavier computational burden, non-separable space-time models can be considered by including suitable parameters that indicate the strength of the interaction between the spatial and temporal components (Guttorp and Schmidt 2013). In addition, non-stationarity and anisotropy are usual characteristics of geophysical data, since the statistical property of a field can often undergo a time change (Appice *et al.* 2014b), to be taken into account in the modeling phase. If data changes are present and unaccounted for in the geo-statistical model development, they can result in poorly specified models, as well as in inappropriate spatial-temporal inferences and predictions.

In view of the need to compute accurate geophysical interpolation models, we adopt a computational approach that reaches a compromise between: (1) deriving a model of the spatio-temporal correlation and (2) accounting for the possible temporal variation of a physical field distribution. To this end, we use a window-based computation methodology (Omitaomu *et al.* 2009, Appice *et al.* 2014a) that allows us to compute local, spatio-temporal interpolation models from geophysical data, observed at a few backward time points. In this way, we are able to compute window-aware correlation functions, which may change over time as new geophysical data windows are processed. This use of a window operator is here supported by the observation that, although geophysical data can be subject to the temporal variation of their distribution, the property of temporal correlation suggests that the stability of the field statistical property can at least be seen at a local (window) level (Appice *et al.* 2014a).

In particular, we illustrate an interpolation scheme, called $\mathsf{CoST}^{\mathsf{K}}$ (CoKriging Spatio-Temporal interpolator), that uses the Window operator to learn a time-evolving, spatio-

temporal interpolation model in a geosenor network. It computes one interpolation model for each (target) time point. Each timestamped model accounts for the geophysical data observed at the target time point, as well as for the data selected by the Window operator at the backward time points. A multi-variate interpolation model, namely CoKriging (Wackernagel 2003), can be used to construct a continuous map of the Principal Component of the spatio-temporal correlated, windowed data. This map is used to interpolate the data at the unobserved sites of the target time point.

CoKriging is a multi-variate interpolation model belonging to the stochastic Kriging-family of optimal linear interpolation models. It is considered here since, similarly to its univariate Kriging counterpart (Krige 1951), it has the attractive properties of being unbiased and having minimal variance of the prediction error. Historically it is used to explore the influence of a secondary co-field as a function of the purely spatial correlation with the primary field (e.g. Knotters *et al.* (1995), Rocha *et al.* (2012), Triantakonstantis and Stathakis (2014)). In this study, CoKriging is examined in the univariate scenario (i.e. data are collected for a single field), but with the goal of dealing with multi-time data (i.e. data that are routinely collected at various time points). In particular, the field under study, observed at a target time point, is the primary variable. On the other hand, the principal component representation of the primary field observations over the reference window is the secondary co-variable. We note that this construction of a secondary variable by combining Principal Component Analysis and a Window operator, as well as its processing via CoKriging, allows us to account for a characterization of the spatio-temporal information.This contributes to modeling the correlation of the primary field in space and time simultaneously.

The paper is organized as follows. The background and the actual contribution of this paper is clarified in the next Section. The relevant related work is briefly reviewed in Section 3. The learning problem and the proposed approach are introduced in Section 4. Experiments with real data are presented in Section 5. Finally, Section 6 draws conclusions and illustrates future directions of this research.

2. Background and contribution

Principal Component Analysis (PCA), Window operator and CoKriging have already attracted valuable research attention in the GISience literature.

Several studies explore the use of PCA, often in combination with Kriging. For example, Jiang *et al.* (2009), as well as Nazzal *et al.* (2015), employ PCA in an exploratory study of various natural and anthropogenic factors affecting the groundwater quality. They extract Principal Components to reduce the dimensionality of the system of measured fields. In particular, the Principal Components are extracted in a purely spatial context, in order to discriminate between the factors with the highest/lowest contribution on the groundwater. They use Ordinary Kriging, in order to interpolate the map of the spatial distribution of each Principal Component score separately. Following this research direction Aversano *et al.* (2017) apply PCA and Kriging in a spatial setting, in order to build accurate surrogate models in reacting flows for predictive purposes and engineering design. Although the above studies describe promising results achieved by combining PCA and Kriging, they neglect the temporal information. They use Kriging instead of CoKriging and compute an interpolation model for each derived Principal Component score (rather than a direct interpolation model for each original observational field). This contributes to making the approach presented in this study different, since it

employs PCA as a means to derive a temporal characterization of the spatial correlation of the observed field. Specifically, the proposed combination of PCA and CoKriging explores a strategy to directly add a temporal perspective of the spatial correlation in the interpolation model of the field under study.

On the other hand, there are also various studies that explore the use of a Window operator as a means to select surrounding (spatial, temporal, spatio-temporal) data for spatio-temporal analysis. In particular, Weiss *et al.* (2014) present an overview of several algorithms that consider information present within surrounding data to interpolate missing data in a remotely sensed imagery time-series. Gafurov and Bárdossy (2009) describe a temporal interpolation algorithm to fill-in the missing data, using data from window-defined earlier and/or later dates. Interestingly, the Window operator is sometimes investigated also in combination with CoKriging. For example, Zhang *et al.* (2009) explore the idea of augmenting primary data collected at a specific time point with secondary data collected, in the same area of interest, at an alternative user-defined date. These primary and secondary data are used to define both the primary and secondary (co-)variables of CoKriging respectively. Similarly, Sideris *et al.* (2014) handle data observed at the previous time point as the secondary co-variable of CoKriging. Finally, Rouhani and Wackernagel (1990), as well as Skoien and Bloschl (2007), construct various secondary co-variables with data observed at consecutive (background) windowed time steps. These multiple co-variables are again handled via CoKriging. The promising results achieved in these studies mainly inspire our idea of pursuing a combination of a Window operator and CoKriging. However, our consideration of CoKriging here is slightly different from that experimented in the background studies. In fact, our study resorts to PCA to derive a characterization of the spatial correctional across time and uses this characterization (instead of the observed data) as the secondary co-variable. This is different from the studies reported above, which define the secondary co-variable(s) of CoKriging based on the observed (window-selected) data without any data transformation. Our feeling is that observations of a field collected at several consecutive time points may be strongly contemporaneously correlated with each other. This may lead to creating a system of strongly redundant variables. Collinearity among variables may lead to a series of problems, such as unreliable coefficients and predictions, as well as aggravated data redundancy and computational complexity (Chen *et al.* 2016). The idea that we promote here is that the reduction of the collinearity in the set of covariates may improve the accuracy (as well as the efficiency) of CoKriging. In particular, to the best of our knowledge, this is the first study that employs PCA as a means to derive a collinearity-free characterization of spatial correlation across time within CoKriging. We have chosen PCA since it is a simple orthogonal transformation, commonly used in regression. It requires no parameter, removes variable correlations and thus reduces collinearity (Dormann *et al.* 2013).

Considering the background, our specific proposal of combining PCA, a Window operator and CoKriging in a single interpolation algorithm represents one of the main contributions of this work. Although the proposed algorithm should be considered a heuristic, another contribution is the empirical demonstration that our spatio-temporal system is often more efficacious (in terms of accuracy and/or efficiency) than existing spatial and spatio-temporal interpolation models. In short, the specific contributions of this paper can be highlighted as follows: (1) We describe a schema to characterize the correlation of geophysical data along the spatial and temporal dimensions simultaneously. (2) We define a system that supplements any observation site with a multivariate system. This includes a field observed from the geosensor network at the same time

with window-coupled data observed in the near past. (3) We apply a multivariate interpolation solution, in order to analyze the structure of this system of variables and construct an accurate spatio-temporal interpolate, with the help of a spatio-temporally coupled covariable. (4) We demonstrate the importance of dealing with both temporal and spatial correlation, in order to yield an accurate interpolate of geophysical data and validate the accuracy of the proposed window-based methodology in the computation of a time-evolving spatio-temporal interpolate.

3. Related work

Historically the challenge of predicting a geophysical field by looking at observational geosensor data has led to a variety of spatial deterministic interpolation algorithms like Inverse Distance Weighting (Shepard 1968) and Radial Basis Functions (Lin and Chen 2004), as well as stochastic, like Kriging (Krige 1951, Cressie 1990) and its multi-variate extension CoKriging (Wackernagel 2003). These algorithms determine field estimates based upon actual measures of the physical field which are spatially sampled in a Geographic Information System (GIS). They account for the property of data correlation. Specifically, they take into account a stronger correlation between data points which are spatially closer than data points that are farther apart.

Although the spatial interpolation theory has a long history, some studies have been recently made to inject traditional temporal data mining techniques into spatial interpolation models. These studies aim at handling both spatial and temporal correlation of geophysical data within a joint interpolation scenario. They still consider physical data collected by a geosensor network as the outcome of a stochastic process with random noise. Therefore, they are founded on the idea that geophysical processes can be modeled by means of relatively few parameters (Shumway and Stoffer 2010).

Following the spatio-temporal direction, Romanowicz *et al.* (2006) describe a two-step spatio-temporal methodology for the air data quality analysis. In the first step, they resort to non-stationary time series analysis methods, in order to supplement data sets over periods with missing measurements. The time series are decomposed into trend and harmonic components. In the second step, they analyze the spatial relations within the data sets. They derive a spatio-temporal model of log-transformed data. This model consists of the trend, while the noise describes the spatio-temporal variations in the data. It is used to predict variations at un-sampled points across time and space. Appice *et al.* (2013) describe a regressive time dynamic model of the physical field. They modify the IDW interpolation algorithm by exploiting a time-evolving spatial model and use the final model to estimate data at un-sampled points across time and space. Finally, Sherman (2010) performs the theoretical work that investigates how the temporal and spatial correlation can interact with each other. This seminal study contributes to the formulation of the empirical estimators of space-time variograms and covariances, which pave the way for learning a spatio-temporal interpolation model with Kriging-family algorithms. Following this research direction, Cressie and Wikle (2011) discuss the importance of a characterization of the joint spatio-temporal co-variance structure of a spatio-temporal process for an optimal Kriging prediction. However, both these studies lack implementations. This issue is overcome by Pebesma (2012), who defines the space-time package to deal with spatio-temporal data in R. Gräler *et al.* (2016) have recently used the space-time classes to estimate spatio-temporal covariance/variogram models and to perform spatio-temporal interpolation. Their implementation handles various types of spatio-temporal

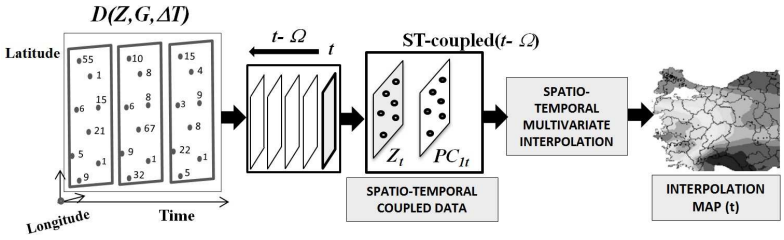

Figure 1. Block diagram of CoST$^{\mathsf{K}}$ - Co (Spatio-Temporal) Kriging. Let $\mathcal{D}(Z, G, \Delta T)$ be a geophysical data set and Ω be a window size. A spatio-temporal interpolation model is computed for each time point $t \in \Delta T$, in order to predict unseen data at time t. Let Z_t be the primary variable representing the data snapshot at time t. The window operator selects the data snapshots with the time varying between $t - \Omega$ and $t - 1$. The Principal Component Analysis is applied to the selected data window and the top ranked Principal Component PC_{1_t} is selected as a secondary variable (see details in Section 4.1). Finally, the CoKriging system of the equation is computed with primary variable Z_t and secondary variable PC_{1_t} (see details in Section 4.2).

covariance structures and facilitates spatio-temporal interpolation. In any case, their formulation assumes the temporal stability of the geophysical data, thus it creates a global covariance function that fits the entire set of spatio-temporal observational data. Then it neglects possible time changes in the statistical property of the examined field, so that the inferred interpolation models may also result in poorly specified models. Finally, Rouhani and Wackernagel (1990) and Skoien and Bloschl (2007) explore the use of CoKriging in the discovery of an interpolation model that considers data observed at the backward time points as secondary co-variables of the interpolation system of equations. Similarly, Zhang *et al.* (2009) and Sideris *et al.* (2014) use CoKriging in spatio-temporal interpolation, by considering a system of equations with a secondary co-variable associated with data observed at a certain backward time point.

4. Spatio-temporal interpolation via CoKriging

Let $\mathcal{D}(Z, G, \Delta T)$ be a geophysical data set. Z is a numeric physical field. G is a network of distinct geolocations over a given spatial domain. Each geolocation is represented by its spatial coordinates (e.g. latitude and longitude). A fixed-to-ground active geosensor is installed at each geolocation of G, in order to routinely observe data for Z. ΔT is the observation (sampling) time line. This time line is discretized in equally-spaced time points denoted as $\Delta T = t_1, t_2, \ldots, t_T$, so that all geosensors of the network observe new data for Z simultaneously at each acquisition time point $t_i \in \Delta T$. According to this formulation, the geophysical data set can be viewed as a sequence of timestamped data snapshots, that is, $\mathcal{D}(Z, G, \Delta T) = \langle G, z_{t_1}(\cdot)\rangle, \langle G, z_{t_2}(\cdot)\rangle, \ldots, \langle G, z_{t_T}(\cdot)\rangle$. A data snapshot $\langle G, z_t(\cdot)\rangle$ is the set of observations sampled for Z at a certain acquisition time point $t \in \Delta T$ by geosensors installed in network G. $z_t \colon G \mapsto Z$ is the field function that assigns a geosensor location $(x, y) \in G$ to the value here observed for Z at time t. $z_t(g)$ denotes the georeferenced, timestamped value of Z, collected at geolocation $g \in G$ for time point $t \in \Delta T$. A geosensor may also be temporally inactive and does not register data at a certain time point (due to synchronization, sensor faults, communication malfunctions or malicious attacks). However, the missing observation can be preprocessed on-the-fly during the learning stage and replaced by an aggregate (e.g. inverse distance weighted mean) of data actually sampled across a (spatio-temporal) neighborhood.

In this study, we present a spatio-temporal interpolation algorithm (see Figure 1),

which inputs data set $\mathcal{D}(Z, G, \Delta T)$ and uses a temporal window to consume its data snapshots. It derives a characterization of the spatio-temporal correlation in the windowed data snapshots, and uses this in-window spatio-temporal information to compute a time-evolving, spatio-temporal-aware interpolation model. Let Ω be the window size. At time t, the Window operator selects Ω backward geophysical data snapshots timestamped between $t-\Omega$ and $t-1$. The interpolation process is a pipeline of three algorithmic steps:

(1) The primary variable is identified. This variable represents the geophysical data in the data snapshot observed at time t.
(2) A data expansion mechanism is triggered by synthesizing the secondary covariable. It borrows from the spatial- and temporal-aware information that is enclosed in the window, by resorting to PCA (see Section 4.1).
(3) A CoKriging system of equations (see Section 4.2) is constructed, in order to derive the interpolate of the primary field at time t. The spatio-temporally coupled co-variable defined in the previous step is exploited.

The computed interpolation model can be used to predict (out-of-sample) unknown data points $\hat{z}_t(g')$ for any new suitable geolocation $g' \notin G$ at present time t.

We note that the above formulation contributes to addressing the traditional spatial interpolation task in a *local* spatio-temporal-aware setting. In fact, it allows us to compute a sequence of spatio-temporal-aware local models, which change over time, by naturally fitting the model to time-drifting spatial data. Specifically, a local interpolation model is computed for each time point t. This model is spatio-temporal-aware, meaning that it is computed from spatial data observed at time t, as well as spatial data temporally observed in the past (before t).[1]

4.1. *A spatio-temporally coupled covariable*

We introduce a Window operator, in order to select the geophysical data snapshots collected over a backward time window. Let Ω be the window size and t be a time point of the time line. The Window operator selects the data snapshots acquired by the geosensor network with the time varying between $t-\Omega$ and $t-1$. Let us introduce variables $Z_{t-\Omega}, Z_{t-\Omega+1}, \ldots, Z_t$, in order to denote the measurements of Z collected in the data snapshots $\langle G, z_{t-\Omega}(\cdot)\rangle, \langle G, z_{t-\Omega+1}(\cdot)\rangle, \ldots$ and $\langle G, z_t(\cdot)\rangle$, respectively. We expect that window variables $Z_{t-\Omega}, Z_{t-\Omega+1}, \ldots, Z_t$ may be strongly contemporaneously correlated with each other. At each time point, they can be interpreted as a linear combination of the same set of observations (i.e. the observations from data sharing the same statistical property). An illustration of this phenomenon is shown in Figure 2(a). It shows the correlation matrix, extracted at one time point, including both the target data (Z_t) and the backward data ($Z_{t-\Omega}, Z_{t-\Omega+1}, \ldots Z_{t-1}$). In general, there is a strong contemporaneous correlation between temporally close data (and such a correlation pattern is independent of the particular data set).

In view of the above discussion (see Section 2), after defining the multiple variables associated with the windowed time points, a natural postprocessing step consists of converting the backward secondary data $Z_{t-\Omega}, \ldots, Z_{t-1}$ into uncorrelated Principal Components $PC_{1_t}, \ldots, PC_{\Omega_t}$, using the PCA standard orthogonal transformation. An example of the effect of this transformation can be seen on the right side of Figure 2(b). The

[1]The investigation of the forecasting potentiality of this interpolation model (i.e. processing past data to predict unseen future data) is out of the scope of this study.

	$\mathbf{Z_t}$	$\mathbf{Z_{t-1}}$	$\mathbf{Z_{t-2}}$	$\mathbf{Z_{t-3}}$	$\mathbf{Z_{t-4}}$
$\mathbf{Z_t}$	1	.889	.934	.862	.651
$\mathbf{Z_{t-1}}$	.889	1	.882	.793	.591
$\mathbf{Z_{t-2}}$	.934	.882	1	.930	.679
$\mathbf{Z_{t-3}}$	.862	.793	.930	1	.775
$\mathbf{Z_{t-4}}$	.651	.591	.679	.775	1

(a) Correlation (Z)

	$\mathbf{Z_t}$	$\mathbf{PC_{1_t}}$	$\mathbf{PC_{2_t}}$	$\mathbf{PC_{3_t}}$	$\mathbf{PC_{4_t}}$
$\mathbf{Z_t}$	1	.919	.198	.030	0.084
$\mathbf{PC_{t_1}}$	.919	1	$-2.8e-16$	$5.4e-16$	$5.3e-16$
$\mathbf{PC_{t_2}}$	.198	$-2.8e-16$	1	$-1.6e-16$	$8.5e-17$
$\mathbf{PC_{t_3}}$	.030	$5.4e-16$	$-1.6e-16$	1	$-4.2e-16$
$\mathbf{PC_{t_4}}$	.084	$5.3e-16$	$8.5e-17$	$-4.2e-16$	1

(b) Correlation (PCA)

Figure 2. NCDC Air Climate geosensor network (see Section 5.1 for further details): physical field Solar Radiation (ncdcS), acquisition time point $t = 45$ and window size $\Omega = 4$. Figure 2(a): Pearson correlation matrix of variables Z_t and backward variables $Z_{t-1}, \ldots, Z_{t-\Omega}$. Figure 2(b): Pearson correlation matrix of variable Z_t and Principal Components $PC_{1_t}, \ldots PC_{\Omega_t}$ expressing a partially orthogonal representation of $Z_{t-1}, \ldots, Z_{t-\Omega}$.

Principal Components are contemporaneously uncorrelated with each other and, on average, are weakly correlated with the primary variable Z_t. This set of almost orthogonal components may have nonzero lagged cross-correlations. The top ranked component (PC_{1_t}) can be plausibly intended as a local characterization of the property of correlation in the surrounding space-time system. Consequently, it is exploited to derive the spatio-temporally coupled scondary co-variable in the CoKriging system of equations to interpolate the primary field.

4.2. *A (spatio, temporal) CoKriging system of equations*

The actual learning process considers, for each time point $t = t_1, t_2, \ldots, t_T$, a multivariate data system defined as $\hat{Z}_t = (Z_t, PC_{1_t})$, where Z_t denotes the primary data observed at time t. PC_{1_t} denotes the secondary co-data associated with the first Principal Component score synthesized from the data observed over the window $[t - \Omega, t - 1]$. The motivation behind the decision of selecting the top-ranked Principal Component, that accounts for the greatest quota of the total variability, is that the Principal Components can be interpreted as latent orthogonal factors useful to aggregate information observed across space and time over a window of near backward data. However, given the strong contemporaneous correlation existing between data observed at close time points, the first Principal Component is expected to explain a very high proportion of the total variability exhibited by the original windowed secondary data. For example, Figure 2(b) shows that the scores associated with the smallest eigenvalues, PC_{2_t}, PC_{3_t} and PC_{4_t}, are unlikely to contain useful information. This can most likely be interpreted as a small irregular component, containing the variation uncommon to the windowed data, which has not been captured by PC_{1_t}. For these reasons, we can reasonably apply this noise reduction, thus shrinking the smallest eigenvalues.

The CoKriging estimate is computed from this multivariate system.[1] This estimate is a linear combination of both the primary variable (variable of interest $Z(t)$) and the secondary variable (co-variable PC_{1_t}). It is unbiased, with a minimization of the variance estimation, and requires the covariance model of the primary variable, the covariance model of the secondary variable and the cross-covariance model of both the primary and secondary variable (see Figure 4.2). In particular, the ordinary CoKriging equation considered in this study[2] is in the form:

[1]We consider here an isotopic data configuration, where both the primary and the secondary data are observed at all the geosensor sites, but both are not available at the estimation site (Subramanyam and Pandalai 2008).
[2]Ordinary CoKriging usually is preferred to simple CoKriging, since it requires neither knowledge nor stationarity of the primary and secondary means over the entire area (Goovaerts 1998).

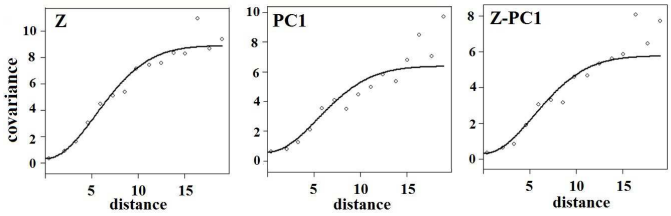

Figure 3. NCDC Air Climate geosensor network (see Section 5.1 for further details): physical field Solar Radiation (ncdcS), acquisition time point $t = 45$ and window size $\Omega = 4$. Covariance (axis Y) of primary variable Z and secondary variable PC, as well as cross-covariance of Z-$PC1$, by varying the distance (axis X) between the observed values. The Gaussian model is used to fit the theoretical variogram to each sample (cross-)variogram. Here the Gaussian model is chosen as it is the best model among Linear, Gaussian, Exponential, Spherical and Matern models fitting on the sample variogram of primary variable Z (see details in Section 4.2.1). The Gaussian model is also fitted on the sample covariance of PC and the sample cross-covariance of Z-PC.

$$\hat{z}_t(g') = \sum_{g \in G} \alpha_t(g) z_t(g) + \sum_{g \in G} \beta_t(g) pc_{1_t}(g), \tag{1}$$

where $\hat{z}_t(g')$ is the estimate computed at an unknown geolocation g', $\alpha_t(g)$ is the weight assigned to the primary value $z_t(g)$ and $\beta_t(g)$ is the weight assigned to the primary value $pc_{1_t}(g)$. The unbiasedness of this ordinary estimator is ensured by forcing all primary data weights to sum up to one (i.e. $\sum_{g \in G} \alpha_t(g) = 1$) and all secondary data weights to sum up to zero ($\sum_{g \in G} \beta_t(g) = 0$) (Isaaks and Srivastava 1989, Goovaerts 1998). Specifically, $\alpha_t(g)$ and $\beta_t(g)$ (for each $g \in G$) are obtained by minimizing the error variance under the two unbiasedeness constraints. This is equivalent to solving the ordinary CoKriging system expressed in terms of the following correlograms:

$$\begin{cases} \sum_{g_j \in G} \alpha_t(g_j) \mathcal{C}_{Z,Z}(g_i - g_j) + \sum_{g_j \in G} \beta_t(g_j) \mathcal{C}_{Z,PC}(g_i - g_j) + \mu_1(g) = \mathcal{C}_{Z,Z}(g_i - g') \text{ with } g_i \in G \\ \sum_{g_j \in G} \alpha_t(g_j) \mathcal{C}_{PC,Z}(g_i - g_j) + \sum_{g_j \in G} \beta_t(g_j) \mathcal{C}_{PC,PC}(g_i - g_j) + \mu_2(g) = \mathcal{C}_{PC,Z}(g_i - g') \text{ with } g_i \in G \\ \sum_{g_i \in G} \alpha_t(g_i) = 1 \\ \sum_{g_i \in G} \beta_t(g_i) = 0 \end{cases}, \tag{2}$$

where $C_{XY}(x, y)$ generally represents the (cross-)covariance (or variogram) function γ between a variable X and a variable Y at geolocations x and y. $\mu_1(u)$ and $\mu_2(u)$ are the two Lagrange parameters accounting for the unbiasedness constraints and g' is the estimation site. To determine γ, we consider the sample variogram. This is constructed from the sample data as a description of how they are (cor)related with distances. The sample variogram is $2\gamma(h)$ with semi-variogram $\gamma(h)$ representing half the average squared difference between points separated by a distance h (Cressie 1993) (see the example of a sample variogram in Figure 4.2). In particular, the sample semi-variogram is calculated as $\gamma(h) = \frac{1}{2|N(h)|} \sum (g_i, g_j) \in N(h)(z(g_i) - z(g_j))^2$, where $N(h)$ is the set of all pairwise Euclidean distances $d(g_i, g_h) = h$, $|N(h)|$ is the number of pairs on $N(h)$, $z(g_i)$ and

$z(g_j)$ are the data values at spatial locations g_i and g_j, respectively. As the measured data are commonly affected with noise, a theoretical variogram (see the theoretical variogram in Figure 4.2) is fitted on the the sample data (Isaaks and Srivastava 1989), in order to diminish the effect of variability. The theoretical model is fitted on the sample variogram values with three parameters, namely nugget, range and sill. The nugget is the intercept of the variogram. It includes variance at scales smaller than the minimum separation distances between points and variance, attributed to changes as data are sampled. The range defines the distance of the spatial dependence in the data: at distances greater than the range, the data are considered to be spatially independent. The sill is the asymptotic value of $\gamma(h)$ as h becomes very large.

4.2.1. Technical details

We solve CoKriging in the R environment for statistical computing. Initially, we determine the best theoretical model fitting the sample variogram of the primary field. The Weighted-Least Mean Square method is used to fit a theoretical model on the sample variogram and determine the nugget, range and sill of this fitting. As the fitting process can be influenced by the initial values of range, nugget and sill, default parameters are initialized according to the guidelines suggested by Pebesma and Graeler (2017). Range is taken as 1/3 of the maximum sample variogram distance, nugget is taken as the mean of the first three values of the sample variogram and sill is taken as the mean of the last five values of the sample variogram. The fitting method is used with weights equal to $N_j/(h_j^2)$, where N_j is the number of point pairs and h_j is the distance. As theoretical models, we consider various model families, namely: Linear, Gaussian, Exponential, Spherical and Matern (Cressie 1993). For each theoretical model, the described procedure may provide a certain numerical estimate of range, nugget and sill. The best theoretical model that minimizes the fitting root mean squared error is selected (Turner and Gardner 2015). Subsequently, the model family selected on the primary variable is employed, in order to fit the sample variogram of the secondary variable, as well as the cross-variogram of the primary-secondary variables. Determining the cross-variogram is a complex step, due to the necessity of giving an estimate that is conditionally positive definite (Huang *et al.* 2009). We follow the technical note of Rossiter (2012) and rely on a cross-variogram that can be fitted using the linear model of co-regionalization. This linear model fits a single spatial structure to both direct and cross-variogram, by optimizing the partial sills and the nugget by least squares. Any negative eigenvalues are set to zero and the eigenvectors are recomputed to ensure positive semi-denite matrices of the partial sills and the nugget (Pebesma and Graeler 2017). This ensures a valid linear model of co-regionalization, providing a quick approximation for iterative methods of adjusting partial sills and nuggets (Kinoshita *et al.* 2016).

5. Empirical evaluation and discussion

The spatio-temporal interpolate presented in this paper is evaluated on several geophysical data sets by considering both accuracy and efficiency metrics.

5.1. *Geophysical data sets*

We consider eleven geophysical data sets measured at equally-spaced discrete time points across six (regular and irregular) geosensor networks. We observe that this study does

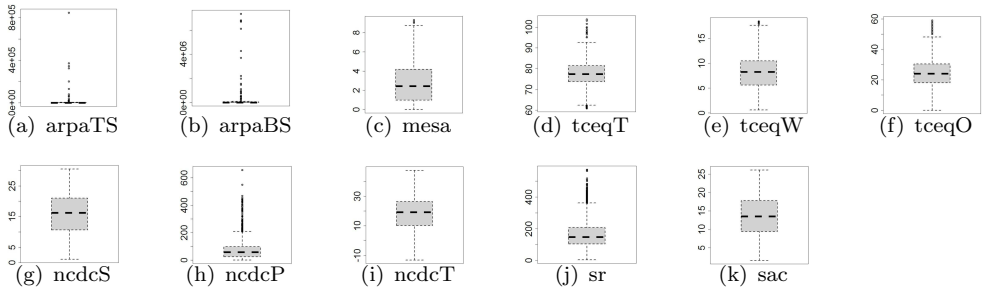

Figure 4. Box plot representation of the geophysical data sets.

not attempt to perform any exploratory data analysis based on spatial and/or temporal patterns hidden in the data. It relies upon automatic interpolation algorithms, in order to estimate parameters of accurate interpolation equations. Both small and large networks are considered for this empirical study, in order to investigate the accuracy of the computation also when few data are collected, as well as when more data feeds the learning process. Several networks (e.g. ARPA, MESA, TCEQ, NCDC, SR) measure data, which contain significant outliers (see the box plots in Figures 4(b)-4(k)). The presence of outliers commonly contributes to modifying the mean squared error of the interpolation model by affecting the model parameters and augmenting generally the difficulty of the interpolation task (Buzzi-Ferraris and Manenti 2010).

ARPA Ostreopsis Ovata geosensor network (`http://www.arpa.puglia.it/web/guest/algatossica`) was used to measure the *Density of Ostreopsis Ovata both in the top water column* (arpaTS) and *in the bottom sea* (arpaBS), through $K = 20$ geosensors, installed in Apulia (Italy) by ARPA-Apulia Agency (Regional Agency for Environmental Prevention and Protection). Geosensors were irregularly distributed in the area with latitude varying between 39.8501 and 41.919 and longitude varying between 15.3418 and 18.4846. For each field data were measured every two weeks from June 15th to September 30th, 2016 ($T = 8$). Upper data vary between 0 and 848066 *cells per liter* (see Figure 4(a)). Bottom datavary between 0 and 7362000 *cells per liter* (see Figure 4(b)).

MESA Air Pollution geosensor network (`http://depts.washington.edu/mesaair/`) was used to measure the NO_x *Concentration*, through $K = 20$ geosensors, installed in California. Geosensors were irregularly distributed in the area with latitude varying between -118.43 and -117.751 and longitude varying between 33.7861 and 34.176. Data were measured every two weeks from January 13th, 1999 to January 12th, 2000 ($T = 48$). They vary between 0.002 and 9.329 *ppb* (see Figure 4(c)).

TCEQ Air Climate geosensor network (`http://www.tceq.state.tx.us/`) was used to measure the *Air Temperature* (tceqT), *Wind Speed* (tceqW) and *Ozone Concentration* (tceqO), through $K = 26$ geosensors, installed in Texas. Geosensors were irregularly distributed in the area with latitude varying between 26.13083 and 33.13222 and longitude varying between -106.5011 and -93.98778. For each field, data were measured hourly from May 5th (00:00) to 7th (00:00) 2009 ($T = 48$). Air Temperature data vary between 60.9 and 103.8 F^o (see Figure 4(d)), Wind Speed data vary between 0.6 and 18.5 *mph* (see Figure 4(e)) and Ozone Concentration data vary between 0 and 59 *ppb* (see Figure 4(f)).

NCDC Air Climate geosensor network (`http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/climate-normals`) was used to measure the *Solar Energy* (ncdcS), *Precipitation* (ncdcP) and *Air Temperature* (ncdcT), through $K = 72$ geosensors, installed in the United States. Geosensors were irregularly distributed in the area with latitude varying between -122.6068 and -67.8833 and longi-

tude varying between 26.5258 and 48.7412. Data were measured monthly from August 2005 to July 2009 ($T = 48$). Solar radiation was recorded as a monthly-averaged measure of the total daily solar radiation. Precipitation was the total amount of precipitation measured in the month. Air Temperature was the maximum air temperature registered in the month. Solar Radiation data vary between 1.1 and 30.5 $MJ/meter^2$ (see Figure 4(g)), Precipitation data vary between 0 and 657.2 mm (see Figure 4(h)) and Air Temperature data range between -13.1 and 47.5 C^o (see Figure 4(i)).

SR US Solar geosensor network SR geosensor network (`http://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/solar-radiation/`) was used to measure the Diffuse Solar Radiation, through $K = 1071$ geosensors, installed across the United States. Geosensors were regularly distributed in a grid 0.5 degrees by 0.5 degrees of latitude/longitude of the area with latitude varying between 37.05 and 41.95 and longitude varying between -123.95 and -122.05. Data were hourly measured on July 17th 2004 from 7:00 to 18:00 ($T = 12$). They vary between 4 and 573 W/m^2 (see Figure 4(j)).

SAC Air Climate geosensor network (`http://climate.geog.udel.edu/~climate/`) was used to measure the *Air Temperature*, through $K = 900$ geosensors, installed in South America. Geosensors were regularly distributed in a grid 0.5 degrees by 0.5 degrees of latitude/longitude of the area with latitude varying between -75.75 and -53.75 and longitude varying between -55.75 and -34.25. Data were monthly-averaged measures from January 1999 to December 2000 ($T = 12$). They vary between 1.6 and 26.1^o C (see Figure 4(k)).

5.2. *Methodology and metrics*

We evaluate the performance of the interpolation process along with its accuracy and efficiency. In the accuracy evaluation, we use the k-fold cross validation methodology to guarantee a prediction phase with out-of-sample sets of unknown data (see Section 5.2.1). To safely compare various algorithms, they are run with the same k-fold cross validation for each data set. In the efficiency evaluation we consider the time spent computing the interpolation model, processing the entire data set (see Section 5.2.2).

5.2.1. Accuracy evaluation

The k-fold cross validation is a commonly used model validation methodology in machine learning, in order to assess how the results of a machine learning predictive method will generalize to an independent (out-of-sample) data set. It allows us to learn an interpolation model from a data set of known data (training data set), while an out-of-sample data set of unknown data (testing data set) can be used to evaluate the performance of the learned model. We note that by resorting to this methodology our evaluation limits problems like overfitting and gives an insight on how the interpolation model will generalize to an out-of-sample data set. Specifically, we use the k-fold cross validation, in order to randomly partition geosensors into k equally-sized complementary sub-samples (folds), so that the holdout evaluation methodology can be repeated k times across the derived folds. Each time the hold-out method consists in performing the training phase on one subset (called the training set) and using the learning model to predict unused data in the other subset (called the testing set).[1] Procedurally, we repeat the learning phase on k

[1]One of the main reasons for using cross-validation instead of using the conventional validation (e.g. repeatedly partitioning the data set into two disjoint sets for training and for testing) is that, in several cases, there is

trials. At each trial one of the k folds is used as the testing set producer and the hold-out $k-1$ folds are put together to form a training set producer. The interpolation model is learned from the data measured from the training geosensors. This interpolation model is then used to predict out-of training data measured from the corresponding testing geosensors. In this way, the testing predictions are computed across all the k trials. The advantage of this methodology is that it matters less how the data are divided. Every data point is in a testing set exactly once and is in a training set $k-1$ times. Finally, we note that, according to the described use of the k-fold cross validation, the learning phase of the interpolation process is never performed by considering cross-sectional data pooled over time. This satisfies our problem formulation, where the PCA is computed by treating the full series of the windowed data of each training geosensor as a single example of the training set. We note that this formulation is coherent with the main goal of this paper, that is, illustrating an interpolation algorithm to supplement unseen data, where no geosensor is installed to acquire them. This scope is slightly different from the case inherent to the cross sectional data, where we consider geosensors that may miss data acquisitions at certain times and the interpolation algorithm can be used to supplement these missed data.

We perform the five-fold cross validation ($k = 5$) with NCDC, SR and SAC and the leave-one-out cross validation with ARPA, MESA and TCEQ. The leave-one-out cross validation is a special case of the k-fold cross validation, which splits a network of K geosensors into K folds of size 1 ($k = K$). We use the leave-one-out cross validation to guarantee a sufficiently large training set when a small network (ARPA with $K = 20$, MESA with $K = 20$ and TCEQ with $K = 26$) is used to collect the geophysical data. In this study, the accuracy of the interpolation process is measured over the k-fold cross validation of a data set in terms of the Root Mean Squared Error and the Akaike information criterion. Formally, let $\mathcal{D}(Z, G, \Delta T)$ be a geophysical data set, $G1, \ldots, G_k$ a k-fold cross validation that partitions the geosensor network G into k disjoint geosensor folds ($\bigcup_{h=1}^{k} G_h = G$ and for each $1 \leq h_1, h_2 \leq k$, $h_1 \neq h_2$ $G_{h_1} \cap G_{h_2} = \oslash$), then RMSE and AIC are computed as follows:

$$RMSE(\mathcal{D}) = \sqrt{\frac{RSS(\mathcal{D})}{size(\mathcal{D})}}, \tag{3}$$

$$AIC = size(\mathcal{D}) \ln \frac{RSS}{size(\mathcal{D})} + 2H, \tag{4}$$

where $RSS(\mathcal{D}) = \sum_{i=1}^{N} \sum_{j=1}^{T} \left(z_{t_j}(g_i) - \hat{z}_{t_j}(g_i)\right)^2$ is the residual sums-of-squares. $size(\mathcal{D})$ is the data set size, that is, $size(\mathcal{D}) = N \times T$, with N the number of geosensors in G and T the number of time points in ΔT. $z_{t_j}(g_i)$ is the field value measured by geosensor g_i at time t_j. $\hat{z}_{t_j}(g_i)$ is the field value predicted for geosensor g_i at time t_j by the interpolation model learned from the complementary training set of g_i. Let G_h be the fold comprising g_i ($g_i \in G_h$). The interpolation model to predict $\hat{z}(g_i)$ is learned from the training set

not enough data available to partition a data set into separate training and test sets without losing significant modeling or testing capability. In these cases, a fair way to properly estimate model prediction performance is to use cross-validation as a powerful general technique.

observed by the geosensors belonging to $G - G_h$. Finally, H is the number of estimated parameters (variogram family, nugget, sill and range for each theoretical variogram fitted to a sample variogram) in the interpolation models. We note that AIC is a criterion which includes a penalty in the error estimate for all computed parameters. The lower the RMSE and AIC, the better the interpolation model.

5.2.2. Efficiency evaluation

The efficiency of the interpolation process is measured as the computation time (in seconds) spent in learning the interpolation models of each data set. For each data set, the computation times are measured by running the interpolation algorithm (without k-fold cross validation) on an Intel(R) Core(TM) i7-4720U CPU@2.60GHz and 16G RAM running Microsoft Windows 8.1 (64 bit).

5.3. *Compared algorithms*

The implementation of the algorithms compared in this experimental study is written in R (version Rx64 3.4.0). Both the code and the data can be downloaded at `http://www.di.uniba.it/~appice/software/COSTK/index.htm`. CoST$^{\mathsf{K}}$ is the local spatio-temporal algorithm that implements the interpolation strategy illustrated in this study. To investigate the effectiveness of PCA we also evaluate a variant (named CoKALL) that excludes PCA, constructs one secondary co-variable for each backward time point in the window and injects windowed co-variables in CoKriging.[1] In addition, we evaluate a variant (named CoST$^{\mathsf{K}}$PC) that employes PCA, selects the top-ranked Principal Components, which explain 90% of the variance in the backward windowed data, and considers the selected Principal Components as distinct secondary co-variables in CoKriging. The technical set-up of these algorithms is described in Section 4.2.1.

Kriging (Krige 1951, Cressie 1993) is the purely spatial interpolate baseline. It learns a local, spatial Kriging interpolation model for each time point t, so that the model learned from the training data of the t-stamped data snapshot is used to predict testing data in the same t-stamped snapshot. For the technical set-up of this algorithm we consider the Linear, Gaussian, Exponential, Spherical and Matern models as candidate theoretical models. We use the Weighted-Least Mean Square method with weights equal to $N_j/(h_j^2)$, starting from default initial parameters (see details in (Pebesma and Graeler 2017)). We select the best theoretical model minimizing the fitting error.

STKriging (Gräler *et al.* 2016) is the global spatio-temporal baseline. It learns the global spatio-temporal Kriging interpolation model from the entire data set. The same global model is then used to predict testing data at each time point of the observational time interval. We use the separable model to represent each spatio-temporal covariance function as the product of a spatial and temporal term with a global sill. We select this model based upon the consideration that the separable model has a strong computational advantage in the setting, where each spatial location has an observation at each temporal instance (Pebesma 2012, Gräler *et al.* 2016), that is exactly the data scenario that we consider in this study. We account for the note of Gräler *et al.* (2016) and select the best spatio-temporal covariance function performing the search over the Linear, Gaussian, Exponential, Spherical and Matern families for both space and time (looping on twenty-five configurations). We use the Weighted-Least Mean Square method with weights equal

[1] CoKALL represents the baseline CoKriging scheme as it is investigated by Rouhani and Wackernagel (1990), Skoien and Bloschl (2007), Zhang *et al.* (2009) and Sideris *et al.* (2014) in the spatio-temporal context.

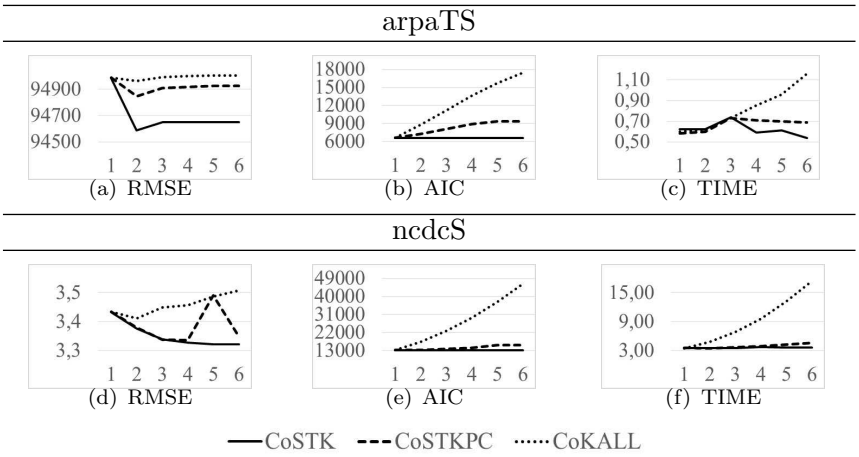

Figure 5. CoKALL, CoST^{K}PC and CoST^{K} (arpaTS and ncdcT): RMSE (axis Y, Figures 5(a) and 5(d)), AIC (axis Y, Figures 5(b) and 5(e)) and computation time (in seconds, axis Y, Figures 5(c) and 5(f)) by varying Ω between 1 and 6 (axis X).

to $N_j/(h_j^2)$. We assess the default initial parameters for both space and time according to the guideline in Pebesma and Graeler (2017), by considering the spatio-temporal surface and fixing the counterpart (space for time, and time for space) at zero (as suggested in (Gräler *et al.* 2016)). We also account for a suggestion of Gräler *et al.* (2016) and deduct the overall spatio-temporal sill of the separate model from the plateau that a nicely behaving sample variogram reaches for large spatial and temporal distances.

Finally, we note that all compared algorithms perform the theoretical model selection and the parameter fitting each time a new interpolation model is learned in the experiment. The local methods (CoST^{K} and Kriging) repeat the learning phase at each new time point, while the global method (STKriging) performs a single learning phase for the entire data set (for each trial of the considered cross validation).

5.4. *Results and discussion*

5.4.1. CoKriging-family algorithms

We analyze CoKALL, CoST^{K}PC and CoST^{K}, in order to find a baseline for the configuration of the Window operator in these algorithms, as well as to look for the empirical evidence of the heuristic considerations reported in Section 4.1. Specifically, we intend to show that removing collinearity in co-variates by PCA can generally improve the efficacy of the computed Co-Kriging-based interpolation model (i.e. CoST^{K}PC and CoST^{K} outperform CoKALL). Considering only the top-ranked Principal component is sufficient to explain a very high proportion of the total variability exhibited by the secondary data without decreasing the accuracy of the model, while reducing the complexity of the interpolation process (as a lower number of parameters is estimated) and diminishing the time spent computing the model (i.e. CoST^{K} outperforms CoST^{K}PC).

We start investigating how the efficacy of these algorithms may change with window size Ω. For this sensitivity study we focus on arpaTS and ncdcT. These data sets are selected for their different distribution properties. Specifically, arpaTS collects irregular data with a high number of outliers (see the box plot in Figure 4(a)), while ncdcT collects

regular data without significant outliers (see the box plot in Figure 4(i)). We vary Ω between 1 and 6. Both the RMSE, AIC and computation time of the compared algorithms are plotted in Figures 5(a)-5(c) for arpaTS and in Figures 5(d)-5(f) for ncdcT, respectively. We note that for these data sets $\mathsf{CoST}^{\mathsf{K}}$ performs better than its variants CoKALL and $\mathsf{CoST}^{\mathsf{K}}$PC. It exhibits the lowest RMSE, AIC and computation time independently of Ω. In all algorithms, the RMSE generally diminishes as Ω increases, by achieving an asymptotic maximum around $\Omega = 4$. This suggests considering this size as a baseline configuration for the Window operator. On the other hand, focusing the attention on the AIC and the computation time, we can also note that only the performance CoKALL is particularly sensitive to the window size. In fact, both the AIC and computation time of CoKALL exhibit a clear monotonic crescent trend with Ω. This is a natural consequence of the penalization of the error, due to the augmented number of parameters to estimate, as well as to the increased complexity of the system of CoKriging equations to be solved.

We proceed by extending the analysis of the efficacy of the considered algorithms along all data sets. For this analysis we consider window size $\Omega = 4$. The RMSE, AIC and computation time collected with $\Omega = 4$ are reported in Table 1 for all data sets. Analyzing RMSE we note that both algorithms with PCA ($\mathsf{CoST}^{\mathsf{K}}$PC and $\mathsf{CoST}^{\mathsf{K}}$) generally yield more accurate interpolations than their baseline without PCA (CoKALL). On the other hand, the consideration of the top-ranked Principal Component as the unique co-variate of the CoKriging system of equations ($\mathsf{CoST}^{\mathsf{K}}$) is often sufficient to achieve the most accurate CoKriging-based interpolation model (see arpaBS, mesa, tceqT, ncdcS, ncedT and sr). In any case, even when the lowest error is produced by either CoKALL (ncdcP and SAC) or $\mathsf{CoST}^{\mathsf{K}}$PC (arpaTC, tceqW and tceqO), the gain in accuracy achieved by the outstanding variant is negligible. Again this is supported by the analysis of the AIC. $\mathsf{CoST}^{\mathsf{K}}$ still outperforms its variants when introducing the number of parameter-based penalization of the error. Finally, analyzing the computation time spent learning the interpolation model, we verify how the time cost of the learning process is actually increased with the number of co-variates. This computation analysis is completed by also accounting for considerations deserved by the theoretical time complexity of both PCA and CoKriging. In the $\boldsymbol{O}$ notation, the time complexity of performing PCA is $\boldsymbol{O}(Nm^2)$ with N the number of examples and m the number of variables, while the time cost of computing a CoKriging system of equations with M variables is $\boldsymbol{O}(N^3M^3)$. Employing PCA allows us to run CoKriging with a number of variates $M \leq m$. The computation times collected via this empirical study show that, although some time is spent performing PCA, the derived data reduction diminishes the time spent computing the CoKriging system independently of the size of processed data. In fact, CoKriging without PCA is always more time-consuming than CoKriging with PCA.

5.4.2. CoKriging vs Spatial and Spatio-temporal Kriging

We analyze the performance of $\mathsf{CoST}^{\mathsf{K}}$ (run with $\Omega = 4$), its spatial counterpart Kriging, as well as its spatio-temporal counterpart STKriging. We intend to investigate how the proposed learning schema is able to take advantage of the temporal dimension of the data, by gaining in accuracy and/or efficiency with respect to the state-of-the-art baselines. The RMSE, AIC and computation time are reported in Table 2.

We start evaluating the efficacy of these algorithms along the RMSE. The results reveal that Kriging yields the most accurate interpolations in three out of eleven data sets, STKriging yields the most accurate interpolations in three out of eleven data sets, while $\mathsf{CoST}^{\mathsf{K}}$ yields the most accurate interpolations in five out of eleven data sets. These

Table 1. CoKALL, CoST^{K}PC and CoST^{K}: RMSE and AIC (computed according to the cross validation methodology), as well as computation time (in seconds) (see details in Section 5.2). The best results are underlined (and runner-up marked with *). Results are collected with $\Omega = 4$.

data set	RMSE			AIC			Computation time (secs)		
	CoKALL	CoST^{K}PC	CoST^{K}	CoKALL	CoST^{K}PC	CoST^{K}	CoKALL	CoST^{K}PC	CoST^{K}
arpaTS	94999.70	94918.3*	94648.3	13587.721	8877.447*	6626.535	0.97	0.71*	0.59
arpaBS	1322293	1319134*	1303050	14430.361	9707.595*	7465.669	0.85	0.57*	0.54
mesa	2.50683	2.50013*	2.49480	85284.516	27439.37*	20715.28	6.21	3.65*	3.25
tceqT	2.98377	2.95119*	2.95092	111304.59	28969.19*	27348.96	6.14	3.18*	3.23
tcqW	2.04403	2.01963	2.02763*	110360.45	44678.48*	26412.34	6.17	3.66*	3.09
tceqO	6.635131*	6.60317	6.65559	113299.37	51499.32*	29379.06	6.23	3.96*	3.11
ncdcS	1.66475	1.63856*	1.62782	24402.886	12743.289*	8107.841	10.85	5.75*	3.95
ncdcP	48.17996	49.11907	48.2383*	47663.60	41251.03*	31531.98	9.61	6.96*	3.46
ncdcT	3.45759	3.33634*	3.32707	29454.85	14508.11*	13048.87	9.63	3.97*	3.73
sr	50.6590	50.6271*	50.5818	105211.2	103665.0*	101992.0	1468.25	703.46*	131.16
sac	0.36490	0.36549*	0.36549	-17455.5	-20600.4*	-20600.4	852.39	80.5*	75.65

results show that accounting for temporal information commonly gains in accuracy. In a few cases (arpaTS, tceqT and sr) the spatial process is not improved by the addition of the temporal dimension. In all these cases data exhibit a massive presence of outliers that, distributed in time, weakens the property of the temporal correlation assumed in temporal modeling. In any case, in these data sets CoST^{K} appears more robust to this phenomenon than the spatio-temporal competitor, yielding a lower error than STKriging. General evidence of the higher robustness of the spatio-temporal scheme proposed here is given by the fact that CoST^{K} is more accurate than STKriging in eight out of eleven data sets. This is mainly due to the difficulty encountered by the procedure in STKriging that is in charge of determining a "good" characterization of the joint spatio-temporal co-variance structure (Gräler *et al.* 2016).

We complete this discussion by comparing the efficacy of the algorithms along the AIC and the computation time. The AIC results show that determining a global spatio-temporal model has the advantage of diminishing drastically the number of parameters to be estimated, although the complexity of the estimation for a few parameters leads, in any case, to a time-consuming learning process. Although STKriging achieves the lowest AIC in the majority of data sets, its learning is the slowest in all data sets.

Based upon these considerations we draw the conclusion that the combination of CoKriging and PCA performed by CoST^{K} via the windowing mechanism can be considered a viable interpolation schema to deal with the temporal dimension of geophysical data. This achieves an acceptable compromise between the accuracy of predictions and the efficiency of learning.

6. Conclusion

This paper describes a new algorithm, called CoST^{K}, that combines spatial and temporal information, in order to enhance the accuracy of computed interpolation models. It uses the window-based computation methodology, in order to derive a model of the spatio-temporal correlation by also accounting for the possible temporal variation of a physical field distribution. A secondary co-variable is computed through the Window operator and PCA, in order to summarize the dynamic structure of spatial correlation over time. CoKriging is employed, in order to analyze the dynamic structure of these two variables and apply it as a time-evolving, spatio-temporal interpolator of the geosensor data. The

Table 2. CoST^{K} (with $\Omega = 4$), Kriging, STKriging: RMSE, AICc and computation time. The best results are underlined (and runner-up marked with *).

data set	RMSE			AIC			Computation time (secs)		
	Kriging	STKriging	CoST^{K}	Kriging	STKriging	CoST^{K}	Kriging	STKriging	CoST^{K}
arpaTS	94509.7	102159	94648.3*	4946.066*	4050.973	6626.535	0.37	23.18	0.59*
arpaBS	212331	125367	1303050*	5941.916*	4853.308	7465.669	0.92	53.31	0.54*
mesa	2.48583*	2.36383	2.49480	9428.370*	2011.744	20715.28	1.85	114.36	3.25*
tceqT	2.93755	59.4807	2.95092*	12673.63*	10665.78	27348.96	1.89	47.25	3.23*
tcqW	2.03057	1.98096	2.02763*	11751.95*	2174.223	26412.34	1.75	80.22	3.09*
tceqO	6.81207*	6.88940	6.65559	14773.06*	5285.240	29379.06	1.67	55.08	3.11*
ncdcS	1.65516*	6.31096	1.62782	5402.968	12823.89	8107.841*	2.03	21.78	3.95*
ncdcP	55.9077	53.4231*	48.2383	29731.83*	27587.62	31531.98	1.75	186.49	3.46*
ncdcT	3.34437*	6.68736	3.32707	10264.71	13224.31	13048.87*	2.00	29.64	3.73*
sr	50.5214	50.7794	50.5818*	101301.3*	101042.2	101992.0	36.46	357.14	131.16*
sac	0.36897*	0.43994	0.36549	-21055.8	-17645.6	-20600.4*	18.55	241.28	75.65*

empirical evaluation investigates the viability of the proposed algorithm in various applications. We compare CoST^{K} to the purely spatial baseline algorithm Kriging, as well as to its spatio-temporal competitor STKriging. The results show that CoST^{K} generally yields more accurate estimates than the spatial and spatio-temporal competitors. The proposed algorithm has advantages (in terms of predictive accuracy robustness) when it is applied to geo-physical phenomena which exhibit patterns that are time-evolving. In this case, the defined spatio-temporal co-variable, computed with the Window operator, is effectively informative of the actual correlation property across the space and time. The accuracy of the computed interpolation models benefits from accounting for this local model of the spatio-temporal correlation in the modeling phase.

A difficulty which is still unsolved arises from the complexity of the learning phase. In this study, the entire learning process (i.e. co-variable synthesis and CoKriging) is repeated from scratch, as a new data window is processed. This is performed without considering that the learning phase is triggered on sliding windows (i.e. as a new window is considered, the oldest data snapshots are discarded, while the newest data snapshots are considered), so that windows at consecutive time points may share data. An interesting research task will consist in formulating an incremental version of the proposed learning process (such as that usually deployed in a streaming scenario (Appice *et al.* 2014b)). This will avoid learning a new spatio-temporal interpolation model at each time point and will account for the existing interpolation model (learned in the past) that will be incremented only if there are actual changes in the data distribution. Additionally, further investigations lead to the design of the proposed algorithm by resorting to the use of some big data technologies (e.g MapReduce), in order to evaluate the performances when large volumes of geosensor data are processed via parallel processing architectures.

7. Acknowledgment

Sonja Pravilovic's research was supported by the Ministry of Science of Montenegro, Higher Education and Research for Innovation and Competitiveness (INVO/HERIC). She received the national scholarship for excellence (1/10/2016-1/10/2017) funded by the proceeds of a loan from the International Bank for Reconstruction and Development. The authors wish to thank Lynn Rudd for her help in reading the manuscript and Benedikt Graler for answering our questions on STKriging.

References

Appice, A., *et al.*, 2014a. *Data Mining Techniques in Sensor Networks - Summarization, Interpolation and Surveillance*. Springer Briefs in Computer Science Springer.

Appice, A., *et al.*, 2013. Using trend clusters for spatiotemporal interpolation of missing data in a sensor network. *Journal of Spatial Information Science*, 6 (1), 119–153.

Appice, A., *et al.*, 2014b. Dealing with temporal and spatial correlations to classify outliers in geophysical data streams. *Information Sciences*, 285 (1), 162–180.

Aversano, G., Gicquel, O., and Parente, A., 2017. Surrogate Models based on the combination of PCA and Kriging. *In*: *Conference presentation, SIAM Sixteenth International Conference on Numerical Combustion*.

Buzzi-Ferraris, G. and Manenti, F., 2010. *Interpolation and Regression Models for the Chemical Engineer*. Wiley-WCH.

Chen, Y., *et al.*, 2016. Enhanced Statistical Estimation of Air Temperature Incorporating Nighttime Light Data. *Remote Sensing*, 8 (8).

Cressie, N., 1993. *Statistics for spatial data*. Wiley Interscience.

Cressie, N., 1990. The origins of Kriging. *Mathematical Geology*, 22 (3), 239–252.

Cressie, N. and Wikle, C.K., 2011. *In*: *Spatio-Temporal Statistical Models*. Wiley.

Dormann, C.F., *et al.*, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36 (1), 27–46.

Gafurov, A. and Bárdossy, A., 2009. Cloud removal methodology from MODIS snow cover product. *Hydrology and Earth System Sciences*, 13 (7), 1361–1373.

Goovaerts, P., 1998. Ordinary Cokriging Revisited. *Mathematical Geology*, 30 (1), 21–42.

Gräler, B., Pebesma, E., and Heuvelink, G., 2016. Spatio-Temporal Interpolation using GSTAT. *R Journal*, 8 (1), 204–218.

Guttorp, P. and Schmidt, A.M., 2013. Covariance structure of spatial and spatiotemporal processes. *Wiley Interdiscip. Rev. Comput. Stat.*, 5 (4), 279–287.

Huang, C., *et al.*, 2009. Multivariate Intrinsic Random Functions for Cokriging. *Mathematical Geosciences*, 41 (8), 887.

Isaaks, E. and Srivastava, R., 1989. *An introduction to applied geostatistics*. Oxford Univ. Press.

Jiang, Y., *et al.*, 2009. Natural and anthropogenic factors affecting groundwater quality in the Nandong karst underground river system in Yunan, China. *Journal of Contaminant Hydrology*, 109, 49–61.

Kinoshita, R., *et al.*, 2016. Large topsoil organic carbon variability is controlled by Andisol properties and effectively assessed by VNIR spectroscopy in a coffee agroforestry system of Costa Rica. *Geoderma*, 262, 254 – 265.

Knotters, M., Brus, D., and Voshaar, J.O., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, 67 (3), 227 – 246.

Krige, D., 1951. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52, 119–139.

Lin, G. and Chen, L., 2004. A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. *Journal of Hydrology*, 288, 288–298.

Mitas, L. and Mitasova, H., 1999. Spatial Interpolation. *In*: P. Longley, M. Goodchild, D. Maguire and D. Rhind, eds. *Geographical Information Systems: Priciples, Techniques, Management and Applications.*, Vol. 1 Wiley, 481–492.

Nazzal, Y., *et al.*, 2015. The combination of principal component analysis and geostatis-

tics as a technique in assessment of groundwater hydrochemistry in arid environment. *Current science*, 108 (6), 1138–1145.

Omitaomu, O.A., *et al.*, 2009. Knowledge discovery from sensor data (SensorKDD). *SIGKDD Explorations*, 11 (2), 84–87.

Pebesma, E., 2012. spacetime: Spatio-temporal data in R. *Journal of Statistical Software*, 51 (7), 1–30.

Pebesma, E. and Graeler, B., 2017. *Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation*. CRAN.

Pravilovic, S., *et al.*, 2017. Using multiple time series analysis for geosensor data forecasting. *Information Sciences*, 380, 31–52.

Rocha, M., *et al.*, 2012. Studying the influence of a secondary variable in collocated cokriging estimates. *Anais da Academia Brasileira de Ciencias*, 84, 335–346.

Romanowicz, R., *et al.*, 2006. A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. *Environ. Model. Softw.*, 21 (6), 759–769.

Rossiter, D.G., 2012. *Technical note: co-kriging with the gstat package of the R environment for statistical computing*. Cornell University.

Rouhani, S. and Wackernagel, H., 1990. Multivariate geostatistical approach to space-time data analysis. *Water Resources Research*, 26 (4), 585–591.

Shepard, D., 1968. A two-dimensional interpolation function for irregularly-spaced data. *In*: *Proceedings of the 1968 23rd ACM national conference*, ACM '68 New York, NY, USA: ACM, 517–524.

Sherman, M., 2010. *In*: *SpaceTime Data.*, 123–148 John Wiley and Sons, Ltd.

Shumway, R. and Stoffer, D., 2010. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics Springer.

Sideris, I.V., *et al.*, 2014. Real-time radarrain-gauge merging using spatio-temporal cokriging with external drift in the alpine terrain of Switzerland. *Quarterly Journal of the Royal Meteorological Society*, 140 (680), 1097–1111.

Skoien, J.O. and Bloschl, G., 2007. Spatiotemporal topological kriging of runoff time series. *Water Resources Research*, 43 (9).

Subramanyam, A. and Pandalai, H.S., 2008. Data Configurations and the Cokriging System: Simplification by Screen Effects. *Mathematical Geosciences*, 40 (4), 425–443.

Triantakonstantis, D. and Stathakis, D., 2014. Cokriging Areal Interpolation for Estimating Economic Activity Using Night-Time Light Satellite Data. *In*: B. Murgante, S. Misra, A.M.A.C. Rocha, C. Torre, J.G. Rocha, M.I. Falcão, D. Taniar, B.O. Apduhan and O. Gervasi, eds. *Computational Science and Its Applications – ICCSA 2014: 14th International Conference, Guimarães, Portugal, June 30 – July 3, 2014, Proceedings, Part IV*. Springer International Publishing, 243–252.

Turner, M.G. and Gardner, R.H., 2015. *Landscape Ecology in Theory and Practice: Pattern and Process*. Springer-Verlag New York.

Wackernagel, H., 2003. *Multivariate Geostatistics: an Introduction with Applications, 3rd edition*. Springer.

Weiss, D.J., *et al.*, 2014. An effective approach for gap-filling continental scale remotely sensed time-series. {*ISPRS*} *Journal of Photogrammetry and Remote Sensing*, 98, 106 – 118.

Zhang, C., Li, W., and Travis, D.J., 2009. Restoration of Clouded Pixels in Multispectral Remotely Sensed Imagery with Cokriging. *Int. J. Remote Sens.*, 30 (9), 2173–2195.