Computers and Electronics in Agriculture 142 (2017) 440–449

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Kriging-based approach to predict missing air temperature data

Anastasiya Shtiliyanova, Gianni Bellocchi*, David Borras, Ulrich Eza, Raphaël Martin, Pascal Carrère

INRA, UMR Ecosystème Prairial, VetAgroSup, 63000 Clermont-Ferrand, France

ARTICLE INFO

Keywords:
Air temperature
Daily and hourly resolutions
Data-driven machine learning
Infilling
Kriging-based temporal interpolation

ABSTRACT

The geo-statistical Kriging method is conventionally used in the spatial dimension to predict missing values in a series by utilizing information from neighbouring data, supported by the hypothesis that mathematical expectation is a function of distance between observations. By using a data-driven machine learning-based inferencing and exploration framework, this research applies a Kriging-based interpolation in the temporal dimension to fill in data gaps in time-series of air temperatures. It assesses its performance for artificial gap scenarios (ranging in length from single one to six consecutive data points) generated using data with both daily and hourly resolutions from five sites in Europe (Laqueuille, France; Grillenburg, Germany; Monte Bondone, Italy; Oensingen, Switzerland; Rothamsted, United Kingdom) and one in France overseas (Sedael, Réunion Island). Results show that the method is capable of predicting missing temperatures with acceptable accuracy, especially with the hourly resolution and for non-high elevation sites: modeling efficiency ($EF \leq 1$, optimum) > 0.8, with the exception of Monte Bondone, placed at > 2000 m a.s.l. ($EF < 0$). With daily data, maximum temperature was correctly predicted at all sites ($0.6 \leq EF < 0.9$), while some less accuracy (down to $EF < 0.4$) was noted when predicting missing daily minimum temperatures. In conclusion, the method appears suitable to be applied to fill in hourly temperature gaps, requiring more stringent hypotheses concerning daily data and mountain sites (but further studies are required to draw concluding recommendations).

1. Introduction

Long records of weather data are needed for evaluating scenarios in natural resource studies (e.g. Shenk and Franklin, 2001; Mavi, 2004). In many of these studies, surface weather observations are the fundamental forcing data of simulation models, for which the absence of a particular station's data at a given time can introduce biases into responses or generate spurious trends (e.g. Hoogenboom, 2000). Missing data are a common problem in meteorological observational datasets. Station history records are often incomplete (e.g. Menne et al., 2010) because instruments may break in malfunction, or data transmission may be interrupted. Meteorological phenomena such as precipitation, snow or ice (but also destruction by animals) may even cause the temporary failure to record observations. Likewise, corrosion is a common problem linked to system failure in weather stations and data loggers. Examples of problems that occur when weather stations are not regularly inspected and maintained can be easily found (e.g. http://www.surfacestations.org/odd_sites.htm). Ensuring regular maintenance of equipment and continuity in the collection of station metadata is critical to long-term station operation and interpretation of the data (e.g. NRCS, 2009). Although field maintenance and calibration of instrumentation is expected to be performed by the respective agencies, budgetary and staffing limitations may prevent routine inspections of weather stations. Consequently, inattention to maintenance has often been identified as the greatest source of failure in weather stations and networks (Davey et al., 2006, 2007).

The process of replacing missing data with substituted values (imputation) is known in fields like meteorology, ecology, climatology or geosciences (e.g. Zhang and Schultz, 1990; Simolo et al., 2010; McCandless et al., 2011; Taugourdeau et al., 2014). In climate change impact studies, in particular, a good resolution of weather data in time and space is needed for creating complete datasets of past climate as a basis for projecting future climate realizations (Ruane et al., 2015).

Missing data pose the challenge (both conceptual and technical) of finding a suitable way for replacing the missing values with a prediction (Schneider, 2001; Gelman and Hill, 2006). Indeed, the development of approaches for filling gaps in weather time series is a key challenge for improving data usability, either for statistical description of the meteorology of a given weather station or in the view of complex climatological and ecological studies (e.g. Linacre, 1992). At the same time, there is a need to preserve the existing dataset by replacing missing data with a "probable value". The latter depends on other available

* Corresponding author.
E-mail address: gianni.bellocchi@inra.fr (G. Bellocchi).

http://dx.doi.org/10.1016/j.compag.2017.09.033
Received 23 March 2017; Received in revised form 18 September 2017; Accepted 23 September 2017
Available online 11 October 2017
0168-1699/

information, including historical statistics and the climate conditions of the place of interest. This is where some understanding of the process under study is required because the density, distribution and range of the variable used to create an imputation function help determine which technique is most suitable for the gap-filling purpose. Techniques for filling missing observations in a time series include temporal interpolations, and rely on the relationship between available and missing data (Henn et al., 2013). There are three general explanations for missingness of the data (Rubin, 1987): Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not At Random (MNAR). If values are missing at random (MAR), then the available data may be representative of the population. MCAR is a special case of MAR and occurs when the events leading to data missingness are independent on both observable and unobservable factors. MNAR happens when the missing values depend on other missing values, that is, one or more factors are impossible to quantify and identify (Schafer and Graham, 2002). The prediction of missing weather data can rely on the MAR mechanism, because the exact moment when one device stops functioning is generally unknown (here, we exclude situations in which recurrent problems are identified, e.g. weather stations suffering power supply overload or voltage violation during long periods, e.g. Cheng et al., 2009). In this study, imputation is applied to a weather variable, i.e. air temperature (sampled at both hourly and daily resolutions), an important factor controlling most physical, chemical and biological processes on Earth, which are key to many environmental studies and the management of Earth surface resources. For instance, air temperature is a common indicator of ecological efficiency as it is related to fundamental characteristics of natural and agricultural systems, e.g. plant processes such as growth, development, sugar partitioning and stress sensing and response (e.g. White et al., 2005). Most agro-climatic indicators are based on air temperature (alone or in combination with other meteorological variables) in order to predict conditions of drought stress based on water balance calculations (e.g. Matthews et al., 2008). Energy-based equations have also been put forward for climate characterization, where the balance of incoming radiation (playing a major role in field ecology) can be predicted by air temperature data (e.g. Bellocchi, 2011; Bojanowski et al., 2013). In urban studies, the difference in air temperature between urban and rural locations within a given time period is a frequently used metric to describe heat islands (Fabrizi et al., 2010).

The above examples reflect the importance of air temperature in applications of wide scientific and engineering interest. Although this primary variable is available in many locations, temperature data series can contain gaps ranging from several hours to several days. Aim of this study is to assess a methodology for gap-filling of hourly and daily temperature series based on a learning process contingent on historical data series. Geo-statistical techniques, in particular, for which near spatial data values are more related to each other than distant data values (Tobler, 1970), are sufficiently versatile to be extended to the temporal dimension (e.g. Rossi and Posa, 1991; Guccione et al., 2012; Zehn and Cai, 2015). This study provides an approach for exploring and assessing the skill of a geo-statistically based method (1-dimensional case Kriging interpolation) by randomly removing available temperature data from time series at a variety of sites and then evaluates the ability of the technique to reconstruct the gaps, with interest in the impact of the length of missing periods and the fraction of overall missing data.

2. Methodology

2.1. Kriging-based imputation of missing data

In Supplementary material, we provide a brief review of the main approaches for the treatment of MAR data in general (also taking into account developments in data missingness made in other domains, e.g. medical surveys, Blankers et al., 2010). Through this review, we gained the understanding that the Kriging interpolation method (whose parameterization is developed on the available observations) holds great potential to reconstruct missing temperature data (the focus of this study). Moreover, comparatively to other MAR methods, it can also be relatively easily implemented and flexibly applied.

There are several versions of the Kriging method (Supplementary material). Here, we use the Ordinary Kriging (OK), which is computationally practical and easier to implement than other Kriging variants, that is, it is not necessary neither external data nor an additional knowledge to determine its parameters. Kriging permits to compute an unsampled value (z), knowing its coordinates (x, y) and neighbours. The Kriging methods are generally applied for data spatialization, where the spatial dimension concerns the position of the data and the distance between location points is taken into account. To give the prediction of $\hat{z}(s_0)$ to the unknown value$z(s_0)$, the general equation of the OK is a linear combination of n known sample values at points s_i around s_0, based on the vector of n observations at primary locations and the vector of Kriging weights. The latter reflect the spatial structure of the method, taking into account the neighbour observations. To compute the Kriging weights we followed Matheron (1963) and Gandin (1963), who introduced correlation functions between neighbouring values, also called semi-variances (Hengl, 2007, 2009). The experimental semi-variance is computed based on the lag vector representing separation between two spatial locations, the vector of spatial sample coordinates, and the number of sample pairs separated by lag. The formula is applied on the pairs of points. Using this formula, a semi-variogram is produced to describe the spatial auto-correlation of the variable. Compared to other methods (e.g. covariance), the key advantages of semi-variograms are: the method is based on raw data and requires no pre-calculated indicators (e.g. means, minima or maxima); both linear and non-linear changes can be detected; the identified change is in relation to expected dynamics, which represent the whole series considered, not just the start and end of a given series; and the dynamics of the series can be analysed, as changes in semi-variogram parameters relate to changes in different aspects of the series. In general, the semi-variogram is fitted using authorized variogram models like linear, spherical, exponential, Gaussian, etc. A semi-variogram is described by its sill (the semi-variogram upper bound), practical range (the distance at which the semi-variogram reaches the sill) and nugget effect (a discontinuity of the semi-variogram that can be present at the origin, typically attributed to micro-scale effects or measurement errors). An important hypothesis is that the Kriging assumes some form of stationarity in the variable under study. This could be a limitation in its use and generalization because time series are often statistically described as a random process in which extremes and seasonalities are more dominant than autocorrelation and stationarity (e.g. Cressie and Wikle, 2011). Different types of non-stationarity exist (e.g. Meul and Van Meirvenne, 2003; De Benedetto et al., 2012) and, in practice, stationarity is often not guaranteed for large distances (Rivoirard, 2005). For instance, a phenomenon may appear stationary locally, whereas it may be non-stationary over longer distances, e.g. in the presence of a large-scale trend. Stationarity is an important condition that is relevant to gap-filling in general (Seitchik, 2012).

In this study, interpolations are based on the gradients between points of the nearest neighbours in the time series. Missing values are filled in based on all the characteristics inherent to the historical set. They are used within a geo-statistical interpolation method to calculate the best linear unbiased predictor in the geo-statistical context (e.g. Journel, 1986) by accommodating records by means of a memory-based process (after concepts by Enzi and Camuffo, 1996; Diodato and Bellocchi, 2014). Using the memory-based concept, any temperature value in a time series is affected by its previous neighbours and, in turn, affects its next-neighbour values. It is based on these principles that interpolation techniques have been developed to handle continuous values using data points on either side of the data gap (Koehler, 1977; Pielke, 1984; Miller, 1990).