Environmental Modelling & Software 66 (2015) 69–77

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A geostatistics-assisted approach to the deterministic approximation of climate data

Maria Lanfredi [a, *], Rosa Coppola [a], Mariagrazia D'Emilio [a], Vito Imbrenda [a], Maria Macchiato [b], Tiziana Simoniello [a]

[a] Institute of Methodologies for Environmental Analysis, Italian Research Council, C.da S. Loja, Tito, PZ, I-85050, Italy
[b] Department of Physics, "Università Federico II", via Cinthia, Napoli, I-80126, Italy

ARTICLE INFO

Article history:
Received 12 August 2014
Received in revised form
7 December 2014
Accepted 11 December 2014
Available online

Keywords:
Geostatistics
Geographical model
Climatic surface

ABSTRACT

We propose a nonconventional application of variogram analysis to support climate data modelling with analytical functions. This geostatistical technique is applied in the theoretical domain defined by each model variable to detect the systematic behaviours buried in the fluctuations determined by other driving factors and to verify the ability of candidate fits to remove correlations from the data. The climatic average of the atmospheric temperature measured at 387 European meteorological stations has been analysed as a function of geographical parameters by a step-wise procedure. Our final model accounts for non-linearity in latitude with a local-scale residual correlation that decays in approximately ten kilometres. The variance of the residuals from the fitted model (approximately 3% of the total) is mostly determined by local heterogeneity in transitional climates and by urban islands. Our approach is user-friendly, and the support of statistical inference makes the modelling self-consistent.

1. Introduction

Recent satellite remote sensing technologies for Earth Observation (EO) have supplied a large amount of spatial data that are promising for improving our understanding of the climate system. Contextually, the sparse and uneven data provided by ground stations are still an essential source of information on many key variables characterizing climate dynamics. Currently, the collection of data obtained from meteorological networks, which are generally regarded as valid for spatial inferences of the state of the low atmosphere (Geiger et al., 2003), are also used within climatic studies at the planetary scale. As an example, the series of global datasets, HadCRUT, gridded on a $5^\circ \times 5^\circ$ latitude–longitude box grid, has been widely exploited for the evaluation and attribution of climate change (e.g., Brohan et al., 2006; Jones and Stott, 2011; Jones et al., 2012).

Multi-resolution, both in time and in space, provides the standard hierarchical framework for studying the dynamics of the climate system. Because details at different resolutions generally characterize different physical structures, a coarse-to-fine descriptive strategy is used to separate the broad scale context that is properly climatic from the local contexts of weather dynamics.

In regional studies, numerical models (Rummukainen, 2010; Feser et al., 2011) add details to global-scale climate models, thus improving simulations and forecasts. Within projects focused on long-term simulations or projections, Regional Climate Models (RCMs) currently operate at horizontal grid resolutions between 25 and 50 km [e.g., PRUDENCE (Christensen and Christensen, 2007), ENSEMBLES (Hewitt, 2005) and NARCCAP (http://www.narccap.ucar.edu/)]. On the whole, there is an increasing demand of fine-scale data (e.g., Jeffrey et al., 2001; Huld et al., 2006; Hancock and Hutchinson, 2006; Daly et al., 2008; Tang et al., 2012) that can be useful to understand any environmental process linked to climate. Within the proper climatic context, a horizontal resolution of 7–10 km is currently recognized as a good target (e.g., Suklitsch et al., 2011).

Our research activity (e.g., Lanfredi et al., 2004; Simoniello et al., 2008, 2011) examines complex processes linking climate and the land surface (Piao et al., 2006; Cleland et al., 2007; Prieto-Blanco et al., 2009). Such studies use remote sensing observations of land and require realistic and accurate climatic surfaces obtained by interpolating data from meteorological stations to be interfaced with remote information. In particular, we need to construct air temperature surfaces that can be linked to land surface maps to

* Corresponding author. Tel.: +39 0971427284; fax: +39 0971427271.
E-mail address: lanfredi@imaa.cnr.it (M. Lanfredi).

http://dx.doi.org/10.1016/j.envsoft.2014.12.009
1364-8152/

better understand biosphere spatio-temporal patterns and to characterize exchange processes (e.g., carbon emission–absorption) that actively involve climatic fluctuations (e.g., Cox et al., 2000; Yuan et al., 2010).

Surface data are mostly obtained by the pure interpolation of sampled observations (e.g., by Thin Plate Smoothing Splines; Hopkinson et al., 2012). More complex strategies combine human-expert knowledge and statistical methods to satisfy the increasing demand for spatial climate data sets in digital form (Daly et al., 2008). All of these methodologies directly supply end-users with gridded data; the underlying physical mechanisms that shape the climatic surfaces are not singled out and thus remain encapsulated within the complexity of the gridding algorithms. Nevertheless, modelling the relationships between a given variable and the factors that generate its spatial patterns is crucial in many scientific frameworks. In our case, we have to consider that the spatial variability of both the land surface and low atmosphere variables is influenced by geography and topography. Any study focussing on fluctuations generated by mutual interactions between these two environments needs to discriminate geographic-induced background patterns that could distort correlation analyses.

This requirement led us to work on the development of a regressive approach that can account for causal linkages between geographic factors and temperature. General non-linear regression implies that functional form selection, estimation of best-fit parameters, and evaluation of fit performances are rather difficult. In contrast to linear regression, there is no closed-form expression for the best-fitting parameters and departures from the optimal approximation can occur, which could not be accounted for by global cost functions and require weighty goodness-of-fit tests (Caouder and Huet, 1997; Crainiceanu and Ruppero, 2004; Demidenko, 2006).

Here, we focus on a simpler approach by developing an additive regression model that is non-linear in the explanatory variables. The ability of such a model to generate random errors starting from spatially structured patterns can be considered as an *a posteriori* criterion to evaluate its performance. The main idea of our proposal is that we can use variogram analysis (Cressie, 1993; Wackernagel, 2003) to characterize the scale properties of the response variable along pseudo-directions that are defined by the explanatory variables of the model within an identification–estimation-checking iterative approach to model building. This analysis can be particularly useful in the diagnostic checking phase to verify the ability of the fit to remove correlation structures from the data and thereby randomize residuals from the fitted model ("whitening"). Efficient best fits should flatten the variogram at the right variance level; improper best fits should result instead in residual correlation between the response and explanatory variables over large scales. This validation is also important because it enables us to evaluate if the prediction error is actually the minimum allowed by the intrinsic degree of randomness of the data. Rigorously speaking, long-range correlation could also be observed in the case of fractal data, but this peculiar circumstance is recognizable due to the typical power law dependence that characterizes them (e.g., Brown and Liebovitch, 2010). Thus we are limited to consider determinism against stationary randomness. Of course, differently from the standard geostatistical applicative framework, the model variables are not necessarily spatial coordinates.

We illustrate our strategy by building up a geographical model for the climatic average of atmospheric temperature over Europe. Data from 387 meteorological stations were recorded over the 30-year period from 1961 to 1990, where the latest global "Normals" are currently defined for climate reference (http://www.wmo.int/pages/themes/climate/statistical_depictions_of_climate.php) according to the World Meteorological Organization. Although air surface temperature is one of the most continuous and studied variables within climate analyses, not only its deep dynamical features in time are still discussed (e.g., Lanfredi et al., 2009 and references therein) but also in truly applicative contexts there is no single strategic approach to the modelling, as observed above for climatic variables in general. We refer to the 8 km × 8 km resolution of the GIMMS-AVHRR (Global Inventory Modelling and Mapping Studies-Advanced Very High Resolution Radiometer) data, which are usually exploited for monitoring land cover in climatic studies (e.g., Zeng et al., 2013). This resolution corresponds well to the typical finest scales of RCMs (e.g., Suklitsch et al., 2011) and, as it will be shown in the following, emerges naturally from scale analyses as a reasonable boundary between locality and globality. The main variables shaping the basic structural part of the spatial variability of near-surface temperature at that resolution in a climatic context are latitude, longitude and elevation. We have also included the distance from the coastline to illustrate our approximation process step by step. The final part of the paper concerns a detailed discussion of the residuals from the fitted model and the comparison between the performances of our model against a standard multi-regressive linear model.

2. Data and study area

The annual mean air temperature data concerning the 30 years climatic period from 1961 to 1990 were obtained from 387 meteorological stations located in the European part of the Eurasian continent (Fig. 1) by averaging daily data. Most of the data were provided by the European Climate Assessment & Dataset (ECA&D) project (Klein-Tank et al., 2002); few stations (<2%) were integrated from local databases to introduce additional information in poorly represented areas.

Differences in latitude and elevation are expected to play a major role in determining the mean annual value of the air temperature, but the longitude and distance to the sea could also be significant parameters. In particular, from the point of view of general atmospheric circulation, the investigated area falls in the Ferrel cell of the Northern hemisphere where prevailing winds are westerlies. Because the west coast of Europe is located on the Atlantic Ocean, whereas the eastern part is continental, the westerlies move hot air masses inland from the sea in the direction of increasing longitude during winter. As a consequence, non-stationary behaviours are expected in the West(Sud)/East(Nord) direction. This variability should prevalently concern annual excursions, but the annual mean values could also be affected. Moreover, sea proximity, in general, modifies the minimum temperature in coastal swaths, which is why this parameter is included in the set of geographical parameters potentially involved in determining air temperature spatial variability.

3. Method

3.1. Variogram analysis

In this Section, we provide some basic definitions and concepts concerning the variogram analysis (for a detailed discussion see Cressie, 1993).

If $Z(\mathbf{s})$ is a regionalized stationary variable with a constant mean μ and variance σ^2 in a d-dimensional Euclidean space D, the quantity $2\gamma(.)$, which has been called a *variogram* by Matheron (1962), is defined as:

$$2\gamma(\mathbf{s_1} - \mathbf{s_2}) \equiv \mathrm{var}(Z(\mathbf{s_1}) - Z(\mathbf{s_2})) \quad \text{for all } \mathbf{s_1}, \mathbf{s_2} \in D \tag{1}$$

Due to the stationary assumption, this is a function of the increments $\Delta\mathbf{s} = \mathbf{s_1} - \mathbf{s_2}$ only and $\gamma(\Delta\mathbf{s}) \propto \sigma^2$ for large values of $\Delta\mathbf{s}$ asymptotically.

When the mean is assumed to be a constant, this equality holds:

$$\mathrm{var}(Z(\mathbf{s} + \Delta\mathbf{s}) - Z(\mathbf{s})) = E(Z(\mathbf{s} + \Delta\mathbf{s}) - Z(\mathbf{s}))^2 \quad \forall \mathbf{s}, \Delta\mathbf{s} \tag{2}$$

where $E(.)$ indicates the expected value, and we can estimate the variogram as: