Automatic processing of bioclimatic data in the space and time domains

Sabino Maggi\textsuperscript{1}, Delia Bruno\textsuperscript{1}, Aimè Lay-Ekuakille\textsuperscript{2}, Rita Masciale\textsuperscript{1} and Giuseppe Passarella\textsuperscript{1}

\textsuperscript{1} CNR-IRSA Water Research Institute, Via F. De Blasio 5, 70132 Bari, Italy
\textsuperscript{2} Department of Innovation Engineering, Università del Salento, Via Monteroni, 73100 Lecce, Italy

E-mail: sabino.maggi@cnr.it

Abstract. This paper presents a software package for the automatic processing of bioclimatic data in the space and time domains whose final goal is to provide land and water management authorities with reliable information about the moisture/dryness level of a region and its water requirements. The current state of development of package is reported, presenting an example of application of the program to a specific case study.

1. Introduction

The bioclimatic characterization of a region is usually based on some well-established indices determined from observed meteorological and climatic data. Some of these indices synthesize the moisture/dryness level of a region and provide precious information to water management authorities about the irrigation requirements of the area under consideration. Bioclimatic indices are also useful tools for land and water management, since they may contribute to assess the degree of hydrologic stress of a given area [1–5]. Such an assessment is particularly important in Mediterranean regions, where climate scenarios predict a trend towards a warmer climate, with the expansion of the arid and semiarid areas.

Bioclimatic indices can be easily calculated from weather data taken at meteorological stations spread over a given region. Standard interpolation methods implemented in GIS packages allow to extend the point-scale information over the whole area, producing continuous (or almost continuous) maps of the bioclimatic indices. These interpolation methods range from fast and easy to use geometrical methods to more complex and demanding interpolators, with very different requirements on input data and accuracy of the results.

Kriging is a complex and computationally expensive task, in particular when used to interpolate large spatio-temporal data sets, but gives better results than simple interpolation methods (or even advanced methods based on neural networks interpolators) [6], provided that the starting data sets satisfy the reasonable assumption of being spatially auto-correlated random variables [7].

We present here a preliminary version of a software package for the accurate automatic determination of bioclimatic data in the space and time domains by universal kriging, written in the R programming language. We report the current state of development of the package, applying it to the same case study of [8,9].
2. Materials and Methods
In previous works we have used several standard commercial tools to process and interpolate bioclimatic data, such as a spreadsheet to pre-process measured data and calculate the bioclimatic indices, a GIS package and specialized tools to carry out the spatial analysis and to produce publication-quality plots [8, 9]. Processing data in this way may seem simple at first, but it is also cumbersome and prone to error, as well as difficult to track and replicate reliably.

To overcome these problems, we have developed from scratch a software package for the automatic processing of bioclimatic data in the space and time domain, capable of giving accurate and reproducible results. The package is written in the R programming language, a free and highly extensible programming language especially useful for statistical and geostatistical analyses. The R language allows to define a series of complex data-analysis steps and to re-use them on the same or similar data set, making also easier for others to reproduce or validate research results, a requirement that should be considered mandatory in all data analysis work.

An important characteristic of the proposed package is that it can be run either in batch mode, where the program processes a well-defined set of measured data and of program parameters (such as the bioclimatic index, the time interval over which to aggregate the data, the variogram function and the spacing of the grid over which to perform kriging) or in interactive mode, exploring the effect that the change of one or more parameters has on the final results.

To illustrate the methodology used here we have selected the annual De Martonne index [10], one of the most widely used bioclimatic indices among those available in the scientific literature. The De Martonne index has been calculated for the 82 weather monitoring stations of the Civil Protection of Apulia, Italy, where precipitation and temperature data from 1931 to 2010 are available. The monitoring stations are spread more or less uniformly across the area of the region, with a density of about 1 station per 250 km². Some of the weather stations are located outside the administrative boundaries of the region but are included in this monitoring network since they belong to the same hydrographic basin. The measurements of these stations have been used for all calculations, even if the maps produced by the package are confined to the Apulia region (Figure 2).

3. Case study
The first stage of the program reads the files containing the temperature and precipitation data and the shapefile of the region under consideration (Apulia in this case). The input data sets can be provided as either Excel or plain comma-separated values (csv) files, where each row contains the data for a measurement station and each column the monthly average temperature or total precipitation. At the moment the handling of missing data is performed separately, but it will be integrated in a future version of the program. To speed-up calculations, the original shapefile is simplified using the Douglas–Peucker iterative reduction algorithm.

The data sets containing the weather data are validated (e.g., by checking for the presence of monitoring stations with equal spatial coordinates and of duplicated or missing values), then the program performs a basic exploratory data analysis step to find patterns and relationships in the data. The results are presented in the form of standard descriptive statistic tables and as a series of plots showing aggregate or specific temporal trends in the data, a few examples of which are shown in Figure 1.

This stage is followed by the calculation of the desired bioclimatic indices, aggregating the data according to one or more different time scales, e.g. by year, quarter, season or month. The time scale aggregation depends in part on the chosen bioclimatic index, as some indices are defined only on a specific time scale while others can be used at different levels of aggregation.

The degree of spatial autocorrelation of the bioclimatic indices is evaluated by suitable spatial statistical tests and, once verified, it is possible to calculate the empirical variogram and the best variogram model that fits the measured data.
Figure 1. Boxplots showing the range of temperatures measured by the monitoring network as a function of the month of the year (top left) or the elevation of the station (bottom left). The multiple plot on the right shows the temperatures measured in the central month of each season over the whole temporal range.

The last step of the spatial analysis is the interpolation of the point-scale index values over the whole study area by ordinary kriging, where the resulting maps can be used to define the bioclimatic zones of the region.

To produce maps that are effective in a managerial perspective, the program allows to reshape the spatially-continuous interpolated data according to different specifications. An example is shown in Figure 2, where the continuous map calculated by ordinary kriging has been aggregated by averaging the bioclimatic index within the area of each municipality of the region.

Another computational component of the proposed software is the temporal trend analysis module, which is still under heavy development. This component implements two different methods, the Mann-Kendall non parametric test recommended by the World Meteorological Organization (WMO) [11], and the parametric F-test, often used to study the significance of the parameters of regressive models. The two methods usually give very similar results, but this approach can be useful to detect special cases where one of the two methods fails or gives meaningless results.

Trend analysis is performed as follows: the program first verifies that the time-series data are statistically homogeneous and, if the test fails, homogenizes the data set. Then it checks for the presence of serial autocorrelation in the data and eventually uses pre-whitening methods to
Figure 2. Median of the De Martonne index interpolated by ordinary kriging and aggregated by municipality. The blue dots are the locations of the weather monitoring stations.

remove it. Last, it verifies the presence of a trend and calculates the slope of the trend when the result is significant, producing summary tables of information and a series of synthetic plots that summarize the results of the analysis.

4. Conclusions
We have developed a software package that processes bioclimatic data in the space and time domains with the goal to provide water delivery and management authorities with reliable information about the level of dryness of a region and its irrigation requirements. Future work will focus on a more modular implementation of the program and on the simplification of the interface to make it more readily usable.

References
[1] Paltineanu C, Tanasescu N, Chitu E and Mihăilăescu I F 2007 International Agrophysics 21 81–93
[2] Masciale R, Emanuele, Barca and Passarella G 2011 Environmental Monitoring and Assessment 177 245–261
[3] Barca E, Calzolari M C, Passarella G and Ungaro F 2013 Water Resources Management 27 5171–5190 URL https://link.springer.com/article/10.1007/s11269-013-0461-6
[4] Zaccaria D, Passarella G, D’Agostino D, Giordano R and Solis S S 2016 CLEAN Soil, Air, Water 44 371–382 URL http://dx.doi.org/10.1002/clen.201400396
[5] Passarella G, Barca E, Sollitto D, Masciale R and Bruno D E 2017 Water Resources Management 31 327–340 URL http://dx.doi.org/10.1007/s11269-016-1527-z
[6] Attorre F, Alfò M, Sanctis M D, Francesconi F and Bruno F 2007 International Journal of Climatology 127 1825–1843 URL http://onlinelibrary.wiley.com/doi/10.1002/joc.1495/full
[7] Matheron G 1971 The theory of regionalised variables and its applications vol 5 (Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau)
[8] Passarella G, Barca E, Bruno D E, Maggi S and Lay-Ekuakille A 2017 Proceedings of the 7th EnelIMEKO, Symposium on Environmental Instrumentation and Measurements, IMEKO TC19 (IMEKO) pp 87–92
[9] Passarella G, Barca E, Bruno D E, Maggi S and Lay-Ekuakille A 2017 Proceedings of the 7th EnelIMEKO, Symposium on Environmental Instrumentation and Measurements, IMEKO TC19 (IMEKO) pp 109–113
[10] de Martonne E 1925-1927 Traité de Géographie Physique (Librarie Arman Colin, Paris)
[11] Mann H B 1945 Econometrica 13 245–259 URL http://www.jstor.org/stable/1907187