A Package for the Homogenisation of Climate Data Using Geostatistical Simulation

Júlio Caineta, Sara Ribeiro, Roberto Henriques and Ana Cristina Costa
NOVA IMS
Universidade Nova de Lisboa
Lisboa, Portugal
Email: {jcaineta, sribeiro, roberto, ccosta}@novaims.unl.pt

***Abstract*—Climate data homogenisation is of major importance in monitoring climate change, the validation of weather forecasting, general circulation and regional atmospheric models, modelling of erosion, drought monitoring, among other studies of hydrological and environmental impacts. The gsimcli package is a work in progress project based on a geostatistical homogenisation method, aiming to make its application easier and more straightforward. It is expected that this contribution will help technicians, researchers and other professionals, to detect and to correct irregularities in climate data.**

***Index Terms*—Climate data; homogenisation; geostatistics; software.**

I. INTRODUCTION

Homogenisation of climate data is a very relevant subject since these data are required as an input in a wide range of studies, such as atmospheric modelling, weather forecasting, climate change monitoring, or hydrological and environmental projects. Often, climate data series include non-natural irregularities which have to be detected and removed prior to their use, otherwise it would generate biased and erroneous results.

In the last two decades, many methods have been developed to identify and remove these inhomogeneities [1][2][3][4]. One of those is based on a geostatistical simulation technique, direct sequential simulation (DSS), proposed by Soares [5], in which local probability density functions (PDFs) are calculated at candidate monitoring stations, using spatial and temporal neighbouring observations, and then are used for detection of inhomogeneities [6]. This approach has been previously applied to detect inhomogeneities in four precipitation series (wet day count) from a network with 66 monitoring stations located in the southern region of Portugal [7]. That study revealed promising results and the potential advantages of geostatistical techniques for inhomogeneities detection in climate time series.

This package is a product of a project (GSIMCLI – Geostatistical simulation with local distributions for the homogenisation and interpolation of climate data (PTDC/GEO-MET/4026/2012)) that aims to improve and develop that geostatistical homogenisation method, deploying its algorithms into a computer program.

The first studies of the method required a lot of time and interaction from its users. That happened mainly for two reasons: i) climate data may come from different sources and in different formats that have to be parsed, and ii) the method makes use of an already existent program, which has its own input and output formats. Handling different file formats among the algorithm's steps may require several transformations, back and forth.

This short paper introduces the method itself, describes the software development and its structure, illustrates an example of its usage, and finally lists some of the features and improvements that are expected in the near future.

II. BRIEF THEORETICAL REFERENCE

A brief theoretical introduction to the related topics will be presented in this section.

A. Climate data homogenisation

The homogenisation of long meteorological time series is of extraordinary interest to the scientific community. The precise quantification of the variability of observed meteorological parameters is essential for many purposes. However, long instrumental records are rarely homogeneous because their values are dictated not only by change in climate but also by non-climatic factors. Irregularities such as relocation of weather stations or changes in measuring instruments, introduces discontinuities in time series, which may lead to data not being representative of real climate change. That may therefore skew the studies' results [1][6].

B. Geostatistical simulation

In geostatistics, it is common to refer to simulation as a stochastic process, opposed to estimation which is regarded as a deterministic process. Besides correlating different samples of a given variable, geostatistics adds their spatial structure to the equation. Therefore, geostatistical simulation is used to reproduce the spatial distribution and uncertainty of variables of different resources in Earth sciences.

One of the geostatistical simulation methods that has been widely used in different contexts (e.g., oil and gas resources, air and water pollutants) is the DSS. One of its main advantages is not requiring the transformation of the original variable, while honouring both the variable's covariance model and histogram.

ISBN: 978-1-61208-383-4

C. Geostatistical simulation for the homogenisation of climate data

A geostatistical approach, using DSS, was proposed by Costa et al. [7] for inhomogeneities detection and correction. The DSS algorithm is used to calculate the local PDF at a candidate station's location, using spatial and temporal observations, only from nearby reference stations, without taking into account the candidate's data. Afterwards, the local PDF from each instant in time (e.g., year) is used to verify the existence of irregularities. A breakpoint is identified whenever the interval of a specified probability p (e.g., 0.95), centred in the local PDF, does not contain the observed (real) value of the candidate station [6]. In Figure 1, the orange areas illustrate the identification of a breakpoint, i.e., the values lying in the orange areas will be classified as inhomogeneous. In practice, the local PDFs are provided by the histograms of simulated maps. If irregularities are detected in a candidate series, the time series can be adjusted by replacing the inhomogeneous records with the mean, or median, of the PDF(s) calculated at the candidate station's location for the inhomogeneous period(s).

This technique accounts for the joint spatial and temporal correlation between observations, and gives greater weight to the nearest stations, both in spatial and correlation terms.

The final goal of the GSIMCLI project is to deliver a complete tool for the homogenisation of climate data, after investigating a new method based on DSS with local distributions [9]. It should result in a procedure that is appropriate for those situations in which the monitoring stations are located in extensive areas with different climatic characteristics, and it should be extensible to situations in which the PDF of the candidate station is different from its neighbours' PDF, which occurs, for example, due to local trends induced by local physiographic features.

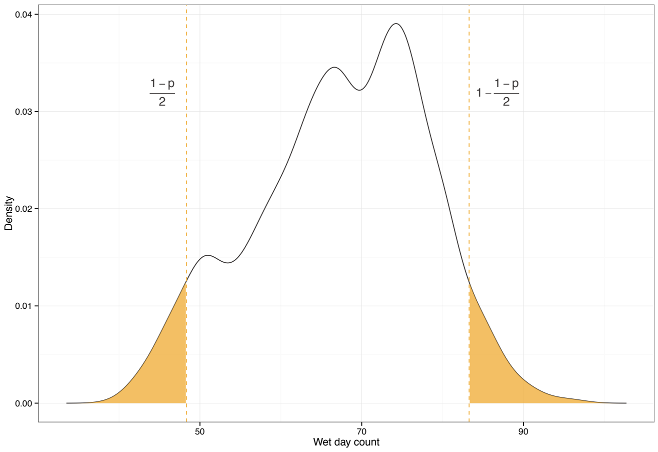


Fig. 1. Example of a probability density function originated from the simulation of the variable wet day count [8].

III. THE GSIMCLI PROGRAM

With the attention drawn into homogenisation methods by the HOME project (COST Action ES0601) [10] and the promising results attained in previous studies [6], the geostatistical approach became an interesting research subject, leading to this computer program.

The homogenisation method referred in Section II-C has been deployed into a computer program, which simplifies its application and assessment. In this section, we will elaborate on the program's details.

A. Software development

The gsimcli computer application (available at http://iled.github.io/gsimcli) is being developed under the object oriented paradigm, with the programming language Python (see http://www.python.org), a cross-platform, open source and general purpose language widely used in the scientific community. Its development is divided in a set of packages described below. The Python language is a good tool for prototyping (it is fast to code and easy to read) and for tasks that are not computationally demanding, still, it has a wide range of libraries that enhance its performance and usability.

The DSS algorithm is currently implemented as a black box element in the algorithm – the binary is launched as a standalone application – and its code is not presently part of this project.

B. Packages

This computer application can be seen as a collection of scripts, which have been divided into four packages. Each package encloses a different number of modules, then each module contains a set of related functions.

1) Parsers: One of the main purposes of the proposed program is minimising the required pre-processing and transformations needed during the homogenisation process, from the user point of view, i.e., the program should manage and transform all the data files automatically. Thus, a set of modules were developed to handle different file types, including reading, writing and the conversion between them:

costhome	files in the format established for the benchmark dataset developed in the HOME project for the comparison of homogenisation algorithms.
dss	DSS parameters file.
gsimcli	parameters file of the gsimcli application.
shapefile	files generated in a Geographical Information System (GIS) environment.
spreadsheet	typical spreadsheet file (e.g., comma separated values).

2) Tools: The definition and handling of objects and the numerical calculations are mainly developed in these modules. There are algorithms implemented to: deal with objects that follow the specification given in the widely used geostatistical library GSLIB [11]; detect and identify breakpoints in time

ISBN: 978-1-61208-383-4

series (the main goal of this program); control parameters; and also to calculate performance metrics for the homogenisation process, i.e., centred root-mean-square error (CRMSE).

3) Launchers: This package operates the main processes execution: DSS and the homogenisation process as a whole.

The DSS related module takes advantage of the fact that different realisations are independent between them, to launch multiple simulations at the same time, in order to achieve a form of parallel execution. In this way, it is possible to use multiple cores, thus reducing the overall processing time.

The module that controls the homogenisation process includes options to make it run in time series separated in decades, and also in time series belonging to different networks (these functions are grouped with the term “batch”). With such functions, all the file handling is automated and user’s perceive the homogenisation as a single process, instead of a repetitive and time-consuming sequence to process different input and output data.

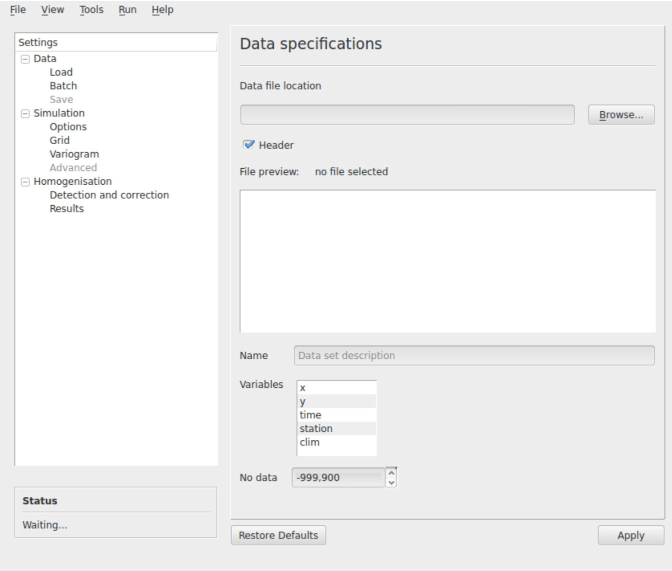
4) Interface: The program offers a graphical user interface (GUI) that was designed to be easy and intuitive to use, having a lot of common structures seen in other programs.

Basically, and at the current point of development, it is a settings pane with three groups of parameters: data, simulation (DSS) and homogenisation. They are organised in the order that the overall process will run and that should match a natural workflow (Figure 2).

5) Documentation: The program development also accounted for its documentation. There is an application programming interface (API) and a user manual for the GUI, both available at http://gsimcli.readthedocs.org.

C. Usage

As stated before, the program usage should be direct and effortless. A set of default and recommended settings is provided, which will help users to start using this application.



Data specifications

Data file location

☑ Header

File preview: no file selected

Name: Data set description

Variables: x, y, time, station, clim

No data: -999.900

Status: Waiting...

Fig. 2. Graphical user interface overview.

If using the default values, a common usage should go as follows:

1) Browse the data file location or, if processing multiple networks and/or decades, browse the data folder location.
2) Confirm the variables order and the place holder number for missing data.
3) Browse the DSS binary file.
4) Set up the simulation grid (if processing multiple networks, the grid details must be included in a spreadsheet file).
5) Provide the theoretical variogram model (if processing multiple decades, the variogram must be included in a spreadsheet file).
6) The settings for the detection and correction if inhomogeneities are also given by default, but it should be simple to try different values.
7) Browse the results file and directory.
8) Run gsimcli.

After that, the program will show the progress of the process. The necessary computational time highly depends on the computer specifications (e.g., frequency and number of cores), but also depends on the data set (number of candidate stations) and on the simulation parameters (e.g., grid size, maximum number of nodes to search for).

The final outcome is a spreadsheet file containing the complete homogenised data set, it will show the previous values of the homogenised samples, as well as a summary table indicating the number of detections and missing data per candidate station.

D. Enhancing the case study

When the geostatistical approach for the homogenisation of climate data was first studied, only 4 candidate stations out of 10 were selected (those 10 had been previously classified as inhomogeneous among a total of 66 stations) [7]. The process was slow and laborious, it required a considerable amount of user interaction and, for that reason, it was not practical to assess a desirable number of candidate stations. Nonetheless, that study revealed promising results and the potential advantages of geostatistical techniques for inhomogeneities detection in climate time series.

The implementation of all the required steps into computer scripts made it feasible to extend the same case study, with the number of candidate stations being increased to 10 (the referred inhomogeneous stations) [8]. Also, it is now practicable to analyse the method sensitivity to any parameter.

E. Performance assessment

The algorithms overall performance can be assessed in two aspects: the computational performance while running the algorithm; and the results of the homogenisation process. Both these perspectives were investigated and considered during the software development.

The computational performance has been an important factor in the design and implementation of the algorithms, both in processing time and required system memory. For example, as already mentioned, running multiple instances of DSS is a way to increase computational efficiency, while the division of time series in decades should help to reduce memory consumption.

In terms of quality of homogenisation, the performance has been measured in two ways: homogenising the southern Portugal case study and comparing with the results obtained in a previous study [7]; and measuring the obtained CRMSE in the homogenisation of some of the benchmark network of the HOME project, against other results presented by the same project [10]. In the first case, the results were promising and consistent with what had been obtained in the mentioned study [8]. In the latter, the results were considerably worse than those obtained by other homogenisation algorithms, given the small size of the monitoring networks and their low spatial dependence, thus highlighting the importance of the implementation of the DSS with local distributions [12].

IV. Conclusion and future work

The resulting computational application reproduces an existing and tested homogenisation method based on a geostatistical simulation technique, while having the advantage of the entire process being conducted in a seamless and practical manner, requiring less user interaction. This is highly beneficial to researchers: it is easier to investigate the influence of any parameter, and it allows the addition of new methods in a given step of the overall process, enhancing the research and development of new techniques and knowledge.

In the near future, it is planned to extend the GUI to include more options to the user, as well as to provide a graphical interface for other operations related to the homogenisation process (e.g., variography, conversion between file types, calculation of performance metrics).

Acknowledgement

The authors gratefully acknowledge the financial support of ”Fundação para a Ciência e Tecnologia” (FCT), Portugal, through the research project PTDC/GEO-MET/4026/2012 (”GSIMCLI - Geostatistical simulation with local distributions for the homogenisation and interpolation of climate data”).

References

[1] H. Caussinus and O. Mestre, “Detection and correction of artificial shifts in climate series,” *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 53, no. 3, pp. 405–425, Aug. 2004.

[2] A. T. DeGaetano, “Attributes of several methods for detecting discontinuities in mean temperature series,” *Journal of Climate*, vol. 19, no. 5, pp. 838–853, 2006.

[3] P. Domonkos, “Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods,” *Theoretical and Applied Climatology*, vol. 105, no. 3, pp. 455–467, 2011.

[4] T. Szentimrey, “Multiple analysis of series for homogenization (MASH),” in *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data*, ser. WMO-TD No. 962, WCDMP No. 41, Budapest, Hungary, 1999, pp. 27–46.

[5] A. Soares, “Direct sequential simulation and cosimulation,” *Mathematical Geology*, vol. 33, no. 8, pp. 911–926, 2001.

[6] A. C. Costa and A. Soares, “Homogenization of climate data: review and new perspectives using geostatistics,” *Mathematical geosciences*, vol. 41, no. 3, pp. 291–305, Nov. 2009.

[7] A. C. Costa, J. Negreiros, and A. Soares, “Identification of inhomogeneities in precipitation time series using stochastic simulation,” in *geoENV VI - Geostatistics for Environmental Applications*, A. Soares, M. J. Pereira, and R. Dimitrakopoulos, Eds. Netherlands: Springer, 2008, pp. 275–282.

[8] J. Caineta, S. Ribeiro, A. C. Costa, R. Henriques, and A. Soares, “Inhomogeneities detection in annual precipitation time series in Portugal using direct sequential simulation,” in *Geophysical Research Abstracts*, ser. EGU General Assembly Conference Abstracts, vol. 16, no. EGU2014-7849, Vienna, Austria, 27 Apr.–2 May 2014, p. 7849.

[9] A. Horta and A. Soares, “Direct Sequential Co-simulation with Joint Probability Distributions,” *Mathematical Geosciences*, vol. 42, no. 3, pp. 269–292, Feb. 2010.

[10] V. K. C. Venema, O. Mestre, E. Aguilar, I. Auer, J. a. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, J. Viarre, G. Müller-Westermeier, M. Lakatos, C. N. Williams, M. J. Menne, R. Lindau, D. Rasol, E. Rustemeier, K. Kolokythas, T. Marinova, L. Andresen, F. Acquaotta, S. Fratianni, S. Cheval, M. Klancar, M. Brunetti, C. Gruber, M. Prohom Duran, T. Likso, P. Esteban, and T. Brandsma, “Benchmarking homogenization algorithms for monthly data,” *Climate of the Past*, vol. 8, no. 1, pp. 89–115, Jan. 2012.

[11] C. V. Deutsch and A. G. Journel, *GSLIB: Geostatistical Software Library*, ser. Applied Geostatistics Series. Oxford University Press, 1998.

[12] J. Caineta, S. Ribeiro, R. Henriques, A. Soares, and A. C. Costa, “Benchmarking a geostatistical procedure for the homogenisation of annual precipitation series,” in *Geophysical Research Abstracts*, ser. EGU General Assembly Conference Abstracts, vol. 16, no. EGU2014-7605, Vienna, Austria, 27 Apr.–2 May 2014, p. 7605.