Analysing the detection and correction parameters in the homogenisation of climate data series using gsimcli

Sara Ribeiro
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide, Lisboa, Portugal
sribeiro@novaims.unl.pt

Júlio Caineta
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide, Lisboa, Portugal
jcaineta@novaims.unl.pt

Ana Cristina Costa
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide, Lisboa, Portugal
ccosta@novaims.unl.pt

Roberto Henriques
NOVA IMS, Universidade Nova de Lisboa
Campus de Campolide, Lisboa, Portugal
roberto@novaims.unl.pt

Abstract

Homogenisation of climate data series is the process of detection and correction of non-natural irregularities present in the data. Such process is extremely important due to the use of climate data in many hydrological and environmental projects. Several homogenisation methods have been developed in the last decades. In the geostatistical field, studies already showed an approach based on the direct sequential simulation algorithm as a very promising technique for the detection and correction of irregularities. This approach, called gsimcli, uses the probability distribution function (estimated from simulated values) to identify the presence of irregularities, with a specific probability p. The correction of the identified irregularity can be done through the replacement of that value by a given percentile value of the probability distribution function. The present work depicts an analysis undertaken in order to assess two parameters, the probability p of detection and the percentile for correction in the homogenisation using gsimcli. Two networks of the HOME benchmark data set were used and the performance metrics were calculated to compare this analysis with other homogenisation methods. Results show gsimcli as a favourable homogenisation method for monthly precipitation data, and reveal the most efficient detection and correction parameters for the homogenisation procedure.

Keywords: Irregularities; precipitation; performance metrics; sequential simulation.

1 Introduction

It is generally recognised that only by using homogenised data series the long-term climatic trends can be accurately detected [11]. Consequently, the homogenisation of climate data series has gained particular importance in the last decades and led to the development of various methods to detect and correct non-natural irregularities [1, 7]. Homogenisation methods typically depend on the type of climate variable (e.g., temperature or precipitation), the temporal resolution of the observations (annual, monthly or sub-monthly), the availability of metadata (station history information), and also the weather station network density or spatial resolution [5].

Those non-natural irregularities are due to sudden or gradual artificial changes in the environment or in the process of measurement of the climate variable [6]. Examples of the former are station relocations, repositioning at different heights, and changes in the instrumentation. The latter may be exemplified by the urban development slowly growing around a weather station, contributing to the phenomenon known as urban heat island effect [8]. A high number of non-natural irregularities are also introduced during the process of collecting, digitising, processing, transferring, storing and transmitting climate data series [3].

Many authors carried out comparison tests among homogenisation methods [2, 6, 8] in order to assess their efficiency.

The COST Action ES0601 "HOME" (Advances in Homogenisation Methods of Climate Series: an Integrated Approach, 2008-2011) prepared a benchmark data set, composed of temperature and precipitation data from networks of weather stations located in Europe, where irregularities were inserted. Knowing the correct location of irregularities allows the measurement of the efficiency of the homogenisation methods, through the comparison between the homogenised and the original data series [10]. The benchmark is composed of network data sets of surrogated, synthetic and real data. Fifteen networks of surrogated data were generated for temperature and precipitation, using original climate data of real weather stations in Europe. For precipitation, those weather stations are located in France and Austria. Each data set comprises 5, 9 or 15 stations. The efficiency is measured through the calculation of some performance metrics.

Participants of the COST Action "HOME" provided homogenised contributions, whereas performance metrics were calculated and referred by Venema et al. [10]. The described values of the metrics can be used also as a comparison indicator to evaluate the efficiency of other homogenisation methods.

A methodology based on a geostatistical simulation technique was proposed by Costa et al. [4] and was implemented in a software package named gsimcli. This work analyses the influence of parameter values for irregularities detection and correction in order to improve the process of homogenisation using gsimcli.

Section 2 presents the study area and data, Section 3 depicts the methodological framework. Section 4 describes the results

achieved. Finally, conclusions and future work are presented in Section 5.

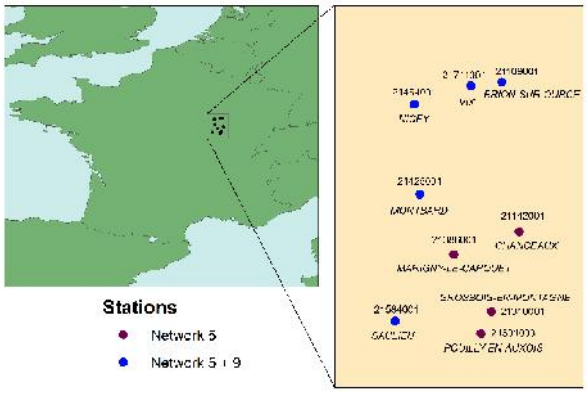
2 Study area and data

This analysis uses the surrogated precipitation data prepared by COST Action "HOME" [10], namely the networks 5 and 9, located in France. Those two networks are composed of 9 and 5 weather stations, respectively, containing monthly values for a period of 100 years (1900 – 1999).

Both data sets also contain missing data in some of the stations. The presence of missing data intends to mimic the absence of weather stations in the beginning of the 20th century (between 1900 and 1924), since the intensification of the weather network was only consolidated later, and also the destruction of some of the existing weather stations during the period of the Second World War (between 1941 and 1945). Stations composing network 9 are all part of network 5 (Figure 1). However, the time series data sets from the two networks are different.

Networks 5 and 9 cover a rectangular area of approximately 4000 km^2 (50 km x 80 km). For the purpose of simulation, a regular grid was defined, comprising 9882 cells (81 x 122 cells), each with an area of 1 km^2.

Figure 1: Location of networks 5 and 9.



3 Methodological framework

Costa et al. [4] proposed a new homogenisation methodology based on direct sequential simulation (DSS) [9]. That methodology was improved and turned into a software package, gsimcli, aiming to make its application easier and more straightforward.

3.1 DSS algorithm

The DSS algorithm is used to calculate the local probability density function (PDF) at a candidate station's location (station to be homogenised), using spatial and temporal observations, only from nearby reference stations (neighbour 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, centred in the local PDF, does not contain the observed (real) value of the candidate station [5]. If irregularities are detected in a candidate series, the time series are corrected by replacing the inhomogeneous records with the mean, median or a given percentile of the PDF(s) calculated at the candidate station's location for the inhomogeneous period(s).

3.2 Gsimcli software

Gsimcli software allows to perform homogenisation tests in a seamless and practical manner. The user adds the inhomogeneous data and the parameters of the semivariogram model and sets up the different parameters for the homogenisation, namely the candidates order, the probability of detection of irregularities and the correction method. The probability of detection is the given value to identify a breakpoint, as referred in Section 3.1. The correction of irregularities can be done through the replacement of those values by the value of the mean, median or a specified percentile provided by the PDF. Missing data values are also completed by the same value indicated as the correction method. In this work, percentile values of the PDF were used as the correction method. Gsimcli software also calculates the performance metrics according to Venema et al. [10].

3.3 Variography

Prior to the simulation by gsimcli, the variography of the precipitation data was studied, dividing the precipitation time series into 10 decades by month for network 5 (9 stations). Due to the variability of precipitation monthly data and the short number of available stations, this revealed to be a challenging task. The correlation between stations' data is lost at very short distances. A way to overcome this drawback is the use of additional data provided by other weather stations located in the surrounding study area. Such task could not be performed in this case, since only the provided data sets by "HOME" can be used in the process.

Due to missing data, a unique semivariogram model was prepared for the first, second and third decades (1900 – 1929). For the same reason, the fourth and fifth decades' data were also joint in order to set another single semivariogram model. Seven semivariogram models were prepared for each monthly series, in a total of 84. The estimated semivariogram models estimated for network 5 were also used in network 9, since the reduced number of stations in this network did not allow a representative variography.

3.4 Performed tests

The tests are performed with 500 simulations, considering 10 data sets (one data set per decade) for each month. For each test, specific values of probability of detection (pdet) and percentile of correction (pcor) are established (Table 1). It was previously decided to test only the percentile option of the correction method, since it provided better results in previous analysis with the same networks using annual time series. Those previous analyses compared the homogenisation performance metrics when correcting inhomogeneities with the mean, median and percentile, where the latter led to a decrease in the performance metrics of 32% and 50%, for Station and Network, respectively.

Two pdet values are tested: 0.95 and 0.975. The percentile of correction is set as 0.95, 0.975 and 0.99. Five combinations of pdet and pcor are tested.

Table 1: Parameters of the performed tests: probability of detection (pdet) and percentile of correction (pcor)

Tests	pdet	pcor
Test 1	0.95	0.95
Test 2	0.95	0.975
Test 3	0.95	0.99
Test 4	0.975	0.95
Test 5	0.975	0.975

3.5 Performance metrics

Following the homogenisation of the networks 5 and 9 using the implementation of each of the five tests, performance metrics are calculated. Those metrics are the Station CRMSE (Centred Root Mean Square Error), the Network CRMSE the Station Improvement and Network Improvement, as defined by Venema et al. [10]. The Station CRMSE quantifies the homogenisation efficiency for each station individually and it is obtained by the mean CRMSE, by station. The Network CRMSE measures the efficiency of the homogenisation of the network, as a whole. It is calculated using the mean CRMSE, by network. The Improvement metrics assess the enhancement over the inhomogeneous data and is computed as the ratio of the Station (Network) CRMSE of the homogenised networks and the Station (Network) CRMSE of the same inhomogeneous networks.

4 Results and discussion

Results are compared with the contribution MASH Marinova submitted to "HOME", since this contribution homogenised the same networks 5 and 9, for surrogated precipitation data (Table 2).

Table 2: Performance metrics (StaC – Station CRMSE; StaImp – Station Improvement; NetC – Network CRMSE; NetImp – Network Improvement; MASH – MASH Marinova contribution [10])

Tests	StaC	StaImp	NetC	NetImp
Test 1	10.450	1.026	4.595	1.238
Test 2	10.375	1.023	4.196	1.130
Test 3	10.812	1.066	4.111	1.107
Test 4	10.467	1.032	4.535	1.222
Test 5	10.201	1.006	4.187	1.128
MASH	8.5	0.84	3.8	1.03

In Table 2, tests with the lowest values of performance metrics correspond to the tests with the best set of parameters (pdet and pcor). Test 5 and Test 2 present the lowest Station CRMSE and Station Improvement metrics from the five performed tests. Those two tests were executed with pcor as 0.975. Also for this metric, Test 3 presents the highest value. Regarding the Network CRMSE, Test 3, pcor as 0.99, provides the lowest value. However, Test 2 and Test 5 also portray low values for the Network CRMSE. Both values of Station and Network Improvement metrics follow the same trend as the respective CRMSE metrics. Comparing the results with the MASH Marinova contribution, Test 5 provides the most approximated value of Station CRMSE, while Test 3 presents the most similar value of Network CRMSE. On the contrary, the values of performance metrics for Tests 1 and 4 are the highest. Both tests were performed with the pcor of 0.95. Tests 2 and 5 were executed with the same value for the correction method, which may indicate that the correction part is more important for the homogenisation method, rather than the detection. From the five tests performed, it can be concluded that Test 5 has the best performance metrics.

5 Conclusion

Five tests were performed using the gsimcli software package in the homogenisation of monthly precipitation data, belonging to two networks of surrogated data located in France. Those five tests were performed with different values of the probability of detection and percentile of correction parameters. The sensitivity analysis showed a high influence of the correction method in the efficiency of the homogenisation, which achieves the best result with a pcor of 0.975. The detection method does not seem to be crucial in the efficiency of the homogenisation. However, the best results, corresponding to the best performance metrics, are obtained with a pdet of 0.975. The results also confirm gsimcli as an encouraging homogenisation method for precipitation monthly data.

As future work, gsimcli will include a new procedure that is appropriate for those situations in which the monitoring stations are located in extensive areas with different climatic characteristics. HOME data set (surrogated precipitation data) will again be used in order to assess this new procedure.

Acknowledgements

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] E. Aguilar, I. Auer, M. Brunet, T. C. Peterson and J. Wieringa. Guidelines on climate metadata and homogenization. In Paul Llansó, editor. *World Meteorological Organization*, (WMO/TD No. 1186), 2003.

[2] C. Beaulieu, O. Seidou, T. B. M. J. Ouarda, X. Zhang, G. Boulet, and A. Yagouti. Intercomparison of homogenization techniques for precipitation data, *Water Resources Research.*, 44, W02425, 2008. doi: 10.1029/2006WR005615

[3] M. Brunet and P. Jones. Data rescue initiatives: bringing historical climate data into the 21st century. *Climate Research*, *47*(1), 29–40, 2011. doi: 10.3354/cr00960

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

[5] A. C. Costa and A. Soares. Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Mathematical Geosciences*, *41*(3), 291–305, 2009. doi: 10.1007/s11004-008-9203-3

[6] P. Domonkos. Measuring performances of homogenization methods. *IDŐJÁRÁS, Quarterly Journal of the Hungarian Meteorological Service*, 117(1), 91–112, 2013.

[7] S. Ribeiro, J. Caineta, R. Henriques, A. Soares, A. C. Costa. Advantages and applicability of commonly used homogenisation methods for climate data, In: *Geophysical Research Abstracts*, Vol. 16, EGU2014-7725, European Geosciences Union General Assembly 2014, 2014.

[8] S. Sahin and H. K. Cigizoglu. Homogeneity analysis of Turkish meteorological data set. *Hydrological Processes*, *24*(8), 981–992, 2010. doi: 10.1002/hyp.7534

[9] A. Soares. Direct Sequential Simulation and Cosimulation. *Mathematical Geology*, *33*(8), 911–926, 2001.

[10] V. Venema, O. Mestre, E. Aguilar, I. Auer, J. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Štěpánek, P. Zahradníček, J. Viarre, G. Muller-Westermeier, M. Lakatos, C. Williams, M. 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 Past*, 8(1), 89–115, 2012.

[11] L. Zhang, G.-Y.Ren, Y.-Y. Ren, A.-Y. Zhang, Z.-Y. Chu and Y.-Q. Zhou. Effect of data homogenization on estimate of temperature trend: a case of Huairou station in Beijing Municipality. *Theoretical and Applied Climatology*, *115*(3-4), 365–373, 2014. doi: 10.1007/s00704-013-0894-0