Assessing the Performance of the Gsimcli Homogenisation Method with Precipitation Monthly Data from the COST-HOME Benchmark

S. Ribeiro[1] , J. Caineta[2] , A. C. Costa[3]

[1]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, sribeiro@novaims.unl.pt

[1]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, jcaineta@novaims.unl.pt

[1]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, ccosta@novaims.unl.pt

This is the Author Peer Reviewed version of the following chapter/conference contribution published by Springer:

Ribeiro, S., Caineta, J., & Costa, A. C. (2017). Assessing the Performance of the Gsimcli Homogenisation Method with Precipitation Monthly Data from the COST-HOME Benchmark. In J. J. Gómez-Hernández, J. Rodrigo-Ilarri, E. Cassiraga, M. E. Rodrigo-Clavero, & J. A. Vargas-Guzmán (Eds.), Geostatistics Valencia 2016 (pp. 909-918). (Quantitative Geology and Geostatistics; Vol. 16). Springer. https://doi.org/10.1007/978-3-319-46819-8_63

Assessing the performance of the gsimcli homogenisation method with precipitation monthly data from the COST-HOME benchmark

S. Ribeiro[1], J. Caineta[2], A. C. Costa[3]

Abstract Nowadays, climate data series are used in so many different studies that their importance implies the essential need of good data quality. For this reason, the process of homogenisation became a hot topic in the last decades and many researchers have focused on developing efficient methods for the detection and correction of inhomogeneities in climate data series. This study evaluates the efficiency of the gsimcli homogenisation method, which is based on a geostatistical simulation approach. For each instant in time, gsimcli uses the Direct Sequential Simulation algorithm to generate several equally probable realisations of the climate variable at the candidate station's location, disregarding its values. The probability density function estimated at the candidate station's location (local pdf), for each instant in time, is then used to verify the existence of inhomogeneities in the candidate time series. When an inhomogeneity is detected, that value is replaced by a statistical value (correction parameter) derived from the estimated local pdf. In order to assess the gsimcli efficiency with different implementation strategies, we homogenised monthly precipitation data from an Austrian network of the COST-HOME benchmark data set (COST Action ES0601: Advances in Homogenization Methods of Climate Series: an integrated approach – HOME). The following parameters were tested: grid cell size, candidates order in the homogenisation process, local radius parameter, detection parameter, and correction parameter. Performance metrics were computed to assess the efficiency of gsimcli. The results show the high influence of the grid cell size and of the correction parameter in the method's performance.

[1]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, sribeiro@novaims.unl.pt

[2]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, jcaineta@novaims.unl.pt

[3]NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, ccosta@novaims.unl.pt

Introduction

As defined by the Intergovernmental Panel on Climate Change (IPCC), climate change refers to a change in the state of the climate that can be identified by changes in the statistical characteristics of its properties, and that persists for an extended period, typically decades or longer (Bernstein, et al., 2007). In order to assess climate change and to develop impact studies, it is imperative that climate signals are clean from any external factors. Hence, two steps must be performed in climate time series: quality control and homogenisation. Quality control relates to the verification and treatment of extremely high and low values (outliers). The second includes an analysis of the time series that is focused on the detection and correction of inhomogeneities caused by non-climatic factors (Bližňák, Valente, & Bethke, 2014; Vertacnik, et al., 2015).

The non-climatic factors include stations' relocations, changes in the environment, instrumentation, time, and in the methods of measurement (Aguilar, Auer, Brunet, Peterson, & Wieringa, 2003). Since these artificial discontinuities have often the same magnitude as the usual variability of climate data series, they can erroneously influence the analysis of natural climate variations (Hannart, Mestre, & Naveau, 2014).

Homogenisation methods usually depend on the type of climate variable, the temporal resolution of the observations, the weather station network density, and also the availability of metadata (Costa & Soares, 2009). Indeed, metadata plays a very important role in the homogenisation of climate data and should be documented and treated with the same care as the data themselves (World Meteorological Organization, 2010). Direct homogenisation methods employ metadata in order to assess the presence of a breakpoint; however, whenever metadata is absent, the indirect homogenisation methods justify the presence of a breakpoint only with the result of homogenisation tests (Ribeiro, Caineta, Costa, Henriques, & Soares, 2016). The relative homogeneity principle (Conrad & Pollack, 1962) assumes that neighbouring series reveal the same climate variations apart from the inhomogeneities integrated in one of the series (Hannart, Mestre, & Naveau, 2014). Based on this principle, homogenisation methods can be classified in regard to the use of reference stations: absolute and relative homogenisation methods. The use of absolute methods (without using reference stations) should be used with care, and always accompanied with metadata (Venema, et al., 2012), since they may introduce more inhomogeneities in the data series.

Several authors have prepared reviews of homogenisation methods (Aguilar, Auer, Brunet, Peterson, & Wieringa, 2003; Costa & Soares, 2009; Domonkos, 2013; Ribeiro, Caineta, & Costa, 2015). The European initiative COST Action ES0601, Advances in homogenisation methods of climate series: an integrated approach (HOME) intended to review and improve common homogenisation methods, and to assess their impact in climate time series (Chair of the Management Committee of the Action, 2011). In order to achieve such goals, HOME has executed a blind intercomparison and validation study for homogenisation methods.

The methods were tested against a realistic benchmark data set, which included temperature and precipitation data (Venema, et al., 2012). This benchmark data set has three different groups of data: real, surrogate and synthetic. The first group contains real inhomogeneous data, while the other two enclose simulated data with inserted inhomogeneities, outliers, missing data periods, local station trends and a global trend, per network (Venema, et al., 2012). Fifteen simulated networks were prepared, and they are located in different places within Europe. The networks comprise 5, 9 and 15 stations.

The submitted methods were evaluated by the calculation of performance metrics. Based on the performance metrics, the best homogenisation contributions were ACMANT (Domonkos, Poza, & Efthymiadis, 2011), MASH (Szentimrey, 1999, 2007, 2008), PRODIGE (Caussinus & Mestre, 1996, 2004) and USHCN (Menne & Williams Jr., 2009; Menne, Williams Jr., & Vose, 2009). Recently, some of the homogenisation methods were transformed into software packages, in order to become fully automatic procedures, and they are available in (http://www.climatol.eu/tt-hom/).

This study assesses the efficiency of the gsimcli homogenisation method, which is based on a geostatistical simulation approach. To do so, we homogenised monthly precipitation data from an Austrian network of the HOME benchmark data set. The following parameters were tested: grid cell size, candidates order in the homogenisation process, local radius parameter, detection parameter, and correction parameter. Performance metrics were computed to assess the efficiency of gsimcli. Precipitation is the focus on this study, since it is one of the most important variables for climate and hydro-meteorology studies. Changes in precipitation pattern may lead to floods, droughts, and consequentially to the loss of biodiversity and agricultural productivity (Sayemuzzaman & Jha, 2014).

This work is organised as follows. The following section depicts the study domain and data. The methodological framework includes the description of the gsimcli method and the set of performed homogeneity tests. In the results and discussion section, the performance metrics are scrutinised. Finally, the conclusion section brings a summary of the lessons learned and recommendations for future work.

Study domain and data

This study analyses monthly surrogate precipitation data that are part of the HOME data set, namely the network 16. This network comprises 15 stations and it is located in Austria (Figure 1). The data series include 100 years of precipitation values, between 1900 and 1999. It covers a rectangular area of approximately 24640 km^2 (220 km x 112 km). Considering the statistics for the annual series (Table 1), the lowest values were recorded in stations 4313302 (northeast corner) and 4315421 (west area), with 374.3 mm and 384.6 mm, respectively. The maximum value was recorded in station 4319710 (southwest corner). The variability of

both annual and monthly series is very high. For example, the standard deviation of the annual series varies between 131.4 mm (station 4315515) and 287.7 mm (station 4320123).

Table 1 Summary statistics of the annual precipitation series from network 16

Station ID	Mean	Median	Std. dev.	Variance	Range	Min.	Max.
4313116	868.9	812.2	211.2	44582	974.7	415.5	1390
4313302	769.5	759.9	131.7	17343	780.8	374.3	1155
4315343	903.0	885.2	152.4	23217	834.1	563.0	1397
4315421	786.0	795.3	194.4	37777	749.6	384.6	1134
4315515	1051.4	1051.5	131.4	17275	639.8	726.4	1366
4315711	793.5	773.3	136.1	18521	733.5	465.8	1199
4316412	824.3	825.5	188.8	35626	994.9	422.7	1418
4317901	1010.5	1015.0	206.8	42770	942.5	597.9	1540
4318210	916.5	908.8	178.4	31825	866.6	470.1	1337
4318906	1150.4	1158.1	210.39	44265	1132.1	550.5	1683
4319710	1290.6	1237.4	279.8	78299	1498.6	618.2	2117
4320001	1247	1270.5	244.4	59720	1096.4	685.5	1782
4320123	1164.0	1100.1	287.7	82792	1408.2	510.1	1918
4320212	1000.5	1012.4	190.8	36409	881.2	538.1	1419
4321300	1282.4	1297.6	203.3	41330	1000.2	705.7	1706

Variography analysis is required to perform kriging interpolation, in which the gsimcli method is based. Hence, before executing the homogenisation procedure, the semivariogram model must be studied and its parameters defined. Data was divided by month, and then by decade. Due to the missing periods in the beginning of the century (1900-1929) and in the fifth decade (1940-1945), the first three decades were joined into a data set, as well as the fourth and fifth decade, for the purpose of the variography study. Seven semivariograms were modelled for each of the monthly series, in a total of 84.

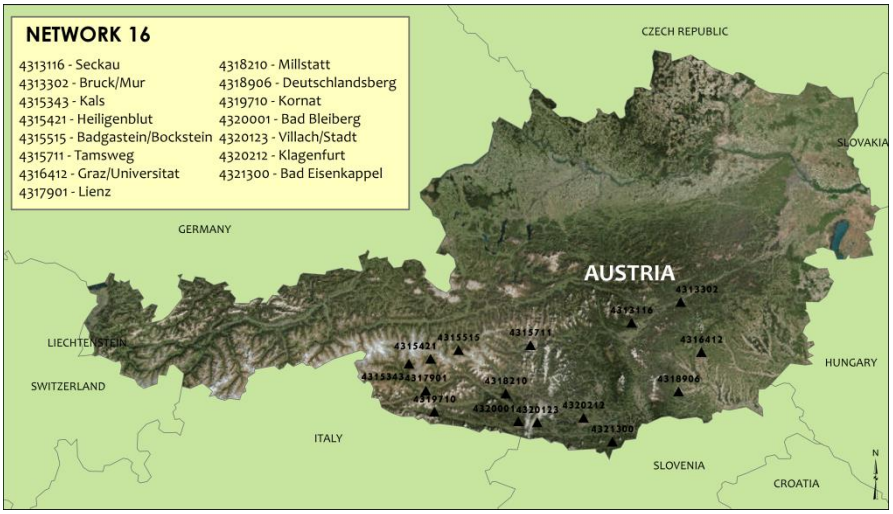

Figure 1. Location of 15 stations from the network 16.

Methodological framework

gsimcli method

This study evaluates the gsimcli homogenisation method, which is based on the direct sequential simulation (DSS) algorithm (Soares, 2001). The gsimcli method uses the DSS in the calculation of the local probability density functions (pdf) (Costa & Soares, 2009) at the location of the candidate station. Such calculation is prepared solely with the temporal and spatial observations of nearby reference stations. A breakpoint is identified whenever the interval of a specified probability p, centred in the local pdf, does not include the real observation of the candidate station. The detected irregular value is then replaced by a statistic value of the local pdf formerly computed (e.g. mean, median or a given percentile).

This method turned into a software package, which allowed the homogenisation process to become direct and quasi automatic (Caineta, Ribeiro, Soares, & Costa, 2015). Two subsets of parameters must be defined before starting the homogenisation procedure: the simulation parameters and the homogenisation parameters. The former define the number of simulations, the Kriging type of the geostatistical method, the maximum number of nodes to be found, number of CPU cores, the simulation grid size and the semivariogram parameters. The latter depict the candidates order, the detection parameter, the local radius and the correction parameter. The simulation grid describing the area where the stations are located must also be analysed, in order to ascertain the cell size, the number of columns

and the number of rows. These values take into account that the bordering stations of the network must be surrounded by a number of cells at least equal to the value of the local radius parameter. The gsimcli method is freely available at http://iled.github.io/gsimcli/.

Homogeneity tests

In order to start the homogenisation procedure, the monthly precipitation surrogate data set was divided in twelve folders, one per month. Each of these folders included the variography file with the semivariogram parameters per decade, the grid settings file containing the grid cell size, and a subfolder with the ten decadal data files. The monthly folders are homogenised separately.

A sensitivity analysis comprising sixteen different strategies was implemented for the following parameters (Table 2): grid cell size, the detection parameter (the probability value to build the detection interval centred in the local pdf), the correction parameter (the statistic value used for the inhomogeneities correction: the inhomogeneities, outliers or missing values, can be replaced by the mean, median, skewness, and percentile), the local radius (sets the radius of a circle centred at the candidate station location where the simulated values of the nodes located within the circle are considered in the calculation of the local pdf), and the candidates order (the order by which the candidate stations are homogenised).

The analysed grid cell sizes are 1000 m, 5000 m, and 10000 m, which correspond to grids of 27709 cells (229 x 121 cells), 2088 (58 x 36 cells), and 792 cells (36 x 22 cells), respectively. The values of the detection parameter analysed are 0.95 and 0.975. The values considered for the correction parameter are the percentiles of 0.90, 0.95, and 0.975. The investigated local radius are 0, 1, 2, and 3. It is noteworthy to mention that the area of the circle centred in the candidate station depend on the grid cell size and the local radius. Regarding the candidates order, Tests #1 to #12 and #14 to #16 used the descending value of the stations' data variance, while Test # 13 used the network deviation (difference between the station and the network average values) to define the sequence of the candidate stations to be homogenised (Table 2).

Table 2 Different homogenisation strategies (grid cell size, detection parameter, correction parameter, and local radius parameter). In all Tests the candidates order was based on the stations' data variance, except in Test #13 (*) that was based on the network deviation.

Test #	**Grid cell size** (*meters*)	**Detection parameter** (*p*)	**Correction parameter** (percentile *p*)	**Local radius parameter** (r)
1	1 000	0.95	0.975	0
2	1 000	0.95	0.95	0
3	5 000	0.95	0.90	1
4	5 000	0.95	0.90	2
5	5 000	0.95	0.90	3
6	5 000	0.95	0.975	0
7	5 000	0.95	0.975	1
8	5 000	0.95	0.975	2
9	10 000	0.95	0.975	0
10	10 000	0.95	0.975	1
11	10 000	0.95	0.975	2
*12**	10 000	0.95	0.90	0
*13**	10 000	0.95	0.90	0
14	10 000	0.95	0.90	1
15	10 000	0.95	0.90	2
16	10 000	0.975	0.975	0

The values defined for the remaining simulation parameters are common to the sixteen strategies, and correspond to default values proposed by Ribeiro, Caineta, Costa, Henriques, & Soares (2016):

- Number of simulations: 500.
- Kriging type: ordinary kriging.
- Maximum number of nodes to be found: 16.
- Number of CPU cores: 4.

Results and discussion

For each homogenisation strategy, four performance metrics are automatically calculated by gsimcli software (Table 3). Tests with the lowest values of performance metrics correspond to tests with the best set of parameters.

Those metrics are the station Centred Root Mean Square Error (CRMSE), the network CRMSE, the station Improvement and the network Improvement, as defined by Venema, et al. (2012). The CRMSE was chosen by the HOME project since the main aim of the homogenisation is not to improve the absolute values but rather the temporal consistency. 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.

The improvement metrics assess the enhancement over the inhomogeneous data. Station (network) Improvement metrics will reflect the quality of the procedure also shown in the station (network) CRMSE. The improvement metrics are computed as the ratio of the station (network) CRMSE of the homogenised networks and the station (network) CRMSE of the same inhomogeneous networks.

Analysing the station CRMSE of the sixteen strategies, it is possible to note that the highest values belong to the strategies where the correction parameter was defined as the percentile of 0.90, regardless of the other parameters. Therefore, it is less appropriate for the correction of irregularities (inhomogeneities, outliers, and missing values). These strategies (Tests #3, #4, #5, #12, #13, #14, and #15) also have the highest values of the network CRMSE.

The most appropriate value for the correction parameter seems to be the percentile of 0.975, because the performance metrics exhibit smaller values.

Comparing the candidates order parameter, focus must be given to Tests #12 and #13. The evaluation of their performance metrics is ambiguous, since the station CRMSE is better for the Test #12 (candidates order by variance), while Test #13 performs better in the network CRMSE (candidates order by network deviation).

Tests #9 and #16 differ in the detection parameter, and their station and network CRMSE vary in opposite directions. In this case, as it happens with the candidates order parameter, it is not possible to make any judgement about the best value for the detection parameter. The value of 0.95 is set as default.

The CRMSE metrics decrease with the increase of the cell size (for e.g. Tests #1, #6, and #9), thus increasing the cell size improves the method performance.

The best metrics are provided by Tests #10 and #11, where the cell size is 10000 m, the detection and correction parameters are set to 0.95 and the percentile of 0.975, respectively. The difference between these two Tests is the local radius (1 and 2, respectively) parameter. Their performance metrics are very similar. The larger the cell size and the local radius are, the greater the quality of the homogenisation results. This fact relates to the area that is considered for the computation of the local pdf.

Table 3 Performance metrics of the16 homogenisation strategies.

Test #	station CRMSE	network CRMSE	station Improvement	network Improvement
1	13.11	5.11	1.10	1.17
2	13.56	5.96	1.13	1.37
3	15.05	7.73	1.26	1.77
4	15.01	7.71	1.26	1.77
5	14.99	7.69	1.25	1.76
6	13.09	5.20	1.09	1.19
7	13.01	5.13	1.09	1.17
8	12.99	5.10	1.09	1.17
9	13.03	5.25	1.09	1.20
10	12.90	5.10	1.08	1.17
11	12.92	5.09	1.08	1.17
*12**	15.02	7.71	1.26	1.77
*13**	15.17	7.59	1.27	1.74
14	14.97	7.65	1.25	1.75
15	14.96	7.65	1.25	1.75
16	12.94	5.33	1.08	1.22

Conclusion

This study aimed at investigating several parameters in the homogenisation of monthly precipitation surrogate series with the gsimcli approach. The analysed parameters were the grid cell size, the detection parameter, the correction parameter, and the local radius parameter.

The analysis has emphasised the importance of the grid cell size and the local radius parameters in the performance of gsimcli. The knowledge of the surrounding area of a candidate station is essential to the improvement of its series quality. The principle of relative homogeneity (Conrad & Pollack, 1962) is again validated.

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 homogenization and interpolation of climate data").

Bibliography

Aguilar, E., Auer, I., Brunet, M., Peterson, T., & Wieringa, J. (2003). Guidelines on climate metadata and homogenization. In P. Llansó (Ed.), *WMO/TD No.1186, WCDMP No. 53*. Geneva: World Meteorological Organization.

Bernstein, L., Bosch, P., Canziani, O., Chen, Z., Christ, R., Davidson, O., et al. (2007). *Climate Change 2007: Synthesis Report.* Cambridge: Cambridge Univeristy Press.

Bližňák, V., Valente, M., & Bethke, J. (2014). Homogenization of time series from Portugal and its former colonies for the period from the late 19th to the early 21st century. *International Journal of Climatology*, 2400-2418.

Caineta, J., Ribeiro, S., Soares, A., & Costa, A. C. (2015). Workflow for the homogenisation of climate data using geostatistical simulation. *Informatics, Geoinformatics and Remote Sensing. 1*, pp. 921-929. Albena, Bulgaria: International Multidisciplinary Scientific GeoConference - SGEM.

Caussinus, H., & Mestre, O. (1996). New mathematical tools and methodologies for relative homogeneity testing. In H. M. Service (Ed.), *Proceedings of the First Seminar for Homogenization of Surface Climatological Data* (pp. 63-82). Budapest, Hungary: World Meteorological Organization.

Caussinus, H., & Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Applied Statistics, 53*, 405-425.

Chair of the Management Committee of the Action. (2011). *Monitoring Progress Report 03/05/2007-01/06/2011. HOME - Advances in Homogenisation Methods of Climate Series: An Integrated Approach (COST Action ES0601).*

Conrad, V., & Pollack, L. W. (1962). *Methods in Climatology.* Cambridge, MA: Havard University Press.

Costa, A. C., & Soares, A. (2009). Homogenization of Climate Data: Review and New Perspectives Using Geostatistics. *Mathematical Geosciences, 117*(1), 91-112.

Domonkos, P. (2013). Measuring performances of homogenization methods. *Idojaras. Quarterly Journal of Hungarian Meteorology, 117*(1), 91-112.

Domonkos, P., Poza, R., & Efthymiadis, D. (2011). Newest developments of ACMANT. *Advances in Sicence Research, 6*, 7-11.

Hannart, A., Mestre, O., & Naveau, P. (2014). An automatized homogenization procedure via pairwise comparisons with application to Argentinean temperature series. *International Journal of Climatology, 34*, 3528-3545.

Menne, M. J., & Williams Jr., C. N. (2009). Homogenization of temperature series via pairwise comparisons. *Journal of Climate, 22*(7), 1700-1717.

Menne, M. J., Williams Jr., C. N., & Vose, R. S. (2009). The U. S. historical climatology network monthly temperature data, version 2. *Bulletin of American Meteorology Society, 90*, 993-1007.

Ribeiro, S., Caineta, J., & Costa, A. C. (2015). Review and discussion of homogenisation methods. *Physics and Chemistry of the Earth*, in press.

Ribeiro, S., Caineta, J., Costa, A., Henriques, R., & Soares, A. (2016). Detection of inhomogeneities in precipitation time series in Portugal using direct sequential simulation. *Atmospheric Research, 171*, 147-158.

Sayemuzzaman, M., & Jha, M. K. (2014). Seasonal and annual precipitation time series trend analysis of North Carolina, United States. *Atmospheric Research, 137*, 183-194.

Soares, A. (2001). Direct Sequential Simulation and Cosimulation. *Mathematical Geology, 33*(8), 911-926.

Szentimrey, T. (1999). Multiple analysis of series for homogenization (MASH). *Proceedings of the Second Seminar for Homogenization of Surface Climatological Data. WCDMP-No. 41, WMO-TD No. 1962*, pp. 27-46. Budapest, Hungary: World Meteorological Organization.

Szentimrey, T. (2007). *Manual of homogenization software MASH v3.02.* Hungarian Meteorological Service.

Szentimrey, T. (2008). Development of MASH homogenization procedure for daily data. In M. Lakatos, T. Szentimrey, Z. Bihari, & S. Szalai (Ed.), *Proceedings of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases. WCDMP-No. 71*, pp. 123-130. Budapest, Hungary: World Meteorological Organization.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J., Domonkos, P., et al. (2012). Benchmarking homogenization algorithms for monthly data. *Climate of the Past, 8*(1), 89-115.

Vertacnik, G., Dolinar, M., Bertalanic, R., Klancar, M., Dvorsek, D., & Nadbath, M. (2015). Ensemble homogenization of Slovenian monthly air temperature series. *International Journal of Climatology, 35*, 4015-4026.

World Meteorological Organization. (2010). *Guide to Climatological Practices, third edition (WMO No. 100).* Geneva, Switzerland: World Meteorological Organization.