Adv. Sci. Res., 6, 167–172, 2011
www.adv-sci-res.net/6/167/2011/
doi:10.5194/asr-6-167-2011

Advances in Science & Research
Open Access Proceedings

Homogenising time series: beliefs, dogmas and facts

P. Domonkos

Centre for Climate Change (C3), Geography Dept., University Rovira i Virgili, Campus Terres de l'Ebre, C. Betánia 5, Tortosa, 43500, Spain

Received: 7 October 2010 – Revised: 25 May 2011 – Accepted: 2 June 2011 – Published: 14 June 2011

Abstract. In the recent decades various homogenisation methods have been developed, but the real effects of their application on time series are still not known sufficiently. The ongoing COST action HOME (COST ES0601) is devoted to reveal the real impacts of homogenisation methods more detailed and with higher confidence than earlier. As a part of the COST activity, a benchmark dataset was built whose characteristics approach well the characteristics of real networks of observed time series. This dataset offers much better opportunity than ever before to test the wide variety of homogenisation methods, and analyse the real effects of selected theoretical recommendations.

Empirical results show that real observed time series usually include several inhomogeneities of different sizes. Small inhomogeneities often have similar statistical characteristics than natural changes caused by climatic variability, thus the pure application of the classic theory that change-points of observed time series can be found and corrected one-by-one is impossible. However, after homogenisation the linear trends, seasonal changes and long-term fluctuations of time series are usually much closer to the reality than in raw time series. Some problems around detecting multiple structures of inhomogeneities, as well as that of time series comparisons within homogenisation procedures are discussed briefly in the study.

1 Introduction

To obtain a precise and reliable picture about the climatic variability of the period with instrumental observation methods, it is necessary to eliminate the influence of technical changes (hereafter: inhomogeneity, IH) in the observation systems. Therefore, together with the collection and archiving the observational data, a special branch of quality control developed for managing this kind of problem, i.e. the so-called time series homogenisation. During this development, a large number of statistical methods were introduced. Recently, enhanced efforts have been devoted to compare and evaluate the efficiency of different methods and this is not an easy task, because in real observed datasets the true statistical properties of IHs are never known exactly. Among other efforts the COST HOME has brought dynamism to these examinations. The present epoch of research on homogenisation methods can be characterized with the following new lines:

(i) Homogenisation methods are tested in simulated databases whose properties approach well the real properties of networks of observed climatic time series; (ii) The performance of homogenisation methods is evaluated by calculating RMSE between corrected time series and the corresponding homogeneous time series, as well as calculating the mean bias of linear trends between corrected time series and perfect time series.

In this study some theoretical problems related to the application of homogenisation methods are briefly described, and an example is shown for demonstrating the superior performance of the detection methods whose algorithm includes a direct identification of multiple structures of change-points.

2 Methods and definitions

Two, frequently appearing forms of IHs are defined here. Note that other forms may also occur, but they are not discussed in this study.

– *Change-point*: A sudden shift in the mean of the observational values. It is the most frequent form of IH, since most technical changes happen abruptly.

Correspondence to: P. Domonkos
(peter.domonkos@urv.cat)

Published by Copernicus Publications.

– *Platform-like inhomogeneity* [Pfm]: Pair of change-points of the same size, but of the opposite direction.

Concepts related to efficiency-evaluation:

– *Correct detection*: When an IH is detected in year j, and an IH really exists in section $[j-2,\ j+2]$.

– *False detection*: In the detection result an IH is included in year j, and no IH exists in section $[j-2,\ j+2]$. Note: If two IHs are detected around a really existing IH (e.g. an IH really exists in year j, but the detection results indicate two IHs, one in $j-2$ and another one in $j+2$) one of them is sorted into the correct detections, but the other one into the false detections.

The total number of correct detections, that of false detections and that of true IHs are denoted by S_R, S_F and S, respectively.

– *Power of detection* (Pw):

$$\mathrm{Pw}=\frac{S_R}{S} \tag{1}$$

In Eq. (1), S stands for the total number of true IHs.

– *Detection skill* (D):

$$D=\frac{S_R-S_F}{S} \tag{2}$$

– *Efficiency of RMSE-reduction* (E): In this study the efficiency is characterised by the improvement of root mean squared error (RMSE) due to homogenisation.

$$E=\frac{\mathrm{RMSE}_{\mathrm{raw}}-\mathrm{RMSE}_{\mathrm{homogenised}}}{\mathrm{RMSE}_{\mathrm{raw}}} \tag{3}$$

3 Purpose of homogenisation

Until the recent years it was common that the efficiency of homogenisation methods was evaluated by some set of simulated time series including a white noise process plus one or a few change-points whose magnitudes are significantly higher than the standard deviation of noise (s_d). In traditional evaluations of usability, most often the detection skill was calculated only. Although this kind of examinations provide valuable information about the properties of homogenisation methods, the results do not give direct information about the effectiveness of the methods, for three reasons: (i) Real properties of observed climatic time series are very different from this simple model, (ii) In the calculation of D the use of some arbitrary parameters is unavoidable, (iii) Detection skill, hit rate, false alarm rate, etc. do not provide direct information about the success in improving the reliability of trends and long-term fluctuations in time series. These ideas had already been taken into account (though in a relatively initial level) in some studies before the COST HOME. Menne and Williams (2005, 2009) analysed the efficiency of homogenisation methods applying test dataset in which the sizes of IHs have normal distribution with zero peak. Domonkos (2008, 2011) built a test dataset whose statistical characteristics are very close to an observed temperature dataset and used various measures for assessing efficiency.

After examining the homogeneity of radiosonde time series, Sherwood (2007) stated: "... detection of change-points is neither realistic nor necessary, ... success should be measured instead by the integrity of climate signals." This thesis is valid also for time series of surface observations. The statistical properties of true and detected IHs often differ markedly. In Figs. 1–2 some experimental results are shown

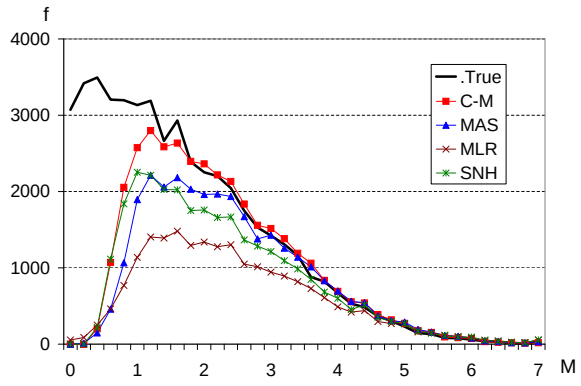

Figure 1. Magnitude-distribution of real and detected IHs (change-points) when the mean frequency of change-points is 5 per 100 yr, and shift-magnitudes have normal distribution with 0 peak and 3.5 times larger standard deviation than s_d. On the abscissa M means magnitude proportioned to s_d, while frequencies (f) are shown with an arbitrary unit. Homogenisation methods: C-M – PRODIGE, MAS – Multiple Analysis of Series for Homogenization, MLR – Multiple Linear Regression, SNH – Standard Normal Homogeneity Test.

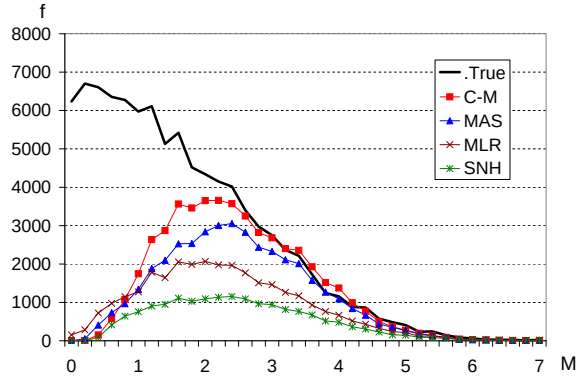

Figure 2. The same as Fig. 1, but 5 Pfm (i.e. 10 change-points of platform-like IHs) per 100 yr are included in time series instead of persistent shifts. Length of platforms has equal distribution between 1 month and 10 yr.

that obtained by examining sets of 10 000 artificial time series including 5 change-points (Fig. 1), or 5 Pfms (Fig. 2). When the sizes of IHs are relatively large, the amount of detected IHs approaches well their true frequency, but for small IHs detection is often impossible by any of the known homogenisation methods. The ratio of undetected IHs is considerably higher when Pfms occur in time series.

Thinking over Sherwood's thesis it can be stated that time series homogenisation is the utilisation of the spatially redundant information for the improvement of reliability of time-variability in data. (Note that instead of or beside the spatial redundancy, other pieces of information or assumptions can also be used for homogenisation, but in climatological studies it is not typical and usually not recommendable.) Characteristics of effective homogenisation methods are: (i) maximal exploitation of spatial information, (ii) high skill in finding timings of IHs, (iii) ability to treat common effects of multiple IHs, (iv) application of an appropriate correction method.

In testing efficiency the most reliable results can be obtained when statistical properties of artificial datasets are close to those of observed datasets. For obtaining test datasets with realistic properties, characteristics of detected IHs should be compared between observed data and artificial data (Domonkos, 2008, 2011). The results of this kind of tests indicate that in observed climatic time series small shifts and Pfms are frequent, thus their direct and indirect effects must be taken into account in the evaluation of homogenisation methods.

4 Common effects of multiple inhomogeneities

In most homogenisation methods used in climatic studies (Standard Normal Homogeneity Test (SNHT), Alexandersson, 1986; Multiple Linear Regression (MLR), Vincent, 1998; Penalised Maximal t-test (PMT), Wang et al., 2007, etc.) a step-by-step procedure is applied, in which methods detect only one IH in a particular step, thereafter time series are cut into two parts at the timing of the detected change-point. This cutting algorithm can be transformed to semi-hierarchic algorithm with supplying the procedure with some other steps (Lanzante, 1996; Moberg and Alexandersson, 1997, etc.), but experimental results indicate that semi-hierarchic algorithms do not provide substantial improvement relative to the cutting algorithm (Domonkos, 2011). Some methods detect multiple structures of IHs in a direct way (Multiple Analysis of Series for Homogenisation, Szentimrey, 1999; PRODIGE, Caussinus and Mestre, 2004, these are referred as "direct methods" hereafter). When the number of IHs is low, cutting algorithm and semi-hierarchic algorithm may function efficiently (Menne and Williams, 2005), but for complex structures of IHs only the direct methods are powerful (Domonkos, 2011).

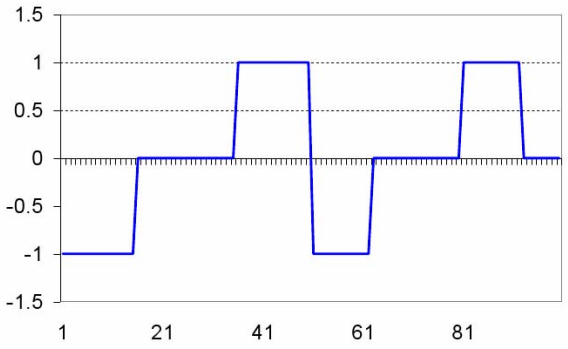

Figure 3. An example for complex structures of IHs. The horizontal axis is for the serial numbers of years, while in the vertical axis cumulative effects of IHs are shown. Noise is excluded for the sake of demonstrativeness.

To demonstrate the limitations of hierarchic methods, a theoretical example is presented here, i.e. a particular structure of IHs without noise (Fig. 3). In this example the largest shift is just in the middle of the time series and the means for the two halves of the time series are the same. For this kind of IH-structure, hierarchic algorithms are often incapable of detecting the largest IH in the first step, particularly when significance-examinations are based on the step-by-step comparison of some statistical characteristics between two parts of the series (SNHT, PMT, extremes of accumulated anomalies, as well as non-parametric methods, as for instance the Wilcoxon Rank Sum Test (WRS), Wilcoxon, 1945). However, a failure in finding the largest IH in the first step might affect the final results of the homogenisation procedure in hierarchic methods, since (i) the other (smaller) IHs can be detected with relatively low certainty because of their small size, (ii) in hierarchic algorithms a possible error in the first step introduces bias for the initial condition of later steps.

The performance of six detection methods (Easterling and Peterson method (E-P, Easterling and Peterson, 1995), MASH, MLR, PRODIGE, SNHT, WRS) is analysed by supplying the IH-structure of Fig. 3 with standard white noise in 10 000 simulation experiments. Detection powers for the different methods in function of the IH-size relative to s_d are calculated for the largest IH in the middle of the time series. The results are shown in Fig. 4, and it can be seen that the direct methods are really more effective than the hierarchic methods in identifying the largest IH of the time series. On the other hand, E-P has an even better Pw, than the direct methods. Note that E-P does not belong to the direct methods, nor to the hierarchic methods. In Fig. 5 the detection skills are shown for the same methods and IH-structures, as they included in Fig. 4. In the calculation of D all the IHs of the time series were taken into account. The results show that for small IHs the E-P still has the best performance, but when the IH sizes are larger than the background noise, the

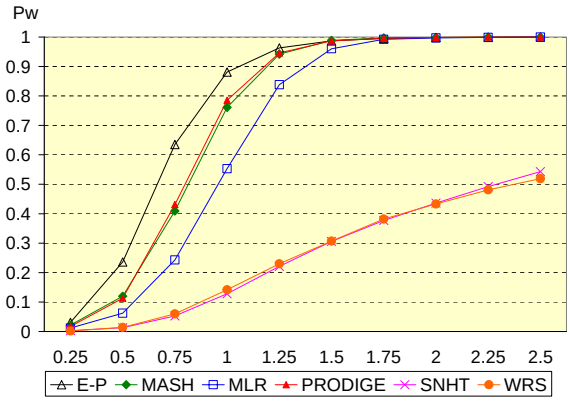

Figure 4. Power of detection for the largest IH of Fig. 3. On the abscissa the unit is s_d.

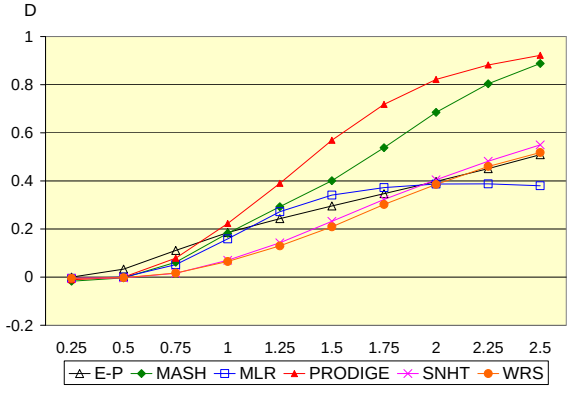

Figure 5. Detection skill for all IHs of Fig. 3. On the abscissa the unit is s_d.

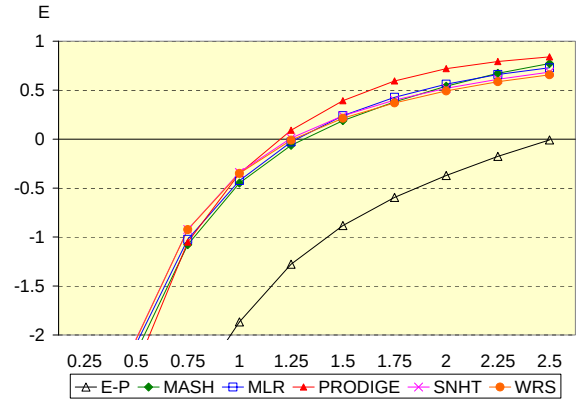

Figure 6. Efficiency in reducing RMSE when the IH-structure is the same as in Fig. 3. On the abscissa the unit is s_d.

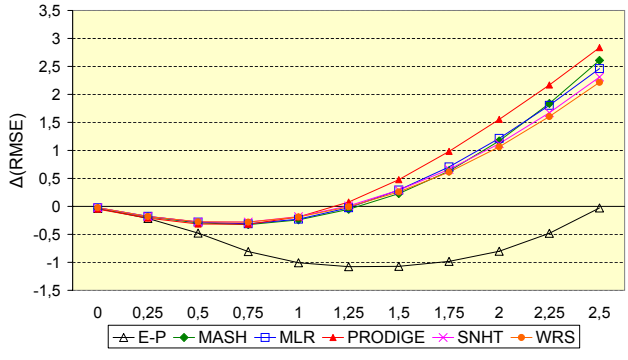

Figure 7. The same as Fig. 6, but the improvement of RMSE-error is presented on an absolute scale. $\Delta(\mathrm{RMSE}) = \mathrm{RMSE}_{\mathrm{raw}} - \mathrm{RMSE}_{\mathrm{homogenised}}$.

PRODIGE and MASH perform best. It is interesting to see that the MLR which has relatively good skill in detecting the largest IH (because the significance test of MLR is based on the autocorrelation of the tested series and not on the comparison of characteristics between two parts of the series), loses this advantage relative to SNHT and WRS when all the detection results are evaluated together. Figure 6 shows the efficiencies in the improvement of RMSE. These results have two striking features, i.e. (a) in case of small-size IHs all the efficiencies are markedly negative, (b) one of the six methods examined, namely the E-P, always performs with negative efficiency when the IH-sizes are lower than 2.5. The latter is the consequence of the fact that in E-P the time-coherence between the pieces of the detection results is less organised than in an hierarchic method (and even less organised than in direct methods). The results of E-P provide clear evidence that the examination of detection skill (detection power, false alarm rate, etc.) is not satisfactory in itself to evaluate the efficiency of homogenisation methods. On the other hand, the negative efficiencies for low-size IHs must not discourage users from applying homogenisation methods: The RMSE error in raw time series of 0.5 characteristic IH-size is only 4 % of the RMSE in raw time series with IHs of size 2.5. Thus, in absolute scale (Fig. 7), time series with low-size IHs (or without IH) might suffer a little corruption, while the quality of time series with large IHs can be improved markedly by the application of homogenisation methods. In the present experiment, PRODIGE performs best for IH sizes of larger than 1.25, while there is very little difference among the efficiencies of MASH, MLR, SNHT and WRS. Other experiments (Domonkos, 2011) confirm that (a) PRODIGE performs best when large- or medium-size IHs occur in time series, (b) PRODIGE and MASH have markedly better detection skills than other methods when complicated IH-structures occur in time series, (c) PRODIGE usually has superior performance in estimating linear trends with homogenised time series, (d) methods that do not consider the connections between the pieces of the detection results for individual IHs, either by direct algorithms or hierarchic algorithms, cannot be recommended for homogenising climatic time series, because the resultant time series often contain large errors, even if the detection skills of the methods are good.

5 Spatial comparison of time series

Since rapid temporal changes in climatic time series might occur also by true climatic variability, homogenisation methods are usually applied to differences of the raw time series. By generating these series the impact of climatic variability is reduced, because it is common for a given climatic region. However, a general problem of the spatial comparisons is that IH-detection results might be affected by the IHs in the series with which the candidate series are compared. Ideally, homogeneous reference series should be found for each candidate series. As this expectation is unrealistic, often a series of pair-wise comparisons is recommended instead of using fixed reference series (e.g. Caussinus and Mestre, 2004). In pair-wise comparisons, change-points existing in any time series of the network are searched and treated individually. However, pair-wise comparison methods also have drawbacks: (i) They use restricted number of series from the neighbourhood (i.e. in one particular comparison only one time series), thus noise and undetected IHs might cause relatively large errors; (ii) Skilled algorithms of multiple pair-wise comparisons can be too complicated for applying them in automatic procedures.

In recent efficiency examinations using the benchmark dataset of COST HOME, it turned out that a traditional creation of reference series may provide competitive efficiency. The method ACMANT (Domonkos et al., 2011) which is a modified and automated version of PRODIGE, performed with efficiency similar to the best of other methods (PRODIGE and MASH, Venema et al., 2010). The role of individual IHs in composites of reference series declines with the increase of the number of the composites. This fact is exploited in ACMANT in the way that the reference series is the weighted average of surrounding time series, as it was recommended by Peterson and Easterling (1994). In the present version of ACMANT the number of reference-composites is unlimited, and the minimum threshold of acceptable spatial correlation is 0.4. Notwithstanding, the author thinks that the optimal way of spatial comparison needs much further examination, because it is hard to find the optimal combination of the following two competitive aspects: On the one hand, impacts of climatic differences (which is indicated by relatively low spatial correlation) should be excluded by using a limited number of composites in building reference series, but, on the other hand, impacts of undetected IHs and noise should be reduced by including as many composites as possible, since with the use of a larger number of composites the effects of individual errors in composites decrease. Naturally, if effective and user-friendly versions of pair-wise comparisons are available, their use will also be recommendable.

6 Discussion and conclusions

The developers and users of homogenisation methods have to bear in mind that the eventual purpose of homogenisation is not to find change-points, but to obtain an improvement in the quality of the observational datasets that gives the opportunity to achieve more precise and more reliable results in climate change and climate variability analyses. Some old rules and recommendations should be re-evaluated. For instance, the performance of homogenisation methods depends on the connections between the pieces of the detection result, thus individual subjective decisions for selected change-points (e.g. using metadata information) may introduce undesired uncertainty to the overall efficiency of the procedure. Further examinations are needed also to find the optimal way of spatial comparison. As examples show, pair-wise comparison technique, but also the classic way of building reference series may both work with high efficiency within some homogenisation procedures.

The selection of the best homogenisation methods has to be based on efficiency tests executed on artificial databases of climatic time series with realistic statistical properties. We should go further on the way that is marked by the COST HOME activity.

Edited by: M. Brunet-India
Reviewed by: B. Trewin and two other anonymous referees

sc | nat The publication of this article is sponsored by the Swiss Academy of Sciences.

References

Alexandersson, H.: A homogeneity test applied to precipitation data, J. Climatol., 6, 661–675, 1986.

Caussinus, H. and Mestre, O.: Detection and correction of artificial shifts in climate series, J. Roy. Stat. Soc. Series, C53, 405–425, 2004.

Domonkos, P.: Testing of homogenisation methods: purposes, tools and problems of implementation, Proceedings of the 5th Seminar and Quality Control in Climatological Databases, edited by: Lakatos, M., Szentimrey, T., Bihari, Z., and Szalai, S., WCDMP-No. 71, WMO/TD-NO. 1493, 126–145, 2008.

Domonkos, P.: Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods, Theor. Appl. Climatol., doi:10.1007/s00704-011-0399-7, 2011.

Domonkos, P., Poza, R., and Efthymiadis, D.: Newest developments of ACMANT, Adv. Sci. Res., 6, 7–11, doi:10.5194/asr-6-7-2011, 2011.

Easterling, D. R. and Peterson, T. C.: A new method for detecting undocumented discontinuities in climatological time series, Int. J. Climatol., 15, 369–377, 1995.

Lanzante, J. R.: Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data, Int. J. Climatol., 16, 1197–1226, 1996.

Menne, M. J. and Williams Jr., C. N.: Detection of undocumented changepoints using multiple test statistics and composite reference series, J. Climate, 18, 4271–4286, doi:10.1175/JCLI3524.1, 2005.

Menne, M. J. and Williams Jr., C. N.: Homogenization of temperature series via pairwise comparisons, J. Climate, 22, 1700–1717, doi:10.1175/2008JCLI2263.1, 2009.

Moberg, A. and Alexandersson, H.: Homogenization of Swedish temperature data. Part II: Homogenized gridded air temperature compared with a subset of global gridded air temperature since 1861, Int. J. Climatol., 17, 35–54, 1997.

Peterson, T. C. and Easterling, D. R.: Creation of homogeneous composite climatological reference series, Int. J. Climatol., 14, 671–679, 1994.

Sherwood, S. C.: Simultaneous detection of climate change and observing biases in a network with incomplete sampling, J. Climate, 20, 4047–4062, doi:10.1175/JCLI4215.1, 2007.

Szentimrey, T.: Multiple Analysis of Series for Homogenization (MASH), WCDMP-41, WMO-TD. 932, Geneva, 27–46, 1999.

Venema, V., Mestre, O., and the COST HOME Team: Benchmark database, EGU General Assembly, Vienna, Austria, 3–7 May 2010, EGU CL4.6-13357, 2010.

Vincent, L. A.: A technique for the identification of inhomogeneities in Canadian temperature series, J. Climate, 11, 1094–1104, 1998.

Wang, X. L., Wen, Q. H., and Wu, Y.: Penalized maximal t test for detecting undocumented mean change in climate data series, J. Appl. Meteor. Climatol., 46, 916–931, doi:10.1175/JAM2504.1, 2007.

Wilcoxon, F.: Individual comparisons by ranking methods, Biometrics Bull., 1, 80–83, 1945.