Review and discussion of homogenisation methods for climate data

Ribeiro, S.\(^a\) (sribeiro@novaims.unl.pt), Caineta, J.\(^a\) (jcaineta@novaims.unl.pt) and Costa, A. C.\(^a\) (ccosta@novaims.unl.pt)

\(^a\) NOVA IMS, Universidade Nova de Lisboa, PORTUGAL

Corresponding author: Sara Ribeiro, NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, PORTUGAL.

E-mail: sribeiro@novaims.unl.pt; Phone: +351 213 828 610

This is the peer reviewed version of the following article:


This work is licensed under a Creative Commons Attribution-NonCommercial-No Derivatives 4.0 International License.
Review and discussion of homogenisation methods for climate data

Ribeiro, S.\textsuperscript{a} (sribeiro@novaims.unl.pt), Caineta, J.\textsuperscript{a} (jcaineta@novaims.unl.pt) and
Costa, A. C.\textsuperscript{a} (ccosta@novaims.unl.pt)

\textsuperscript{a} NOVA IMS, Universidade Nova de Lisboa, PORTUGAL

\textsuperscript{1}Corresponding author: Sara Ribeiro, NOVA IMS, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, PORTUGAL.
E-mail: sribeiro@novaims.unl.pt; Phone: +351 213 828 610

Abstract

The quality of climate data is of extreme relevance, since these data are used in many different contexts. However, few climate time series are free from non-natural irregularities. These inhomogeneities are related to the process of collecting, digitising, processing, transferring, storing and transmitting climate data series. For instance, they can be caused by changes of measuring instrumentation, observing practices or relocation of weather stations. In order to avoid errors and bias in the results of analysis that use those data, it is particularly important to detect and remove those non-natural irregularities prior to their use. Moreover, due to the increase of storage capacity, the recent gathering of massive amounts of weather data implies also a toilsome effort to guarantee its quality. The process of detection and correction of irregularities is named homogenisation. A comprehensive summary and description of the available homogenisation methods is critical to climatologists and other experts, who are looking for a homogenisation method wholly considered as the best. The effectiveness of homogenisation methods depends on the type, temporal resolution and spatial variability of the climatic variable. Several comparison studies have been
published so far. However, due to the absence of time series where irregularities are
known, only a few of those comparisons indicate the level of success of the
homogenisation methods. This article reviews the characteristics of the most important
procedures used in the homogenisation of climatic variables based on a thorough
literature research. It also summarises many methods applications in order to illustrate
their applicability, which may help climatologists and other experts to identify adequate
method(s) for their particular needs. This review study also describes comparison
studies, which evaluated the efficiency of homogenisation methods, and provides a
summary of conclusions and lessons learned regarding good practices for the use of
homogenisation methods.

**Keywords:** Methods comparison; Data quality; Irregularities; Trends; Homogenization
1 Introduction

Success in atmospheric modelling, weather forecasting or climate change monitoring depends on the quality of climate data used as input. Long time series without artificial discontinuities in their statistical characteristics are rare (Alexandersson and Moberg, 1997). Those irregularities can be due to climatic factors, or can be related to facts that happened during the process of collecting or recording climate data. Examples of climatic factors are the eruption of a volcano and the emission of its gases and ashes to the atmosphere contributing to the decrease of solar radiation, or the effect of the North Atlantic Oscillation in extreme temperature and precipitation records across Europe (Gaffen et al., 2000).

Non-climatic factors may introduce abrupt or gradual changes in the time series (Alexandersson and Moberg, 1997). Examples of the former are changes in the method of measuring and calculating climate values, such as the use of different daily times in the calculation of daily mean temperature (Peterson et al., 1998), change of measurement units (K, °C and °F for temperature) without any notice (Aguilar et al., 2003), changes in the formula for calculation of the variable's average (Puglisi et al., 2010), relocation of a station (Venema et al., 2013), or its repositioning to a different height (Auer et al., 2005). Gradual and soft changes can be exemplified by the presence of a tree or bush growing nearby the weather station, or the development of an urban area on its surroundings – the increasing of nocturnal temperature called the “Urban Heat Island Effect” (Brunet et al., 2006; Li et al., 2004; Sahin and Cigizoglu, 2010). A high number of non-natural irregularities are also introduced during the process of collecting, digitising, processing, transferring, storing and transmitting climate data series (Brunet and Jones, 2011).

These non-climatic factors may introduce artificial discontinuities, or inhomogeneities, in the time series. Such discontinuities can lead to misinterpretations of the studied
climate. In order to avoid errors and obtain homogeneous climate time series, non-natural irregularities in climate data series must be detected and removed prior to its use.

Three main types of inhomogeneities can be distinguished: point errors (coming from the observation to transmission and mechanisation processes); breakpoints corresponding to change-points or shifts in the mean (changes of location, instrumentation, observing practices or land use of the surroundings); and trends (sensor decalibration or urban growth) (Guijarro, 2006). Breakpoints are the most frequent form of inhomogeneities, since most technical changes happen abruptly (Domonkos, 2011a). Trend inhomogeneities are generally more difficult to detect, because they may be superimposed on a true climate trend (Easterling and Peterson, 1995).

Homogenisation is known as the process of detecting and correcting inhomogeneities (Aguilar et al., 2003). Another definition is provided by Štěpánek et al. (2006), where homogenisation includes the following steps: detection, verification and possible correction of outliers, creation of reference series, homogeneity testing (various homogeneity tests), determination of inhomogeneities in the light of test results and metadata, adjustment of inhomogeneities and filling in missing values. Mathematics, software and metadata are referred by Szentimrey (2011) as indispensable for homogenisation of climate data.

Recently, the importance of studying extremes of weather and climate required the development of homogenisation methods for climate data series with higher temporal resolution (e.g., daily data) (Brunetti et al., 2012). In case of precipitation, this task became a challenge due to its great variability (Rustemeier et al., 2011). This variability also results in great uncertainty in homogenisation. True climatic fluctuations in daily precipitation may be interpreted as change-points and removed from time series as
inhomogeneities. Moreover, the magnitude of inhomogeneities may differ with varying weather situations (Nemec et al., 2013). Another problem is associated with errors linked to the measuring process, particularly during extreme weather events. For example, larger adjustments are likely to be required for precipitation as its recording is strongly affected by wind strength (Auer et al., 2005). Systematic underestimation of snowfall is also a serious problem in areas where a substantial part of precipitation is collected by rain gauges as snow (Auer et al., 2005; Eccel et al., 2012). To overcome these issues, daily homogenisation methods require complex techniques or the improvement of homogenisation methods previously used for monthly and annual climate series. Those homogenisation methods are of paramount importance as those series are the basis for political decisions with socio-economic consequences (Venema et al., 2013).

The present review provides a description and discussion of homogenisation methods for climate data series, and summarises the conclusions of some comparison studies undertaken to assess their efficiency. Section 2 addresses the classification of homogenisation methods, Section 3 comprises a review of the available homogenisation methods, and Section 4 presents several homogenisation software packages. Comparison studies are briefly described in Section 5, where it is also given focus to the HOME project (COST Action ES0601). Finally, some conclusions are drawn in Section 6.

2 Approaches for detecting and correcting inhomogeneities

Homogenisation methods may have different characteristics, depending on the use of metadata, the subjectivity involved, the use of additional climate time series, the capability of detecting multiple breakpoints, etc. Those characteristics are discussed in the following subsections.
2.1 Direct and indirect homogenisation methods

Some authors define direct methods as those that are only based on metadata and subjective judgements (e.g., Li-Juan and Zhong-Wei, 2012). Direct methods have also been defined as mathematical algorithms that are able to detect multiple breakpoints in a direct way (e.g., Domonkos, 2011a), or that are able to deal with inhomogeneous reference time series (e.g., Venema et al., 2012). In the following, we will consider the definitions of direct and indirect methods provided by Aguilar et al. (2003) and Peterson et al. (1998). For these authors, direct methods include the use of metadata, the analysis of parallel measurements, and statistical studies of instrument changes. The indirect methodologies consider the use of single station data (absolute approaches), the development of reference time series (relative approaches), and include both subjective and objective methods.

2.1.1 Direct methods

Direct methods aim to keep the climate time series homogeneous by anticipating changes in and around a meteorological station and limit their impact on data homogeneity (Aguilar et al., 2003, pp. 30-31). Direct methods rely on registering in the station history a metadata entry describing any change, and on collecting parallel measurements for a long enough period of time or by reproducing the old conditions (Aguilar et al., 2003, pp. 30-31; Peterson et al., 1998). Metadata information can provide precise knowledge of when the discontinuity occurred and what caused it, but correction factors can only be objectively derived from the records of the “new” and “old” conditions or from a plausible correction model.

2.1.2 Indirect methods

Indirect methods use a variety of statistical and graphical techniques to test the homogeneity and adjust the data series (Peterson et al., 1998; Szentimrey, 2006a).
Many of these procedures use metadata for identifying or validating the discontinuities found in a time series, as recommended by Aguilar et al. (2003, pp. 33). Among the indirect methods, Peterson et al. (1998) also distinguish between subjective and objective approaches. Subjective methods rely mostly on experts’ judgments. Subjective judgement can be useful in the exploratory analysis stage to identify discontinuities, for example by plotting the stations’ data, by using the Double-mass analysis (Kohler, 1949), or by assessing the reliability of metadata.

Domonkos and Štěpánek (2009) define objective detection methods as those that can be applied in automatic way, without any subjective step. Objective homogenisation methods (OHOMs) have become increasingly more complex (e.g., Domonkos, 2006, 2011b). Domonkos (2006) discusses the conditions, advantages, and limitations related to the practical application of many of these methods.

OHOMs search and correct significant inhomogeneities of time series. Their procedures are applied in a fully computerised way, so no subjective decision is needed during the application. These methods are appropriated for the homogenisation of large data sets, and their efficiency can be quantitatively determined. The statistical methods applied in recent OHOMs are as follows (Domonkos, 2011b): calculation of extremes of accumulated anomalies; non-parametric methods relying on rank-order of sample elements; comparison of averages for adjacent sub-periods; regression functions and the calculation of residual sum of squares; maximum likelihood methods; and tendency of separation of sample elements into different clusters around change-points.

Aguilar et al. (2003, pp. 32-40) and Peterson et al. (1998) include in the set of objective methods the group of absolute and relative approaches, which will be detailed in the following sections.
2.2 Absolute and relative homogenisation methods

Considering the use of additional climate data series, homogenisation methods can be distinguished in two classes: absolute and relative methods. Absolute methods consider only the time series of a single station to identify and adjust inhomogeneities (candidate station). Relative methods use data from the surrounding stations (reference stations) to homogenise the candidate station. Some relative approaches are based on a pairwise comparison of the candidate time series with the reference stations data, while other methods are based on composite reference series of differences (for temperature or pressure) or ratios (for precipitation) between candidate and reference stations. According to Domonkos (2013), there are three main approaches for time series comparisons: building one reference series from composite series for each candidate series; using multiple reference comparisons for each candidate series; and using multiple comparisons without defining which are the candidate and the reference series.

When detecting a discontinuity, an absolute method cannot distinguish if it is natural or artificial without the support of the station’s history records. Begert et al. (2005) referred a clear limitation in the absolute methods’ capacity to separate discontinuities from true climate signals. Same opinion is shared by Guijarro (2011), advising that absolute homogenisation methods are to be avoided in favour of relative methodologies.

Surrounding stations are exposed to almost the same climate signal. Relative homogenisation is favoured when the spatial density and coherence of the climate data series allows it, because the climatic variation that is common for the study region does not appear in the differences between the candidate and nearby stations (Domonkos, 2013). The difference time series can be used to detect inhomogeneities, but if a break is detected it may be not clear to which of the stations it belongs to. Furthermore, time series typically have more than just one break. These are two of the problems that
homogenisation techniques try to solve. Moreover, the difference time series is useless when the whole network has been simultaneously affected by changes. However, such collective changes are usually well documented, otherwise changes can be detected by comparing multiple networks, and thus this situation is not so problematic.

Most of the relative methods can only be effective if the surrounding weather stations are homogeneous, i.e. if they include natural discontinuities only. This fact raises another question: how to select surrounding stations that are free from artificial discontinuities? According to Reeves et al. (2007), a good reference series should be homogeneous and highly correlated with the candidate series. The use of a reference series that is not homogeneous and/or has different climate signals (trends and periodicities) complicates the problem of change-point detection/adjustment. Peterson et al. (1998) mention the use of metadata to determine which nearby stations would not be expected to have inhomogeneities during specific time periods. Another possible solution is to combine data from different reference stations into a composite reference series assumed as homogenised. Szentimrey (2006a) refers that the spatial covariance structure of data series is very important to develop efficient methods addressing reference series creation, difference series constitution or multiple comparisons of series.

Menne and Williams (2009) discuss the limitations and challenges of many relative homogeneity testing methods, and propose an algorithm that is able to deal with inhomogeneous neighbouring series. Other methods currently address the presence of change points within the reference series (e.g., Caussinus and Mestre, 2004; Domonkos, 2011c, 2014; Mestre et al., 2013; Szentimrey, 1999, 2006b, 2011).

2.3 Multiple breakpoint techniques

One of the fundamental problems of homogenisation is that usually more than one breakpoint is present in the candidate time series (Lindau and Venema, 2013). The
Majority of the statistical homogenisation methods deals with this problem by applying single-breakpoint techniques multiple times. Typically, when a breakpoint is detected, the time series is divided into two subsets of observations at the identified break and the single-breakpoint algorithm is applied separately to each subset of data. This process is repeated until no more breaks are found or the number of observations becomes too small. The disadvantage of this segmentation process is that the same test applied several times on the same observations can increase the risk of false detection (Beaulieu et al., 2009). The most efficient single-breakpoint technique is known as cutting algorithm (Domonkos et al., 2012), which is a hierarchic method for identifying multiple breakpoints proposed by Easterling and Peterson (1995).

Multiple breakpoint methods are those that detect and correct multiple change-points jointly, and not step-by-step. Recent studies indicate that these are the most effective detection procedures (e.g., Domonkos, 2011b; Venema et al., 2012). Multiple breakpoint algorithms use as detection criterion the maximum external variance between the means of constant time segments in between multiple breakpoints (Lindau and Venema, 2015). These methods apply a relatively simple model (step-function) and select the most probable parameters of this model by the examination of all possible combinations of breakpoint positions (Domonkos, 2013).

3 Statistical homogenisation methods and homogenisation procedures

There are many homogenisation methods described in the literature. A chronological review of the development of homogenisation methods for temperature series is provided by Domonkos et al. (2012). This section highlights the most used approaches, as well as the state-of-the-art homogenisation algorithms that are able to handle inhomogeneous reference series and multiple structures of inhomogeneities. The homogenisation techniques are classified by type of approach (Table 1 of the...
Appendix). Statistical techniques were classified based on their characteristics: non-parametric tests, classical tests (traditional techniques), regression models and Bayesian approaches. Techniques that were directly proposed as methods for the homogenisation of climate data series are named “homogenisation procedures”. These procedures may include more than one statistical technique. Moreover, considering the discussion in Section 2, the procedures listed in Table 1 (Appendix) are classified as objective bearing in mind the definition provided by Domonkos and Štěpánek (2009). Several techniques are used in the detection stage only (qualifying tests), thus they are useful for homogeneity diagnosis. A sample of studies where the referred methods were applied is provided as Supplementary Material (Table S1), to illustrate their applicability regarding the study region, climate variable and temporal resolution.

3.1 Non-parametric tests

The most common non-parametric tests used for homogeneity testing are: Von Neumann ratio test (Von Neumann, 1941), Wald-Wolfowitz runs test (Wald and Wolfowitz, 1943), Mann-Kendall test (Mann, 1945; Kendall, 1975), Wilcoxon-Mann-Whitney (Wilcoxon, 1945; Mann and Whitney, 1947), Kruskall-Wallis test (Kruskal, 1952; Kruskal and Wallis, 1952) and Pettitt’s test (Pettitt, 1979).

The Von Neumann ratio test (Von Neumann, 1941) calculates a ratio of the mean square between successive (year-to-year) differences to the variance, which is closely related to the first-order serial correlation coefficient (Talaee et al., 2014). The calculated value of this ratio is an indicator of the presence of irregularities in the series. This test does not provide the information regarding the date of the discontinuity (Costa and Soares, 2009a) and usually it is used together with other homogeneity tests.

The Wald-Wolfowitz runs test (Wald and Wolfowitz, 1943) is a well-known non-parametric test for randomness. It calculates a statistic based on the sum of the
number of changes, by comparing every datum from the time series with the median, over time. This test is sensitive to shifts and trends, but gives little information about the probable dates for breaks. This method is not powerful enough to be used individually in the relative homogeneity analysis and must be supported by graphical analysis so to increase the power of overall analysis, and to obtain the probable date and magnitude of the inhomogeneity, as stated by Tayanç et al. (1998).

The Mann-Kendall (M-K) test (Mann, 1945; Kendall, 1975) has been popularly used for assessing the significance of trend in hydrological time series, such as stream flow and precipitation. This test has proved to be a valuable tool on trend detection, since it provides useful information on the possibility of change tendency of the variables in the future (Yue and Wang, 2004). It has the advantage of not assuming any special form for the data distribution function, while having a power nearly as high as their parametric competitors. For this reason, it is highly recommended by the World Meteorological Organization (WMO) (Mourato et al., 2010).

The Wilcoxon-Mann-Whitney test (Mann and Whitney, 1947; Wilcoxon, 1945) is based in the use of rank order change-point detection (Aguilar et al., 2003). This approach is advisable when the normality of data is in doubt, such as precipitation data. For this variable, normality is easier to achieve in yearly averaged or in accumulated quantities than in monthly data.

The Kruskal-Wallis test (Kruskal, 1952; Kruskal and Wallis, 1952) is used to compare two or more independent groups of data. The Kruskal-Wallis test allows determining if the difference in the average ranks of three or more independent samples is significant. This test verifies if the hypothesis that all the samples came from the same parent population can be safely rejected.

Pettitt’s test (Pettitt, 1979) is a non-parametric rank test that detects single break points. The calculated statistic, derived from the Mann-Whitney, achieves the maximum
value for the year with the most likely break point. The test is capable of locating the period where a break may occur, but is more sensitive to breaks in the middle of the time series (Wijngaard et al. 2003).

3.2 Classical tests

Double mass analysis (Kohler, 1949), Craddock’s test (Craddock, 1979), Bivariate test (Potter, 1981), and Buishand Range test (Buishand, 1982) are classified as (statistical) classical tests as they correspond to traditional homogenisation techniques.

The Double mass analysis (Kohler, 1949) was one of the first techniques specifically proposed for homogeneity testing. The double-mass curve method is performed by plotting the cumulative amounts of the station under consideration against the cumulative amounts of a set of neighbouring stations. The plotted points tend to fall along a straight line under conditions of homogeneity. Cumulative deviations from some average value can alternatively be plotted to verify the time series homogeneity. It is only used during the exploratory analysis of the time series (Costa and Soares, 2009a). For precipitation time series, cumulative deviations are preferred, since changes in the mean amount are easier to be recognised (Buishand, 1982).

The Craddock’s test (Craddock, 1979) is a simple statistical method developed to compare annual precipitation records. This test requires a homogeneous reference series or, in some cases, long enough homogeneous sub-periods. It accumulates the normalised differences between the test series and the homogeneous reference series to determine the inhomogeneities (Aguilar et al., 2003). Craddock’s test is recommended by Venema et al. (2012). This test was included in two homogenisation packages: HOCLIS (software package for homogenisation of climatological time series) and THOMAS (tool for homogenisation of monthly data series) from ZAMG (Central Institute for Meteorology and Geodynamics, Austria) and MeteoSwiss (Federal Office
of Meteorology and Climatology, Switzerland), respectively (Auer et al., 2005; Begert et al., 2005).

Potter (1981) applied the bivariate test, developed by Maronna and Yohai in 1978, to precipitation annual series. This is a test for detecting a single systematic change in the mean of an independent time series, based on a second correlated series which is assumed as unchanged (Aguilar et al., 2003). Potter’s method generates a test statistic for each data value and an estimate of the maximum probable offset, or adjustment, for that year (Plummer et al., 1995). It closely resembles the double mass curve analysis (Aguilar et al., 2003).

Buishand (1982) used the cumulative deviations to perform some statistical tests, which were compared with the Von Neumann ratio test. This author concluded that both methods give nearly the same results. The Buishand Range test is more sensitive to breaks in the middle of the time series (Wijngaard et al., 2003).

3.3 Regression methods

Three methods using regression models are described: Two-phase regression (Easterling and Peterson, 1995), Multiple linear regression (Vincent, 1998), and the Method of cumulative residuals (Allen et al., 1998).

Easterling and Peterson (1995) developed the Two-phase regression (TPR) model, following the work of Solow (1987) who has constrained two regression functions to meet at the point of the inhomogeneity. These authors modified the previous technique so that the two regression lines do not need to meet at the discontinuity. For a given year (or time unit), one regression line is fitted to the reference series for the previous time interval of that year, and the second regression line is fitted to the second part of the time series. This process is repeated for all the years of the time series. The lowest residual sum of squares between the two regression functions will determine the point of discontinuity.
Vincent (1998) proposed the Multiple linear regression (MLR) homogenisation procedure. This technique consists of four linear regression models, applied in a sequence. The first model determines if the candidate series is homogeneous for the tested time interval. If it is homogeneous, the test will end and the remaining models are not used. If inhomogeneities are found, a second model is estimated to ascertain the existence of an overall trend in the candidate series. If the inhomogeneity found in the first model is not an overall trend, the third model is applied to identify the single step change. The fourth model will define the existence of trends before and after that step. If the four models are applied, it indicates that the candidate series have multiple inhomogeneities. In this case, the candidate time series will be divided at the position of the identified step and each segment will be tested separately, starting from the first model. Ducré-Robitaille et al. (2003) classified MLR as one of the most robust homogenisation methods. More recently, efficiency tests have shown that its detection skills are often lower than other objective methods (Domonkos, 2011b).

The Method of cumulative residuals (Allen et al., 1998) provides a way to relate data sets from two weather reference stations. For a given weather station with a homogeneous time series (independent variable), the records of a second station (dependent variable) can be considered to be homogeneous if the cumulative residuals from their simple linear regression model are not biased. This is tested by verifying if the residuals are contained within an ellipsis, which depends on the size of the data set, the standard deviation of the tested sample and the probability used to test the hypothesis (80% is commonly used). Costa and Soares (2006) proposed an extension of the cumulative residuals method that takes into consideration the concurrent relationship between several candidate series from the same climatic region. This technique uses the residuals from a Seemingly Unrelated Regression equations (SUR) model instead of the residuals from a simple linear regression model.
3.4 Bayesian approaches

Bayesian methods have a different approach from classical techniques. Through a prior distribution, the Bayesian approach acquires some knowledge about the climate variable being studied. That information and the observations are combined in a posterior information, which is used to make inference about the parameters. Their advantage is the formal use of non-experimental sources of information to complement the posterior probability distribution function for the studied variable, comprising the position of the shifts, which can be multimodal or skewed. After specifying a loss function, an estimate of the shift’s position can be obtained. Several Bayesian techniques were already used for the homogenisation of climate data series, which are described in this section: Bayesian multiple change-point detection in multiple linear regression (Seidou and Ouarda, 2007), Bayesian change-point in multiple linear regression (Seidou et al., 2007), Bayesian change-point algorithm (Ruggieri, 2013), Bayesian multiple change-points and segmentation algorithm (Hannart and Naveau, 2009), Change-point detection algorithm (Gallagher et al., 2012), and Bayesian Normal Homogeneity Test (Beaulieu et al., 2010).

The Bayesian multiple change-point detection in multiple linear regression (BAMS) (Seidou and Ouarda, 2007) follows a Bayesian linear regression model designed to detect multiple change-points. Its main characteristic is the identification of an unknown number of shifts. This procedure requires two training data sets and a prior distribution on the distance between adjacent change-points, which reveals the assumption of the number of existing change-points (Ruggieri, 2013). Beaulieu et al. (2009) considered this approach effective as it often detects the exact number of shifts in an artificial data set.

The Bayesian change-point in multiple linear regression (BARE) model (Seidou et al., 2007) was designed to infer the position of a single change-point in the parameters of a
multiple linear regression equation. Seidou et al. (2007) considered non informative prior distributions for the regression parameters and the variance. The prior for the change-point position is a uniform distribution. The method can also be applied for multiple change-points using a segmentation approach. Beaulieu et al. (2009) compared BAMS and BARE using synthetic series of total annual precipitation data series from Canada. Both techniques had similar detection skills, but BAMS performed better for the series with multiple shifts.

Ruggieri (2013) introduced a Bayesian Change-point Algorithm, which provides uncertainty estimates both in the number and location of change-points through a probabilistic solution to the multiple change-point problem. Two main differences should be referred, when comparing this method to BAMS: the nature of recursion and the prior distributions on the model parameters. This algorithm follows three steps: calculation of the probability density of the data; forward recursion (dynamic programming) and stochastic back-trace via Bayes rule (by sampling the number of change-points, the locations of change-points and the regression parameters for the interval between adjacent change-points). Ruggieri (2013) studied the performance of this method by analysing the irregularities in annual global surface temperature.

Hannart and Naveau (2009) used Bayesian Decision Theory to minimise a cost function for the detection of multiple change-points, the Bayesian multiple change-point and segmentation algorithm. The method identifies subsequences of the time series that isolate a unique change-point. These authors studied the performance of this method, by comparison with other methods using simulated series, and they also applied the method to annual temperature data from 16 weather stations located in France (1882-2007).

Gallagher et al. (2012) proposed a Bayesian homogenisation method, the Change-point detection algorithm, for daily precipitation series. The model can be described as
a two-state Markov chain with periodic dynamics. The chain serves to induce
dependence in the daily (precipitation) amounts, having two different states (dry or
wet). If the state considered for a specific day is wet, the amount of the precipitation is
modelled as a positive random variable with a seasonally dependent mean (amounts
are distribution-equivalent, but the distribution is not necessarily the same). This
method was used to homogenise daily precipitation data from Alaska and
Massachusetts.

The Bayesian normal homogeneity test (BNHT) enables the detection of a change in
the mean of a single normally distributed time series (Beaulieu et al., 2010). It is
applied to a reference series, similarly to SNHT. This test also allows the integration of
prior information on the date of the change-point (metadata or expert knowledge).
Beaulieu et al. (2010) applied this test to synthetic series of total annual precipitation in
Canada.

3.5 Homogenisation procedures

Techniques that were directly proposed as methods for the homogenisation of climate
data series are summarised in this section: SNHT – Standard Normality Homogeneity
Test (Alexandersson, 1986), SNHT with trend (Alexandersson and Moberg, 1997),
MASH – Multiple Analysis of Series for Homogenisation (Szentimrey, 1999), PRODIGE
(Caussinus and Mestre, 1996, 2004), Geostatistical simulation approach (Costa et al.,
2008), ACMANT – Adapted Caussinus-Mestre Algorithm for homogenising Networks of
Temperature series (Domonkos, 2011c), and ACMANT2 for homogenising daily and
monthly precipitation series (Domonkos, 2014).

The Standard Normal Homogeneity test (SNHT) (Alexandersson, 1986) is one of the
most popular and robust homogenisation methods for climatic variables (Ducré-
Robitaille et al., 2003). The application of SNHT begins with the creation of a
composite (ratio or difference) series between the station values and some regional
reference values assumed homogeneous. This composite series is then standardised. At a given moment, averages are calculated for the previous and the following period of that composite series. If the difference between those averages meets a critical value, a shift is inferred to exist at that moment, and the series is said to be inhomogeneous (Ducré-Robitaille et al., 2003).

Later, Alexandersson and Moberg (1997) improved the SNHT method to extend its detection to trends as well. In this innovative SNHT with trend, the alternative hypothesis is that the change of the mean level is gradual, starting and ending at arbitrary points of time, $a$ and $b$. A test value is computed for all combinations of $a$ and $b$. The pair that maximises this value has the highest likelihood for being the starting and ending of the trend section. When an inhomogeneity occurs as a sudden shift, such inhomogeneity will be determined by the trend test to be an abrupt change. SNHT with trend is suitable for gradual trends in climate time series, like the increasing of the urban heat island effect (Moberg and Alexandersson, 1997).

The Multiple Analysis of Series for Homogenisation (MASH) (Szentimrey, 1999, 2006b, 2011) was one of the first multiple breakpoint techniques. Currently, it is based on mutual comparisons of series within the same climatic area, and does not assume a homogenised reference series. Breakpoints commonly identified in the difference series (or ratio series for multiplicative variables) are attributed to the candidate series, since it is the only series presented in all. It is a step by step procedure: the role of the series (candidate or reference) changes gradually in the course of the procedure. MASH can be applied to yearly, seasonal and monthly time series. In the new multiple breakpoint procedure, significance and efficiency are formulated according to the conventional statistics related to types I and II errors, respectively. Additionally to the breakpoints and shifts, confidence intervals are also determined. MASH has turned into a software, where metadata can be used automatically to detect inhomogeneities. This
method is included in the HOCLIS-system (Auer et al., 2005). Since MASH v3.01, it is possible to homogenise daily datasets (Szentimrey, 2006b).

Caussinus and Mestre (1996, 2004) proposed a new multiple breakpoint technique named PRODIGE, which is based on penalised likelihood methods. The methodology uses a pairwise comparison for preselecting a set of accidents, which are considered within the framework of a multidimensional approach. This method is based on the principle that the series is reliable between two change-points. Those sections will be used as reference series. Instead of comparing a given series with a reference series whose definition is problematic, the comparisons are performed with all other series, by a series of differences. The series of differences is tested against discontinuities through the Caussinus and Lyazrhi (1997) technique. If a change-point (or an outlier) is constantly detected in all the difference series, it can be attributed to the candidate station. The second step of this method is an overall detection and correction. Those two steps are performed by using moving neighbourhoods. The size and the shape of these neighbourhoods are a compromise between the knowledge of the climatologist about the regional climate and the necessity to have enough data, in order to ensure good estimation. Another technique was later developed on basis of PRODIGE, named ACMANT.

The Geostatistical simulation approach proposed by Costa et al. (2008) can be summarised as follows (Costa and Soares, 2009a). The Direct Sequential Simulation (DSS) algorithm (Soares, 2001) generates realisations of the climate variable through the resampling of the global probability density function (pdf), using the local mean and variance of the candidate station, which are estimated through a spatiotemporal model. The local pdf for each time instant 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. When an irregularity is identified, the time series can be adjusted by replacing the
inhomogeneous record by the mean (or the median) of the pdfs calculated at the
candidate station location for the inhomogeneous periods.

Domonkos (2011c) proposed an Adapted Caussinus-Mestre Algorithm for
homogenising Networks of Temperature series (ACMANT), which is a relative
homogenisation technique applicable to monthly temperature series (Domonkos,
2011d). ACMANT is a fully automatic homogenisation method, and its most relevant
characteristics are: (i) harmonisation of examinations in different time-scales (annual or
monthly); (ii) use of optimal segmentation and the criterion proposed by Caussinus and
Lyazrhi (1997) in the detection of inhomogeneities; and (iii) use of ANOVA for the final
corrections of inhomogeneities. ACMANT comprises four main steps: preparation; pre-
homogenisation; homogenisation and final adjustments (Domonkos, 2011d). Recently,
Domonkos (2014) proposed a new unit for the homogenisation of monthly or daily
precipitation series, ACMANT2. This new version takes into consideration the climatic
regions of snowy winters, by making a distinction between rainy season and snowy
season and by searching the seasonal inhomogeneities with bivariate detection.
Another main difference from the previous version of ACMANT is that outlier filtering
and detection of short-term inhomogeneities are not included in the homogenisation of
precipitation series because, in this case, due to the lack of spatial consistency at
short-time scale, a possible identified break is very likely to be a true local extreme and
not an erroneous precipitation record. Currently, ACMANT and its unit ACMANT2 are a
homogenisation software package.

4 Homogenisation software packages

Lately, some of the homogenisation methods already described in the previous
sections were developed into software, in order to diminish the time consumed during
the homogenisation process and to minimise the interaction of users. The examples
described are: Climatol (Guijarro, 2006), RHTest (Wang, 2008), AnClim and
Climatol (Guijarro, 2006) is a set of routines for climatological applications than run under the cross-platform statistical programming language R. Although it may be applied to daily data, it is generally used in the homogenisation of monthly series. This computational application compares each candidate series with a reference series. Once the reference series has been computed, it can be used to determine which variations in the candidate series are due to the climate variability and which are real inhomogeneities that should be corrected. Climatol avoids the use of regression techniques and enables the use of data from surrounding stations when there is no common period of observation. The comparison between the candidate series and their estimated references allows the detection of point errors, shifts and trends through standard statistical tests. The graphical representations of the results can also be shown. Missing values from the candidate series can be directly replaced by the computed reference values. The application of the method to a dense monthly database indicates the importance of using an iterative strategy, thereby detecting and correcting only the coarser errors in the first place, and leaving the less prominent ones to the following iterations. Literature refers this method as robust and simple. However, the final decision on which inhomogeneities to correct must be complemented with visual inspection of the graphical representations.

The RHTest software package (Wang, 2008) is designed to detect multiple step change-points that might exist in a time series. Its recent version, RHTestV3, includes a fully automatic package. This package comprises two penalised maximal tests, PMF (Penalised Maximal F-test) and PMT (Penalised Maximal T-test). The PMF test allows the tested time series to have a linear trend throughout the whole period of the data record, with the annual cycle, linear trend, and autocorrelation of the base series, being estimated one after the other through iterative procedures, while accounting for all the
identified mean shifts (Wang, 2008). No reference series is used in any of these functions. The PMT test assumes the tested time series with zero-trend and Gaussian errors. In this case, a reference series is needed. The base-minus-reference series is tested to identify the position(s) and significance of change-point(s), but a multi-phase regression (MPR) model with a common trend is also fitted to the anomalies of the base series in the end to obtain the final estimates of the magnitude of shifts (Wang, 2008). In the MPR fit, the annual cycle, linear trend, and autocorrelation are estimated sequentially through iterative procedures, while accounting for all the identified mean-shifts.

AnClim (Štěpánek, 2008a) and ProClimDB (Štěpánek, 2008b) were developed as a combination of several features from methods mentioned above. ProClimDB is used for processing whole datasets (finding outliers, combining series, creating reference series, preparing data for homogeneity testing, etc.). AnClim works with one station at a time for homogeneity testing, but automated processing of many stations is enabled as well. Results from homogeneity testing produced by AnClim are imported back to ProClimDB and further processed. Two main steps are carried out (Štěpánek et al., 2009): data quality control and homogenisation. The first step is performed by several methods: (i) analysing difference series between candidate and neighbouring stations through pairwise comparisons; (ii) applying limits derived from interquartile ranges; and (iii) comparing the series values tested with “technical” series created by means of statistical methods for spatial data. In the homogenisation step, SNHT, Bivariate and Two-Phase Regression tests are applied to the series. The criterion for identifying a year of inhomogeneity is the probability of detection of a given year, calculated by the ratio between the number of detections for a given year from all tests results for a given station and the total of all theoretically possible detections. The correction of the inhomogeneity is given by the value of the instant before the detected break plus a calculated correction factor, which is determined by the reference series. Štěpánek et
al. (2009) applied AnClim and ProClimDB to daily temperature and precipitation data sets.

Menne and Williams (2009) developed an automated homogenisation algorithm for monthly data that builds on efficient change-point detection techniques, named USHCN (United States Historical Climatology Network). The pairwise algorithm proposed by those authors is able to detect undocumented breakpoints and to deal with inhomogeneous neighbouring series. The algorithm conducts a pairwise comparison in order to first identify all evidences of change-points, combining those evidences with information about documented changes. The algorithm relies upon a pairwise comparison of series in order to reliably distinguish artificial changes from true climate variability, even when the changes are undocumented. In addition, the algorithm employs a recursive testing strategy to resolve multiple undocumented change-points within a single time series. Lastly, the procedure explicitly looks for abrupt "jumps" as well as local and unrepresentative trends in the series.

HOMER, HOMogenization softwarE in R, is an interactive semi-automatic procedure that explores the best characteristics of other state-of-the-art homogenisation methods (PRODIGE and ACMANT), as well as from Climatol and the cghseg joint-segmentation method (Mestre et al., 2013). Basic quality control and network analysis are adapted from Climatol. Detection can be performed using a partly subjective pairwise comparison technique (adapted from PRODIGE) or, alternatively, by applying the full automatic cghseg detection. HOMER includes the ACMANT capability to coordinate the operations on different time scales (from multiannual to monthly). HOMER also includes the UBRIS (Urban Bias Remaining in Series) procedure, which allows characterising artificial climatic trends, in most cases related to urbanisation.
5 Comparison of homogenisation methods

A homogenisation method is considered efficient when it is able to overcome two problems: the fact that nearby stations are also inhomogeneous, and the existence of more than one irregularity within the time series (Lindau and Venema, 2013). Depending on the used techniques, some homogenisation methods can be more appropriate for a specific climate variable (e.g., first version of ACMANT for temperature), while others can only be used at a given time scale resolution, providing less efficiency for high temporal resolution data series (e.g., daily observations). In order to assess their efficiency, numerous comparison exercises are described in the literature. This section summarises comparison studies undertaken for homogenisation methods, emphasising the HOME project (COST Action ES0601) in the second subsection.

5.1 Comparison Tests

In the past two decades several comparison studies have been published in order to determine the most efficient homogenisation method. A synopsis of those comparison tests is disclosed as Supplementary Material (Table S2), and describes the location, variable and periodicity of the climate time series, the compared tests, and some of the achieved conclusions. Those comparison tests are described by chronological order.

Comparison studies also proved the difficulty of indicating which method is the most efficient. Some of the studies were performed using a set of common homogenisation methods, achieving different conclusions. Climate variables also have influence on the efficiency of the method, due to their variability and temporal resolution. Venema et al. (2012) provide a valuable discussion on many of these comparison tests. Problems related to the choice of efficiency measures and the creation of appropriate test-datasets are discussed by Domonkos (2011b, 2013).
5.2 HOME project (Advances in Homogenisation Methods of Climate Series: An Integrated Approach)

In 2008, a European Cooperation in the field of Scientific and Technical Research, HOME – Advances in Homogenisation Methods of Climate Series: An Integrated Approach (COST Action ES0601), was released to compare, evaluate and develop homogenisation methods (HOME, 2011). New (or extensions of earlier) methods were proposed as homogenisation techniques to test a benchmark data set comprising temperature and precipitation data. HOME’s main objective was to achieve a general method for homogenising climate and environmental data sets.

The benchmark data set contains real inhomogeneous data as well as simulated data with inserted inhomogeneities, which comprise outliers, break points and local trends. Missing data was also simulated (on those generated data sets) and a global trend was added. This benchmark was composed of three distinct data sets: inhomogeneous (real) climate networks, surrogated and synthetic data sets. The real data set allows comparisons between the different homogenisation methods, since it is comprised of the most realistic type of data and inhomogeneities. Surrogate data was prepared to reproduce the structure of real data in an accurate way so that it could be used as its substitute. Synthetic data is based on surrogate networks. However, the differences between the stations have been modelled as uncorrelated Gaussian white noise. Later, it was concluded that synthetic data is easier to homogenise than the more realistic surrogate data (Venema et al., 2012).

Twenty-five contributions based on 13 algorithms (including MASH, PRODIGE, USHCN, AnClim, Craddock, RH Test V2, SNHT, ACMANT and Climatol) were submitted before the release of the list of known/inserted inhomogeneities in data sets (blind contributions). Different performance metrics and detection skill scores were calculated for monthly, yearly and decadal scales. The blind contributions ((1) for
Temperature, (2) for Precipitation) that had the best metrics considered by HOME are as follows:

- **MASH**: station and network Centered Root Mean Squared Error (CRMSE) (1), trends (1);
- **PRODIGE**: station and network CRMSE (1); CRMSE anomalies (2) and trends (2);
- **USHCN**: station and network CRMSE (1), probability of false detection (1), Heidke skill score (1);
- **Craddock test**: CMRSE anomalies (1), network CRMSE (1), probability of detection (1), Heidke skill score (1);
- **Climatol**: Heidke special skill score (2).

From the climatologists’ point of view, the most important factor to account for in homogenisation is the methods capability to improve the temporal consistency of the climate time series. In this sense, the CRMSE and the trend error metric are more relevant than detection scores such as the Heidke skill score. On the other hand, results also depend on the averaging scale at which the CRMSE is computed and the period under consideration. Domonkos (2013) provides a comprehensive discussion on the problems related to the choice of efficiency measures, and summarises the results of the blind test experiment of the HOME project. For a more thorough discussion on the assessment of the contributions performance see Venema et al. (2012). There was only one contribution (PMFred abs) that performed absolute homogenisation, and it produced much more inhomogeneous data.

After the truth was revealed to the participants, some of the blind contributions were improved in order to address problems revealed by the results. The all-over best blind contributions were MASH and PRODIGE. Although more limited regarding some tasks, Craddock also had an excellent performance. The USHCN contribution had the lowest
probability of false detection and its general performance was only slightly lower than
the other best methods. Hence, besides MASH and PRODIGE, Craddock and USHCN
were also recommended for practical use (Domonkos, 2013; Venema et al., 2012).
However, the updated ACMANT late contribution suggested that ACMANT was the
most accurate method for temperature (Venema et al., 2012). Improved
homogenisation methods were included in software packages and are available at

Some of the conclusions agreed by the participants at the end of the project can be
described as follows (HOME, 2011; Venema et al., 2012):

- There is not one ideal metric for homogenisation, but the use of detection
  scores as sole performance criterion should be discouraged;
- More homogenisation algorithms should implement the automatic use of
  metadata;
- Within the same climatic area, series share a common climate signal;
- Additive structure of the models seems fairly reasonable: temporal and spatial
  behaviours are separable;
- At monthly to annual time scales, models focus on correction of the means only;
- Covariance is time independent; residuals are not serially correlated;
- Spatial covariance can play a role. Techniques for estimation of spatial
covariance are still to be compared. Based on 1st order differentiation of the
series (MASH approach), this simple technique relies on a “smooth climate”
assumption. Many parameters have to be estimated, or based on the
variography analysis of residuals (PRODIGE approach). This technique relies
on the variogram of the residuals. It requires the estimation of few parameters
at the cost of modelling the spatial structure, which may be more complex.
6 Conclusions

The importance of having accurate and precise climate records is the main reason for the development of homogenisation methods. Many techniques proposed in the literature aim to detect artificial discontinuities. However, the correction of time series is a very delicate task, and the availability of stations’ history information is extremely important to assist the homogenisation process. Furthermore, the number of procedures to correct the artificial discontinuities is limited. In fact, some researchers choose to exclude from further analysis the inhomogeneous series and those with no metadata available, or only consider the longest homogeneous period in the analysis (e.g., Buishand et al., 2013; Costa and Soares, 2009b; de Lima et al., 2013; Santos and Fragoso, 2013).

An up-to-date list of the most important homogenisation methods for climate data series has been discussed in the previous sections, as well as several homogenisation software packages. A classification of the methods has also been proposed. An extensive review of applications is disclosed in the Supplementary Material, which may also provide guidance to climatologists and other experts to choose the most appropriate method(s) for a particular climatic region, climate variable and temporal resolution.

Based on the analysis from the comparison studies and on a thorough literature review, it is possible to enunciate the following conclusions:

- Techniques that detect and correct multiple breakpoints and work with inhomogeneous references generally perform better than other methods, namely ACMANT, MASH, PRODIGE and HOMER;
- Relative homogenisation algorithms improve the homogeneity of data;
• Absolute homogenisation methods have the potential of making the data even more inhomogeneous;
• Training of the operator when performing homogenisation is very important;
• Homogenisation algorithms developers should invest more effort into making their software easy to use and to include relevant warnings;
• Currently, automatic and semi-automatic algorithms can perform as well as manual ones;
• The use of metadata and the climatological knowledge of the operator are advantages of manual methods;
• Strengths of automatic methods are their objectivity, reproducibility, and easiness to be applied in large data sets;
• Efficiency tests need the use of simulated test datasets with similar properties to real observational datasets;
• Annual climate data sets achieve better homogenisation results than monthly data sets, which may be due to the increase of variability of data series, when the temporal resolution also increases;
• Given the low number of homogenisation studies for precipitation data and their results, the homogenisation of precipitation should be a priority.

The latter conclusion also meets the consideration provided by Auer et al. (2005), referring that precipitation data requires much greater effort, as their variability is more spatially complex. In other words, the spatial and temporal correlation between neighbouring stations should be included when performing homogenisation (Costa and Soares, 2009a; Eccel et al., 2012), particularly for precipitation.

7 Acknowledgements

The authors gratefully acknowledge the financial support of “Fundação para a Ciência e Tecnologia” (FCT), Portugal, through the research project PTDC/GEO-
Appendix

Table 1 summarises the main characteristics of the homogenisation methods, which are classified by type of approach: statistical techniques (non-parametric tests, classical tests, regression models and Bayesian approaches), homogenisation procedures and homogenisation software packages.

References


Auer, I., Bohm, R., Jurkovic, A., Orlik, A., Potzmann, R., Schoner, W., Uengersbock, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O.,


E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S.,
Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P.,
Bradsma, T., and Willett, K., 2013. Benchmarking homogenization algorithms for


Von Neumann, J., 1941. Distribution of the ratio of the mean square successive

Wald, A., and Wolfowitz, J., 1943. An exact test for randomness in the non-parametric

Wang, X. L., 2008. Accounting for Autocorrelation in Detecting Mean Shifts in Climate
Data Series Using the Penalized Maximal t or F Test. J. Appl. Meteorol. Climatol.,
47(9), 2423–2444. doi: 10.1175/2008JAMC1741.1

century European daily temperature and precipitation series. Int. J. Climatol., 23(6),
679–692. doi: 10.1002/joc.906


Yue, S., and Wang, C., 2004. The Mann-Kendall Test Modified by Effective Sample
Size to Detect Trend in Serially Correlated Hydrological Series. Water Resour. Manag.,
18(3), 201–218. doi: 10.1023/B:WARM.0000043140.61082.60