Computers & Geosciences 106 (2017) 109–117

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Optimal estimation of areal values of near-land-surface temperatures for testing global and local spatio-temporal trends

Hong Wang[a,b], Eulogio Pardo-Igúzquiza[b], Peter A. Dowd[c], Yongguo Yang[a,*]

[a] School of Resources and Geosciences, China University of Mining and Technology, XuZhou, JiangSu Province 221116, China
[b] Geological Survey of Spain, Ríos Rosas 23, 28003 Madrid, Spain
[c] University of Adelaide, Australia

ARTICLE INFO

Keywords:
Constrained spatial clustering
Temperature-altitude correlation
Regression kriging
Time series
Temperature trend detection
Global warming

ABSTRACT

This paper provides a solution to the problem of estimating the mean value of near-land-surface temperature over a relatively large area (here, by way of example, applied to mainland Spain covering an area of around half a million square kilometres) from a limited number of weather stations covering a non-representative (biased) range of altitudes. As evidence mounts for altitude-dependent global warming, this bias is a significant problem when temperatures at high altitudes are under-represented. We correct this bias by using altitude as a secondary variable and using a novel clustering method for identifying geographical regions (clusters) that maximize the correlation between altitude and mean temperature. In addition, the paper provides an improved regression kriging estimator, which is optimally determined by the cluster analysis. The optimal areal values of near-land-surface temperature are used to generate time series of areal temperature averages in order to assess regional changes in temperature trends. The methodology is applied to records of annual mean temperatures over the period 1950–2011 across mainland Spain. The robust non-parametric Theil-Sen method is used to test for temperature trends in the regional temperature time series. Our analysis shows that, over the 62-year period of the study, 78% of mainland Spain has had a statistically significant increase in annual mean temperature.

1. Introduction

Changes in near-land-surface temperatures are perhaps the most common and reliable indicator of global warming (Robeson, 1994). Near-land-surface temperature is usually measured at a finite number of irregularly spaced sampling locations comprising networks of weather stations. Although temperature measurements are affected by many factors, including longitude, latitude, altitude, slope orientation, atmospheric circulation and proximity to the sea, altitude is the most significant variable and explains most of the spatially dependent variance in temperature (Hudson and Wackernagel, 1994). In mountainous areas, altitude is the simplest direct measurement that is most highly correlated with temperature (Dodson and Marks, 1997; Benavides et al., 2007). The correlation is usually linear and negative so that temperature decreases as altitude increases with, in general, a mean gradient of 0.6 °C per 100 m of altitude (Viers, 1975). However, for large areas (several degrees of latitude), the many other factors listed above may affect the temperature in such a way that the linear relationship between altitude and temperature is much weaker because, for example, different climate factors are merged within the large area. For example, for mainland Spain the Mediterranean marine influence is different to the Atlantic marine influence. In addition, temperature measurements are biased because weather stations tend to be located at low altitudes (Rolland, 2002) and areas at high altitudes (for example, mountainous areas) are poorly represented (Robeson, 1994). This under-representation is particularly important as evidence mounts for altitude-dependent global warming (see, for example, Pepin and Lundquist, 2008 and Mountain Research Initiative EDW Working Group, 2015). Fig. 1 shows a histogram of altitudes obtained from a digital elevation model (DEM) of mainland Spain together with a histogram of the altitudes of weather stations for the year 1994. This figure shows that 25% of the surface of mainland Spain has altitudes less than 400 m and 20% of the surface has altitudes greater than 1000 m; whereas, 42% of the temperature monitoring stations (i.e., the data collection points) are located at altitudes less than 400 m and only 10% of the stations are at altitudes greater than 1000 m. This problem can be solved by using the DEM altitude as a secondary variable together with the linear relationship between altitude and temperature. However, the correlation of altitude and temperature over large areas is relatively small because of the influence of other factors such as

* Corresponding author.
E-mail addresses: wanghongcumt@hotmail.com (H. Wang), e.pardo@igme.es (E. Pardo-Igúzquiza), peter.dowd@adelaide.edu.au (P.A. Dowd), ygyang88@hotmail.com (Y. Yang).

http://dx.doi.org/10.1016/j.cageo.2017.06.002
Received 6 February 2017; Received in revised form 25 May 2017; Accepted 5 June 2017
Available online 06 June 2017
0098-3004/

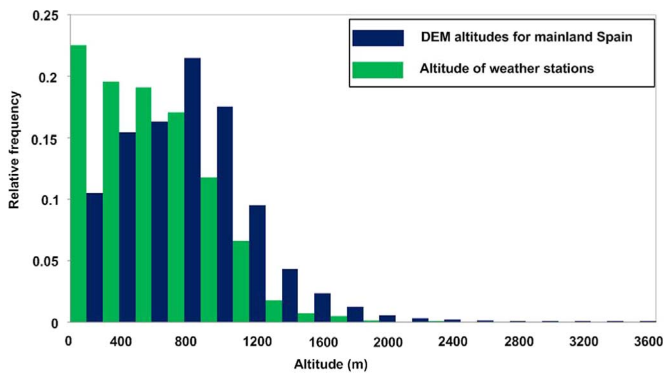

Fig. 1. Histogram of digital elevation model (DEM) altitudes for mainland Spain (blue) and altitudes of weather stations for the year 1994 (green). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

latitude, longitude, proximity to the sea, pressure and wind patterns. Thus, it is useful to identify zones in which the correlation between altitude and temperature is as strong as possible. Cluster analysis is highly suited to this purpose.

Clustering algorithms (Stooksbury and Michaels, 1991; Fovell and Fovell, 1993; Gerstengarbe et al., 1999; DeGaetano, 2001; Unal et al., 2003; Huth et al., 2008; Mahlstein and Knutti, 2010; Cannon, 2012; Tang et al., 2012; Zscheischler et al., 2012) have been used for similar, but not identical, problems to the one dealt with here. In this work we propose a new cluster method that has two novel aspects. The first is the recognition that the problem is a particular form of a constrained cluster analysis problem. The second is accounting for the spatial correlation of the data when testing the spatial correlation of the residuals of the regression of temperature on altitude for the clusters. There is no requirement for the obtained clusters to coincide with climatic regions because the definition of the latter differs from that of the obtained clusters. For example, the Spanish state meteorological agency (la Agencia Estatal de Meteorología) defines climate regions on the basis of the Köppen-Geiger Climate Classification (AEMET, 2011), which is a classification system based on the assumption that native vegetation is the best expression of climate. The purpose of our clustering approach is not to identify climate regions but to obtain regions with a high correlation between temperature and altitude. The regions resulting from the cluster analysis are not interpreted climatologically, they are used solely to obtain optimal estimates of mean areal temperatures. In addition, the regional clusters implicitly take account of secondary variables such as latitude, longitude and proximity to the sea. A detailed explanation of the methodology employed in this study is given in the following section.

2. Methodology

Geostatistical methods are widely used for mapping temperature (Hudson and Wackernagel, 1994) and estimating areal values of temperature (Ishida and Kawashima, 1993). The mean areal value of temperature over a particular area is defined by:

$$\overline{T} = \frac{1}{\chi}\int_{\chi} T(u)du \tag{1}$$

where $\chi \subset \mathcal{R}^2$ is the zone of interest of finite area and $T(u)$ is the temperature at the spatial point location $u \in \chi$.

The integral in Eq. (1) is approximated by summing the temperatures of a discrete pixel or small cell representation of the zone of interest:

$$\overline{\overline{T}} = \frac{1}{k}\sum_{i=1}^{k} T(u_i) \tag{2}$$

where k is the number of discrete cells comprising the zone $\chi \subset \mathcal{R}^2$ and $T(u_i)$ is the temperature at the i^{th} cell.

The value $T(u_i)$ is usually unknown and must be estimated from a finite set of data values. To avoid the bias introduced by the data (because of over-representation of low altitudes) and to account for the correlation between altitude and temperature, the altitude of each cell is determined from a DEM of the zone of interest. For very large zones, such as mainland Spain with a surface area of 492,072 km^2, the relationship between altitude and temperature would be obscured if data from all temperature stations were considered together. This is because the topography of the Iberian Peninsula is complex and there are many specific effects that change with latitude and longitude; for example, the different Atlantic and Mediterranean marine influences, the different frequencies of easterly winds in the Mediterranean area and westerly winds in the Atlantic area, the heating and cooling of hillsides depending on their orientation and perturbation effects such as the incursion of relatively cold air masses from the Atlantic. For these reasons we divide the zone of interest into smaller areas in which the relationship between altitude and temperature is stronger (higher negative correlation between altitude and temperature). These areas, which maximize the correlation between altitude and temperature, are identified by a new cluster analysis procedure.

Classical cluster analysis identifies groups of objects that are similar. It does so by maximising the similarity of objects (in our case, temperature measurements from weather stations) within a group and maximising the dissimilarity of different groups of objects (Gordon, 1996). There are two broad types of clustering methods: hierarchical clustering and non-hierarchical clustering. Among the non-hierarchical clustering algorithms the most widely used is the k-means algorithm. The similarity of objects is usually defined in terms of a distance (e.g., Euclidean, Mahalanobis) according to the measured characteristics of the objects.

For the problem addressed in this paper, the first difference with respect to classical clustering is that, instead of defining the similarity measure as a distance between the objects of a group, it is an objective function to be maximised or minimised. The second difference is that the problem addressed in this paper is a case of constrained clustering in which a contiguity constraint restricts the sets of allowable solutions (Gordon, 1996), i.e., the objects in each group must comprise a spatially contiguous set. Thus, given a number of groups, an object can change its membership from group A to group B if two requirements are met: (i) groups A and B are contiguous and (ii) the value of the objective function is improved. Clustering temperatures into regions with high linear correlation between altitude and temperature can thus be seen as a contiguity-constrained optimisation problem.

The first issue is the definition of clusters and contiguity. The locations of the weather stations are used as the seeds of a Voronoi tessellation of the geographic space covered by the stations. A cluster, or group, of weather stations (or of the corresponding temperature measurements) is a union of contiguous Voronoi cells and the boundary of the cluster is the outermost sequence of its constituent cell boundaries. Two clusters are contiguous if they share a boundary. A member, or object, belonging to cluster A is contiguous with cluster B if its Voronoi cell shares a boundary with the Voronoi cell of any member of cluster B. These definitions are used in the application of the contiguity constraint.

In the proposed algorithm for contiguity-constrained classification of a set of N objects (weather stations) the algorithm starts with an exhaustive classification into M groups. The manner in which this starting classification is obtained is described below. The classification is exhaustive in the sense that the N objects have been classified and each belongs to one of the M groups.

For any given configuration of groups $(G_1, \ldots, G_M)$ the objective function, $OF(G_1, \ldots, G_M)$, of the configuration is defined by: