Spatial Grouping of Swiss Climatological Stations using Cluster Analysis

English extract from "Begert M. 2008. Die Repräsentativität der Stationen im Swiss National Basic Climatological Network (Swiss NBCN). Arbeitsberichte der MeteoSchweiz, 217", describing the hierarchical clustering algorithm used for spatial grouping of climatological stations.

1 Introduction

Climatology often makes use of the spatial grouping of observation stations. Mostly the aim is to divide a major area into different climate regions which serve as a basis for addressing various issues. The description of the different climates of a region is a frequent topic. When grouping stations, several parameters such as temperature, precipitation and sunshine duration are included in the analysis. Recent examples include the climate typology of the People's Republic of China (Strauch, 2004) or of Turkey (Unal et al., 2003). Variability studies of different climate elements are another application field in climatology where spatial grouping can be useful (e.g. Baeriswyl et al., 1997; Matulla et al., 2003; Auer et al., 2007). In such studies, regions are established on the basis of individual parameters and the development of climate elements is analysed and described over a specific period of time. A further important application area is the planning of observing networks. There the question is quite often to what extent a network can be reduced without having to renounce a sufficiently detailed coverage of spatial variability. DeGaetano (2001) for instance uses a cluster analysis technique to determine the number of stations of an existing network necessary to adequately observe the US climate. The statistical method for most studies consists of various types of cluster analysis or the principal component analysis with rotation of the principal components.

MeteoSwiss divides Switzerland into several different sets of subregions, depending on different application purposes. The three most important groups are: firstly, the regions used by the forecasting service to issue regionally focused warnings of severe weather phenomena, secondly, the so-called "OPKO" regions which are used for the monitoring of forecasting quality, and thirdly, the climate regions according to Schüepp et al (1981) used mostly in climatology. The Schüepp regions divide Switzerland up into 12 large and 53 small climate regions; they have been used for the most diverse scientific issues, among them the re-definition of the basic Swiss observation network for climate monitoring (Begert et al., 2007). However, there it became obvious that a grouping according to specific parameters would make sense, since the different variables have different requirements as to network density. In addition, the methodology of the division according to Schüepp can no longer be retraced and the grouping is based on an out-dated station network. The aim of the present study is to develop a largely objective methodology to group the MeteoSwiss stations according to parameters, based on the data series of the recent past. Stations with a similar monthly variability should be grouped together independently of their altitude above sea level. This categorization should provide the basis for defining the climate monitoring network and for supporting other network concepts in the future. Since climatological variables such as temperature and precipitation present mostly gradual variations and no abrupt changes in their spatial distribution, the division should not create fixed groups with strict boundaries. The aim is to determine an "adequate" number of station groups or climate regions which can be adapted depending on the current issue. Thus the methodology has to be flexible enough to take different degrees of specification into account.

The following paper presents the data set used in this study; the selection of stations and data quality is discussed. Using the examples of temperature and precipitation, the methodology of creating groups is developed. The results for temperature are presented in more detail, since the smaller number of stations allows for a more detailed presentation. The last chapter contains a short discussion of the results and the application of the method at MeteoSwiss.

2 Data

Currently, MeteoSwiss operates a network of several hundred ground based stations with different observation programmes. The climate division has a total of 90 stations with long-term measurements at its disposal where various climate-relevant parameters are collected simultaneously. In addition, precipitation is recorded at about 320 stations. Some of the stations have been operative since 1864 when the Swiss Academy of Sciences created the first official observation network. However, this network has been subject to more or less constant change: new stations were added, stations had to be relocated or were closed down. This resulted in long data series being rare and in existing data series containing artificial shifts and trends. These are due to changes in observing conditions such as station relocation, changes in the types of instruments used or in the environment. Within the framework of the project NORM90 (Begert et al., 2003) the data series of the most important variables of all current stations from 1961 were examined with respect to such inhomogeneities, which were then corrected. The homogenized monthly temperature and precipitation series derived from this project form the data basis for the present study.

The data set of NORM90 contains 93 temperature and 439 precipitation series. A number of stations, however, became operative only after 1961, some of them presenting only short observation periods. Since the analysis of the similarity of stations requires a sufficiently long time period to avoid any influence of short-term, random variability on the result, only complete date series are taken into account. Consequently, 55 temperature stations and 297 precipitation stations are used in the analysis. These stations are distributed more or less equally over Switzerland and cover altitudes between 203 m and 2690 m (precipitation) or 3580 m (temperature).

3 Statistical methods and results

3.1 Processing data for measuring distances

The statistical method of cluster analysis – along with the principal component analysis – is best suited for assigning climate stations to similar groups. In an nxp data matrix with n observations (measuring time) and p variables (stations), $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$, the variables with similar observations should be grouped. In the present case, correlation is suitable as a distance measure between the variables. Data series with a high correlation should be subsumed in groups. For greater distances to present greater values the distance between two stations i and k is defined as follows: $d(\underline{x}_i, \underline{x}_k) = 1 - \left| corr(\underline{x}_i, \underline{x}_k) \right|$. We do not use the following definition which is also suggested in literature: $d(\underline{x}_i, \underline{x}_k) = \sqrt{1 - corr^2(\underline{x}_i, \underline{x}_k)}$, since its outcome is identical as demonstrated in sample analyses. To determine correlation, the rank correlation coefficient according to Spearman is calculated in order to minimize the influence of possible outliers. In principle, the grouping technique presented here can also be carried out using the product-moment correlation coefficient according to Pearson or other methods for measuring distance.

The selection of input data has a decisive impact on the results when employing methods of exploratory data analysis such as cluster analysis. In the present case, the exact formulation of the inquiry is essential when processing input data, before correlations are calculated. The results vary according to whether and how initial data have been normalized and/or standardized. In the present case, cluster analysis should produce groups of stations whose variability is identical from one month to the next, independent of their altitude above sea level. Therefore, the monthly values for temperature and precipitation have to be standardized separately for each month before calculating correlations. Subtracting the mean value for each month eliminates the altitude factor of the measurement and evens out differing amplitudes in the course of the year. The monthly standardization of data ensures that seasonal differences in variability are removed which could lead to individual seasons dominating the correlation. The data retain spatial and temporal differences from a mean state whose similarities define their inclusion in a group.

3.2 Clustering method

Scientific literature describes a multitude of clustering methods: most approaches can be assigned to certain main categories. As opposed to partitioning methods, the group of hierarchical methods tries to find a flexible answer to the question of classification without requiring a fixed number of groups at the beginning. In the case of p variables a distribution in k groups is suggested for each number $k \in \{1,..., p\}$. This method is therefore suitable for the present inquiry where it is not clear initially how many regions should be formed. To maintain control on correlations within the group for all k groups, an agglomerative method with “complete linkage” is used. The "complete linkage" method uses the greatest of all distances between two points of two clusters to decide which clusters are joined next in an agglomerative process. Thus the weakest correlation of the newly created group can be recognized immediately from the distance indicated. A graphical representation of all distances indicates how many clusters must be formed to maintain a specific correlation within the groups. Furthermore, this method is referred to as AGNES (agglomerative nesting) according to the function implemented in the statistics package R (www.r-project.org), implying the use of "complete linkage" to join the clusters. The method proposed is based on the work of DeGaetano (2001) and was similarly used in Switzerland for the spatial grouping of snow stations (Laternser, 2002).

4 Results

4.1 Clustering according to temperature

The result of a hierarchical clustering can be displayed as a so-called dendrogram. This display demonstrates the progressive formation of groups in response to the growing distance between the groups created. Figure 1 shows the dendrogram for the temperature series. Observation stations are represented here and subsequently by their codes consisting of three letters. Instead of the usual distances, the y axis displays the relevant correlation coefficients. The figure therefore exemplifies for every joining of subgroups the lowest correlation within the group. On the other hand it is possible – although at some expense – to indicate for each correlation within any group the number of required clusters. This can be done more comfortably using the display in figure 2 which shows correlations directly in relation to the number of clusters. This figure demonstrates that the correlation within groups shows a marked increase at first as the number of clusters grows, however it increases only slowly when the total of 11 clusters is reached. The last significant correlation leap occurs between 10 and 11 clusters. This leads to the conclusion that it would make sense - according to the hierarchical cluster analysis - to form 11 station groups. The lowest correlation within any group is still very high at 0.93, the maximum correlation between existing stations amounting to 0.99. Figure 3 shows the spatial display of the 11 groups created. The individual stations are placed on the map according to their coordinates and coloured to identify their association with a group.

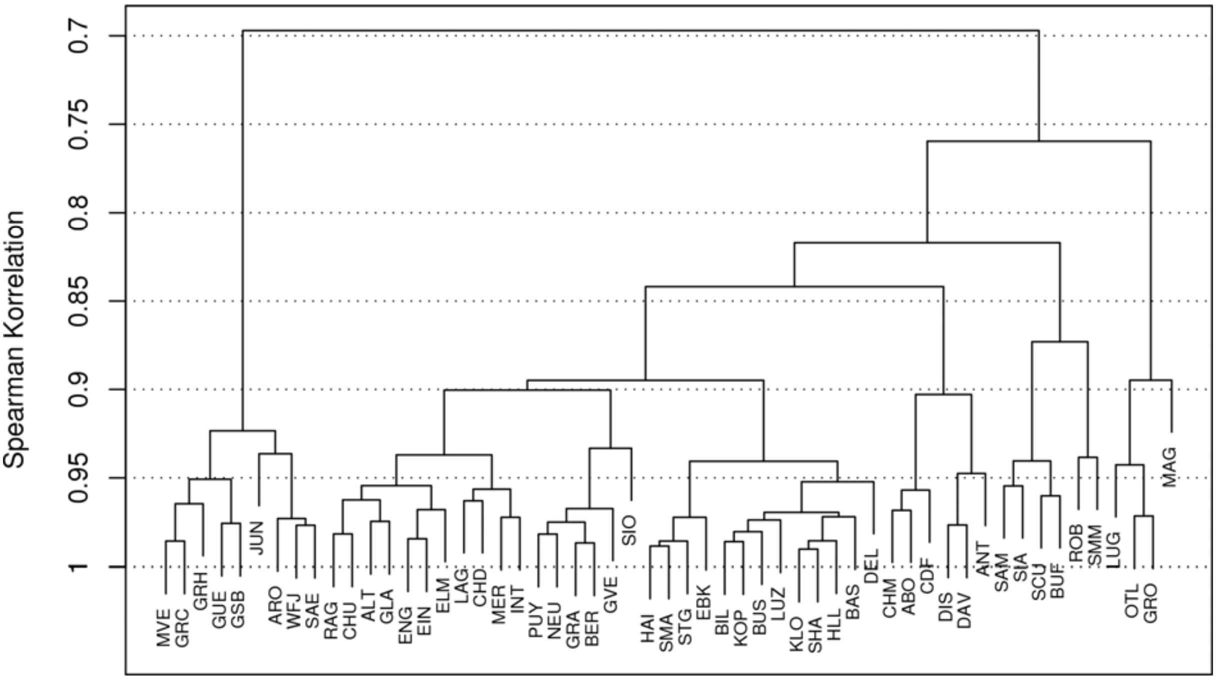

Figure 1: Dendrogram of hierarchical clustering for 55 stations with temperature measurement between 1961 and 2006. Station names are given as national abbreviations.

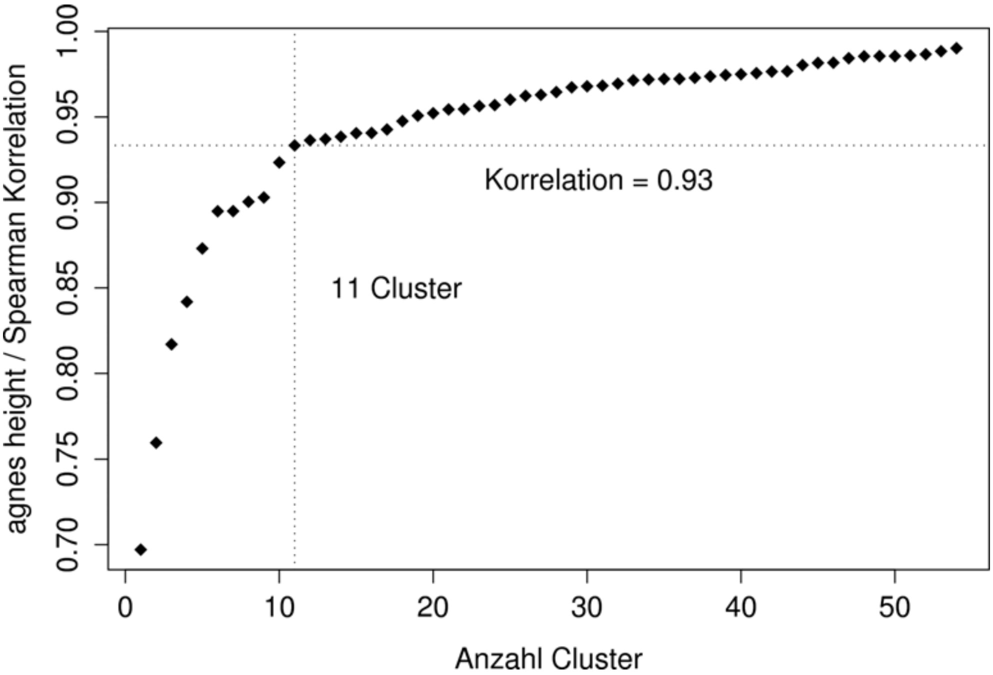

Figure 2: Dendrogram heights and lowest corrlation within the groups, respectively, against the number of clusters in the hierarchical clustering of temperature series. Dashed lines indicate the „optimal“ number of clusters and the according correlation.

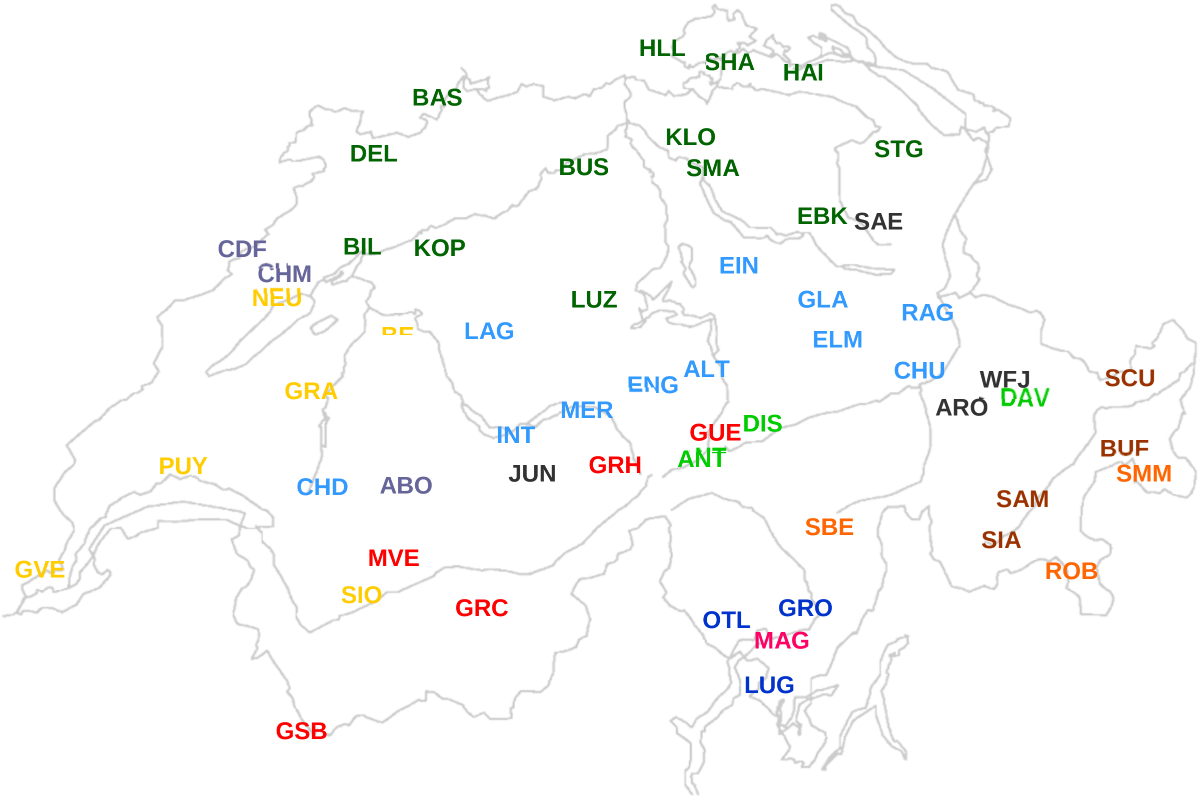

Figure 3: 11 groups of staions with temperature measurements according to the proposed heararchical clustering with complete linkage.

Where temperature is concerned, figure 1 reveals that altitude above sea level on the one hand and position relative to the Alps on the other are central factors for the formation of groups. High-altitude stations such as Col du Grand St-Bernard (GSB), Weissfluhjoch (WFJ) or Säntis (SAE) form a group which takes shape early on in the course of the clustering process and which is united with the rest of the stations only at the very end. The same is true for the low and high altitude stations of the South Side and of the North Side of the Alps. From a climatological point of view these groups are plausible, particularly the formation of 11 groups as shown in figure 3. This sustains confidence in the method selected. Some interesting observations should be briefly discussed here: The fact that Magadino (MAG) forms a group of its own can be attributed to the influence of the cold air pool in the region of Magadino. However, Magadino is basically part of the low-altitude stations of the South Side of the Alps, a fact that is confirmed by the dendrogram. The grouping of Adelboden (ABO) with the Jura stations La Chaux-de-Fonds (CDF) and Chaumont (CHM) seems interesting, since Adelboden could be associated at first sight with the stations of the north slope of the Alps such as Château-d'Oex (CHD) or Elm (ELM). The present result, however, emerges also when using other methods, and experience has proved that Adelboden is more exposed to influences from the south-west than the rest of the stations from the north slope of the Alps. The separation of Davos (DAV) and Arosa (ARO) also comes as a surprise, although the data series show a strong correlation at 0.97. The reason for this is that Davos is first joined with Disentis (DIS), Arosa with the cluster formed by Weissfluhjoch and Säntis. By this process, Arosa is subsequently teamed with the high-alpine, Davos with the lower alpine stations of the Eastern Alps, a correlation difference of $4 \cdot 10^{-4}$ being decisive. As a matter of fact, Arosa is located at a noticeably higher altitude than Davos, which makes the separation acceptable from a climatological point of view.

4.2 Clustering according to precipitation

Where precipitation is concerned, the display of the clustering process by dendrogram is slightly confusing due to the much greater number of 305 available stations. Nevertheless, as with temperature, it is possible to infer the adequate number of groups by looking at the lowest correlations within the groups against the number of clusters. Any decision based on this figure, however, is less unequivocal. There are several possibilities, the last significant correlation leap occurring between 31 and 32 clusters. With a reliable correlation of at least 0.79 within any group, this is lower than for temperature. Initially, this comes as a surprise since the choice of stations is much greater and a higher number of clusters is selected. The explanation lies with the significantly higher spatial variability of precipitation. Figure 4 displays the resulting 32 clusters. As with temperature, the precipitation study also presents very plausible groups, while in this case the geographical position, not altitude above sea level, is the central factor for association with a group. The analysis reveals some remarkable details about Swiss precipitation climatology. For example the group of western stations in the rain shadow of the Jura chain, reaching from Genève (GVE) to Payerne (PAV), the groups with substantial rainfall, located in and around the Alpstein region or the division of the Engadine stations into those more subject to north side barrier effects such as Scuol (SCU) or Martina (MAT) and those which receive most precipitation from the south such as Samedan (SAM) or Segl-Maria (SIA). Small details enhance confidence in the method employed: e.g. the fact that the station Grimsel Hospiz (GRH), situated north of the Grimsel pass, is grouped with Guttannen (GTT) on the north side, or that the stations Binn (BIN), Simplon-Dorf (SIM) and Saas Fee (SAF), frequently receiving their precipitation from the south-east, are separated from the rest of the Valais stations.

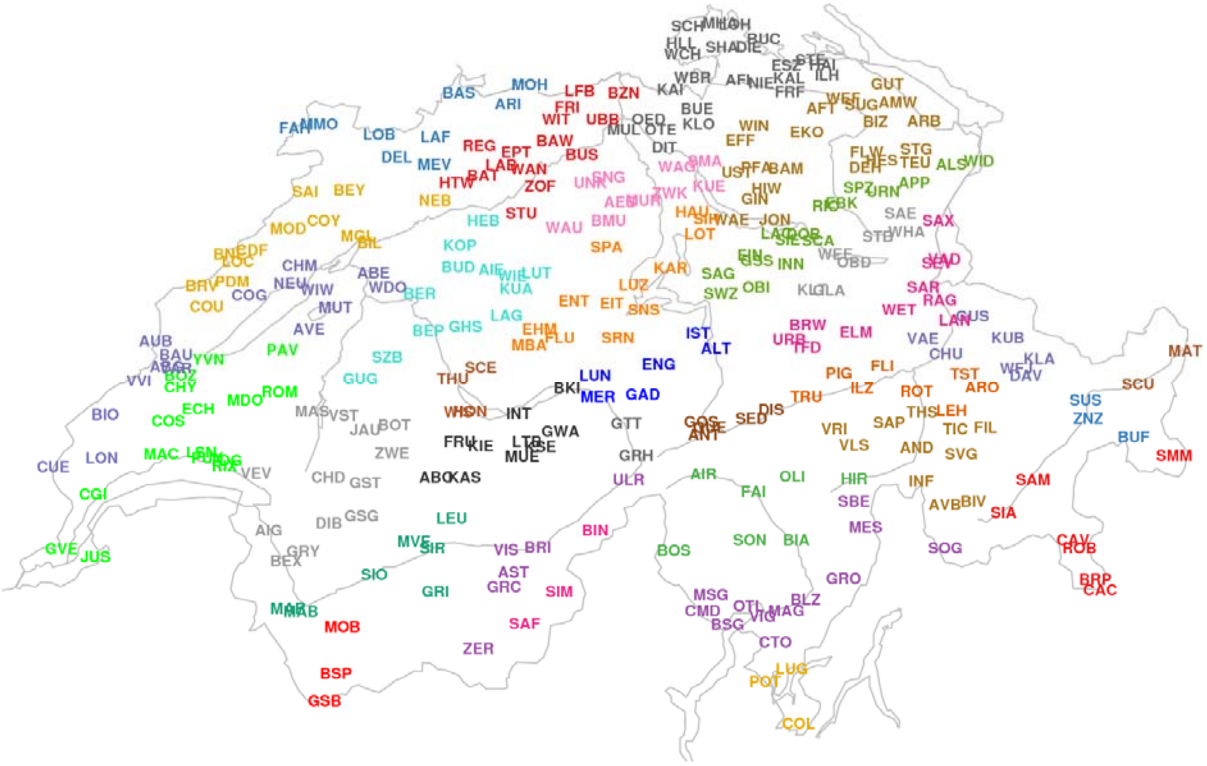

Figure 4: Grouping of 305 stations with precipitation measurements according to the proposed heararchical clustering with complete linkage.

5 Conclusions

A cluster analysis is proposed to divide up the network of climatological stations into similar groups. The process was conducted separately for temperature and precipitation. It could be demonstrated that a hierarchical, agglomerative method with "complete linkage" resulted in plausible station groups for both observation parameters. The groups define regions where variables demonstrate a similar temporal development since 1961. Those regions can therefore be regarded as parameter-specific climate regions with a focus on climate monitoring.

The clustering method leads to plausible station groups, however due to input variables which present only gradual spatial variations, they are not very stable. A modification of the time period examined or of the number of stations included can easily result in groups which are slightly different, especially in larger areas with a high correlation, as on the Swiss plateau. Marginal stations can easily be assigned to another group or individual groups are divided up and joined with others. This fact has to be taken into account when interpreting and using the results.

At MeteoSwiss the results of this study were used to analyse the station distribution of the Swiss climatological reference network (Swiss NBCN; Begert et al., 2007). The results differed depending on the parameter. The network of the climatological stations clearly meets the requirements for observing temperature development. The 11 regions established are in their majority well covered by two or more stations. Within the groups, different altitude levels are represented. Where precipitation is concerned, however, 11 out of the 32 regions established were not covered. The larger spatial variability of this parameter required the stations to be complemented by stations from the precipitation observing network (Begert, 2008).

6 References

Auer et al. 2007. HISTALP – historical instrumental climatological surface time series of the Greater Alpine Region. Int. J. Climatol. 27: 17-46.

Baeriswyl PA, Rebetez M. 1997. Regionalization of precipitation in Switzerland by means of principal component analysis. Theor. Appl. Climatol. 58: 31-41.

Begert M, Seiz G, Schlegel T, Musa M, Baudraz G, Moesch M. 2003.Homogenisierung von Klimamessreihen der Schweiz und Bestimmung der Normwerte 1961-1990. Schlussbericht des Projekts NORM90. Veröffentlichung der MeteoSchweiz, Nr.67, Zürich.

Begert M, Seiz G, Foppa N, Schlegel T, Appenzeller C, Müller G. 2007. Überführung der klimatologischen Referenzstationen ins Swiss NBCN.

Begert M. 2008. Die Repräsentativität der Stationen im Swiss National Basic Climatological Network (Swiss NBCN). Arbeitsberichte der MeteoSchweiz, 217.

DeGaetano AT. 2001. Spatial grouping of United States climate stations using a hybrid clustering approach. Int. J. Climatol. 21: 791-807.

Kaufman L, Rousseeuw PJ. 1990. Finding groups in data: an introduction to cluster analysis. Wiley, N.Y.

Laternser M, Schneebeli M. 2002. Spatial grouping of Swiss snow stations by clustering. Chapter 4 in “Snow and avalanche climatology of Switzerland” by Martin Christian Laternser, PhD thesis ETH Zürich No 14493.

Matulla C, Penlap EK, Haas P, Formayer H. 2003. Comparative analysis of spatial and seasonal variability: Austrian precipitation during the 20th century. Int. J. Climatol. 23: 1577-1588.

Schüepp M, Gensler G. 1980. Klimaregionen der Schweiz. In: Müller G. 1980. Die Beobachtungsnetze der Schweizerischen Meteorologischen Anstalt. Konzept 1980. Arbeitsbericht der Schweiz. Met. Zentralanstalt, Nr. 93, Anhang Ib, Zürich.

Strauch J. 2004. Klimatypisierung der Volksrepublik China auf der Basis einer erweiterten nicht-hierarchischen Clusteranalyse. Diplomarbeit am geographischen Institut der Justus-Liebig-Universität, Giessen.

Unal Y, Kindap T, Karaca M. 2003. Redefining the climate zones of Turkey using cluster analysis. Int. J. Climatol. 23: 1045-1055.