Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 9 (2012) 897 – 906

Procedia Computer Science

International Conference on Computational Science, ICCS 2012

Climate classifications: the value of unsupervised clustering

Jakob Zscheischler[a,b,*], Miguel D. Mahecha[a], Stefan Harmeling[b]

[a] *Max Planck Institute for Biogeochemistry, PO Box 100164, 07701 Jena, Germany*
[b] *Max Planck Institute for Intelligent Systems, Spemannstr. 38, 72076 Tübingen, Germany*

Abstract

Classifying the land surface according to different climate zones is often a prerequisite for global diagnostic or predictive modelling studies. Classical classifications such as the prominent Köppen–Geiger (KG) approach rely on heuristic decision rules. Although these heuristics may transport some process understanding, such a discretization may appear “arbitrary” from a data oriented perspective. In this contribution we compare the precision of a KG classification to an unsupervised classification (k-means clustering). Generally speaking, we revisit the problem of “climate classification” by investigating the inherent patterns in multiple data streams in a purely data driven way. One question is whether we can reproduce the KG boundaries by exploring different combinations of climate and remotely sensed vegetation variables. In this context we also investigate whether climate and vegetation variables build similar clusters. In terms of statistical performances, k-means clearly outperforms classical climate classifications. However, a subsequent stability analysis only reveals a meaningful number of clusters if both climate and vegetation data are considered in the analysis. This is a setback for the hope to explain vegetation by means of climate alone. Clearly, classification schemes like Köppen-Geiger will play an important role in the future. However, future developments in this area need to be assessed based on data driven approaches.

Keywords: clustering, k-means, multivariate statistics, Köppen-Geiger climate classification, PCA

1. Introduction

Discretization techniques are important to get an intuitive understanding of complex geo-spatial data sets. A typical example are global climate classifications like the Köppen-Geiger approach [1, 2, 3] and its recent updates [4]. Classifications of this kind provide an intuitive way to discretize the Earth land surface properties, at the cost of a loss of information. While the Köppen-Geiger (KG) classification certainly reflects several decades of environmental and geographic research, today we can ask the question how well KG performs when being compared to modern clustering techniques. These techniques aim to identify points in the data cloud which could be used as ”predictors” for other samples in a class. Another critical issue is that the original objective of Köppen and Geiger was to classify vegetation zones based on temperature and precipitation patterns only. Later on, Thornthwaite criticized the lack of rational justifications for most of the boundaries separating KG classes [5]. He claimed that the boundaries in a climate classification should relate to truly active climatic factors.[1]. Along these lines one may ask if it is effectively

[*] Corresponding author, *jzsch@bgc-jena.mpg.de*

[1] As a side remark it is worth noting that he questioned the uncritical reception of the KG classification and pointed out that people tend to evaluate climatic classification according to the ease with which they can be applied [5].

1877-0509 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.procs.2012.04.096

possible to delineate vegetation zones based on temperature and precipitation only. This ecophysiological question becomes important nowadays, where additional variables such as radiation patterns are becoming available and could be used in novel classification approaches.

Clearly, the critique *sensu* [5] focuses on the lack of process understanding in the available classifications. However, adopting a data oriented perspective the question is rather if the heuristics in KG could be directly retrieved from the observations. Along these lines, Cannon [6] recently showed that modern data driven approaches may equally allow to derive relevant discretization rules. The author provides another climate classification based on a multivariate regression tree and argues that he can yield a higher performance as measured by a quantity called *expected variance* (*EV*).

The main problem of climate classifications, however, remains unaddressed by [6]: climate classifications seek to classify vegetation exclusively based on climate conditions. It is implicitly assumed that vegetation is a function of climate only; the idea is to identify “climate thresholds” that divide the vegetation space into discrete classes. But how can we find these jumps in an unsupervised fashion directly in the data? Ecological arguments against this idea can be found in the fact that vegetation is also influenced by history preventing e.g. seeds establishment or leading to extraordinary rates of mortality. Similarly, rapid changes in climate conditions may not allow vegetation to adapt. Anthropogenic influence additionally causes shifts in vegetation classes. Finally, there may exist certain climate conditions which favor two or more different vegetation types - a phenomenon known as bistable system [7, 8, 9].

In this paper we investigate a multivariate data cube, consisting of different combinations of climate variables and remote sensing vegetation indices to examine whether climate and vegetation classes coincide. We apply unsupervised clustering techniques and look into the differences between clustering of climate variables versus vegetation variables. In order to compare different clustering results we use a distance measure introduced in [10]. We also discuss the problem of finding the right number of clusters. Our aim here is to discover discrete climate classes in an unsupervised, purely data driven way and to compare the performance skills to the KG approach.

2. Data

We use three climate variables, namely the updated CRU[2] and GPCC[3] data sets which were used for the updated KG classification in [4]. We additionally use short wave radiation from ERA-Interim [11]. Furthermore, we use two vegetation variables. Firstly, we include the *Enhanced Vegetation Index* (EVI) which is known to be responsive to structural variations in the canopy, including leaf area index, canopy type, plant physiognomy, and canopy architecture [12, 13] (EVI data is taken from MODIS[4]). Finally we use the *Fraction of Absorbed Photosynthetically Active Radiation* (FAPAR) which is directly related to primary productivity (the used data is described in [14]).

All data sets are used only over land on a spatial resolution of 0.5 degrees with averaged monthly values. We work with an average year of the biggest intersection of years where all data sets are available, 2001-2007. The techniques presented in this work can easily be applied on larger data sets with both varying spatial and temporal resolution.

Both EVI and FAPAR have difficulties with snow. Taking the minimum intersection of full data coverage we end up with 54580 pixels, excluding some parts of northern Siberia and Greenland. We introduce the following shortcuts:

- 1 Air Temperature T,
- 2 Precipitation P,
- 3 Downward Shortwave Radiation SW,
- 4 Enhanced Vegetation Index EVI,
- 5 Fraction of Absorbed Photosynthetically Active Radiation FAPAR

We denote the data cube by X. With X^v, $v = 1, \dots, 5$, we denote the different variables. $X^v_{p,\cdot}$ denotes a time series on pixel p ($p = 1, \dots, 54580$) and $X^v_{\cdot,t}$ a time step at month t ($t = 1, \dots, 12$) of variable v.

3. Tools

The data sets will be analyzed by several tools, which are listed in Table 1. In the following subsection we briefly introduce all tools used in this paper.

[2] available at `http://www.cru.uea.ac.uk/cru/data/hrg/`

[3] available at `ftp://ftp-anon.dwd.de/pub/data/gpcc/html/fulldata_download.html`

[4] available at `http://modis.gsfc.nasa.gov/`

tool	task	result	interpretation
k-means	how does the data cluster?	assignments	
EV	is a clustering good?	$[0, 1]$	1 is perfect
VI	are two clusterings different?	$[0, \log(k)]$	0 implies identical
Instab	is a clustering for fixed k stable?	$[0, \log(k)]$	0 is stable

Table 1: Tools to cluster and analyze clustering for this paper.

3.1. Preprocessing

We will apply k-means to several subsets of the variables. This will involve calculating distances between vectors, where each coordinate has a different unit (e.g. for temperature and precipitation). To alleviate the effects of different scales, we normalize each variable by its standard deviation, more precisely, we divide every X^v by the standard deviation of X^v taken over all pixels and months,

$$X^v = \frac{X^v}{\mathrm{std}(X^v)} \,. \tag{1}$$

Since the distance calculations involve only differences between variables of the same type, we do not have to remove the mean.

To investigate the differences between climate and vegetation variables we will perform k-means clustering both on the whole data set (including all variables) and on subsets (including only some variables). Thus we combine variables in the following way: e.g. considering temperature and precipitation, we use $[X^1, X^2]$ and perform k-means on the resulting data set of size $12s \times 54580$ where s is the number of variables included ($s = 2$ in the example).

3.2. k-means

Given a set of n data points $x_1, \dots, x_n \in \mathbb{R}^d$ and a fixed number k of clusters to construct, k-means minimizes the clustering objective function:

$$Q(c_1, \dots, c_k) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1,\dots,k} \|x_i - c_k\|^2 \tag{2}$$

where $c_1, \dots, c_k$ denote the centers of the k clusters. In our case, $n = 54580$ and $d = 12s$. We use the implementation of Gehler (2007) [15] based on [16]. A similar fast implementation of k-means which exploits the triangle inequality was introduced in [17, 18] and used in [19]. Note that k-means does not determine the “best” number of clusters k automatically. Instead we compute k-means for the range $k = 3, \dots, 40$ and study Q dependent on k. Additionally, we analyze the stability by considering different subsets of X for varying k (see Section 3.5).

*3.3. Explained predictand variance (*EV*) — a quality measure for clusterings*

[6] introduced the *explained predictand variance EV* which can be interpreted as a measure of performance for clusterings and classifications. Let $WCSS_k$ be the within-cluster sum of squares of a clustering with k clusters,

$$WCSS_k = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \,, \tag{3}$$

where the S_i are disjoint sets containing the data points assigned to the i-th cluster with mean μ_i. Then we can measure the quality of a clustering by EV which is defined by

$$EV_k = 1 - \frac{WCSS_k}{WCSS_1} \,. \tag{4}$$

If EV_k is equal to one, it means that $WCSS_k = 0$. This can only be achieved for a data set with exactly k data points. So on larger data sets we expect EV to be between zero and one.

3.4. Variation of information (VI) — a distance measure for clusterings

[10] introduced the *variation of information* (*VI*), which is an information theoretic index that defines a metric on the space of clusterings. Even for clusterings with different a number of clusters *VI* can provide a distance.

A clustering C is a partition of a set of points, or data set D into mutually disjoint subsets $C_1, C_2, \dots C_K$. Formally

$$C = \{C_1, C_2, \dots, C_K\} \quad \text{such that} \quad C_k \cap C_l = \emptyset \quad \text{and} \quad \bigcup_{k=1}^{K} C_k = D\,. \tag{5}$$

Let n and n_k be the number of data points in D and in C_k, respectively. Then

$$n = \sum_{k=1}^{K} n_k\,. \tag{6}$$

We assume $n_k > 0$, i.e. empty clusters are ignored.

Consider a second clustering $C' = \{C'_1, C'_2, \dots, C'_{K'}\}$ with cluster sizes n'_k. How does *VI* measure the distance between C and C'? Criteria for comparing clusterings are usually based on the so-called *confusion matrix* (also called *association matrix* or *contingency table*) of C and C'. The confusion matrix is a $K \times K'$ matrix, whose (k, k')-th element is the number of points in the intersection of clusters C_k of C and $C'_{k'}$ of C',

$$n_{kk'} = |C_k \cap C'_{k'}|\,. \tag{7}$$

VI is now defined via random variables that are determined by the different clusterings. We start with some well-known information theoretic quantities which we translate into the setting of clusterings. If one picks a point of D at random and each point has an equal probability of being picked, the probability that the chosen point is in cluster C_k is

$$P(k) = \frac{n_k}{n}\,. \tag{8}$$

This defines a discrete random variable taking K values associated to the clustering C. The uncertainty in the picking process is given by the entropy of the random variable,

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k)\,. \tag{9}$$

Furthermore, we can define the joint probability $P(k, k')$ which values how likely it is that a point belongs to C_k in clustering C and to $C'_{k'}$ in clustering C'

$$P(k, k') = \frac{|C_k \cap C'_{k'}|}{n} = \frac{n_{kk'}}{n}\,. \tag{10}$$

The mutual information $I(C, C')$ between two clusterings is defined to be the mutual information between the associated random variables

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P(k')}\,. \tag{11}$$

Then we can define *VI* to be the

$$\begin{aligned} VI(C, C') &= H(C) + H(C') - 2I(C, C') & (12)\\ &= H(C, C') - I(C, C'). & (13) \end{aligned}$$

which is the difference between the joint entropy and the mutual information. [10] has shown that *VI* is a metric on the space of clusterings. Furthermore, *VI* is n-invariant (independent of the number of samples) and bounded by $2 \log K$ if both C and C' have at most K clusters with $K \leq \sqrt{n}$. More properties and their proofs can be found in [10]. The most important property of *VI* is that it can measure distances between clusterings with different number of clusters.

*3.5. Instability of a clustering (*Instab*)*

Another quality measure for clustering is *stability*. Given two samples from the same distribution we expect k-means to cluster the data similarly, if k was chosen appropriately. For wrong k we could get different clusterings for different samples, see [20]. We define *instability* of a clustering by the expected difference between two clusterings $C_k(S_n)$, $C_k(S'_n)$ on different data sets S_n, S'_n of the same size n, i.e.

$$Instab(k,n) = \mathbb{E}\left[d(C_k(S_n), C_k(S'_n))\right] , \tag{14}$$

where $d(\cdot,\cdot)$ is some distance measures between clusterings. The expectation is taken with respect to the drawing of the two samples. There exist various methods to compute stability scores. We will use the procedure described in Algorithm 1 following [20]:

Algorithm 1 Compute instability scores

```
Given: a set of data points $X$, a clustering algorithm $A$ that takes the number of clusters $k$ as input
for $k = 2, \ldots, k_{\max}$ do
    Generate subsamples $X_b$ of the original data set
    for $b = 1, \ldots, b_{\max}$ do
        cluster the set $X_b$ with algorithm $A$ into $k$ clusters to obtain clustering $C_b$
    end for
    for $b, b' = 1, \ldots, b_{\max}$ do
        compute pairwise distance $d(C_b, C'_b)$ between these clusterings
    end for
    compute instability scores as the mean distance between clusterings $C_b$
        $\widehat{Instab}(k) = \frac{1}{b_{\max}^2} \sum_{b,b'=1}^{b_{\max}} d(C_b, C_{b'})$    (15)
end for
```

As distance measure we use variation of information (*VI*) which we described in Section 3.4. We need, however, a protocol to compare clusterings on different data sets X_b. We overcome this by comparing the clusterings on the extended data sets $X_b \cup X'_b$, i.e. we estimate two clusterings one for X_b and one for X'_b and then evaluate them on the union. For a given clustering characterized by k means, we assign new data points to the cluster with the closest mean [20]. The instability score is bounded from above by the maximal distance between clusterings. An instability score of 0 describes a stable clustering which is desired.

3.6. Principal Component Analysis (PCA)

PCA is widely used in multivariate statistics. We want to use it as a method for dimensionality reduction and as a measure for the complexity or nonlinearity of the various data streams. We briefly explain PCA here. Let X be a data matrix with m rows representing different features (e.g. months in our case) and n columns representing the samples. The singular value decomposition (SVD) of X is given by $X = W\Sigma V^T$, with an orthogonal $m \times m$-matrix W containing the eigenvectors of XX^T and a rectangular diagonal matrix Σ containing the singular values of X sorted by decreasing magnitude. V contains the eigenvectors of $X^T X$. The principal components are given by

$$Y = W^T X = W^T W \Sigma V^T = \Sigma V^T \tag{16}$$

where now the first row contains the data projected into the direction of maximum variance, etc. We can obtain a reduced-dimensionality representation by projecting X onto the first l singular vectors, W_l.

$$Y_l = W_l^T X = \Sigma_l V^T \tag{17}$$

Which corresponds to ignoring the lower rows of Y and only keeping the first l rows. To measure the complexity of the data we analyze the spectrum of the covariance matrix of the data, i.e. its eigenvalues. A scaled version of their eigenvalues are contained in $\Sigma\Sigma^T$. We rescale and sort them such that $\lambda_1 \geq, \ldots, \geq \lambda_m$, $\sum_{i=1}^{m} \lambda_i = 1$. To project the data onto the first two components we first have to reorder the columns of V according to the order of the eigenvalues.

4. Results

4.1. Comparing the performance of k-means clustering via EV

First, we compare *EV* (see Section 3.3) of clusterings of *k*-means with the values of *EV* reported in [6]. Note that we do not adjust for the seasons between northern and southern hemisphere which is usually done in KG, *cf.* also [6]. In our case, northern hemispheric summers coincides with southern hemispheric winters and vice versa. Consequently, regions with similar behavior may end up in different clusters depending on their hemispheric location. An adjustment of this issue, however, would lead to discontinuous clusters along the equator which we want to avoid here.

We choose the same number of clusters used in [6], namely 5,13, and 30 to report values of *EV* for different subsets of variables. The analysis is then performed based on different subsets, i.e. {P}, {T}, {P, T}, {EVI, FAPAR} and {P, T, SW, EVI, FAPAR} and summarized in Figure 1 where *Z* denotes the results of our analysis. In terms of statistical prediction, *k*-means yields better results than both Köppen-Geiger and Cannon's multivariate regression tree classification. This results is not very surprising given that *k*-means is precisely optimizing this quantity. On the contrary, it is noteworthy that both other classifications perform relatively well too. Nonetheless, if one is purely interested in the statistical performance, the *k*-means approach is to be preferred.

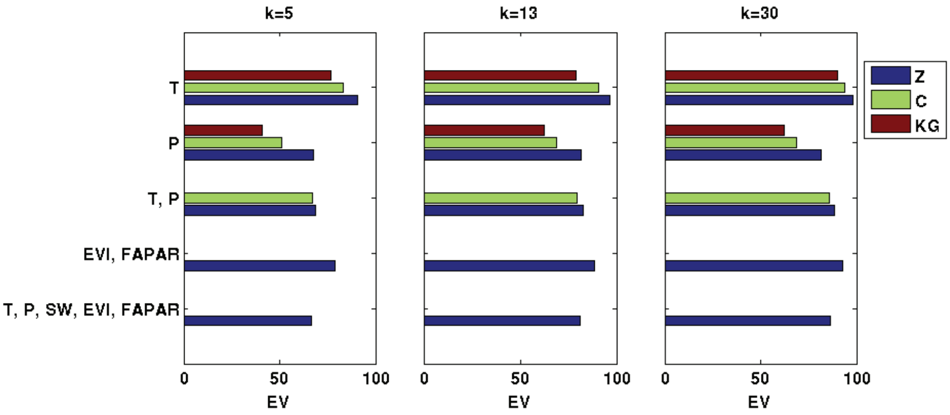

Figure 1: Expected variance (*EV*) for different scenarios. KG=Köppen-Geiger, C=Cannon, Z=our analysis.

4.2. Climate versus vegetation

Second, we investigate whether the clusterings obtained using only the climate variables are very different from the clustering results obtained using only vegetation variables. Both approaches are compared to the clustering containing all variables (Figure 2).

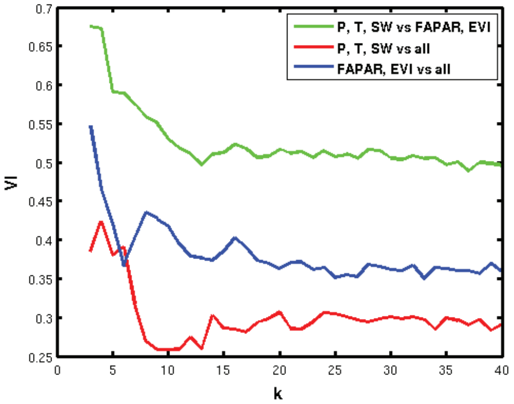

Figure 2: Pairwise comparison of distances (*VI*) between climate, vegetation and all variables.

In this experiment the distance between the clusterings of the climate and vegetation variables is quite large for all k. We conclude that a good partitioning of climate and vegetation space, representing both climate and vegetation properties, cannot be achieved using climate variables alone.

4.3. Finding the right k

A well-known problem of the k-means algorithm is identifying the "right" number of clusters without any additional information. Our first approach is to evaluate $Q(k)$ (see Eq. (2)) for $k = 1, \ldots, k_{\max}$ and search for the first kink in the monotonically decreasing curve. Figure 3(left) shows $Q(k)$ for different subsets of the data with $k_{\max} = 40$ illustrating that it is not easy to find a clear kink in these curves. Inspecting the difference between two consecutive values of Q may reveal more insights as shown in Figure 3(right). One can recognize a first strong decrease of the difference $Q(k) - Q(k-1)$ in the curve of all variables at $k = 13$ pointing on $k = 12$ as the right number of clusters for that set of variables. Such a clear decrease is not there for the other variables. We will investigate further whether we're on the right way.

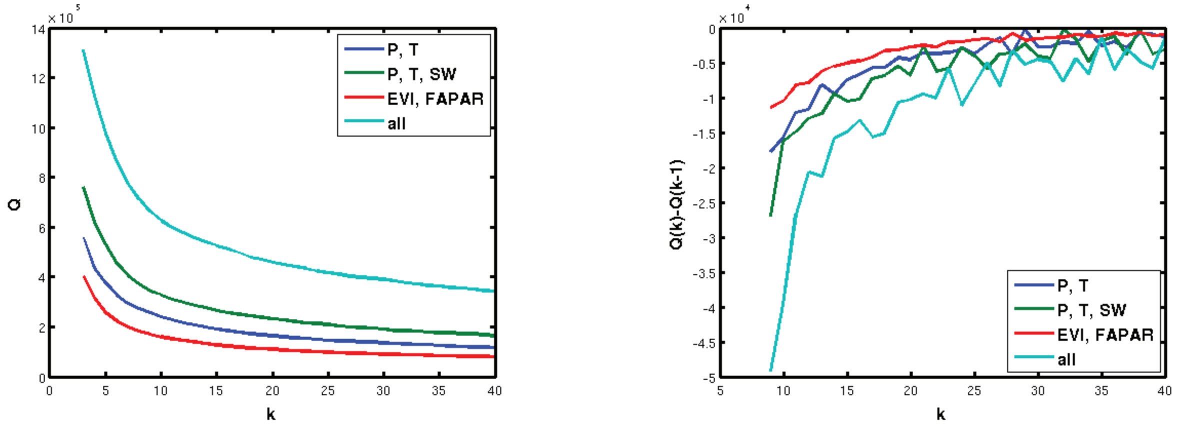

Figure 3: Left: Q for different number of clusters k and different variable subsets. Right: difference between $Q(k)$ and $Q(k-1)$.

A more robust method of detecting the right number of clusters is calculating the cluster stability. As described in Section 3.5, we first need to compute clusterings on subsets of the data. Instability scores are then calculated following Algorithm 1 where we use *VI* (see Section 3.4) as distance measure. We built 20 subsets of 10000 pixels each and compute pairwise distances. Instability scores are shown in Figure 4(left) for different sets of variables (for the sake of clarity we don't show the results for the set T, P, SW; they don't differ much from T, P).

Although it is not possible to determine whether a chosen number k of clusters is too small, it seems feasible to identify cases where k is too large [20]. Surprisingly, however, using either exclusively climate (P and T) or vegetation variables (EVI and FAPAR) does not allow to identify a region of too many clusters. Only if we combine all variables in the analysis we find a big jump between k =12 and k =13, suggesting again that the "right" number of clusters for this configuration is 12[5].

This finding is supported by an analysis over the standard deviation of distances between the different subset clusterings,

$$\mathrm{std}(k) = \sqrt{\mathrm{var}\left(d(C_b^k, C_{b'}^k)_{b,b'=1,\ldots,b_{\max}}\right)}.$$

Standard deviations for the same subsets of variables as in Figure 4(left) are shown in Figure 4(right). For $k = 12$ and using all variables, the standard deviation is very small compared to larger values of k, making the above result even more robust. The resulting map of the k-means clustering with $k = 12$ for all variables is shown in Figure 5.

[5] Note that small stability scores are desirable and that one has to look from the right to find the first significant jump towards zero [20]. It cannot be decided whether the jump between 10 and 11 renders 10 to be stable not (cf. Conjecture 3.8 in [20]).

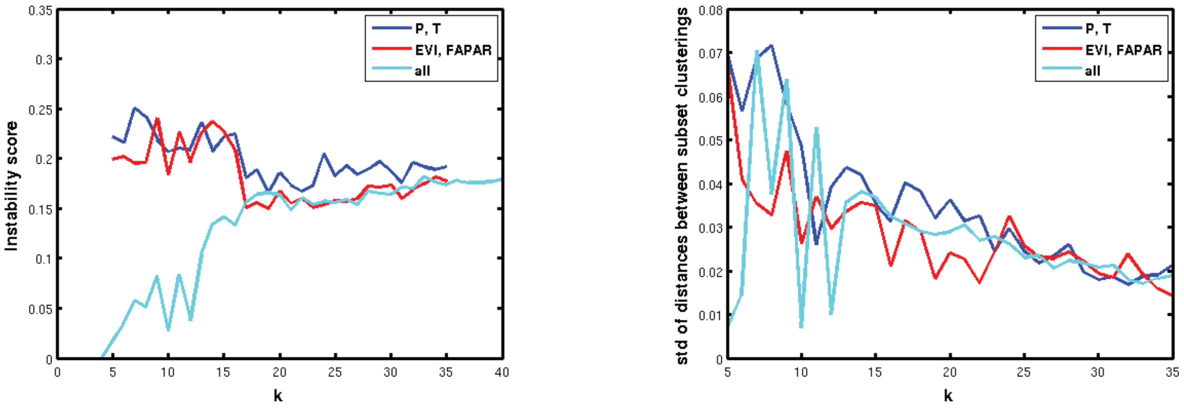

Figure 4: Left: Instability scores for different number of clusters k and different variable subsets. Right: Standard deviation of pairwise cluster distances over subsets of X for different number of clusters k and different variable subsets.

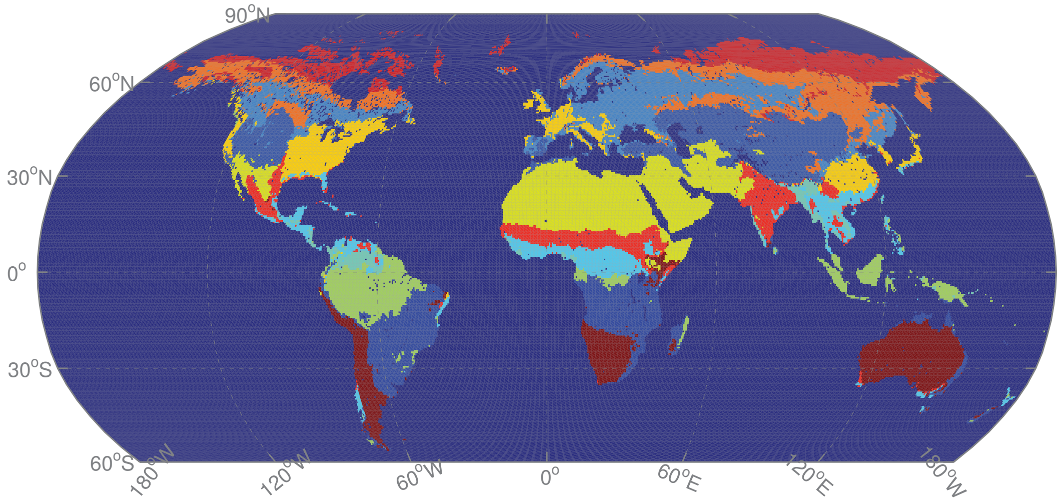

Figure 5: Map of k-means clustering with $k = 12$ for the variables P, T, SW, EVI, FAPAR.

4.4. Comparing k-means clusterings with the Köppen-Geiger classification

We investigate whether a k-means clustering with $k = 12$ sticks out against other values of k in terms of cluster distance to the KG classification. To this end we first classified the used temperature and precipitation data according to KG (we use the decision rules of [4]). We compared the obtained classification with several of our clusterings with k ranging from 3 to 40. As distance measure we again used VI (Sec. 3.4). Figure 6 shows VI as a function of k for different subsets of the investigated variables. Although clusterings of data containing the climate variables are a bit closer to the KG classification in general, a clear change of behaviour at $k = 12$ is not determinable.

4.5. How well does Köppen-Geiger match the the properties of the data?

If we reduce the dimensionality, e.g. by projecting the data onto the first two principal components, k-means quite nicely partitions the data (not shown here). This suggests that the data streams, although 12-dimensional, have low intrinsic dimensionality. Rapidly decreasing eigenvalue spectra of the covariance matrices support that (also not shown here).

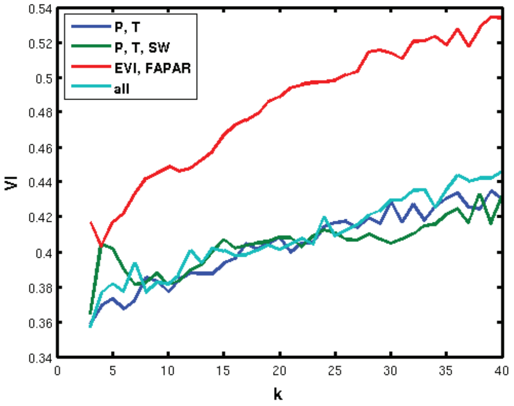

Figure 6: Values of *VI* between the Köppen-Geiger classification of the years 2001-2007 and several clusterings obtained with *k*-means.

To visualize KG on a low-dimensional projection of the data, we colour every sample with the label obtained from the Köppen-Geiger classification. To show the plots clearly, we only use the main climates A to E (see [4]) for the colouring. Interestingly, it seems that Köppen-Geiger classifies the northern hemisphere quite well even if we apply PCA on all variables (see Figure 7 left). The southern hemisphere, on the other hand, is matched rather poorly (Figure 7 right). We conclude that in the northern hemisphere the boundaries for the five main climates coincide well for both climate and vegetation. The difference between north and south maybe explained by the fact that the developers of the classification had much more experience with the climate and vegetation zones in the northern hemisphere, especially in Europe.

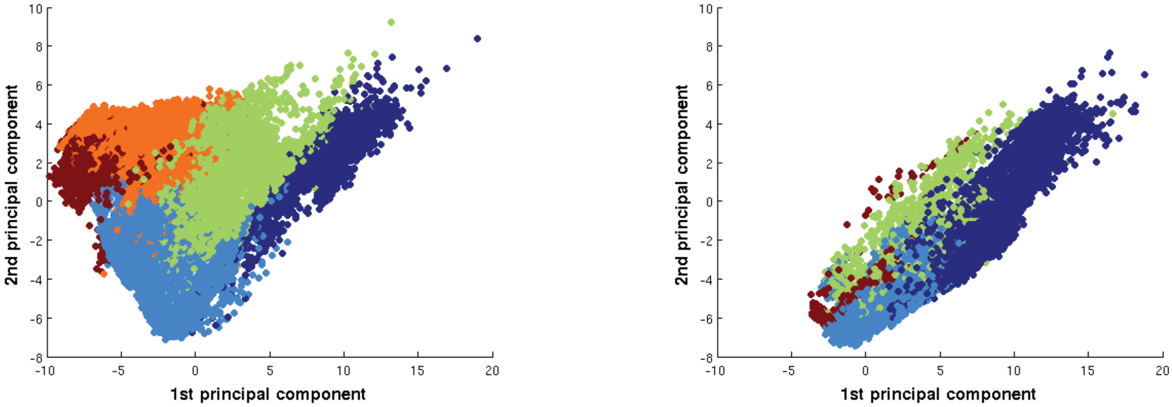

Figure 7: Projection of the data from the northern (left) and southern (right) hemisphere onto the first two principal components. Each point constitutes one pixel and is coloured according to the five main climates of the Köppen-Geiger classification

Figure 7 and other produced plots of this kind also suggests that the data rather represents a continuous manifold than discrete classes.

5. Discussion

With the currently observed constantly increasing amount of data unsupervised learning methods are gaining more and more importance. Hypotheses accepted for a long time can be approved or questioned by data driven analysis and possibly need to be adjusted.

Climate classifications like KG aim for predicting vegetation zones using only climate variables. We have shown that KG may not be optimal regarding this task and we tried to tackle it with an approach using unsupervised clustering. However, the clustering of climate and vegetation variables separately leads to different results in cluster space. Furthermore, we were not able to determine a stable (and thus naturally emerging) number of classes in both cases. Interestingly, using both climate and vegetation variables together, we could identify a stable number of clusters, namely 12. The close entanglement between climate and vegetation which cannot be described by a one-directional function may explain this different behaviour.

With our analysis, we challenge the assumption whether a prediction from climate variables to vegetation zones is possible at all. Consequently, the obtained clustering including climate and vegetation has rather diagnostic than predictive character. With the proposed tools a number of questions could be addressed. Future research should, e.g., analyse the interannual variability of the obtained clustering and whether the found 12 classes remain stable over time. With longer time series an important objective might be to find regions of changing classes over time. This would give information about changing climate-vegetation interaction.

Acknowledgements. JZ is part of the International Max Planck Research School for Global Biogeochemical Cycles. The authors thank Nuno Carvalhais for providing a script computing the KG classification and Björn Reu for useful comments.

References

[1] W. Köppen, Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt, Geogr. Zeitschr. 6 (1900) 593–611, 657–679.

[2] R. Geiger, Landolt-Börnstein – Zahlenwerte und Funktionen aus Physik, Chemie, Astronomie, Geophysik und Technik, Springer, Berlin, 1954, Ch. Klassifikation der Klimate nach W. Köppen, pp. 603–607., Band III, alte Serie.

[3] R. Geiger, Überarbeitete Neuausgabe von Geiger, R.: Köppen-Geiger / Klima der Erde. (Wandkarte 1:16 Mill.), Klett-Perthes, Gotha, 1961.

[4] M. Kottek, J. Grieser, C. Beck, B. Rudolf, F. Rubel, World Map of the Köppen-Geiger climate classification updated, Meteorologische Zeitschrift 15 (3) (2006) 259–263.

[5] C. Thornthwaite, Problems in the classification of climates, Geographical Review 33 (2) (1943) 233–255.

[6] A. J. Cannon, Köppen versus the computer: an objective comparison between the Köppen-Geiger climate classification and a multivariate regression tree, Hydrology and Earth System Sciences Discussions 8 (2) (2011) 2345–2372.

[7] B. Walker, D. Ludwig, C. Holling, R. Peterman, Stability of semi-arid savanna grazing systems, The Journal of Ecology 69 (2) (1981) 473–498.

[8] M. Scheffer, S. Carpenter, J. Foley, C. Folke, B. Walker, Others, Catastrophic shifts in ecosystems, Nature 413 (6856) (2001) 591–596.

[9] P. D'Odorico, F. Laio, L. Ridolfi, Noise-induced stability in dryland plant ecosystems, Proceedings of the National Academy of Sciences of the United States of America (PNAS) 102 (31) (2005) 10819–10822.

[10] M. Meilă, Comparing clusterings – an information based distance, Journal of Multivariate Analysis 98 (5) (2007) 873–895.

[11] D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, Others, The ERA-Interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society 137 (656) (2011) 553–597.

[12] A. Huete, K. Didan, T. Miura, E. Rodriguez, X. Gao, L. Ferreira, Overview of the radiometric and biophysical performance of the MODIS vegetation indices, Remote Sensing of Environment 83 (1-2) (2002) 195–213.

[13] X. Gao, A. Huete, W. Ni, T. Miura, Optical-biophysical relationships of vegetation spectra without background contamination, Remote Sensing of Environment 74 (3) (2000) 609–620.

[14] M. Jung, M. Reichstein, H. Margolis, A. Cescatti, A. Richardson, M. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, Others, Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res 116 (2011) G00J07.

[15] P. Gehler, Mpikmeans, `http://mloss.org/software/view/48/` (2007).

[16] C. Elkan, Using the triangle inequality to accelerate k-means, in: Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), 2003, pp. 147–153.

[17] S. Phillips, Reducing the computation time of the Isodata and K-means unsupervised classification algorithms, Geoscience and Remote Sensing Symposium (IGARS 2002) 3 (2002) 1627–1629.

[18] S. Phillips, Acceleration of k-means and related clustering algorithms, Revised Papers from the 4th International Workshop on Algorithm Engineering and Experiments (2002) 166–177.

[19] J. Kumar, R. T. Mills, F. M. Hoffman, W. W. Hargrove, Parallel k-Means Clustering for Quantitative Ecoregion Delineation Using Large Data Sets, Procedia Computer Science 4 (2011) 1602–1611.

[20] U. von Luxburg, Clustering stability: An overview, Found. Trends Mach. Learn. 2 (3) (2010) 235–274.