World climate search and classification using dynamic time warping similarity function

P. Netzel, T. F. Stepinski

Space Informatics Lab, Department of Geography, University of Cincinnati, Cincinnati, OH 45221-0131, USA
Telephone: +1 513 .556.3583
Fax: +1 513.556.3370
Email: pawel@netzel.pl , stepintz@uc.edu

Abstract

We present a data-mining approach to climate classification. Local climates are represented as time series of climatic variables and a similarity between two local climates is calculated using the dynamic time warping (DTW) function which allows for scaling and shifting of the time axis to model the similarity more appropriately than the Euclidean function. A 30 arc second resolution global grid of climatic data is clustered into 31 climatic classes and the resultant world-wide map of climate types is compared to the empirical Köppen–Geiger classification. We also present a concept of climate search – an interactive, internet-based application that allows retrieval and mapping of world-wide locations having climates similar to a user-selected location query.

Keywords: climate classification; dynamic time warping; climate search; clustering; segmentation.

1. Introduction

Climate classification (CC) schemes discretize the local Earth land surface climatic properties to enable an assessment of global climate models, to help in identification of ecologically similar regions across continents, and to analyze global issues in hydrology, agriculture, and biology.

Classical CCs such as the Köppen–Geiger (KG) approach (Kottek et al. 2006) rely on heuristic decision rules reflecting a body of environmental and geographical research, but they appear rather arbitrary from a modern, data-oriented perspective. Data-mining approach to CC reveals inherent spatial patterns in global distribution of climatic variables by means of unsupervised regionalization (spatial classification) – a process that divides the land surface into regions in a way that maximizes intra-region homogeneity and inter-region heterogeneity.

Previous applications of unsupervised regionalization to CC (Zscheischler et al. 2012, Metzger et al. 2012) relied on the following set of techniques: representing local climates by vectors of climatic variables, using Euclidean distance to calculate dissimilarity between local climates, and using clustering in data space in lieu of regionalization. Each of these techniques is not an optimal approach to the problem of CC. Here we present a data mining approach to CC that uses a different, more appropriate set of techniques: local climates represented as 12-months-long time series, dissimilarity between local climates measured by the Dynamic Time Warping (DTW) distance (Berndt and Clifford, 1994), and grid segmentation as the means of regionalization. In addition, we introduce a new analytic tool – climate search – as means to study a similarity of climates across the world. Climate search (CS) is an interactive, internet-based application that retrieves locations having climate similar to a user-identified query.

2. Data and Methods

We use climate data from the WordClim project (Hijmans et al. 2005). The following variables are used: monthly mean air temperature (T_{ave}), monthly maximum air temperature (T_{max}), monthly minimum air temperature (T_{min}), and monthly total precipitation (P). All data are long term averages calculated from measurements taken between 1950 and 2000. The climate data is given on a 30 arc second global grid obtained by interpolating measurements from a world-wide network of climate stations.

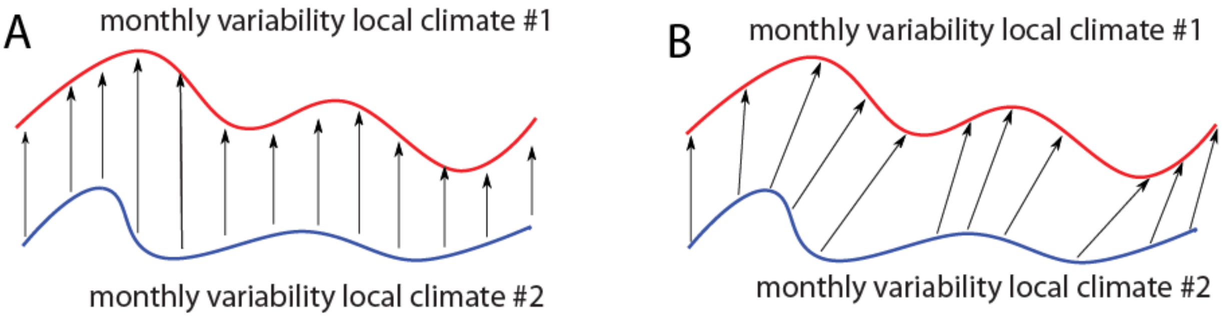
A
monthly variability local climate #1
monthly variability local climate #2
B
monthly variability local climate #1
monthly variability local climate #2

Figure 1. Difference between Euclidean distance (A) and DTW distance (B).

A local climate is defined as a 12-months-long time series at a single grid cell. The data is corrected to remove phase shift caused by sun position change during the year. The DTW algorithm calculates the distance between any two local climates. DTW is calculated using normalized data and modified to yield results in a range 0 to 1. The difference between Euclidean distance and DTW distance (Fig.1) is that DTW synchronizes two time series thus finding two local climates similar even if there is some time shift in their variability. DTW powers both CS and CC.

3. Climate Search

CS works on the principle of query-by-example, a user identifies a location, and a local climate at this location (a query) is compared to all other local climates in a grid using the DTW. The result is a grid of the same size as the data grid but storing the values of similarity between a query and local climates. Visualizing these values reveals a map showing similarity relations between the climate in the query location and climates in all other sites in the world. We have developed an internet application, Climate Explorer (ClimateEx), available at http://sil.uc.edu, which performs CS calculations and returns the climate similarity map. Fig.2 shows examples of climate search using ClimateEx for the following queries: New York, NY, Death Valley, CA, and Minneapolis, MN. ClimateEx takes about 40 sec to execute a query on the 30 arc second grid and to display the map.

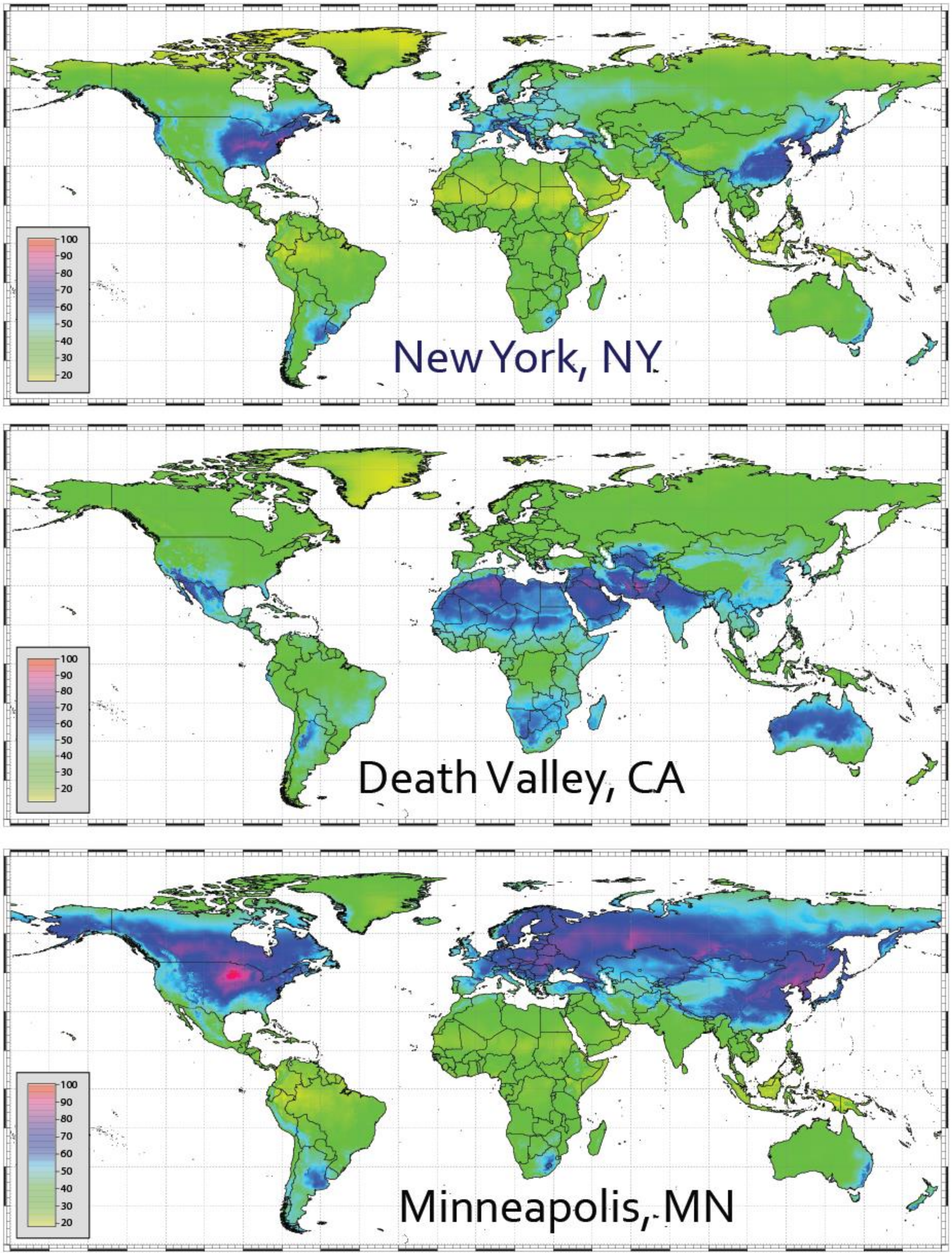

Figure 2. Examples of climate search for indicated locations. Climate similarity is indicated by color gradient from red (high similarity) to yellow (low similarity).

4. Climate Classification

To obtain a CC we first segment a grid into about 30,000 segments using a region growing algorithm. This is done in order to reduce the number of objects to be clustered and to assure spatial cohesion of climatic classes. We calculate a 30,000 by 30,000 distance matrix between all segments using the DTW distance. Using this matrix a hierarchical clustering with Ward linkage is calculated. A number of clusters (climate classes) is a free parameter that needs to be selected. We have selected 31 classes, the same number as in the KG classification. Fig.3 (top panel) shows the resultant CC with different colors (see insert legend) indicating spatial extent of each climatic class. The meaning of these classes has to be assigned on the basis of the average values of T_{ave}, T_{max}, T_{min}, and P over an extent of each class. This data mining-based CC could be compared with the KG classification (Fig.3 bottom panel). KG classes have been a priori interpreted by the design of the classification; for their meaning see the legend in Kottek et al. (2006).

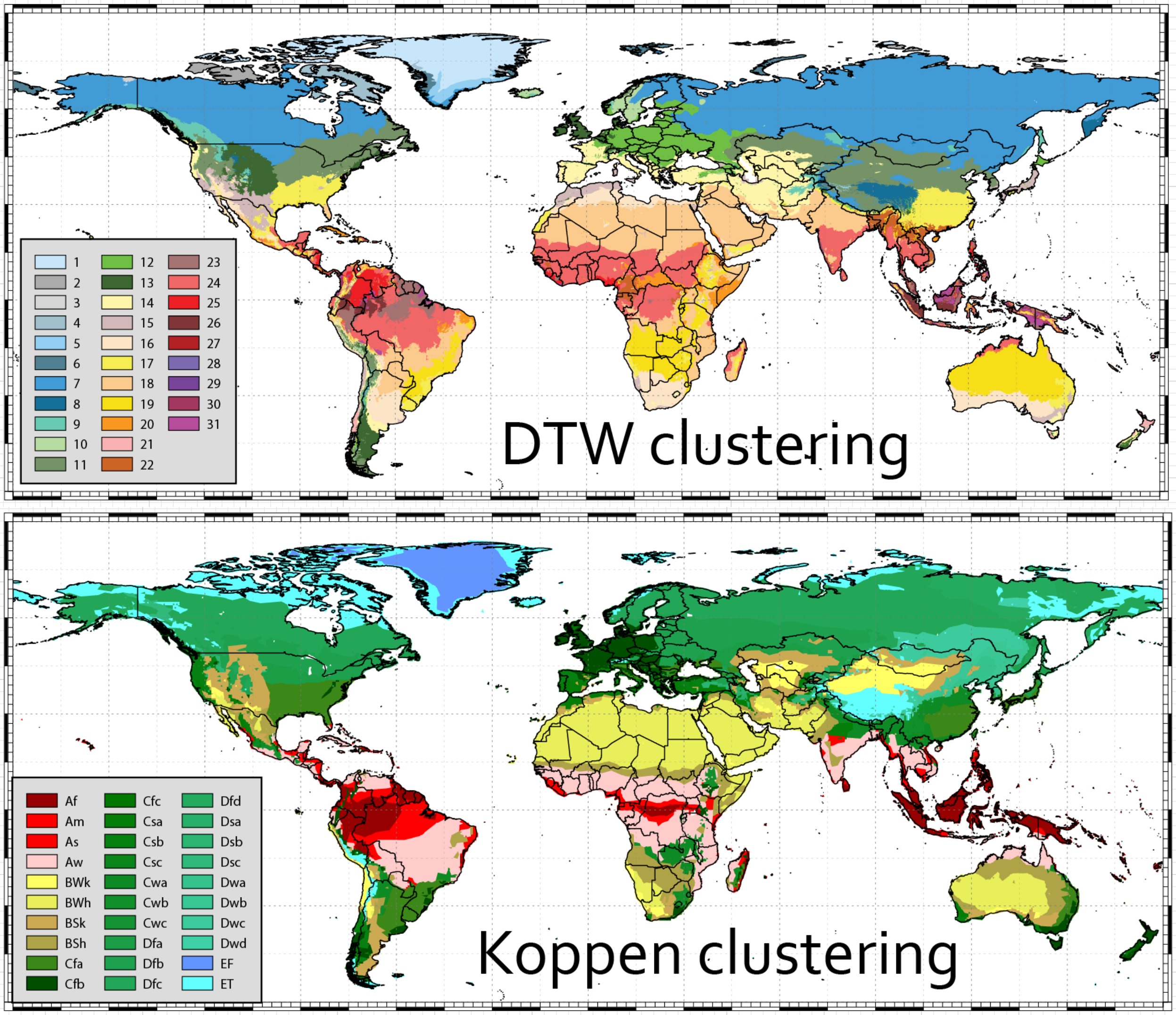

Figure 3. (Top) Climate classification using a data-mining method presented in this paper. (Bottom) Climate classification using the Köppen–Geiger method based on heuristics.

5. Conclusions

Our proposed method of CC is a step forward in an effort to study spatial regionalization of world climates using data mining approaches. The result of the classification (Fig.1 top panel) shows an overall similarity to a classic KG classification but also displays some notable differences. This is not surprising as the assumptions behind each classification are different. KG classification implicitly assumes that vegetation is a function of climate so a design of the KG classifier is tied to observed spatial variability of vegetation types. Our classification is based purely on climatic data with no ties to vegetation. The observed differences in climate maps suggest that spatial changes in vegetation types cannot be explained by means of climate variability alone.

CS is a new analytical tool for custom made investigations of climate similarities. It yields results which are much more sensitive than a global CC which digitizes all climates in just few categories. CS is not restricted to global exploration (like in the examples given in Fig.1) but can also be used to explore regional differences in climate. Examples include delineation of climate changes due to rising elevation (valleys vs. ridges) and delineation of local microclimates.

In the present implementation of our method all four climatic variables (T_{ave}, T_{max}, T_{min}, and P) contribute to the DTW with the same weight which skews the results toward temperature-centric classification and search. Further research will experiment with different ways including giving equal weight to precipitation and all temperature variables taken together.

6. Acknowledgements

This work was supported by the University of Cincinnati Space Exploration Institute and by the National Science Centre (NCN) under Grant DEC-2012/07/B/ST6/01206.

7. References

Berndt, DJ, Clifford, J, 1994, Using Dynamic Time Warping to Find Patterns in Time Series. In *KDD workshop*, vol. 10(16): 359-370.

Hijmans, RJ, Cameron, SE, Parra, JL, Jones, PG, Jarvis, A, 2005, Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology* 25(15): 1965-1978.

Kottek M, Grieser J, Beck C, Rudolf B, Rubel F, 2006, World map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift* 15(3): 259-263.

Metzger, MJ, Bunce, RGH, Jongman, RHG, Sayre, R, Trabucco, A, Zomer, R, 2012, A high-resolution bioclimate map of the world: a unifying framework for global biodiversity research and monitoring. *Global Ecology and Biogeography* 22(5): 630-638.

Zscheischler, J, Mahecha, MD, Harmeling S, 2012 Climate classifications: the value of unsupervised clustering. *Procedia Computer Science*, 9: 897-906.