Predicting Climate Types For The Continental U.S. Using Unsupervised Clustering Techniques

D. Sathiaraj

Dept. of Geography and Anthropology and NOAA Southern Regional Climate Center, Louisiana State University, Baton Rouge, USA

X. Huang

NOAA Southern Regional Climate Center, Louisiana State University, Baton Rouge, USA

J. Chen

Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, USA

Research Article

Author Note

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/env.2524

Correspondence: Dr. D. Sathiaraj, E335 Howe Russell Bldg, Louisiana State University, Baton Rouge, LA, 70803

Contact: davids@srcc.lsu.edu

Abstract

The problem of clustering climate data observation sites and grouping them by their climate types is considered. Machine learning based clustering algorithms are used in analyzing climate data time series from more than 3000 climate observation sites in the United States, with the objective of classifying climate type for regions across the U.S. Understanding the climate type of a region has applications in public health, environment, actuarial science, insurance, agriculture and engineering.

In this study, daily climate data measurements for temperature and precipitation from the time period 1946-2015 have been used. The daily data observations were grouped into three derived datasets: a monthly dataset (daily data aggregated by month), annual dataset (daily data aggregated by year), and a threshold exceeding frequency (TEF) dataset (TEF provides frequency of occurrence of certain climate extremes). Three existing clustering algorithms from literature: k-means, DBSCAN and BIRCH were each applied to cluster each of the datasets and the resulting clusters were assessed using standardized clustering indices. The results from these unsupervised learning techniques revealed the suitability and applicability of these algorithms in the climate domain. The clusters identified by these techniques were also compared with existing climate classification types such as the Köppen classification system. Additionally, the work also developed an interactive web and map-based data visualization system that uses efficient Big Data management techniques to provide clustering solutions in real-time and to display the results of the clustering analysis.

Keywords: computational geosciences, climate data, machine learning, Big Data,

clustering

Predicting Climate Types For The Continental U.S. Using Unsupervised Clustering Techniques

Introduction

Data analytics and data-driven machine learning techniques are important for the discovery of valuable insights from large high-dimensional datasets. Data mining techniques such as Clustering algorithms (Duda et al., 2012; Jain et al., 1999) help in discovering clusters that may have gone undetected. The clusters can offer important clues on similarities, correlations and relationships within a large dataset. It can also help in outlier detection.

Climate is a critical component of the Earth's ecosystem and is an important aspect of people's daily lives, industrial production and agricultural output. Climate is defined as long-term averages and variations in weather measured over a period of several decades (Melillo et al., 2014). The Continental U.S. has the advantage of a having a rich source of voluminous climate data observations. The daily climate data records go back about a hundred years. These large climate data repositories are housed in national, operational data centers such as the NOAA Regional Climate Centers (the NOAA Southern Regional Climate Center (`http://www.srcc.lsu.edu`) is housed at Louisiana State University (LSU)) and the National Center for Environmental Information (`http://www.ncei.noaa.gov`). The programatic access to these large, distributed data repositories is made available using enterprise grade web-services and an applications programming interface (API, `http://www.rcc-acis.org`, (DeGaetano et al., 2015)) that was developed by IT staff at the Regional Climate Centers. Recently, a new dataset, the

Threshold Exceeding Frequency (TEF) (Huang et al., 2017) was created from such observational data archives. Using historical climate data measurements at individual sites, climate scientists have developed more generic, broader, region-based climate classification types. The rich climate data source is a good foundation for the application of clustering algorithms to derive unsupervised learning based climate types or clusters.

Background

Clustering analysis is an unsupervised learning technique that finds applications in several fields including: pattern recognition, machine learning, image processing and information retrieval. Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (Jain et al., 1999). Clustering does not need prior knowledge about the dataset. Clustering techniques partition a dataset into clusters so that objects in the same cluster are more similar to each other than objects in different clusters. These objects are also referred to as data points or points. A commonly used clustering algorithm is the k-means clustering method (MacQueen et al., 1967) that tries to minimize a squared error criterion (Jain et al., 1999). Nonparametric methods for density-based clustering have also been developed (Jain et al., 1999). A widely used density based clustering method is the DBSCAN (Density Based Spatial Clustering of Applications with Noise) (Ester et al., 1996) which groups dense and nearby points to form clusters. BIRCH (Zhang et al., 1996) is another hierarchical clustering technique. This study evaluates the performance of 3 clustering algorithms - k-means, DBSCAN and BIRCH - in grouping climate measurement sites into climate types or clusters.

In climate science, the Köppen classification (KC) is used to spatially group regions

based on the predominant climate of that region. The spatial distribution of this classification is shown in Figure 1. Using empirical observations, Köppen (Köppen and Geiger, 1930) pioneered and established a climate classification system which uses monthly air temperature and rainfall to define boundaries (clusters) of different climate types around the world (Chen and Chen, 2013).

In this work, unsupervised, predictive learning techniques are used to predict climate classification types and each of the clustering solutions are evaluated using the KC system. Prior work in climate classification include work such as (Fovell and Fovell, 1993) that used principal component analysis to generate climate zones. Work such as (Zscheischler et al., 2012) have explored the use of unsupervised clustering techniques such as k-means to explore the feasibility of such techniques as compared to empirical rule based KC system. Some differentiating factors between (Zscheischler et al., 2012) and our work were that the former used gridded datasets and remotely sensed vegetation indices as compared to this work that uses actual ground truth data from climate observation sites. The time period studied was also a shorter time span from 2001-2007, whereas this work spans a 70 year window from 1946-2015. Gridded datasets are typically derived using interpolation processes and hence can be prone to inaccurate estimations due to lack of accounting for topographical climate variations (Mourtzinis et al., 2017) (such as in mountainous regions) or approximating in areas that have a poor coverage of measurement sensors. Since this work uses observational datasets, the data being fed to the clustering algorithms is of better quality and void of any inaccurate estimations or approximations. Some of the challenges involved in this work included the extensive data pre-processing and management of the Big Data sets (processing nearly 50 million points of information) and

the ability to derive clusters in real-time as new observations come in.

Recent work such as (Netzel and Stepinski, 2016) have used similar clustering approaches to derive climate classification types on a global scale. Again, in this case, gridded datasets were used as compared to actual climate station measurements as in this work. There was difference also in the clustering algorithms deployed to generate the climate type classes and the area of focus (this work focussed on continental US as compared to the global study by (Netzel and Stepinski, 2016)). A similarity between work by (Netzel and Stepinski, 2016) and our work is that the proposed clustering will not predict KC types but instead alternative climate type classifications are provided. The reason for this is that unsupervised clustering techniques are initiated using a random seed and the clusters obtained do not have a cluster type label. Our map-based visualization system, described later, provides a spatial mapping of the clusters that enables the effective inferring of clusters. These predicted climate classifications can also be used as a more dynamic representation of a climate type in a changing climate scenario as compared to the rule based, empirical KC technique. The visualization system developed also provides the ability for a spatial comparison between the predicted clusters and KC types. Work by (Zhang and Yan, 2014) highlights the importance of cluster analysis in classifying climate types in the context of climate change especially when KC is less able to reflect variations in climate classifications. The work by (Zhang and Yan, 2014) also used a global gridded dataset and only the k-means algorithm to derive climate types. The focus was to look for temporal and geographical shifts in climate types. In (Liss et al., 2014), machine learning techniques were used to define climate regions but the focus was solely in the context of public health. Work by Mahlstein and Knutti (2010) involved using cluster analysis to look

at regional climate change patterns but the analysis did not involve comparing the clusters with KC types.

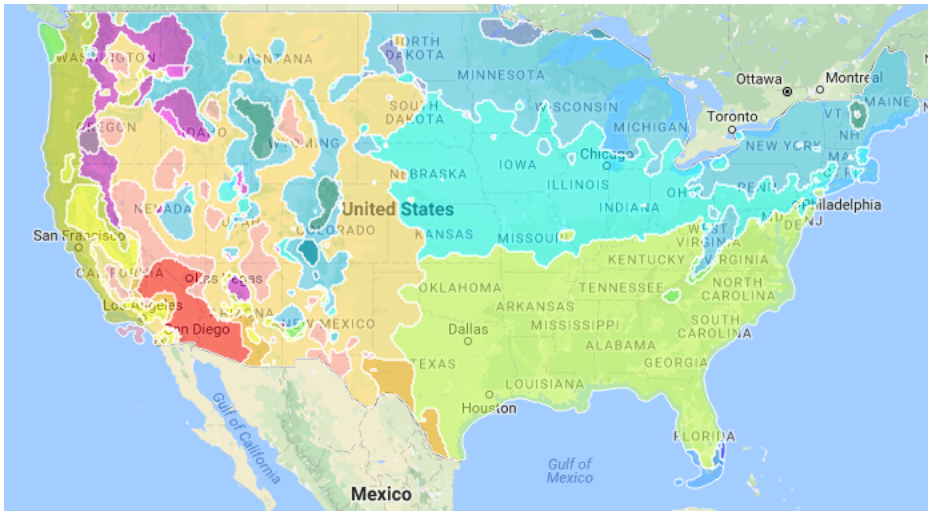


Figure 1. A screenshot of Köppen Climate Classification from `goo.gl/2mU2qr`

The KC classification consists of five major climate type groups and a number of climate type sub-types under each major group, as listed in Table 1. The continental US, which is our study area, has 4 main types. The tropical climate type A is determined when the monthly mean temperature for every month is equal to or higher than 18°C and the four sub-types under A are determined based on their annual and monthly mean precipitation values. The dry climate type B is defined by very low precipitation. The monthly mean temperature of the coldest month lies between -3°C and 18°C for mild temperate C type and the different seasonal precipitations result in 4 sub-types. The continental climate D has at least one month with temperatures equal or lower than -3°C, and the sub-types are decided according to the seasonal precipitation. Finally the polar climate E, is for regions that have their warmest monthly mean temperature of any month to be equal or lower than 10°C (Chen and Chen, 2013).

Table 1

As provided from (Chen and Chen, 2013): Main characteristic of Köppen climate major groups and sub-types

Major group	Sub-types
A: Tropical	Tropical rain forest: Af Tropical monsoon: Am Tropical wet and dry savanna: Aw or As
B: Dry	Desert (arid): BWh, BWk Steppe (semi-arid): BSh, BSk
C: Mild temperate	Mediterranean: Csa, Csb, Csc Humid subtropical: Cfa, Cwa Oceanic: Cfb, Cfc, Cwb, Cwc
D: Snow	Humid: Dfa, Dwa, Dfb, Dwb, Dsa, Dsb Subarctic: Dfc, Dwc, Dfd, Dwd, Dsc, Dsd
E: Polar	Tundra: ET Ice cap: EF

This climate classification system has been widely used by geographers and climatologists around the world. The system is powerful in linking climate and natural vegetation (Bailey, 2009) and to evaluate climate change impacts (Fovell and Fovell, 1993; Kottek et al., 2006; Liss et al., 2014; Diaz and Eischeid, 2007).

This work uses machine learning based clustering algorithms to generate clusters or climate classification types. These clusters are evaluated using standardized metrics for clustering algorithms and also compares the classification types revealed by the empirical, deterministic KC technique.

The work is structured as follows: First, information regarding sources for the daily climate data and metadata for the 3000 climate stations in the continental U.S. is provided. The data preprocessing routines used to transform raw climate data observations into derived annual, monthly, and TEF datasets (Huang et al., 2017) are explained and how machine learning techniques are applied to generate clusters. Secondly, a real-time

data visualization system and a user-interface to visualize the clustered solutions is described. Lastly, computational results and analysis are provided.

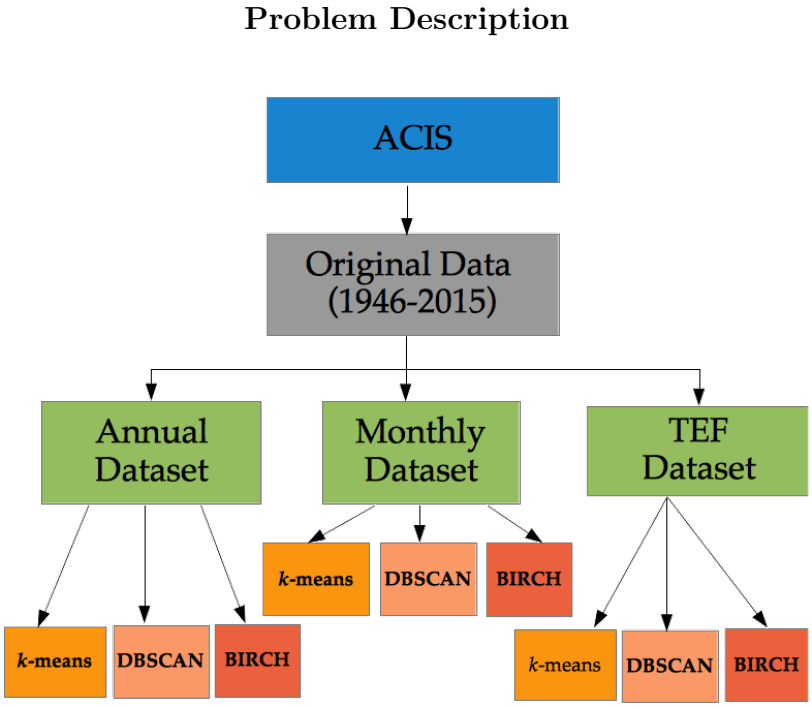


Figure 2. Flowchart for data processing

Figure 2 depicts a flowchart for the data processing involved. First, the daily climate datasets from 1946 to 2015 are obtained from ACIS. The data are then grouped by month or year to generate the annual climate dataset, monthly climate dataset and threshold exceeding frequency (TEF) climate dataset. The 3 derived time series (annual, monthly and TEF) are stored in a memory-based cache (Redis, `http://redis.io`) for fast access and analysis. Then, these three climate time-series datasets are utilized for unsupervised learning. 3 unsupervised learning based clustering algorithms: *k*-means, DBSCAN, and BIRCH are used to derive clusters. The clusters formed (9 sets of clusters) are evaluated using clustering metrics and also compared with existing KC types.

Data Sources

The study used the daily climate data from the Applied Climate Information System (ACIS) (DeGaetano et al., 2015), an Internet-based system designed to facilitate the generation and dissemination of climate data products to users. ACIS is developed by the NOAA Regional Climate Centers (RCCs) (DeGaetano et al., 2010) to manage the complex flow of information from climate data collectors to end users of climate data information.

ACIS accepts and returns climate information in JavaScript Object Notation (JSON), which uses structures that are similar to those used in many coding languages, including C, C++, Java, JavaScript, Perl, and Python. For each call, users specify a set of parameter to describe the data being requested. After passing these parameters to the server and accessing these climate data, a climate data product is returned to users.

For vegetation and ecosystems, climate of a region involves the following elements: temperature, precipitation, humidity, wind and radiation. But most climate analyses use near-surface air temperature and precipitation as the two major variables (Kottek et al., 2006; Fovell and Fovell, 1993; Liss et al., 2014). In this study the climate elements used were: maximum temperature, minimum temperature and precipitation.

There are more than 26000 GHCN climate data measurement sites. However not all span the entire time period of 1946-2015. Additional criteria used for this data analysis included the following: allow for less than 10% missing values for a station per year and every climate division should have at least 3 climate measurement sites. Once these criteria were applied, the number of valid stations that fit these rules reduced to 3210, as shown in Figure 3. These 3210 stations were well-distributed and covered most parts of the

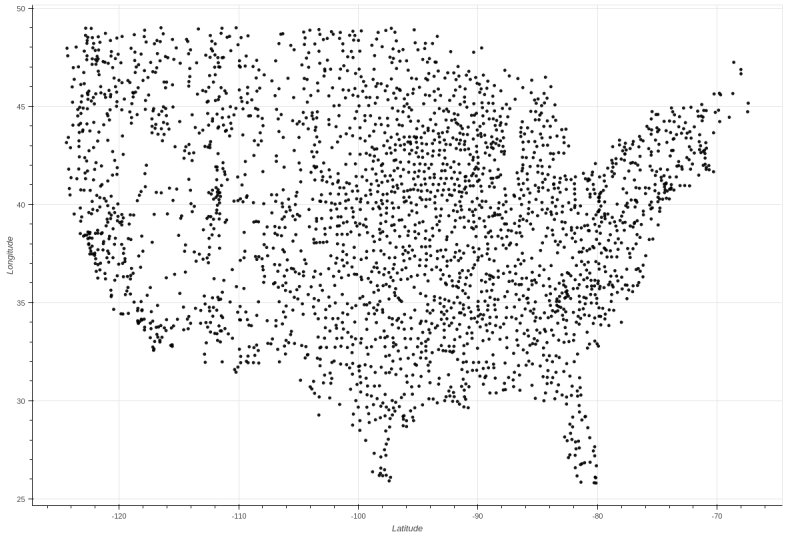


Figure 3. Spatial distribution of climate data observing stations continental United States.

Dataset Generation

Data from ACIS is on a daily time scale. Data was processed to derive 3 datasets. The 3 datasets derived were annual, monthly and the Threshold Exceeding Frequency (TEF) dataset.

Annual and monthly climate datasets were derived by grouping and averaging by year and month respectively. The TEF provides frequency of occurrence of extreme climate events. To establish the TEF dataset, some climate extremes thresholds were chosen. Generation of the TEF dataset was first described in (Huang et al., 2017). The thresholds used in generating the dataset were based on a combination of thresholds used in the *CLIMDEX - Datasets for Indices of Climate Extremes* (Donat, 2013) and the Southeast chapter of the *US National Climate Assessment* document, released in 2014 (Melillo et al., 2014).The chosen extremes are provided in Table 2.

Table 2

Threshold values to generate TEF dataset

Element	Thresholds
Maximum Temperature (in °F(°C))	≥105(41), ≥100(38), ≥95(35), ≥85(29)
Minimum Temperature (in °F(°C))	≥80(27), ≥75(24), ≥70(21), ≥65(18), ≤36(2), ≤32(0), ≤28(-2), ≤24(-4), ≤15(-9), ≤10(-12), ≤5(-15), ≤0 (-18)
Precipitation (in inches(mm))	≥2 (50mm), ≥4 (100mm)

The TEF dataset was generated by using the defined thresholds and the daily climate dataset to estimate the annual frequency of days that either exceeded or fell below a certain threshold. This was carried out for each climate measurement site. As an example, for threshold *Minimum Temperature* ≤ 75, the number of occurrences where the minimum temperature was less than or equal to 75 was tallied by year and calculated for each of the years in the time period, 1946-2015. This process was repeated for all 3210 climate measurement sites and for each of the thresholds to obtain the derived TEF dataset. This threshold based dataset provides a representation of climate extremes at each site and is a useful application resource for clustering sites that experience similar climate trends and extremes.

The input data structure for the clustering algorithm is described in Figure 4. The input data of any climate measurement site is divided into three arrays: maximum temperature array, minimum temperature array, and precipitation array (each array comprising of data from 1946-2015). Each row corresponds to a climate measurement site and all rows correspond to all the climate measurement sites. In addition, to ensure a serially complete dataset, climate stations had no more than 10% missing values per year

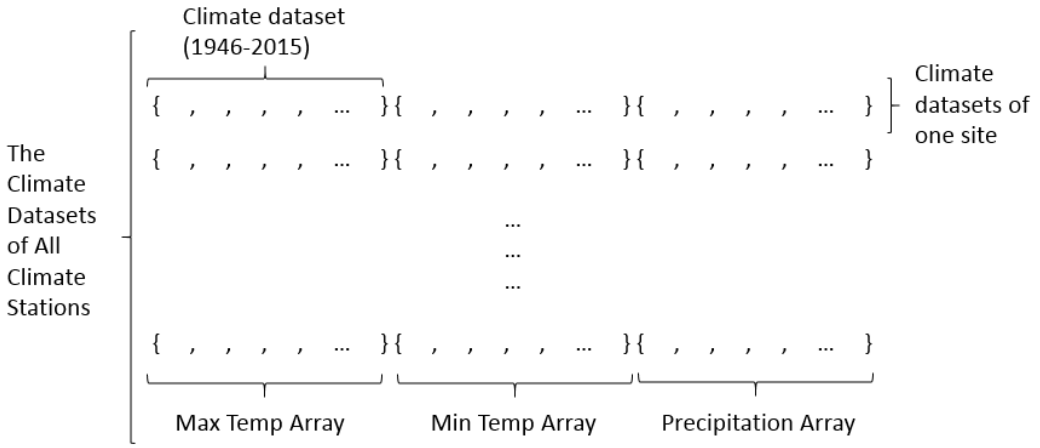


Figure 4. Input data structure after preprocessing

or month. A climate observation site that contained more than 10% missing values was removed from the analysis. In cases where missing values were less than 10%, missing data was imputed using an average value of nearby climate measurement sites. Also to eliminate the effects of different units between precipitation, maximum temperature and minimum temperature, the pertinent data columns were normalized as follows: $X' = (X - E)/\sigma$, where X is the original value, E is the mean of the data column and σ is the standard deviation of the data column.

Clustering Algorithms and Metrics Used

***k*-means.** k-means clustering algorithm (MacQueen et al., 1967), is a commonly used clustering algorithm and comes under the category of partitioning-based clustering techniques. In k-means, an *a priori* knowledge of number of clusters is important. The k-means algorithm can be applied over continuous data as noted in (Fukunaga, 2013) and (Duda et al., 1973). k-means algorithm calculates its centers iteratively. A detailed survey of data clustering algorithms - particularly k-means - can be found in (Jain, 2010). The

formation of clusters using this methodology can be mathematically defined as: Given a dataset $D = \{d_1, ..., d_N\}$ with N points that need to be clustered into say K clusters. Clusters can be considered as $C = c_1, c_2, ..., c_K$. k-means algorithms forms clusters such that a clustering criterion is optimized. A commonly used clustering criterion is the sum of squared Euclidean distances between each data point d_N and the centroid m_K (cluster center) of the subset c_K which contains d_N. This criterion is called clustering error E and depends on the cluster centers $m_1, ..., m_M$ (this corresponds to a priori defined number of clusters). The error for each cluster c_K can be denoted as: $E(c_K) = \sum_{d_N \in c_K} ||d_i - m_K||^2$.

The goal is to minimize the sum of the squared errors across all clusters. The k-means algorithm is a locally optimal algorithm. It is an iterative algorithm that initially selects arbitrary cluster centers and a set of clusters around each center. It then iteratively reassigns cluster centers to make the sum of squared errors minimal. The main disadvantage of the method lies in its sensitivity to initial starting positions of the cluster centers. Therefore, in order to obtain near optimal solutions using the k-means algorithm, several runs of the algorithm are conducted by varying the initial positions of the cluster centers.

DBSCAN. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering technique (Ester et al., 1996) that can identify clusters of arbitrary shapes. Density-based clustering techniques define a cluster as a region comprised of high density objects. A cluster is described as a linked region that exceeds a given density threshold (Biçici and Yuret, 2007). DBSCAN uses 2 predefined parameter values: the size of the neighborhood denoted as ε and the minimum number of points in a cluster as N_{min}. In DBSCAN, a random point x is chosen and it finds all the points that

are density-reachable from x while ensuring that the minimum points criteria of N_{min} is met. If x is a core point, then the formation of a cluster is completed with respect to ε and N_{min}. If x is a border point, then no points are density-reachable from x. DBSCAN then begins with an unclassified point to repeat the same process. The two parameters, ε and N_{min}, direct the operation of DBSCAN and ensure the quality of clusters. The 2 parameters are adopted universally in functioning of DBSCAN. In most cases, DBSCAN explores each point of the database multiple times.

BIRCH. BIRCH (stands for a Balanced Iterative Reducing and Clustering using Hierarchies) algorithm was introduced by Zhang et al. (Zhang et al., 1996). As compared to DBSCAN (density-based clustering) and k-means (a partition-based clustering), BIRCH is a hierarchical clustering algorithm that scans the data once and develops an in-memory tree representation for clusters. 2 key features in BIRCH include a data structure for each cluster called *Clustering Feature (CF)* and an in-memory tree representation denoted as *CF-tree*.

The *CF* is a compact data structure that allows for easy merging of sub-clusters. A *CF* is a tuple that summarizes information about the cluster. Given N d-dimensional data points in a cluster, $\{\vec{X}_i\}$ where $i = 1, 2, ..., N$, CF is defined as a triple: $CF = (N, \vec{LS}, \vec{SS})$ where N is the number of data points in the cluster, $\vec{LS}$ is the sum of the N data points and $\vec{SS}$ is the square sum of the N data points. A *CF-tree* has two additional parameters: branching factor B for non-leaf nodes in the tree and threshold T for the leaf nodes.

Homogeneity and V-Measure

The notion of Homogeneity and V-Measure were first introduced by (Rosenberg and Hirschberg, 2007), with the objective that any clusters formed will have 2 properties - Homogeneity and Completeness. Homogeneity, H is an objective that the cluster contains only members of a single class and the score is value between 0 and 1, with higher values (closer to 1) representing higher Homogeneity. Completeness, C is the notion that all members of a given class are assigned to the same cluster. V-Measure is then defined as: $2 \times \frac{H*C}{H+C}$. V-Measure also yields scores in between 0 and 1, with higher scores indicating greater completeness.

Data Visualization

To demonstrate the computational analysis, a data visualization system was developed. The data visualization system contains a low latency database, a real-time interactive machine learning system for generating clusters and an interactive user interface for map-based visualizations and setting clustering parameters. The visualization system provides a web-based visual interface to compare predicted clusters (from k-means, DBSCAN and BIRCH) with the classification types of the KC system. The site is hosted at `http://climext.srcc.lsu.edu`. The technologies used in this study include **Scikit-learn** (Pedregosa et al., 2011) (a Python module for machine learning, particularly clustering and classification), **Tornado** (a Python web framework and asynchronous networking library), **Redis** (an in-memory database cache), **Pandas** (a data analysis toolkit for Python), **React** (a JavaScript library for building user interfaces), **GraphQL** (a query language for APIs), **WebGL** (a JavaScript library for rendering 2D and 3D

graphics) and **D3** (a JavaScript visualization library).

Computational Results and Analysis

The computational analysis involves evaluating the performance of the 3 clustering algorithms on the 3 climate time series datasets (annual, monthly and TEF). The evaluation used the above outlined metrics, Homogeneity and V-measure (Rosenberg and Hirschberg, 2007), to evaluate the quality of the clusters formed. A diagnostic explanation between KC and the derived clusters is also provided. It is important to note that the predicted clusters cannot be directly compared with KC types (as outlined in Table 1). Clustering algorithms generate clusters in a random order and provide cluster labels that are numeric. However, when the climate measurement sites along with their predicted cluster labels are plotted on a map, the spatial visualization provides specific groupings or polygons that closely resemble the spatial distribution of the KC system. These visual groupings provide a strong case for the use of unsupervised clustering models that rely on purely climate data and are computationally nimble as compared to KC that rely on additional vegetation and remotely sensed data to derive climate types.

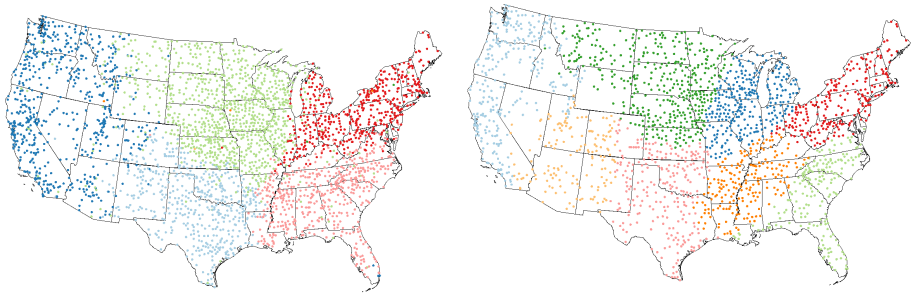
Analyzing Clusters Formed by *k*-means

The k-means method worked well when applied to all 3 datasets. Table 3 provides Homogeneity and V-Measure scores for varying cluster sizes (ranging from 4 to 12). One can observe from Table 3 that the Homogeneity Scores for the TEF dataset are higher than those of monthly and annual datasets. Secondly, as the number of clusters increase, the Homogeneity Scores show an increasing trend.

Table 3

Homogeneity and V-Measure Scores of k-means for different number of clusters and the 3 datasets (higher score is better)

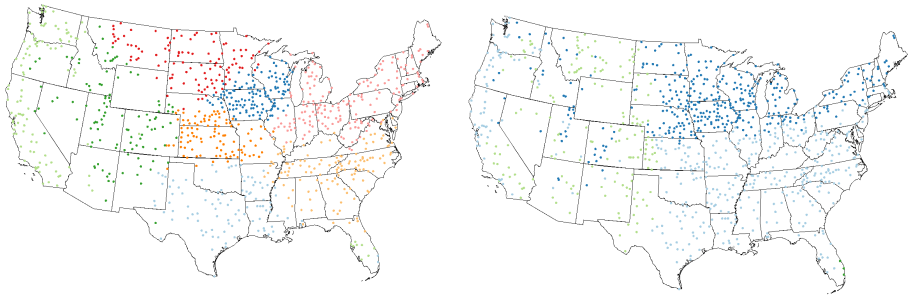
	Homogeneity			V-Measure		
# of Clusters	Annual	Monthly	TEF	Annual	Monthly	TEF
4	0.202	0.231	0.315	0.180	0.211	0.279
5	0.201	0.274	0.306	0.180	0.228	0.250
6	0.207	0.277	0.344	0.184	0.217	0.265
7	0.240	0.326	0.425	0.201	0.244	0.310
8	0.217	0.379	0.464	0.180	0.270	0.325
9	0.266	0.382	0.433	0.208	0.262	0.293
10	0.269	0.422	0.476	0.219	0.281	0.311
11	0.235	0.425	0.533	0.185	0.274	0.340
12	0.328	0.432	0.574	0.232	0.274	0.357



(a) Clustering Using k-means on Annual Time Series

(b) Clustering Using k-means on Monthly Time Series

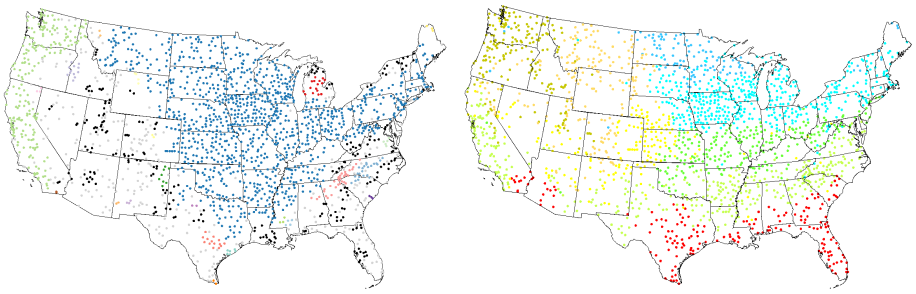
Figure 5. k-means on Annual and Monthly time series



(a) Clustering Using k-means on TEF Dataset

(b) Köppen classification types or labels

Figure 6. k-means on TEF and Spatial distribution using Koppen Classification Labels



(a) Clustering using DBSCAN with parameters $\varepsilon = 21$ and $N_{min} = 2$

(b) Clustering using BIRCH on monthly dataset

Figure 7. Clustering Result using BIRCH and DBSCAN on monthly dataset

Figure 5(a), Figure 5(b),and Figure 6(a) display 8 predicted clusters after applying k-means clustering on annual, monthly and TEF datasets, respectively. In the map, the different colors of points represent different clusters. Figure 6(b) displays spatial distribution for Köppen classification.

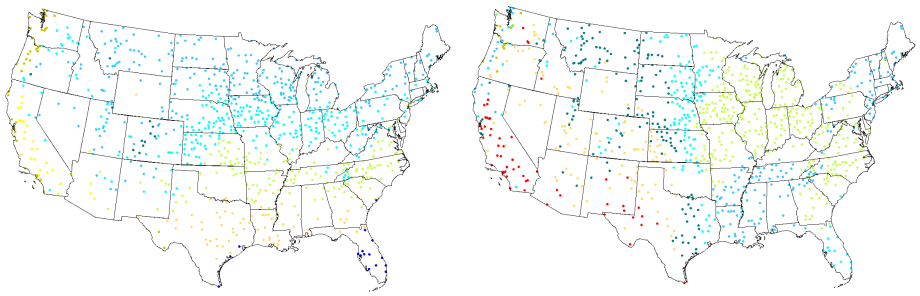
Analyzing Clusters formed with DBSCAN

When DBSCAN was applied to each of the 3 datasets, degenerative results emerged, such as all climate sites were merged into 1 or 2 clusters. After a gridded search of optimal parameters, the parameters used in DBSCAN were: $\epsilon = 21$ and the minimum points in a cluster set, N_{min} set to 2. Only one example of such a degenerative solution is provided as Figure 7(a), when DBSCAN was applied on the TEF dataset. The grey points represent the noisy samples.

Analyzing clusters formed with BIRCH

Table 4 depicts Homogeneity and V-Measure scores for BIRCH.

BIRCH has approximately the same results as k-means, and most of Homogeneity and V-Measure scores of BIRCH are slightly lower than k-means.



(a) Clustering using BIRCH on Annual dataset (b) Clustering using BIRCH on TEF dataset

Figure 8. Clustering Result using BIRCH on annual and TEF dataset

Figures 7(b), 8(a) and 8(b) shows the spatial distribution of 8 predicted clusters using BIRCH clustering on monthly, annual and TEF datasets respectively. The parameters used were a threshold of 1 and a branching factor of 50.

Summary of Analysis

The clustering results from k-means and BIRCH is compared with KC types using the climate stations that are part of each cluster. (Results using DBSCAN were not reported due to its poor performance in producing degenerative clusters). Since clustering algorithms are unsupervised by nature they yield clusters in a random order. They also do not predict the KC types. So a new approach is explored by grouping the climate stations by KC types (using inherent spatial information such as latitude and longitude) and then comparing the groups to the clusters generated. Percentages of stations that are part of a generated cluster will provide clues on the similarities of the cluster and a KC type. For 8 cluster solution, a summary comparison has been provided in Table 5. To interpret the table, 8 KC types are represented in rows and the columns represent combinations of the 3 datasets and 2 clustering algorithms. Along each row, each value represents the percentage

Table 4

Homogeneity and V-Measure Scores using BIRCH with different number of clusters and datasets (BIRCH parameters: threshold of 1 and branching factor of 50) (higher scores are better)

	Homogeneity			V-Measure		
# of Clusters	Annual	Monthly	TEF	Annual	Monthly	TEF
4	0.099	0.187	0.264	0.107	0.174	0.234
5	0.142	0.236	0.289	0.133	0.201	0.238
6	0.142	0.241	0.384	0.133	0.197	0.298
7	0.142	0.246	0.438	0.133	0.192	0.323
8	0.193	0.324	0.460	0.168	0.233	0.324
9	0.194	0.342	0.471	0.168	0.235	0.322
10	0.242	0.384	0.492	0.191	0.257	0.328
11	0.242	0.388	0.542	0.191	0.252	0.350
12	0.242	0.409	0.544	0.191	0.261	0.337

of climate stations that are part of a cluster as compared to the number of stations that are part of a KC category.

From table 5, we can infer that k-means and BIRCH provide dissimilar or alternative climate types as compared to the KC system. There are a few observations with high similarities, such as: for classification type *Dfb* and BIRCH on TEF yielding a 90% match rate, *Dfb* and k-means on TEF yielding a 72.8% rate, *Csb* and k-means on TEF yielding a match rate of 70.2%, and *Dfa* on BIRCH and monthly yielding a match rate of 84.4%. However, most match percentage numbers in table 5 hover below 50%. On the whole, the TEF and Monthly datasets yielded more similar clusters to the KC clusters. The level of dissimilarity also suggests value in considering alternative classification types that may be more pertinent to climate data records as compared to the empirical KC types.

Table 5

Comparing Clustering Results with Köppen Classification (KC)

	k-means			BIRCH		
Köppen (KC) Type	Annual	Monthly	TEF	Annual	Monthly	TEF
Cfa	37%	52.8%	39.2%	45.5%	35.3%	47.8%
BSk	54.1%	38.8%	34.8%	45.9%	31.5%	10.0%
Dfb	42.0%	67.4%	72.8%	42%	31.3%	90.0%
Dfa	32.8%	65.3%	50%	36.7%	84.4%	72.7%
Csb	19.3%	58.3%	70.2%	26.3%	59.0%	70.2%
Cfb	9.5%	4.5%	5%	4.7%	29.5%	5.3%
Dsb	20%	48.7%	33.3%	20%	48.7%	44.4%
Csa	5%	43.2%	66.7%	38.9%	15.4%	66.7%

Conclusions and Future Work

Overall, by analyzing the climate data from more than 3000 climate stations in the continental U.S. between 1946 and 2015, *k*-means and BIRCH offered better clustering solutions as compared to DBSCAN. This conclusion was formed after evaluating clusters using standardized metrics for clustering such as V-measure and Homogeneity Scores and by comparing the cluster compositions (climate stations) with the original KC types. *k*-means and BIRCH had approximately similar results. DBSCAN failed to provide effective clusters partly due to its known difficulties of scaling to high-dimensional databases (Gan et al., 2007). The derived dataset, TEF (Huang et al., 2017), provided similar clustering solutions to the KC system as compared to the monthly and annual climate datasets. A possible reason for this is that the empirical KC technique (used for evaluation) is designed around extreme climate data values and hence the TEF dataset, which is a climate extremes based dataset, fits that description. On the whole, a new clustering based approach based on actual climate station data has been proposed and implemented and the clusters generated provided for alternative climate type

configurations. A intuitive, dynamic climate clustering web-based tool has also been created to interactively generate climate clusters for the Continental U.S. and to visually compare the clustering solution with that proposed by the KC system. In addition, a number of possible directions for future research can be stated. One can apply the same clustering algorithms to evaluate how the climate classification types are changing over time and space. This can help geographers and meteorologists evaluate and assess a changing climate using machine learning algorithms. One can also extend the area of study beyond the continental U.S. to North America and to the entire World. This will enable comparisons on a large scale between machine learning derived clusters and those derived from empirical climatology based classification techniques. The work can be extended to incorporate future climate model scenarios spanning the time period 2020-2070 and evaluate how clusters and climate classification types change over time.

References

Bailey, R. G. (2009). *Ecosystem geography: from ecoregions to sites*. Springer Science & Business Media.

Biçici, E. and Yuret, D. (2007). Locally scaled density based clustering. In *International Conference on Adaptive and Natural Computing Algorithms*, pages 739–748. Springer.

Chen, D. and Chen, H. W. (2013). Using the köppen classification to quantify climate variation and change: An example for 1901–2010. *Environmental Development*, 6:69–79.

DeGaetano, A. T., Brown, T. J., Hilberg, S. D., Redmond, K., Robbins, K., Robinson, P., Shulski, M., and McGuirk, M. (2010). Toward regional climate services. *Bulletin of the American Meteorological Society*, 91(12):1633–1644.

DeGaetano, A. T., Noon, W., and Eggleston, K. L. (2015). Efficient access to climate products using acis web services. *Bulletin of the American Meteorological Society*, 96(2):173–180.

Diaz, H. F. and Eischeid, J. K. (2007). Disappearing alpine tundra of köppen climatic type in the western united states. *Geophysical Research Letters*, 34(18).

Donat, M. e. a. (2013). Climdex: Climate extremes indices. `https://www.climdex.org/indices.html`. (Accessed on 05/02/2018).

Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.

Fovell, R. and Fovell, M. (1993). CLIMATE ZONES OF THE CONTERMINOUS UNITED-STATES DEFINED USING CLUSTER-ANALYSIS. *JOURNAL OF CLIMATE*, 6(11):2103–2135.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Academic press.

Gan, G., Ma, C., and Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Huang, X., Sathiaraj, D., Wang, L., and Keim, B. (2017). Deriving data-driven insights from climate extreme indices for the continental US. In *2017 IEEE International Conference on Data Mining Workshops, ICDM Workshops 2017, New Orleans, LA, USA, November 18-21, 2017*, pages 303–312.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Köppen, W. and Geiger, R. (1930). *Handbuch der klimatologie*, volume 3. Gebrüder Borntraeger Berlin, Germany.

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

Liss, A., Koch, M., and Naumova, E. N. (2014). Redefining climate regions in the United States of America using satellite remote sensing and machine learning for public health applications. *GEOSPATIAL HEALTH*, 8(3):S647–S659.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mahlstein, I. and Knutti, R. (2010). Regional climate change patterns identified by cluster analysis. *CLIMATE DYNAMICS*, 35(4):587–600.

Melillo, J. M., Richmond, T. T., and Yohe, G. (2014). Climate change impacts in the united states. *Third National Climate Assessment*.

Mourtzinis, S., Edreira, J. I. R., Conley, S. P., and Grassini, P. (2017). From grid to field: Assessing quality of gridded weather data for agricultural applications. *European Journal of Agronomy*, 82:163 – 172.

Netzel, P. and Stepinski, T. (2016). On Using a Clustering Approach for Global Climate Classification. *JOURNAL OF CLIMATE*, 29(9):3387–3401.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel,

M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rosenberg, A. and Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

Zhang, T., Ramakrishnan, R., and Livny, M. (1996). Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM.

Zhang, X. and Yan, X. (2014). Spatiotemporal change in geographical distribution of global climate types in the context of climate warming. *CLIMATE DYNAMICS*, 43(3-4):595–605.

Zscheischler, J., Mahecha, M. D., and Harmeling, S. (2012). Climate classifications: the value of unsupervised clustering. *Procedia Computer Science*, 9:897 – 906. Proceedings of the International Conference on Computational Science, ICCS 2012.