DEPICTION OF CLIMATE ZONES IN THE CONTERMINOUS UNITED STATES USING REMOTE SENSING: APPLICATION TO VULNERABILITY ASSESSMENT AND PUBLIC HEALTH

Alexander Liss [a,b,d*], Magaly Koch [c,d], Elena N. Naumova [a,d]

[a] Department of Civil and Environmental Engineering, Tufts University
200 College Avenue, Medford, MA 02155, USA
Tel: +1 (617)-627-2653, Email: alexander.liss@tufts.edu

[b] Boston Financial Research, LLC
196 Boston Ave, Medford, MA 02155, USA
Tel: +1 (617)-807-0360, Email alexander.liss@bfinr.com

[c] Center for Remote Sensing, Boston University
725 Commonwealth Avenue, Boston, MA 02215, USA
Tel:+1 (617)-353-7302, Email: mkoch@bu.edu

[d] Tufts Initiative for Forecasting and Modeling of Infectious Diseases
196 Boston Avenue, Medford, MA 02155, USA
Tel: +1 (617)-627-2273, Email: elena.naumova@tufts.edu

KEY WORDS: Remote Sensing, NDVI, extreme weather, cold spells, vulnerability, human health, hypothermia hospitalization, classification, unsupervised, Köppen-Geiger, climate

Abstract: Identification of areas with comparable weather patterns has the potential to substantially improve forecasting of adverse health outcomes, disturbances in ecosystems, and infrastructure damages caused by extreme meteorological events. We aimed to design informative climate classification for assessing vulnerability and health risks. We proposed to define climate zones for the Conterminous United States using remote sensing and to compare it with the Köppen-Geiger divisions. Using the nationwide database of hospitalizations, maintained by the Centers of Medicare and Medicaid Services, we examined the utility of the unsupervised classification based on a time series of Normalized Difference Vegetation Index (NDVI) to assess vulnerability to extreme weather, defined by the likelihood of extreme cold weather, the number of residents prone to adverse effects, and the rate of severe health outcomes relevant to exposure to cold among older adults.
Individual biweekly MODIS NDVI image composites from July 2002 to July 2012 were aggregated, masked, and compiled into a 229-dimensional dataset. Longitudinal and temporal data redundancies were reduced by extracting the first 12 principal components. A "majority" 15x15 convolution was applied to the classification output. Locations of climate zones were determined by *k*-means unsupervised classification. NDVI-based zones exhibited a high degree of similarity with Köppen-Geiger divisions and a well-defined separation by the annual and seasonal temperature and precipitation values.
We determined that relatively warm and humid south eastern region had the largest population at risk (10.9 million) and the highest rate of hospitalizations due to exposure to cold (28.3 cases per 10,000 persons-at-risk). The dry south-western region had the lowest rate of hypothermia hospitalizations (14.6 cases per 10,000 persons-at-risk). Although vulnerability assessment and public health study was limited to a relatively short period 1991-2006 and exposures to cold weather among adults aged 65 and older, the proposed method demonstrates strong potentials for public health applications.

INTRODUCTION

The extreme weather events cause large number of adverse health outcomes. Heat waves and cold spells cause an increase in mortality and hospitalizations of vulnerable population, hurricanes and tornado cause an increase in trauma admissions.(Kalkstein and Davis 1989; Kalkstein and Smoyer 1993; Kalkstein 2000; Tan, Kalkstein et al. 2004). It has been shown that there could be a substantial difference in the rate of hospitalizations for similar events conditioned on the climate patterns at patient's locale. The climate differences play a major role in assessing vulnerability scores for different locations and efficiency of an early warning systems and public health interventions.(O'Neill and Ebi 2009)

Climate classification allows separating a large continuous territory into smaller regions with similar prevailing weather conditions. There are several climate classification methods in use today. Köppen-Geiger climate classification is one of generally accepted climate classification systems and was first published by a Russian/German climatologist Wladimir Köppen in 1884 (Köppen, Volken et al. 2011).

This system is based on a fundamental premise that a climate of the area can be defined by the type of prevailing vegetation in that area. Due to the complexity of a process required to determine a prevailing vegetation type at the time*, relatively simple proxies, such as temperature and precipitation were used to determine areas with similar prevailing vegetation and, therefore, climate. Another climate classification method, which is based on estimating soil water budget using the concept of potential evaportranspiration was suggested by C.W. Thornthwaite in 1948 (Thornthwaite 1948). It is, however, derived from the same set of meteorological proxies.

Considering the same concept of defining a climate based on vegetation type it is feasible to derive new proxies from emerging data sources, such as remote sensing. Plants reflect sunlight strongly in the Near Infrared (NIR) part of the spectrum (wavelengths of 700 to 1000 nanometers), while strongly absorbing in the visible spectrum (400 to 700 nanometers). The clouds and the bare soil, including snow, have an opposite reflectance properties, reflecting strongly in all visible bands, and absorbing NIR part of the spectrum. The Normalized Difference Vegetation Index (NDVI) is defined as a ratio between the difference and the sum of reflectance in these bands.

$$NDVI = \frac{NIR - Red}{NIR + Red}$$

The greater values of the index indicate healthier, growing vegetation cover. The spectral characteristics of the NDVI index allow the differentiation of individual types of vegetation. MODerate-resolution Imaging Spectroradiometer (MODIS) on board NASA's Terra and Astra satellites produces worldwide NDVI composites. The MODIS NDVI data set is available on a 16 day schedule.

A climate classification system can provide a valuable insight for potential vulnerability to extreme weather events in a uniform manner at refined temporal and spatial resolution for large regions. It is known that exposure to thermal extremes might produce detrimental health effects especially in vulnerable populations, such as children, elderly, and people with pre-existing health condition. The effects of extreme weather on human health may depend on build in infrastructure, mitigation strategies, and overall health status of affected population. By understanding the differences in those effects the preventive strategies can be better tailored to local need and climatic conditions.

The objectives of this study were to determine the climate zones of the conterminous United States using remote sensing data, specifically the Normalized Difference Vegetation Index and to perform the vulnerability assessment. We defined vulnerability based on the likelihood of extreme cold weather, the number of residents prone to adverse effects, and the rate of severe health outcomes relevant to exposure to cold among older adults and utilized the nationwide database of hospitalizations maintained by the Centers of Medicare and Medicaid Services (CMS).

DATA AND METHODS

Data sources

NDVI: The remote sensing data were downloaded from an FTP depository as a Hierarchical Data Format – Earth Observation Satellites (HDF-EOS) files. Each file contains 13 layers: 16 days NDVI, 16 days Extended Vegetation Index (EVI), Vegetation Index (VI) quality information, individual MIR, NIR, red and blue reflectance channels, average Sun Zenith Angle, NDVI and EVI pixels' standard deviation, number of pixels used, and pixel's reliability information. The average size of each file is 100 Mb. The files contain version 5 of the data set. The data is available for the period from July 4, 2002 till present. In our analysis we use data set for the period from Jul 4, 2002 to July 3, 2012. There were 229 individual snapshots in this data set. Each NDVI snapshot has a worldwide coverage, with -180 to +180, -90 to +90 degrees extent, placed on 0.05x0.05 degree grid. There are 3600 x 7200 pixels in each image. For the climate analysis of the conterminous United States we clipped the worldwide NDVI snapshot to a bounding box of -125.00 to -65.00, 24.00 to 50.00 degrees longitude and latitude respectively. A resulting image covers conterminous Unites States with resulting pixel dimensions of 521 x 1201 pixels. Two NDVI individual snapshots taken on September 2002 and May 2003 are presented in Figure 1 and Figure 2 respectively as examples.

Water Mask: The water reflectance pattern should produce very low values in NDVI data. Therefore open water in places such as oceans and large lakes should appear as very dark areas in NDVI composite. However, shallow water and water that is covered with water plants, such as algae, can produce uneven reflectance. In order to avoid

*End of 19th century-beginning of 20th century

the misclassification due to uncharacteristic reflectance, the water bodies needed to be masked in our analysis. Using the vector map of coast lines, lakes and large water bodies over the Conterminous United States we built a raster water mask. The water mask is a raster in the same geographic coordinates as an original NDVI composites from MODIS, with cell value of 1 indicating a land mass while value of 0 indicating large bodies of water as presented in Figure 3.

Meteorological parameters: We extracted minimum daily temperature, maximum daily temperature, mean daily temperature, mean daily dew point temperature, mean sea level atmospheric pressure, average and maximum daily wind speed, and maximum daily wind gusts, and amount of daily precipitation from major meteorological stations in the USA and its territories for the period from January 1st, 1991 to December 31st, 2006. We only considered those stations that provided meteorological data for a given day, therefore the number of stations varied between 1,058 stations in 1991 to 1,411 in 2006. For each meteorological station we collected information on its spatial location, such as latitude and longitude. Data was scrubbed of errors and verified for the absence of improbable values such as temperature readings outside of a conceivable range between 80F below zero (- 62C) to 130F above zero (+54.5C) or sudden intraday jumps of more than 75F.

Hospitalization data: To estimate population at risk we abstracted the annual counts of Medicare recipients aged 65 years and older from the CMS Denominator files for each zip code. We aggregated number of Medicare recipients per individual zip code and computed an aggregate number of recipients per county and state that zip code belongs to for analysis and visualization. To estimate the rate of hospitalizations due to cold weather, 74,030 records with ICD codes 991.0-991.8 were abstracted. Each record contained date of admission and zip code of residence. Using zip code of residence the average number of elderly people (65 y.o. or older) living in each climate zone for 16 years between 1991 and 2006 were computed. Similarly, zip-code based number of hospitalizations due to hypothermia and rates of hypothermia were estimated.

Analysis

The NDVI data is presented as a 229-dimensional data set. The data set represents biweekly time series of NDVI snapshots. By its nature the data set is redundant and heavily correlated across time and space. In order to reduce dimensionality of the data set, and extract as much as possible useful information, a principal component analysis was performed on the entire 229-dimensional data set. We used first 12 principal components in the subsequent analysis. We applied the *k*-means unsupervised classification algorithm to 12 principal components. The convolution with a "majority" 15x15 kernel applied to the output of the *k*-means algorithm with eight clusters. The "majority" kernel assigns value to each point in the data set equal to the value of the majority points surrounding the point, excluding those that do not have class values (ex: points over water). This transformation makes inter-cluster edges less jagged, more smoothly defined, and removes smaller clusters fully contained within larger areas.

To assign a conventional name category to determined clusters as climatic zone, each meteorological station within conterminous US was assigned to the closest cluster. Geographic coordinates (latitude and longitude) associated with each station determined the zone membership for each of the stations. For each cluster the basic statistics for meteorological parameters, such as precipitation and temperature, are computed for different seasons. The analysis included median annual temperature, median monthly temperature over the hottest month, and median monthly temperature over the coldest month, average temperature over the hot season (six summer months), median monthly temperature over the cold season (six winter months), average precipitation over the wettest month and average monthly precipitation over the coldest month. Each cluster was assigned a conventional category based on these results, that reflects generalized climate pattern in the particular zone, e.g. "Hot, humid summer; Warm, dry winter."

As each point in the original data set was assigned to a specific cluster (or NoData value) the data were exported via geotiff file to an ArcGIS ArcMap (version 10.0). Using Spatial Analyst and Zonal Analysis every zip code was assigned a majority class. The number of people at risk was limited to older adults, aged 65 and older residing in a given zip code, thus the rate of hypothermia in a particular climate zone was defined as number of hospitalizations divided by the average number of elderly people (65 y.o. or older) living in each climate zone in 1991-2006.

ENVI 4.7 is used for exploratory image analysis and visualization. Principal component decomposition, *k*-means analysis, majority smoothing and climate analysis is performed in MATLAB version 7.13.0.564. Spatial and zonal analysis and cartography executed in ESRI ArcMap and Spatial Analyst v.10.0. Health data were analyzed in MATLAB and SQL Server 2008.

RESULTS

We transformed an original 229 dimensional data set to principal components and extracted the 12 top components that collectively explain 92.6% of the variance in an original data set (Figure 4). The pseudo-color composite image of the first three components shows well-defined areas and a good separation of colors (Figure 5). Conversely, the last three principal components demonstrate a lot of irregular noise, banding and other image artifacts in large quantities (Figure 6).

To generate zones suitable for an assessment of health risks due to extreme weather, we performed an analysis on eight zones, which provided sufficient spatial and climate resolution. A large number of small specks scattered around edges and within larger zones present in the initial *k*-means* output (Figure 7). Smoothing an output with 15x15 majority kernel produced smooth areas, compatible with empirical data for climate distribution (Figure 8).

The produced eight zones also allowed us to identify a distinct trend in health-related outcomes. Meteorological and hospitalization parameters per each climate zone were summarized in Table 1. Meteorological parameters per defined climate zone are summarized in Table 1. The largest population was in Zone 3 with 10.5 million persons at risk. The smallest population was in Zone 8 with 440 thousand people at risk. The highest hospitalization rate due to hypothermia was observed in Zones 5 and 3 with 28.3 and 27.9 hospitalizations per 100,000 persons at risk, respectively. The lowest rate was in Zones 4 and7 with 14.7 and 15.3 hospitalizations per 100,000 persons at risk, respectively.

DISCUSSION

Climate classification is a challenging problem. The large and mostly smoothly changing spatial distribution of data, an absence of continuous measurements across the entire spatial extent, and time varying properties create a very difficult problem to tackle. Historically there were several attempts to classify global climate based on measuring or interpreting various proxy variables, however none of them were assessed for usability to mitigate the harmful effects of extreme weather on human health. This study directly addresses this gap. We demonstrated that eight selected classes differ by the potential vulnerability and the risk of hypothermia and can be utilized in epidemiological studies and public health practice. We implemented the *k*-means unsupervised classification algorithm, which groups data points based on the level of their "similarity", using inter- and intra-cluster distance measures. This algorithm is fast and efficient and the number of clusters was matched to a structure suggested by hierarchical and consensus cluster analysis by Fovell (Fovell and Fovell 1993; Fovell 1997). In our future work we plan to assess an optimal number of clusters suitable for public health research.

Widely used classification systems rely on sets of predetermined rules and a limited set of proxies. The Köppen-Geiger classification system uses temperature and precipitation as proxies for the vegetation types. It is clear that one of the major deficiencies of this system is an arbitrary assignment of its selection criteria. Another fundamental deficiency of the Köppen-Geiger climate classification system is the fact that it uses only limited subset of proxies, such as temperature and precipitation, to describe the particular climate. There are many other possible characteristics that can be used, such as soil type and its properties, average cloud cover, and solar radiation among many others. The advantage of using vegetation type is that it itself can be viewed as an aggregate proxy to many environmental characteristics that define a climate of each region. Obviously, in 1884 it was impossible to obtain reliable measures for vegetation type directly and their spatial distribution. With expanded availability of remotely sensed data it is now conceivable to better measure properties of vegetation cover on a regular temporal scale over a long duration.

To increase objectivity in defining climate characteristics based on massive volumes of data, the use of predetermined rules has to be supported by computer-intensive classification methodology. The clustering methods derive climate characteristics from statistical properties of collected data. The problem of data classification, or separation of a large data set into a number of distinct classes, is a well-known problem and often arises in many areas of data analysis, including data mining and pattern recognition, image processing and target acquisition. There are two major groups of algorithms employed – a supervised and an unsupervised (Theodoridis and Koutroumbas 2009). Supervised classification consists usually of two major steps. At first step an expert assigns some data regions to known classes, and then an algorithm identifies similar unmarked areas. There are many different method of determining "similarity." The supervised classification works best when an "expert" is available and it is possible to determine a "true" group membership before the analysis. Unsupervised classification, on the other hand, uses similarity and dissimilarity measures derived from the data itself, to separate it into similar groups.

* Classification results using *k*-means algorithm for 8 clusters

Usually the goal of a classification algorithm is to maximize similarity of the clusters, while increasing dissimilarity between the clusters. In our case the goal is to determine the best allocation of climate zones, without assigning any prior knowledge. We do not know "true" climate beforehand. Therefore, the unsupervised classification is the most appropriate method.

Several algorithms exist for unsupervised classification, such as *k*-means, isodata, fuzzy methods (allowing some areas to belong to several clusters at the same time with various degrees of certainty), and hierarchical clustering (Wilks 2011). The hierarchical clustering method is an attractive classification method for climate research.. Fovell has classified US climate based on hierarchical clustering and consensus clustering (Fovell and Fovell 1993; Fovell 1997). The classification was performed based on temperature and precipitation reading from climate stations. The classification produced 8, 14 and 25 clusters with various levels of details. Climate zones of Turkey were redefined by using the hierarchical cluster analysis based on temperature and precipitation data (Unal, Kindap et al. 2003). Climate data from 117 climate stations in Iran was used to determine 6 climate zones using cluster analysis (Alijani, Ghohroudi et al. 2008).

It is very unlikely that a small area would have a climate pattern that is sharply distinct from surrounding areas. K-means algorithm produces jagged edges and a lot of small "specks" within larger homogenous areas (Figure 7). The convolution with a "majority" 15x15 kernel applied to the output of the *k*-means algorithm with selected number of clusters assigned value to each point in the data set equal to the value of the majority points surrounding the point, excluding those that do not have class values (ex: points over water). This transformation makes inter-cluster edges less jagged, more smoothly defined, and it removes smaller clusters placed within large areas.

To facilitate classification process we applied a principal component analysis, a mathematical orthogonal transformation, which creates a set of variables from an original data set that is linearly independent. The transformation is defined in such a way that the first principal component contains the largest possible variance. Each of the following components also has the largest possible variance given that it is orthogonal to the already defined components. This causes the majority of useful information to concentrate in the first few components. The number of principal components to include in the following analysis is determined by the marginal explained total variance, which changed rapidly with an inclusion of the first few components and then slowly diminishes with the inclusion of additional components. The NDVI data by its nature are redundant and heavily correlated across time and space. In order to reduce dimensionality of the data set, and to extract as much as possible useful information, we performed a principal component analysis on the entire 229-dimensional data set and included first 12 principal components.

CONCLUSIONS & RECOMMENDATIONS

Objective clustering method produced sensible climate divisions for the conterminous US. The hierarchical clustering method is an attractive classification method for climate research. It produces intuitive results where one large cluster is subdivided into smaller ones when the number of classes increases. However, the hierarchical clustering requires building a large distance matrix. The distance matrix grows very fast and, with large number of data points, quickly becomes intractable. The problem we have been working with would have required building a distance matrix with approximately 150 billion elements – clearly a difficult task. We adapted a number of algorithms to define climate zones suitable for public health applications. The utility of climate classification based on remotely sensed data in mitigating the adverse effects of severe weather on human health has a strong potential and need to be further explored by both meteorological and public health communities. For example, the heavily populated areas in a warm south-east Sunbelt area should be carefully explored and evaluated for developing preventive strategies to reduce hypothermia hospitalizations in vulnerable populations. The analysis of cluster optimality and validity and extension of health-based climate classification to other regions is recommended.

TABLES AND FIGURES

Figure 1. NDVI composite, Sep 22, 2002, Conterminous US

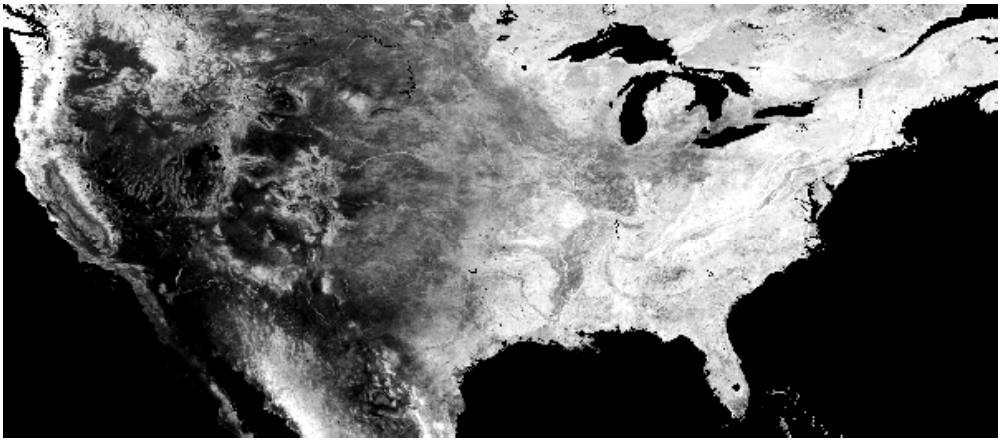

Figure 2. NDVI composite, May 01, 2003, Conterminous US

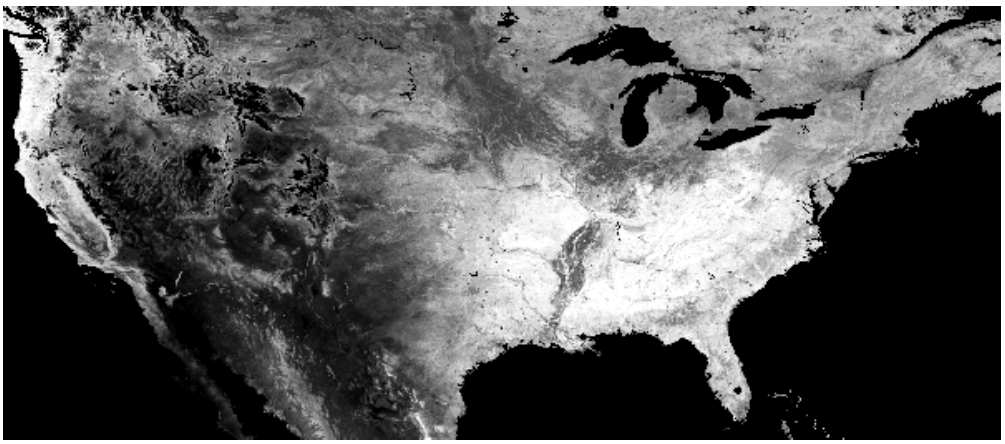

Figure 3. Water mask raster

Figure 4. Principal components cumulative variance

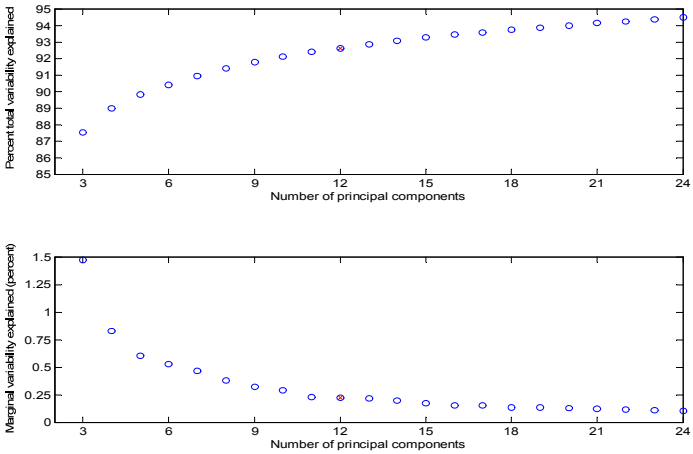

Figure 5. Pseudo-color composite of the first three principal components

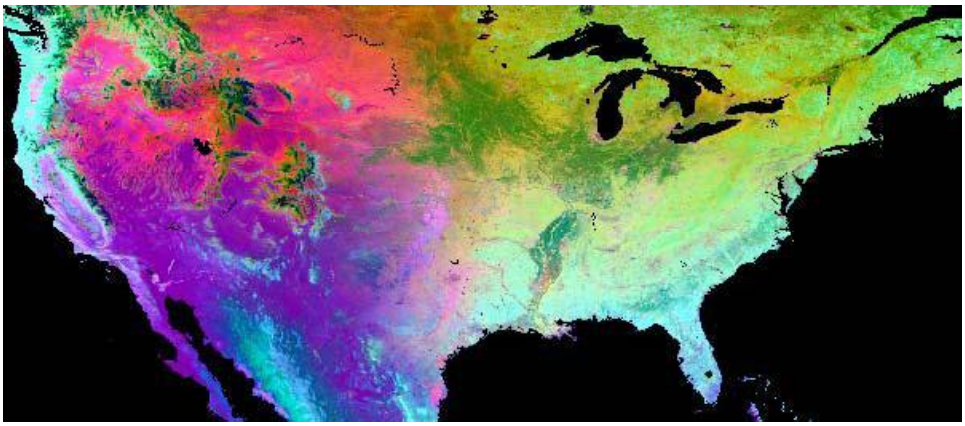

Figure 6. Pseudo-color composite of the last three principal components

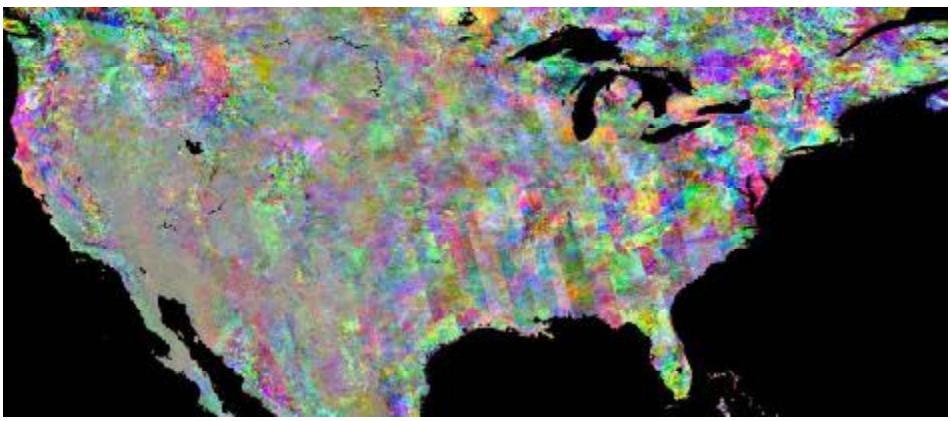

Figure 7. An output of *k*-means algorithm with eight zones defined

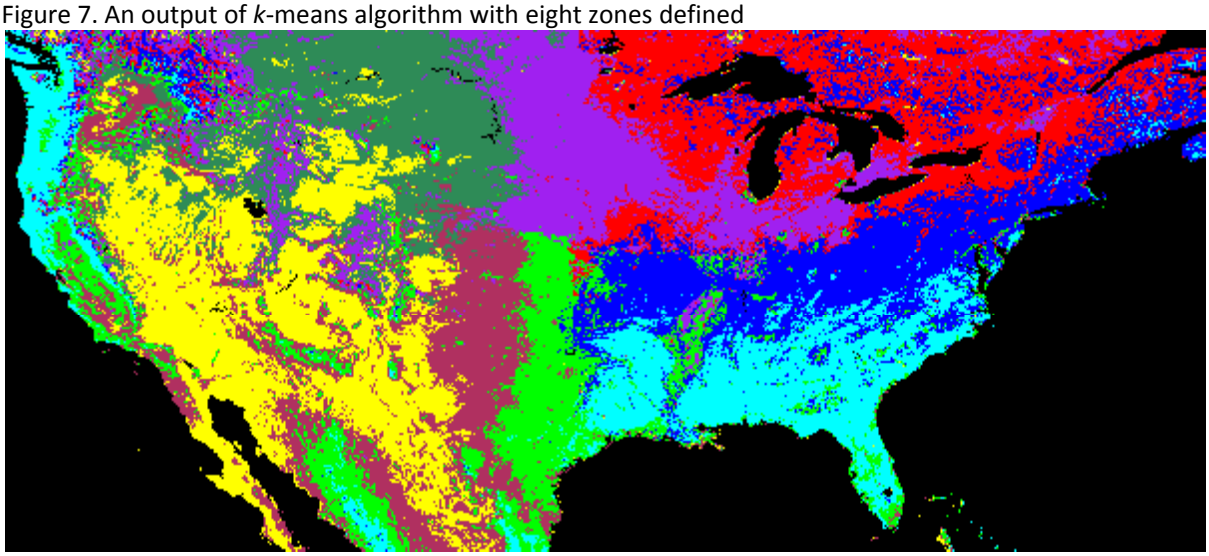

Figure 8. A smoothed (15x15 majority kernel) version of *k*-means output with superimposed political borders

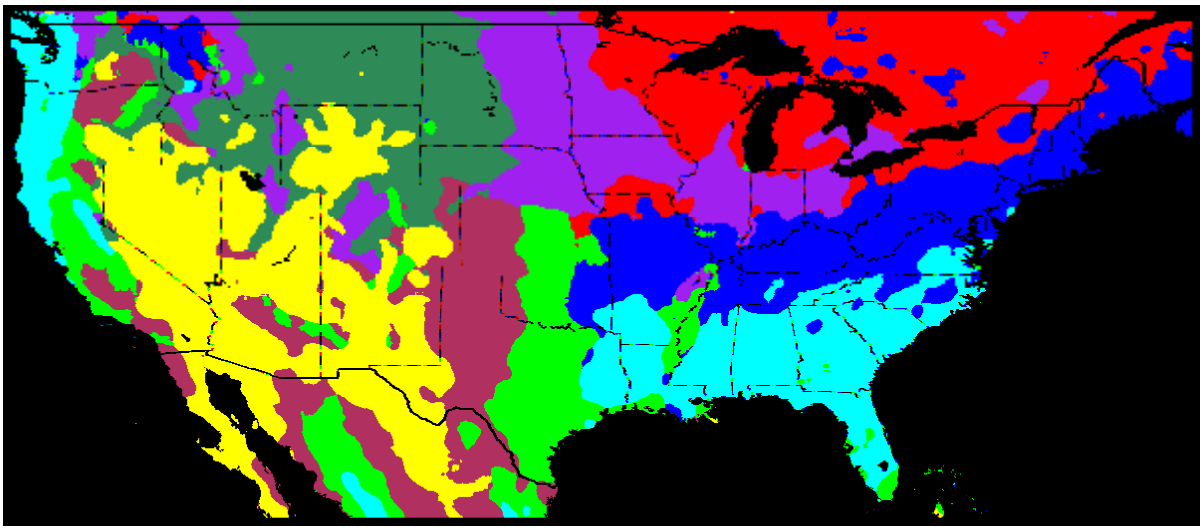

Table 1. Meteorological parameters per defined climate zone

	Abbreviation	Description	Median Annual Temperature, °C	Median Annual Precipitation, mm	Hot Season Temperature, °C	Cold Season Temperature, °C	Hot Month Temperature, °C	Cold Month Temperature, °C	Elevation, m	Population at Risk MM	Hospitalizations	Hospitalizations Per 10K population at-risk
1	CCd	Cool, wet summers; Cold, moderately dry winters	7.5 (5.9;9.1)	845 (768;923)	15.7 (14.4;17.0)	-0.7 (-2.6;1.3)	21 (19.7;22.3)	-6.3 (-8.7;-4.0)	271 (204;339)	3.28	7,509	22.88
2	HHd	Hot, wet summers; Hot, moderately wet winters	16.1 (13.4;18.9)	737 (484;990)	22.9 (20.1;25.8)	10.7 (7.4;14.0)	27.6 (25.0;30.1)	6.9 (2.4;11.4)	193 (7;379)	5.13	10,378	20.22
3	TTw	Temperate, wet summers; Temperate, wet winters	12.1 (10.2;14)	1088 (994;1183)	19.5 (17.6;21.4)	4.5 (2.6;6.5)	24.1 (22.3;25.8)	-0.4 (-2.4;1.6)	183 (87;280)	10.93	24,803	22.70
4	TTa	Temperate, arid summers; Temperate, arid winters	11.1 (7.6;14.7)	253 (184;322)	18.7 (15.5;21.9)	3.9 (.3;7.5)	24.2 (21.5;26.8)	-0.8 (-5.3;3.6)	1370 (1110;1631)	1.23	1,790	14.58
5	HHw	Hot, wet summers; Hot, wet winters	16.6 (14;19.1)	1254 (1115;1393)	22.9 (19.7;26.1)	10.5 (7.9;13.1)	26.4 (24.1;28.6)	6.7 (4.1;9.2)	69 (1;149)	5.34	15,109	28.28
6	TCd	Warm, wet summers; Cold moderately dry winters	8.8 (6.8;10.8)	699 (540;859)	17.7 (16.3;19.1)	0 (-2.5;2.4)	22.9 (21.7;24.1)	-6.1 (-9.2;-3.0)	335 (206;465)	3.35	8,156	24.31
7	TTa	Warm, moderately dry summers; Moderate, arid winters	12.8 (9.9;15.7)	432 (351;513)	20.3 (18.0;22.7)	5.5 (2.2;8.9)	25.3 (23.1;27.5)	0.9 (-2.7;4.5)	720 (446;994)	1.95	2,976	15.23
8	CCa	Cool, moderately dry summers; Cold, arid winters	6.8 (5.4;8.1)	358 (304;413)	14.8 (13.6;16.0)	-1.1 (-2.9;.7)	20.7 (19.1;22.3)	-6.5 (-8.8;-4.2)	1169 (766;1573)	0.44	1,240	27.89

REFERENCES

Alijani, B., M. Ghohroudi, et al. (2008). "Developing a climate model for Iran using GIS." Theoretical and Applied Climatology **92**(1-2): 103-112.

Fovell, R. G. (1997). "Consensus clustering of U.S. temperature and precipitation data." Journal of Climate **10**(6): 1405-1427.

Fovell, R. G. and M. Y. C. Fovell (1993). "Climate zones of the conterminous United States defined using cluster analysis." Journal of Climate **6**(11): 2103-2135.

Kalkstein, L. S. (2000). "Saving lives during extreme weather in summer: Interventions from local health agencies and doctors can reduce mortality." British Medical Journal **321**(7262): 650-651.

Kalkstein, L. S. and R. E. Davis (1989). "Weather and human mortality: an evaluation of demographic and interregional responses in the United States." Annals - Association of American Geographers **79**(1): 44-64.

Kalkstein, L. S. and K. E. Smoyer (1993). "The impact of climate change on human health: Some international implications." Experientia **49**(11): 969-979.

Köppen, W., E. Volken, et al. (2011). "The thermal zones of the Earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world." Meteorologische Zeitschrift **20**(3): 351-360.

O'Neill, M. S. and K. L. Ebi (2009). "Temperature extremes and health: Impacts of climate variability and change in the United States." Journal of Occupational and Environmental Medicine **51**(1): 13-25.

Tan, J., L. S. Kalkstein, et al. (2004). "An operational heat/health warning system in shanghai." International Journal of Biometeorology **48**(3): 157-162.

Theodoridis, S. and K. Koutroumbas (2009). Pattern recognition. London, Academic Press. **1:** 961.

Thornthwaite, C. W. (1948). "An approach toward a rational classification of climate." Geogr. Review **38**: 55-94.

Unal, Y., T. Kindap, et al. (2003). "Redefining the climate zones of Turkey using cluster analysis." International Journal of Climatology **23**(9): 1045-1055.

Wilks, D. S. (2011). Cluster Analysis. Statistical Methods in the atmospheric sciences. USA, Academic Press. **100:** 603-616.