Geospatial Health 8(3), 2014, pp. S647-S659

Redefining climate regions in the United States of America using satellite remote sensing and machine learning for public health applications

Alexander Liss[1,2*], Magaly Koch[2,3], Elena N. Naumova[1,2]

[1]*Department of Civil and Environmental Engineering, Tufts University, Medford, USA;* [2]*Tufts Initiative for Forecasting and Modeling of Infectious Diseases, Medford, USA;* [3]*Center for Remote Sensing, Boston University, Boston, USA*

Abstract. Existing climate classification has not been designed for an efficient handling of public health scenarios. This work aims to design an objective spatial climate regionalization method for assessing health risks in response to extreme weather. Specific climate regions for the conterminous United States of America (USA) were defined using satellite remote sensing (RS) data and compared with the conventional Köppen-Geiger (KG) divisions. Using the nationwide database of hospitalisations among the elderly (≥65 year olds), we examined the utility of a RS-based climate regionalization to assess public health risk due to extreme weather, by comparing the rate of hospitalisations in response to thermal extremes across climatic regions. Satellite image composites from 2002-2012 were aggregated, masked and compiled into a multi-dimensional dataset. The conterminous USA was classified into 8 distinct regions using a stepwise regionalization approach to limit noise and collinearity (LKN), which exhibited a high degree of consistency with the KG regions and a well-defined regional delineation by annual and seasonal temperature and precipitation values. The most populous was a temperate wet region (10.9 million), while the highest rate of hospitalisations due to exposure to heat and cold (9.6 and 17.7 cases per 100,000 persons at risk, respectively) was observed in the relatively warm and humid south-eastern region. RS-based regionalization demonstrates strong potential for assessing the adverse effects of severe weather on human health and for decision support. Its utility in forecasting and mitigating these effects has to be further explored.

Keywords: remote sensing, LKN-regionalization, machine learning, morbidity, climate, classification, decision support, United States of America.

Introduction

The widely adopted Köppen-Geiger (KG) climate classification system, developed in 1884 by the Russian/German climatologist Wladimir Köppen, is based on the concept that regional climate can be defined by a prevalent phenology (Geiger and Pohl, 1954; Koppen et al., 2011). Due to the complexities associated with a reliable determination of phenology over large and remote areas in the pre-satellite era, available proxies, such as temperature and precipitation, were used to determine regions with similar climate. While the KG climate classification is still actively used to quantify climate variation (Chen and Chen, 2013), the arbitrary nature of suggested parameters in this classification system has been criticised (Thornthwaite, 1943). Although an alternative method, based on a concept of potential evapotranspiration and soil water budget, has been introduced, this common climate classification approach is also derived from the same set of meteorological proxies (Thornthwaite, 1931, 1948). Emerging data sources, such as vegetation indices, spectral radiation patterns, surface albedo and other measures, available with the advent of satellite remote sensing (RS) technology, allow for a definition of a prevailing phenological pattern at virtually every place worldwide. It is now feasible to derive such patterns directly from RS data using one of the existing vegetation indices, e.g. the normalized difference vegetation index (NDVI) (Carroll et al., 2000). The spectral characteristics of the NDVI allow the differentiation of phenology and states of vegetation. The moderate-resolution imaging spectraradiometer (MODIS) on board NASA’s Terra and Aqua satellites generate data for worldwide NDVI composites with 16 days overlapping temporal resolution and various spatial resolutions (LPDAAC-NASA, 2000-2013). The temporal feature offers better understanding of local weather variability and its implications for spatial regionalization.

Corresponding author:
Elena N. Naumova
Tufts Initiative for Forecasting and Modeling of Infectious Diseases
196 Boston Avenue, Medford, MA 02155, USA
Tel. +1 617-627-2273
E-mail: elena.naumova@tufts.edu

Spatial regionalization can be viewed as a special case of a classification problem that aims to assign a finite set of labels (i.e. climate categories or classes) to a very large number of multidimensional objects (i.e. pixels, representing a defined area on the ground) based on their similarity. Classification algorithms are typically divided into supervised classification, which seeks to assign labels known *a priori* and unsupervised classification, which derives classes based on properties of the data (Arnold, 2003). Spatial regionalization is usually discussed in the context of post office service areas, school districts and political jurisdictions (Barkan et al., 2006; Tasnádi, 2011; Tong and Murray, 2012). It was shown that *p-Regions* problem's complexity belongs to a non-deterministic polynomial-hard (also known as *NP-hard*) group of problems and quickly becomes computationally intractable in finite time (Duque et al., 2011). Several attempts to use machine learning and objective climate regionalization methods based on intrinsic properties of the data have been made in the past. Consensus clustering and hierarchical clustering, an attractive classification method for climate research that produces intuitive results where one large cluster is subdivided into smaller ones when the number of required classes increases (Fovell and Fovell, 1993; Fovell, 1997), have been used for this task. The hierarchical clustering was performed using only temperature and precipitation proxies, obtained from climate stations, to reduce computational complexity. Fovell's regionalization produced 8, 14 and 25 nested clusters within the conterminous United States of America (USA) with various levels of detail, e.g. Unal et al. (2003) used hierarchical clustering to redefine climate zones of Turkey and Arbabi (2011) classified the climate in Iran based on meteorological observations into 4 clusters. While attractive and intuitive, hierarchical clustering requires building a complete distance matrix. With increased numbers of data points and dimensions in the dataset this quickly becomes infeasible in finite time and requires climate proxies to reduce complexities. One of the possible solutions to the problem of the exponentially growing size of the complete distance matrix is to relax a contiguity requirement and to define a number of desired distinct regions *a-priory.*

Spatial climate regionalization is essential for large-scale epidemiological studies to properly address geographic heterogeneity. Administrative boundaries are typically used to design relatively homogeneous administrative units, which are further adjusted in models by including information on population composition and density. Our own research highlights the need to consider both socio-economic and climatic characteristics in understanding temporal variability of weather-sensitive health conditions (Cohen et al., 2008; Chui et al., 2009; Jagai et al., 2009, 2012).

Exposure to extreme weather events is associated with a wide range of adverse health consequences. Heat waves and cold spells cause an increase in mortality and morbidity (Goldberg et al., 2011; Anderson et al., 2013; Xie et al., 2013; Margolis, 2014). The relationships between low environmental temperature, vulnerability of specific population groups, common co-morbidities, susceptibility and resilience of an individual have been a subject of intense research for the last quarter of the century. The recent report from the U.S. Global Change Research Program (National Climate Assessment, 2014) emphasises that certain people and communities in the USA are especially vulnerable climate change vulnerable with respect to climate change, including children, the elderly, the sick, the poor and some communities of colour (Watts, 1972; Rango, 1980, 1985, Bragagni et al., 2012). Individuals aged 75 years and over are five times more likely than a younger person to suffer mortality as a result of lowered or raised core body temperature due to exposure to cold or heat (Rango, 1980). According to a study on accidental hypothermia, 73% of elderly patients died as a result (Seman et al., 2009). The last canicule of summer 2003 remains a searing memory for Europe, France, in particular. Record highs were broken in many French cities when relentless heat dragged on for over a week. It was the hottest and the deadliest summer Europe had seen in 150 years with 14,802 lives claimed by heat wave, the vast majority of which were people over the age of 75. Concurrent illnesses, common among individuals of advanced age, exacerbate the health risks associated with exposure to extreme weather events. Older people are more susceptible to the heat or cold injury due to physiological and psychological changes attributed to aging (Horvath and Rochelle, 1977).

The effects of extreme weather on human health may depend on built infrastructure, mitigation strategies and overall health status of the affected population. Multiple studies demonstrate that prevailing climate patterns in the patient's geographic region may account for substantial differences in health outcomes from similar weather conditions (Kalkstein and Davis, 1989; Montero et al., 2012; Ebi and Mills, 2013). Climate differences play a major role in assessing population vulnerability scores and in developing efficient early warning systems and public health interventions (O'Neill and Ebi, 2009; Johnson et al., 2012; Aubrecht et al., 2013; Chebana et al., 2013;

Pascal et al., 2013; Xie et al., 2013). This paper links health consequences due to the weather events as short term atmospheric conditions and illustrates that such impact may differ in different locations based on their climate patterns. These findings are especially valuable for predicting short-term health effects due to extreme weather as climate changes. The hospitalisations and adverse health consequences are measured in relatively short intervals, usually days but sometimes several days or weeks, if lag structure or incubation period/vector cycle is considered. The long-term climate patterns modify these effects. 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. However systematic investigations of the utility of these classification systems for public health applications and their ability to reflect population at risk, although needed for epidemiological applications, are lacking. By understanding the differences in those effects the preventive strategies can be better tailored to local needs and climatic conditions.

The objectives of this study were two: dividing the to divide the contiguous area of the conterminous USA into spatial regions based on the prevailing phenology using machine learning algorithms; then performing an assessment of suitability of proposed regionalization for epidemiological studies using a nationwide database of hospitalisations of elderly Americans maintained by the Centers of Medicare and Medicaid Services (CMS).

Materials and methods

Remote Sensing

To accomplish the first objective, version 5 of MODIS NDVI mesurements and pixel quality (QA) data from July 4, 2002 to July 3, 2012 were obtained through the online Data Pool at the NASA Land Processes Distributed Active Archive Center (LP DAAC) and the United States Geological Survey (USGS)/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (LPDAAC-NASA, 2000-2013). Each file, downloaded in hierarchical data format from Earth observation satellites (HDF-EOS) contained composite NDVI, vegetation index (VI-QA) quality information and other datasets. Each NDVI composite has a worldwide coverage that was clipped to a bounding box covering longitude 65.00 to 125.00 and latitude 24.00 to 50.00 for the climate analysis of the conterminous United States. The resulting 521 by 1,201 pixel images covered the entire extent of the USA with 8 days temporal resolution. Therefore vegetation index data was aggregated in a layered space-time series containing 430 individual time steps. The water reflectance pattern differs significantly from almost any other surface material by absorbing most of the incoming radiation. In order to avoid misclassification due to the water reflectance pattern, the water bodies needed to be masked. There are many different ways water pixels can be excluded from NDVI composite for this analysis, e.g. TerraSarX radar data, the normalized difference water index (NDWI) or the normalized humidity index (NHI). For the continental USA land features, such as roads or water bodies, are well mapped. Consequently, the existing geographical information systems (GIS) water body features were used to mask out open water. In other cases, when the region is remote, less populated and less well charted, the other methods can be used. Using the vector map of coast lines, lakes and large water bodies over the conterminous United States, a corresponding raster mask was built in the same geographic coordinates and spatial resolution as the original multilayer composites and applied to the original composites to exclude pixels associated with water bodies from the analysis.

Spatial regionalization

Using satellite imagery, the spatial regionalization was performed in the following three major delineated steps aiming to limit noise and collinearity (the L-step), classify the regions using k-means algorithm (the K-step), and nominate the selected area for future analysis (the N-step), LKN in short. First, dimensionality and space-time collinearity were reduced using principal components analysis (PCA), whose goal was to remove additive noise and orthogonalise the dataset. The data in the original time series, consisting of 430 individual NDVI composites constructed with 8 days interval naturally, have a very high degree of spatial and temporal collinearity. The first 12 components out of the total of 430 from PCA analysis collectively maintain more than 90 percent of the variability within the original time series. After applying PCA, the resulting amount of variance was significant enough to capture small difference in weather patterns and, at the same time, to remove noise and unwanted autocorrelation and time variance. Next, given a set of n data points distributed over time t $x_{n,t} \in R^{n*t}$ the clustering algorithm was applied to minimise the clustering objective function:

$$(c_1, \ldots, c_k) = \frac{1}{n} \sum_{i=1}^{n} \min_{k=1, \ldots, k} D\ (x_i, c_k),$$

where $c_1, \ldots, c_k$ represent the centres of the respective clusters 1 to k, and $D\ (x_i, c_k)$, is a distance measure between each point and centre of the clusters, and k is the number of classes that the data set needs to be partitioned into. To implement a k-means classification algorithm, a target number of regions k was determined by maximising cluster validity index. The Calinski-Harabasz Cluster Validity Index (CH-index) also known as the Variance Ratio Criterion (VRC) (Caliński and Harabasz, 1974) was used to measures within-group and between-group dispersion using the following equation:

$$VRC = \frac{\sum_i n_i\, D^2\ (c_i, c)}{\sum_i \sum_x c_i\, D^2\ (x_i, c)} * \frac{n - NC}{NC - 1}$$

where n is the total number of points, NC the number of clusters, $D^2\ (c_i, c)$ the within group distances and $D^2\ (x_i, c_i)$ the distance between cluster centres. VRC assigns a higher score to a partition that is more compact, well defined and contains well separated clusters. Based on the results of these steps, the area of conterminous USA was partitioned with an unsupervised classification algorithm and the preliminary assignment of regions was performed. Finally, kernel convolution to the output of the classification algorithm was applied to create regions for nomination (the N-step). The convolution made the inter-cluster edges less jagged and more smoothly defined. This step also helps to remove discontinuities and very small clusters fully contained within larger regions.

Climate zone properties

Minimum and maximum daily temperatures and amounts of precipitation were extracted from major ground based meteorological stations in the United States and its territories, as well as from water buoys provided by the National Oceanic and Atmospheric Administration National Climatic Data Center, Global Historical Climate Network (NOAA-NCDC GHCN v.2) (NOAA, 2014). For each day during the study period, gauges were included that provided valid weather data for that day. The number of meteorological stations included in the dataset each day varied from a minimum of 1,058 to the maximum of 1,411 stations throughout this period. For each station, information on the station's spatial location, such as its latitude and longitude, was recorded in addition to the meteriological data. Data was scrubbed of errors and checked for the absence of improbable values, such as temperature readings outside of a conceivable range between 80F below zero (-62 °C) to 130F above zero (+54.5 °C), sudden intraday jumps of more than 75F or other outliers. Surface temperature was interpolated to the same scale as NDVI data using a modified Shepard's Inverse Distance Weighting method (Franke and Nielson, 1980; Renka, 1988).

For each of the climatic regions, defined by the spatial regionalization method described above, basic statistics were computed for daily minimum and maximum temperature and amount of precipitation for different seasons. The analysis included mean annual daily temperature and precipitation, mean temperature over the hot season (June-August) and the mean daily temperature and precipitation over the cold season (December-February). Based on these results, each region was assigned a unique name reflecting the generalised climate pattern in this particular region, e.g. hot, humid summer; warm, dry winter, etc., which is used as the base for the LKN regionalization.

Analysis of suitability for epidemiological studies

To complete the second objective, population at risk within each region and the rate of hospitalisations due to cold weather and hot weather were estimated and compared across climate regions. The population at risk was limited to Medicare subscribers aged 65 and older and residing in a given region and estimated from the annual counts of eligible Medicare beneficiaries separately for each zip code, age group and gender. The number of Medicare beneficiaries was aggregated per each individual zip code and an aggregate number of recipients per county and state belonging to that zip code was computed for spatial analysis and visualisation.

Rates of hypothermia for each of the suggested regions were estimated using the number of hospitalisations due to hypothermia and population at risk per spatial unit. Hospitalisation records (n=74,030) was abstracted that contained medical diagnoses recorded as "Effects of reduced temperature" by the International Classification of Diseases, 9th revision (ICD-9 codes 991.0-991.8) from approximately 219 million hospitalisation records obtained from the Center for Medicare and Medicaid Services for the 16-year period between January 1st, 1991 and December 31st, 2006. We abstracted 41,927 hospitalisation

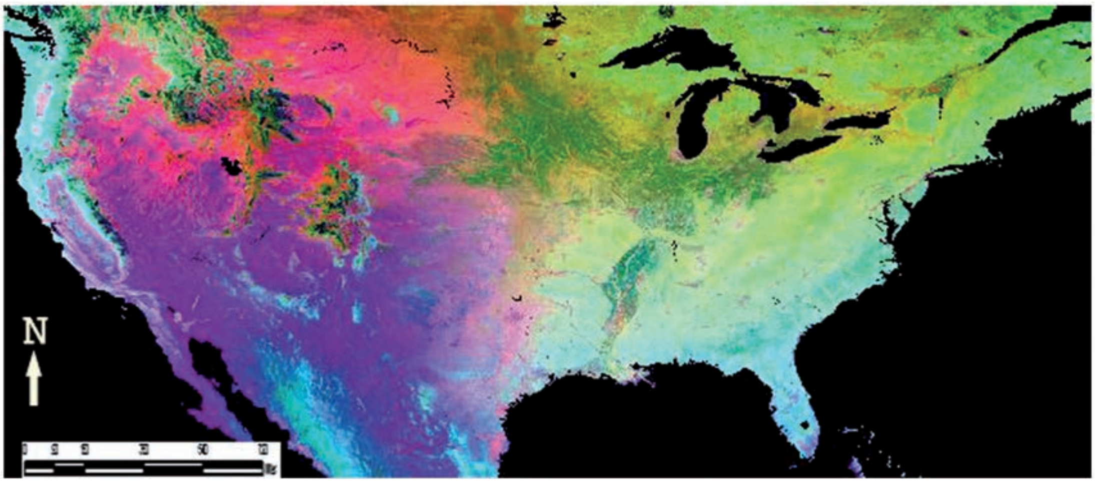

Fig. 1. Pseudo-colour composite of the first three principal components.

records that contained medical diagnoses recorded as “Effects of heat and light” (ICD-9 codes 992.0-992.9) from the same set of hospitalisation records. Each record contained the date of admission and zip code of the patient’s residence. Thus, each hospitalisation record and diagnose in the original dataset could be assigned to a specific region based on patient’s place of residence.

The rate of a disease in a particular region was defined as the number of hospitalisations divided by the average number of elderly persons living in this region. For hospitalisations related to exposure to heat or cold exposure was defined as a number of people-day residents were exposed to five degree F temperature band in each zone. For exposure to cold the range of minimum daily temperatures was bounded between -35 °F (-37.2 °C) and 65 °F (18.3 °C). For exposure to heat the range of minimum daily temperatures was similarly bounded between 40 °F (4.5 °C) and 120 °F (48.9 °C). In case a region did not experience the full range of temperatures described above, the actual range of temperatures was used.

Results

The first 12 principal components collectively explained 92.6% of the variance in the original dataset. The pseudo-colour composite image of the first three components showed low noise and the absence of image artefacts. Fig. 1 presents a pseudo-colour composite image of the first three components each assigned to a separate colour channel. This figure demonstrates smooth changing colours without any noticeable noise in the image. Each colour combination represents a different combination of values for the first three principal components. Conversely, an image derived from the last three principal components demonstrates a lot of irregular noise, banding and other image artefacts in large quantities. Fig. 2 presents a pseudo-colour composite image of the last

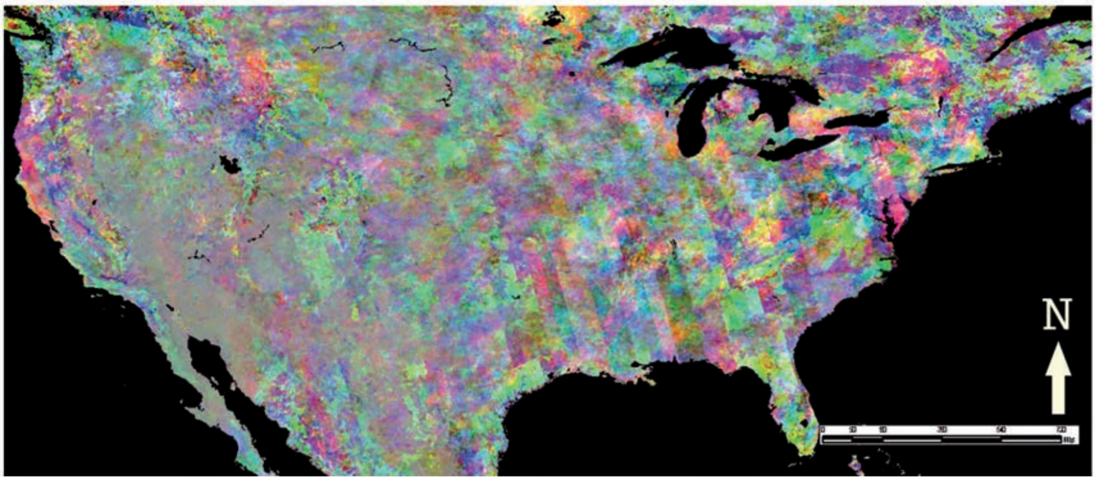

Fig. 2. Pseudo-colour composite of the last three principal components.

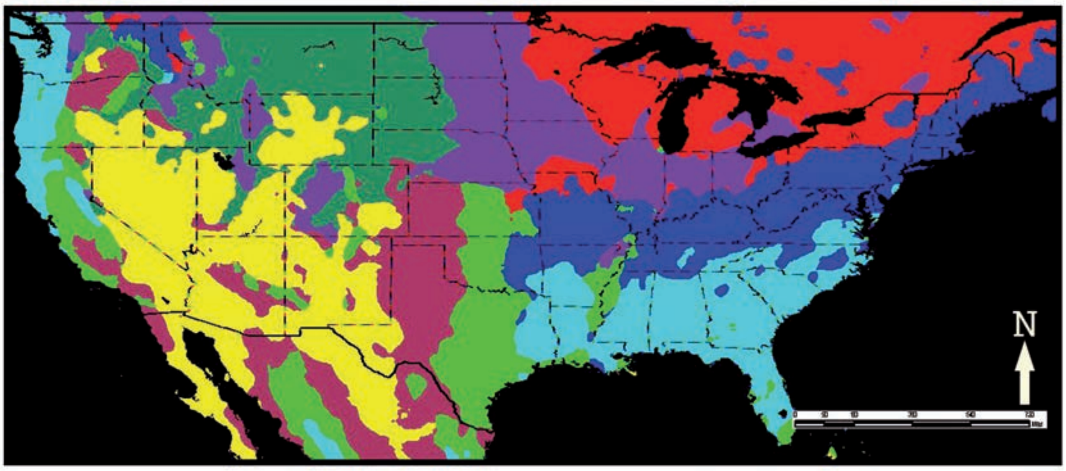

Fig. 3. Regions over conterminous USA as defined by the proposed non-supervised LKN regionalization prior to nomination. Colors represent distinct labels assigned by the classification procedure, but prior to climate categories nomination.

three components each assigned to a separate colour channel. This figure demonstrates that PCA decomposition removed significant amount of random noise and image artefacts prior to the classification. It has been shown that the PCA automatically projects to the subspace where the global solution of K-means clustering lies, and thus facilitates K-means clustering to find near-optimal solutions (Ding and He, 2004). Eight climate regions maximised the VRC. Consequently the entire spatial extent of the conterminous USA was classified into eight distinct regions using K-means classification. Subsequently a 15 x 15 majority kernel filter was applied that produced smooth, mostly contiguous regions. Fig. 3 presents the political boundaries of the conterminous USA overlaid on the output from the classification algorithm and the geographic distribution of different regions, as defined by the method described above.

Basic statistics of climate characteristics for the study period for each of the eight regions were computed. The proposed method achieves good separation between regions for customary weather proxies, such as temperature and precipitation, for both cold and hot seasons. Figs. 5 and 6 show climatic separation by

Table 1. Average annual and seasonal, daily minimum ambient temperatures and precipitation levels per each of the eight defined zones in conterminous USA.

Prevailing climate with LKN colour code	Elevation (m)	Annual		Hot Season*		Cold Season**	
		Temperature (°C)	Precipitation (mm)	Temperature (°C)	Precipitation (mm)	Temperature (°C)	Precipitation (mm)
Cool, wet summers with cold, moderately dry winters (CwCd)	271 (204; 339)	7.5 (5.9; 9.1)	845 (768; 923)	15.2 (13.4; 17.0)	1.03 (0.35; 1.71)	-7.5 (-11.9; -3.1)	0.68 (0.26; 1.09)
Hot, wet summers with warm, moderately wet winters (HwHd)	193 (7; 379)	16.1 (13.4; 18.9)	737 (484; 990)	20.4 (19.6; 21.3)	1.03 (0.57; 1.49)	4.3 (2.6; 6.0)	1.17 (0.44; 1.89)
Temperate, wet summers with temperate, wet winters (TwTw)	183 (87; 280)	12.1 (10.2; 14)	1088 (994; 1183)	17.6 (16.3; 18.9)	1.31 (0.43; 2.18)	-2.8 (-5.9; 0.4)	1.34 (0.73; 1.94)
Temperate, arid summers with emperate, arid winters (TaTa)	1370 (1110; 1631)	11.1 (7.6; 14.7)	253 (184; 322)	18.3 (17.0; 19.6)	0.29 (0.10; 0.47)	-0.7 (-2.6; 1.3)	0.30 (0.00; 0.65)
Hot, wet summers with warm, wet winters (HwHw)	69 (1; 149)	16.6 (14; 19.1)	1254 (1115; 1393)	19.9 (19.2; 20.7)	1.42 (0.66; 2.19)	4.3 (2.1; 6.4)	2.07 (1.16; 2.99)
Temperate, wet summers with cold moderately dry winters (TwCd)	335 (206; 465)	8.8 (6.8; 10.8)	699 (540; 859)	16.4 (14.6; 18.1)	0.91 (0.11; 1.71)	-7.2 (-11.3; -3.2)	0.50 (0.07; 0.92)
Temperate, moderately dry summers with arid, temperate winters (TdTa)	720 (446; 994)	12.8 (9.9; 15.7)	432 (351; 513)	17.5 (16.5; 18.5)	0.40 (0.14; 0.67)	2.2 (0.5; 4.0)	0.63 (0.00; 1.37)
Cool, moderately dry summers with arid, cold winters (CdCa)	1169 (766; 1573)	6.8 (5.4; 8.1)	358 (304; 413)	12.1 (10.6; 13.7)	0.62 (0.28; 0.95)	-8.2 (-11.1; -5.2)	0.23 (0.06; 0.40)

* June-August; **December-February.

minimum daily temperature and precipitation during the cold season between December and January and warm season between June and August separately across eight regions defined by the proposed regionalization method with the colour representing separate regions and the colour transparency proportional to each quintile of the data distribution. Based on the distribution of resulting climate parameters intuitive, climate description for each region reflecting the prevailing weather patterns in each region during cold and warm periods of the year was suggested. Climate parameters for each of the eight defined zones and their intuitive climatological descriptions are summarised in Table 1 and shown on the map overlaid with political borders (Fig. 4).

Population at risk was computed for each region for each 10 degree F minimum daily temperature range. The largest population at risk, with 10.9 million elderly residents, was in a region (number 3) best characterised by temperate, wet summers and winters (TwTw). The smallest population at risk resided in a region (number 8) best characterised by cool, moderately dry summers and cold, arid winter (CdCa) with 445 thousand elderly residents. The highest hospitalisation rates due to heat and cold temperatures was observed in region 5 best characterised by hot, wet summers and warm, wet winters (HwHw) with 9.6 and 17.7 hospitalisations per 100,000 persons at risk, respectively. The lowest hospitalisation rate due to heat was observed in region 8 characterised by cool, moderately dry summers and cold arid winters (CdCa) with 4.5 hospitalisations per 100,000 persons at risk. The lowest hospitalisation rate due to cold was observed in region 4 characterised by temperate, arid summers and winters (TaTa) with 9.1 hospitalisations per 100,000 persons at risk (Table 2).

Discussion

This study proposes a new approach to spatial partitioning, a three-step LKN regionalization method built upon a fundamental idea, suggested by Wladimir Köppen nearly two centuries ago, that climate of a region can be distinguished by the phenology of that

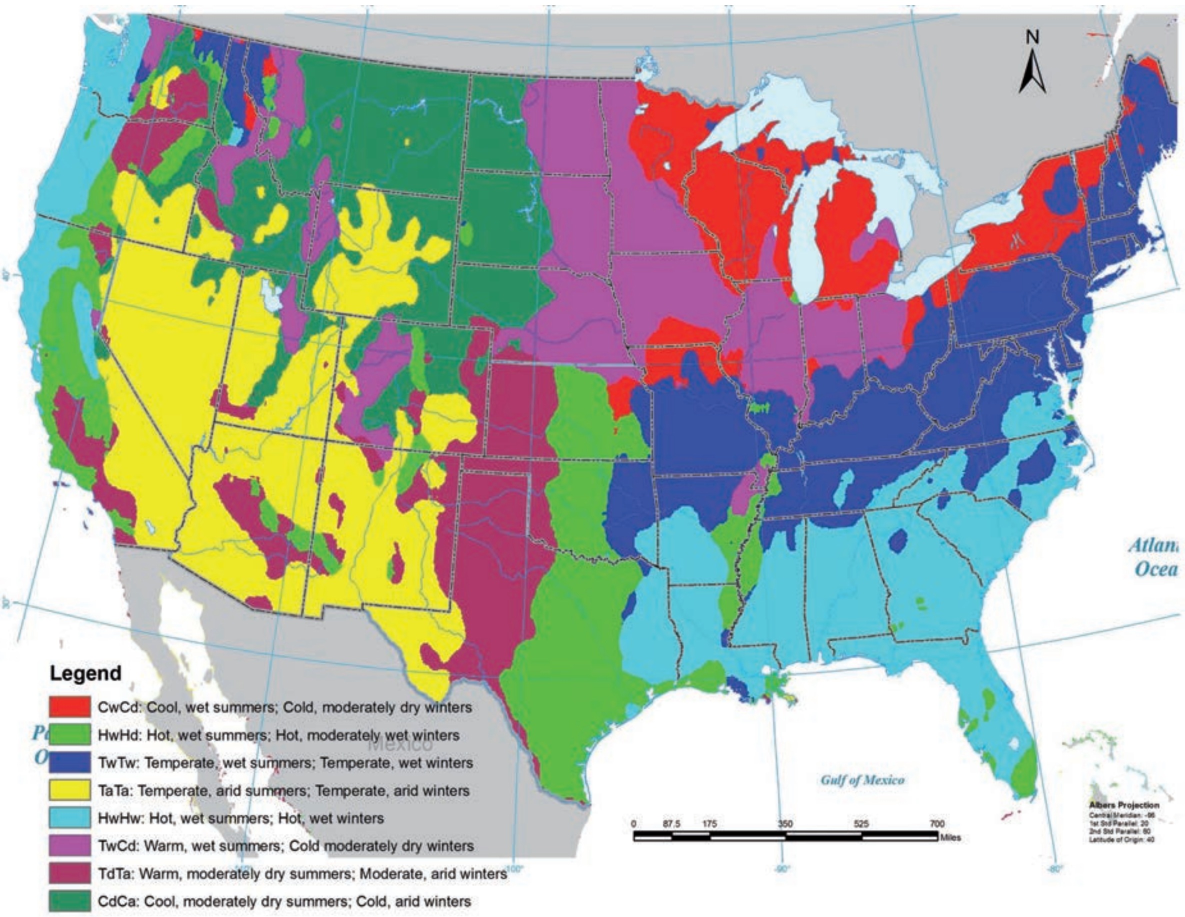
N
Atlan
Ocea
Gulf of Mexico
Mexico
Legend
🟥 CwCd: Cool, wet summers; Cold, moderately dry winters
🟩 HwHd: Hot, wet summers; Hot, moderately wet winters
🟦 TwTw: Temperate, wet summers; Temperate, wet winters
🟨 TaTa: Temperate, arid summers; Temperate, arid winters
🟦 HwHw: Hot, wet summers; Hot, wet winters
🟪 TwCd: Warm, wet summers; Cold moderately dry winters
🟥 TdTa: Warm, moderately dry summers; Moderate, arid winters
🟩 CdCa: Cool, moderately dry summers; Cold, arid winters
0 87.5 175 350 525 700
Miles
Albers Projection
Central Meridian: -96
1st Std Parallel: 20
2nd Std Parallel: 60
Latitude of Origin: 40
-100°
-90°
-80°

Fig. 4. Eight climate regions of conterminous USA as defined by the proposed non supervised LKN regionalization colour coded with overlaid political borders.

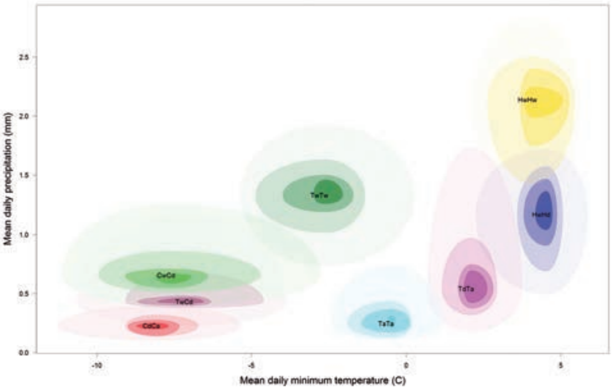

Fig. 5. Climatic separation by minimum daily temperature and precipitation during the cold season between December and January across eight regions defined by the proposed regionalization method with the colour representing separate regions and the colour transparency proportional to each quintile of the data distribution.

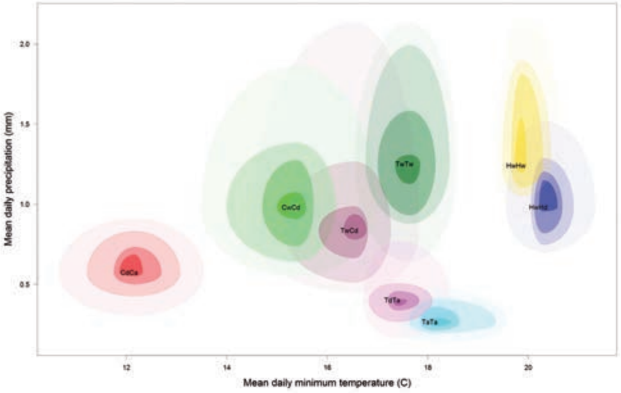

Fig. 6. Climatic separation by minimum daily temperature and precipitation during the hot season between June and August across eight regions defined by the proposed regionalization method with the colour representing separate regions and the colour transparency proportional to each quintile of the data distribution.

region. Phenology is very sensitive to minor climate nuances and changes in profile, which allows a researcher to differentiate regions with varying climate patterns and detect effects of climate changes. Limited technological capacity at his time forced Köppen and his contemporary researchers to use proxy variables and simplified decision rules. An alternative objective method of defining climate regions has now been introduced based on actual phenology and annual vegetation cycle derived directly from the worldwide measurements by a constellation of NASA's earth-observing satellites. Furthermore, the utility of the suggested regionalization was assessed for public health research and that of defined regions was evaluated for the assessment of population vulnerability due to exposure to ambient heat and cold temperatures.

Climate classification is a subset of the larger group of 'p-Region' problems, which collectively belong to a 'NP-hard' set of optimisation problems. Several attempts to classify global or regional climate based on measuring or interpreting various proxy variables have been made (Fovell, 1997; Unal et al., 2003; Chen

Table 2. Populations at risk and mean hospitalisation counts for hypothermia and hyperthermia for the climate regions defined by the LKN regionalization method.

Prevailing climate types according to the LKN code	Population per class	Number of hypothermia hospitalisations		Number of hyperthermia hospitalisations	
	Thousands per year	Mean annual count	Annual rate*	Mean annual count	Annual rate*
Cool, wet summers with cold, moderately dry winters (CwCd)	3,282	469	14.3	199	6.06
Hot, wet summers with warm, moderately wet winters (HwHd)	5,130	649	12.7	478	9.32
Temperate, wet summers with temperate, wet winters (TwTw)	10,928	1,550	14.2	841	7.70
Temperate, arid summers with temperate, arid winters (TaTa)	1,227	112	9.1	98	7.98
Hot, wet summers with warm, wet winters (HwHw)	5,343	944	17.7	512	9.58
Temperate, wet summers with cold moderately dry winters (TwCd)	3,355	510	15.2	240	7.15
Temperate, moderately dry summers with arid, temperate winters (TdTa)	1,954	186	9.5	112	5.73
Cool, moderately dry summers with arid, cold winters (CdCa)	445	78	17.5	20	4.50

*per 100,000 at risk.

Table 3. Populations at risk and mean hospitalization counts for hypothermia and hyperthermia for the climate regions defined by the Köppen-Geiger regionalization method.

Prevailing climate types according to the Köppen-Geiger code	Population per class	Number of hypothermia hospitalisations		Number of hyperthermia hospitalisations	
	Thousands per year	Mean annual count	Annual rate*	Mean annual count	Annual rate*
Tropical monsoon climate (Am)	274	7	2.6	6	2.2
Tropical, wet-dry climate (Aw)	297	9	3.0	16	5.4
A-category: "*Tropical*"	571	16	2.8	22	3.9
Mid-latitude steppe and desert (Bsh)	487	24	4.9	43	8.8
Tropical and subtropical steppe (Bsk)	2,742	290	10.6	136	5.0
Tropical and subtropical desert (Bwh)	470	28	6.0	46	9.8
B-category: "*Arid*"	3,699	342	9.2	225	6.1
Humid, subtropical climate (Cfa)	9,360	1,586	16.9	997	10.7
Maritime, temperate climate (Cfb)	91	22	24.1	5	5.5
Dry summer, Mediterranean climate (Csa)	318	39	12.3	20	6.3
Dry summer, subtropical climate (Csb)	1,089	141	12.9	40	3.7
C-category "*Temperate*"	10,858	1,788	16.5	1,062	9.8
Hot summer, continental climate (Dfa)	11,724	1,617	13.8	955	8.1
Warm Summer, Continental climate (Dfb)	4,444	666	15.0	223	5.0
Continental subarctic climate (Dfc)	64	10	15.7	2	3.1
Warm summer, continental climate (Dsb)	292	54	18.5	12	4.1
Continental subarctic or boreal, i.e. taiga (Dsc)	0.038	0	0	0	0.0
Hot summer, continental climate Dwa	9	2	21.2	1	10.6
Warm summer, continental climate (Dwb)	1	0	0	1	96.2
D-category: "Cold"	16,534	2,349	14.2	1,194	7.2

*per 100,000 at risk.

and Chen, 2013). The most often used proxies were ambient temperature and precipitation, with a quite complicated set of decision rules. An example of climate zone determinants for a popular KG climate classification system is listed in Table 4. Objective research, that identifies climate characteristics based on data properties, should not, at least in theory, depend on a set of predefined parameters. The machine learning methods should be able to divide regional climate characteristics based on the properties of the data itself. Therefore, these methods are especially suitable for researchers who want to detect small but persistent changes in regional climate.

The complex problem of separating a large data set into a number of distinct cohesive classes is well known and often arises in many areas of scientific data analysis and machine learning, including data mining and pattern recognition, image processing and target acquisition. There are two major groups of algorithms employed: supervised and unsupervised (Theodoridis and Koutroumbas, 2009). The supervised method works best when an expert is available and it is possible to determine a "true" group membership of a data subset before the analysis. Unsupervised classification, on the other hand, uses similarity measures derived from the data itself to assign class labels. Many different methods exist for defining similarity metrics. These methods are often domain-specific and use the concept of various "distance measures". Usually the goal of an unsupervised classification algorithm is to maximise similarity within the clusters, while increasing dissimilarity between the clusters. In the case of regionalization based on climate, the "true" climate is not known beforehand. The objective of this analysis was to avoid an involvement of expert opinions in regionalization process whenever possible, making unsupervised classification the most appealing method. The hierarchical clustering method, as suggested by Fovell and Fovell (1993) and Fovell (1997) is very attractive. It creates an intuitive, hierarchical structure of nested regions. However, it is very computationally expensive. A full distance matrix required for this method to analyse remote sensing data described above would exceed 150 billion entries - a challenging problem even by today's computing standards. The suggested method groups regions based on the "similarity" of the phe-

Table 4. Description of Köppen-Geiger climate classification definitions and decision criteria for 16 regions in conterminous USA.

1st	2nd	3rd	Description	Criteria
A			Tropical	Tcold ≥18C
	f		Rainforest	Pdry ≥60
	m		Monsoon	Not (Af) & Pdry ≥100–MAP/25
	w		Savannah	Not (Af) & Pdry <100–MAP/25
B			Arid	MAP <10×Pthreshold
	W		Desert	MAP <5×Pthreshold
	S		Steppe	MAP ≥5×Pthreshold
		h	Hot	MAT ≥18
		k	Cold	MAT <18
C			Temperate	Thot >10 & 0 <Tcold <18
	S		Dry Summer	Psdry <40 & Psdry <Pwwet/3
	W		Dry Winter	Pwdry <Pswet/10
	F		Without dry season	Not (Cs) or (Cw)
		a	Hot Summer	Thot≥22
		b	Warm Summer	Not (a) & Tmon10 ≥4
		c	Cold Summer	Not (a or b) & 1 ≤Tmon10 <4
D			Cold	Thot >10 & Tcold ≤0
	S		Dry Summer	Psdry <40 & Psdry <Pwwet/3
	W		Dry Winter	Pwdry <Pswet/10
	F		Without dry season	Not (Ds) or (Dw)
		a	Hot Summer	Thot ≥22
		b	Warm Summer	Not (a) & Tmon10 ≥4
		c	Cold Summer	Not (a, b or d)
		d	Very Cold Winter	Not (a or b) & Tcold <–38
E			Polar	Thot <10
	T		Tundra	Thot >0
	F		Frost	Thot ≤0

Criteria description: MAP = mean annual precipitation; MAT = mean annual temperature; Thot = temperature of the hottest month; Tcold = temperature of the coldest month; Tmon10 = number of months where the temperature is above 10C; Pdry = precipitation of the driest month; Psdry = precipitation of the driest month in summer; Pwdry = precipitation of the driest month in winter; Pswet = precipitation of the wettest month in summer; Pwwet = precipitation of the wettest month in winter; Pthreshold = varies according to the following rules (if 70% of MAP occurs in winter then Pthreshold = 2 x MAT, if 70% of MAP occurs in summer then Pthreshold = 2 x MAT + 28, otherwise Pthreshold = 2 x MAT + 14). Summer (winter) is defined as the warmer (cooler) six month period of October to March and April to September.

nology signature based on a vegetation index as recorded by Earth observation satellites. Thus this clustering method derives regional climate assignment from statistical properties of collected data and does not require an input of additional decision rules.

To determine an optimal number of distinct partitions (climatic regions) we applied the Caliński-Harabasz Cluster (1974) Validity Index, a quantitative measure specifically designed to evaluate the quality of data partitioning. Determination of the quality of data partitioning is an important, but challenging, step to insure quality unsupervised classification. There are a number of other cluster validation measures proposed in the literature, e.g. the DB criterion (Davies and Bouldin, 1979), which is based on a ratio of within-cluster and between-cluster distances, the Dunn's index (Dunn, 1973) with multiple variations that are based on geometrical measures of cluster compactness and separation. There is an ongoing discussion about the applicability of cluster validity indices in the presence of noise, outliers and other properties often encountered in environmental, spatial and public health datasets (Kim and Ramakrishna, 2005). Dunn's index, for example, is sensitive to noisy clusters and the presence of outliers as well as being computationally intensive (Bezdek and Pal, 1998). As a variance ratio measure it is usually more stable in the presence of white noise than Dunn's index or the DB criterion.

It was found that the number of regions selected using VRC methodology matches the number of major regions as determined by Fovell's hierarchical clustering methodology, which indirectly confirmed the appropriateness of the selected cluster validity measure.

Climate properties of each of the eight zones, such as average high and low temperatures, average range of temperatures in cold and hot seasons and amounts of precipitation in these seasons demonstrate contrasting meteorological characteristics across the eight clusters. The suggested regionalization method is not based on expert opinions but rather intricate properties of the data and allows detection of small changes in the climate characteristics, which is especially important in the context of climate change. It is not practical to frequently modify a complex set of expert rules, like KG classification trees (see Table 4). In fact, these rules have remained practically unchanged for more than a century. By using a frequently updated worldwide data stream from the EOS constellation, the proposed decision support method can capture small changes in the climate patterns over regions of various size in a relatively short time frame. In the future, quantification of cluster differences based on a customary set of proxies, such as temperature and precipitation, and comparison of the resulting assignments with existing climate classification systems needs further study.

Climate and human health are deeply interconnected (McMichael, 2012). Several studies have demonstrated the impact of cold and hot weather on human mortality and morbidity in the presence of a geographic and climatic heterogeneity over a large, spatial extent (Keatinge and Donaldson, 2001; Basu and Samet, 2002; Curriero et al., 2002; Yu et al., 2012). Separation of a large, contiguous area into smaller regions with similar prevailing climate conditions allows quantification of the effects of short-term exposure to extreme weather events and long-term adaptations to the prevailing climate conditions. This work shows that hospitalisations due to exposure to cold in warm regions with wet summers and winters (HwHw) were almost twice higher than in temperate regions with arid summers and winters (TaTa). Similarly, for hospitalisations related to heat exposure in the hot wet region HwHw also almost twice higher when compared to cold, dry region (CdCa). The substantial fluctuations of hospitalisation rates across regions might be indicative of adaptation processes. People adapt to climatic conditions by creating protective infrastructure, including location-specific heating and cooling systems, by designing urban environments in which buildings and transit systems are connected, by establishing economic and social policies, such as availability and amount of heating assistance, and by changing behaviour to reduce an effect of climatic stress. The effect of adaptation requires further exploration.

A systematic approach to defining climatic conditions is essential for conducting studies to examine differential effects of temporal extremes in various climates. While generally accepted, the KG classification produces a much skewed distribution of population across USA. The majority of the susceptible population (about 80%) is concentrated just in 3 out of 16 defined regions for this territory (Table 3). Even an aggregation of smaller regions into four main categories is insufficient to resolve this issue. The uneven distribution may result in high instability of disease rates and risk estimation, which is one of the main questions for epidemiological studies of the effects of extreme weather and/or climate change on human health. The proposed regionalization represents a relatively even distribution of vulnerable populations across defined regions to offer a more suitable approach for national-scale epidemiological studies, which are often sensitive to large variations in population sizes.

Conclusions

The proposed objective clustering and decision support method produce sensible, spatial divisions for the conterminous USA. The suggested method provides an attractive, computationally efficient alternative to the hierarchical clustering method to perform a spatial regionalization based on local climate conditions. This work adapts a number of machine learning algorithms, such as clustering and kernel filtering, spatial interpolation and principal component analysis suitable for decision support frameworks in a public health applications. The use of frequently updated remote-sensing satellite data streams and machine learning algorithms allow for frequent revisions. This empowers researchers to capture temporal climate changes and their effect on human health. The advantage of climate classification based on remotely sensed data with regard to mitigating adverse effects of severe weather on human health has a strong potential and needs to be further explored by both meteorological and public health communities. The heavily populated regions in the warm South-east (the "Sunbelt") should be carefully explored and evaluated for developing preventive strategies to reduce thermal stress related hospitalisations in vulnerable populations. The analysis of cluster optimality and validity, extension of health-based climate classification to other regions in the world, further assessment of a regional resilience to climate change and applications to infectious and other diseases and health conditions are recommended future steps.

Acknowledgements

The authors would like to express their gratitude to Dr. Kurt Pennell and Dr. Gilbert E. Metcalf (Tufts University, Medford, MA, USA), two anonymous reviewers for their insightful comments and suggestions to Alexandra Kulinkina (Tufts University) and Daniel Aronov (Oxford University, Oxford, UK) for editorial assistance. Tufts University School of Engineering, Tufts Institute for the Environment (TIE), Tufts Initiative for Modeling Infectious Diseases (InForMID), University of Mississippi Medical Center, Boston Financial Research, Inc and NASA ROSES 2013 grant provided financial support.

References

Anderson GB, Dominici F, Wang Y, McCormack MC, Bell ML, Peng RD, 2013. Heat-related emergency hospitalizations for respiratory diseases in the medicare population. Am J Respir Crit Care Med 187, 1098-1103.

Arbabi A, 2011. Cluster-based method for understanding the climactic diversity of Iran. Afr J Agric Res 6, 6525-6529.

Arnold GM, 2003. Cluster Analysis. J R Stat Soc Series D 52, 407-408.

Aubrecht C, Steinnocher K, Köstl M, Züger J, Loibl W, 2013. Long-term spatio-temporal social vulnerability variation considering health-related climate change parameters particularly affecting elderly. Nat Hazards 68, 1371-1384.

Barkan JD, Densham PJ, Rushton G, 2006. Space matters: designing better electoral systems for emerging democracies. Am J Pol Sci 50, 926-939.

Basu R, Samet JM, 2002. Relation between elevated ambient temperature and mortality: a review of the epidemiologic evidence. Epidemiol Rev 24, 190-202.

Bezdek JC, Pal NR, 1998. Some new indexes of cluster validity. Trans Sys Man Cyber Part B 28, 301-315.

Bragagni G, Alberti A, Castelli G, Lari F, 2012. Ipotermia accidentale nell'anziano - Hypotermia in the elderly. Ital J Med 6, 47-51.

Caliński T, Harabasz J, 1974. A dendrite method for cluster analysis. Commun Stat 3, 1-27.

Carroll ML, DiMiceli CM, Sohlberg RA, Townshend JRG, 2000. MODIS normalized difference vegetation index. University of Maryland.

Chebana F, Martel B, Gosselin P, Giroux JX, Ouarda TBMJ, 2013. A general and flexible methodology to define thresholds for heat health watch and warning systems, applied to the province of Québec (Canada). Int J Biometeorol 57, 631-644.

Chen D, Chen HW, 2013. Using the Köppen classification to quantify climate variation and change: an example for 1901–2010. Environ Dev 6, 69-79.

Chui KK, Webb K, Russell RM, Naumova EN 2009. Geographic variations and temporal trends of Salmonella-associated hospitalization in the U.S. elderly, 1991-2004: a time series analysis of the impact of HACCP regulation. BMC Public Health 9, 447.

Cohen SA, Egorov AI, Jagai JS, Matyas BT, DeMaria Jr A, Chui KKH, Griffiths JK, Naumova EN, 2008. The SEEDs of two gastrointestinal diseases: socioeconomic, environmental, and demographic factors related to cryptosporidiosis and giardiasis in Massachusetts. Environ Res 108, 185-191.

Curriero FC, Heiner KS, Samet JM, Zeger SL, Strug L, Patz JA, 2002. Temperature and mortality in 11 cities of the eastern United States. Am J Epidemiol 155, 80-87.

Davies DL, Bouldin DW, 1979. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1, 224-227.

Ding C, He X, 2004. K-means clustering via principal component analysis. Banff, ICML 2004, 225-232 pp.

Dunn JC, 1973. A Fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cyb 3, 32-57.

Duque JC, Church RL, Middleton RS, 2011. The p-Regions problem. Geogr Anal 43, 104-126.

Ebi KL, Mills D, 2013. Winter mortality in a warming climate: a reassessment." Wiley Interdiscip Rev Clim Change 4, 203-212.

Fovell RG, 1997. Consensus clustering of U.S. temperature and precipitation data. J Clim 10, 1405-1427.

Fovell RG, Fovell MYC, 1993. Climate zones of the conterminous United States defined using cluster analysis. J Clim 6, 2103-2135.

Fowlkes EB, Mallows CL, 1983. A method for comparing two hierarchical clusterings. J Am Stat Assoc 78, 553-569.

Franke R, Nielson G, 1980. Smooth interpolation of large sets of scattered data. Int J Numer Methods Eng 15, 1691-1704.

Geiger R, Pohl W, 1954. Eine neue wandkarte der klimagebiete der erde nach W. Köppens klassifikation (a new wall map of the climatic regions of the world according to W. Köppen's classification). Erdkunde 8, 58-61.

Goldberg MS, Gasparrini A, Armstrong B, Valois MF, 2011. The short-term influence of temperature on daily mortality in the temperate climate of Montreal, Canada. Environ Res 111, 853-860.

Horvath SM, Rochelle RD, 1977. Hypothermia in the aged. Environ Health Perspect 20, 127-130.

Jagai JS, Castronovo DA, Monchak J, Naumova EN, 2009. Seasonality of cryptosporidiosis: a meta-analysis approach. Environ Res 109, 465-478.

Jagai JS, Sarkar R, Castronovo D, Kattula D, McEntee J, Ward H, Kang G, Naumova EN, 2012. Seasonality of rotavirus in south asia: a meta-analysis approach assessing associations with temperature, precipitation, and vegetation index. PLoS One 7, 1-14.

Johnson DP, Stanforth A, Lulla V, Luber G, 2012. Developing an applied extreme heat vulnerability index utilizing socioeconomic and environmental data. Appl Geogr 35, 23-31.

Kalkstein LS, Davis RE, 1989. Weather and human mortality: an evaluation of demographic and interregional responses in the United States. Ann Assoc Am Geogr 79, 44-64.

Keatinge WR, Donaldson CG, 2001. Mortality related to cold and air pollution in London after allowance for effects of associated weather patterns. Environ Res 86, 209-216.

Kim M, Ramakrishna RS, 2005. New indices for cluster validity assessment. Pattern Recognit Lett 26, 2353-2363.

Koppen W, Volken E, Brönnimann S, 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. Meteorol Z 20, 351-360.

LPDAAC-NASA, 2000-2013. MODIS vegetation data (MOD13 and MYD13). NLPDAA Center. Sioux Falls, South Dakota, NASA.

Margolis H, 2014. Heat waves and rising temperatures: human health impacts and the determinants of vulnerability. In: Global climate change and public health. Pinkerton KE, Rom WN (eds), Springer, New York 7, 85-120.

McMichael AJ, 2012. Insights from past millennia into climatic impacts on human health and survival. Proc Natl Acad Sci U S A 109, 4730-4737.

Montero JC, Mirón IJ, Criado-Álvarez JJ, Linares C, Díaz J, 2012. Influence of local factors in the relationship between mortality and heat waves: Castile-La Mancha (1975–2003). Sci Total Environ, 414, 73-80.

National Climate Assessment, 2014. Washington DC, USGCR Program. Available at: http://nca2014.globalchange.gov (accessed on May 2014).

NOAA, 2014. Global historical climatology network daily dataset overivew. Available at: https://gis.ncdc.noaa.gov/geoportal/catalog/search/resource/details.page?id=gov.noaa.ncdc:C00838 (accessed on March 2014).

O'Neill MS, Ebi KL, 2009. Temperature extremes and health: impacts of climate variability and change in the United States. J Occup Environ Med 51, 13-25.

Pascal M, Wagner V, Le Tertre A, Laaidi K, Honoré C, Bénichou F, Beaudeau P, 2013. Definition of temperature thresholds: the example of the French heat wave warning system. Int J Biometeorol 57, 21-29.

Rango N, 1980. Old and cold: hypothermia in the elderly. Geriatrics 35, 93-96.

Rango N, 1985. The social epidemiology of accidental hypothermia among the aged. Gerontologist 25, 424-430.

Renka RJ, 1988. Multivariate interpolation of large sets of scattered data. ACM Trans Math Softw 14, 139-148.

Seman AP, Golim V, Gorzoni ML, 2009. [Study on accidental hypothermia in institutionalized elderly]. Rev Assoc Med Bras 55, 663-671.

Tasnádi A, 2011. The political districting problem: a survey. Society and Economy 33, 543-554.

Theodoridis S, Koutroumbas K, 2009. Pattern recognition. London, Academic Press, 961 pp.

Thornthwaite CW, 1931. The climates of North America: according to a new classification. Geographical Review 21, 633-655.

Thornthwaite CW, 1943. Problems in the classification of climates. Geographical Review 33, 233-255.

Thornthwaite CW, 1948. An approach toward a rational classification of climate. Geogr Review 38, 55-94.

Tong D, Murray AT, 2012. Spatial optimization in geography. Ann Assoc Am Geogr 102, 1290-1309.

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

Watts AJ, 1972. Hypothermia in the aged: a study of the role of cold-sensitivity. Environ Res 5, 119-126.

Xie H, Yao Z, Zhang Y, XuY, Xu X, Liu T, Lin H, Lao X, Rutherford S, Chu C et al., 2013. Short-term effects of the 2008 cold spell on mortality in three subtropical cities in Guangdong province, China. Environ Health Perspect 121, 210-216.

Yu W, Mengersen K, Wang X, Ye X, Guo Y, Pan X, Tong S, 2012. Daily average temperature and mortality among the elderly: a meta-analysis and systematic review of epidemiological evidence. Int J Biometeorol 56, 569-581.