Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

What weather variables are important in predicting heat-related mortality? A new application of statistical learning methods

Kai Zhang [a,*], Yun Li [b], Joel D. Schwartz [c], Marie S. O'Neill [d]

[a] *Division of Epidemiology, Human Genetics and Environmental Sciences, University of Texas School of Public Health, Houston, TX 77030, USA*
[b] *Department of Statistics, University of Michigan, Ann Arbor, MI, USA*
[c] *Departments of Environmental Health and Epidemiology, Harvard School of Public Health, Boston, MA, USA*
[d] *Departments of Environmental Health Sciences and Epidemiology, University of Michigan School of Public Health, Ann Arbor, MI, USA*

ARTICLE INFO

Article history:
Received 27 March 2013
Received in revised form
8 April 2014
Accepted 9 April 2014

Keywords:
Absolute humidity
Heat
Mortality
Random forests
Temperature
Weather

ABSTRACT

Hot weather increases risk of mortality. Previous studies used different sets of weather variables to characterize heat stress, resulting in variation in heat–mortality associations depending on the metric used. We employed a statistical learning method – random forests – to examine which of the various weather variables had the greatest impact on heat-related mortality. We compiled a summertime daily weather and mortality counts dataset from four U.S. cities (Chicago, IL; Detroit, MI; Philadelphia, PA; and Phoenix, AZ) from 1998 to 2006. A variety of weather variables were ranked in predicting deviation from typical daily all-cause and cause-specific death counts. Ranks of weather variables varied with city and health outcome. Apparent temperature appeared to be the most important predictor of heat-related mortality for all-cause mortality. Absolute humidity was, on average, most frequently selected as one of the top variables for all-cause mortality and seven cause-specific mortality categories. Our analysis affirms that apparent temperature is a reasonable variable for activating heat alerts and warnings, which are commonly based on predictions of total mortality in next few days. Additionally, absolute humidity should be included in future heat-health studies. Finally, random forests can be used to guide the choice of weather variables in heat epidemiology studies.

1. Introduction

Heat waves are projected to occur more frequently, more intensely and to last longer as a consequence of climate change (Meehl and Tebaldi, 2004). Epidemiological studies have shown that heat waves are associated with elevated risk of mortality, hospital admissions, heat stroke, heat exhaustion, cardiovascular and respiratory diseases (Kovats and Hajat, 2007). Previous heat-related epidemiological studies have characterized heat or heat waves by using a single temperature metric (e.g., daily mean/minimum/maximum temperature), or a composite index combining temperature and relative humidity, or a more sophisticated index requiring substantial meteorological knowledge (e.g., spatial synoptic classification) (Hajat et al., 2010; Barnett et al., 2010). However, these weather metrics may not characterize human exposures to extreme heat very well since biometeorological studies have shown that human body temperature is related to many weather variables, e.g., temperature, relative humidity, solar radiation, barometric pressure, wind speed, etc. (Steadman, 1979a, 1979b, 1984). Also, people usually spend majority of their time indoors, e.g., Americans spend 86.9% of their time indoors on average (Klepeis et al., 2001). Some variables (e.g. absolute humidity) penetrate better than others. Moreover, several metrics are typically used for each weather variable mentioned above, e.g., daily mean, minimum, and maximum temperature, and no consensus exists on which measure of temperature has the most influence on mortality. Two likely reasons are that there is no single measure and that using temperature alone is not sufficient to characterize heat exposures. This fact contributes to the difficulty of comparing various studies and inconsistencies in the heat-health associations found in addition to differences in culture, housing and exposure across regions and populations. Identifying which variables are most consistently predictive of health outcomes across multiple cities could aid epidemiologic research. Furthermore, identifying the local weather conditions most predictive of heat-related mortality could inform design of heat wave and heat health warning systems by reducing the number of triggering metrics considered. Such information may guide local public and weather service authorities to more effectively mobilize resources to prevent adverse health effects during hot weather.

A small number of studies have examined the performances of different weather-related exposure metrics in estimating

[*] Corresponding author. Fax: +1 713 500 9264.
E-mail address: kai.zhang@uth.tmc.edu (K. Zhang).

http://dx.doi.org/10.1016/j.envres.2014.04.004
0013-9351/© 2014 Elsevier Inc. All rights reserved.

heat–mortality relationships; we describe two here. A multi-city study examined the performance of mean, minimum and maximum temperature with and without humidity, and apparent temperature and the Humidex (a function of temperature and relative humidity) in predicting mortality using mortality and weather data from 107 U.S. cities during 1987–2000 (Barnett et al., 2010). The measure of temperature most associated with mortality varied with city, season and age groups, but these different temperature measures had the same predictive ability, on average. Another multi-city study evaluated maximum temperature, dew point temperature and a few combinations of these two variables in 105 U.S. cities during 1987–2005 (Bobb et al., 2011). It was reported that the best temperature measure varied by city.

All these studies used either temperature predictors or temperature-humidity indices within the regression framework, and did not examine additional weather conditions simultaneously (e.g., absolute humidity and barometric pressure). Also, the generalized linear model (GLM) or generalized additive model (GAM) used in these prior studies does not have the ability to account for high-order interaction among covariates. Our prior work proposed a hybrid clustering method to classify potentially 'dangerous' heat based on four daily weather conditions: maximum/minimum temperature and maximum/minimum dew point (Zhang et al., 2012). Yet, even that approach did not take many weather variables into consideration simultaneously. Like studying multi-pollutant mixtures, properly accounting for the multiple weather conditions to which humans are exposed is a challenge for assessing heat-related health effects.

This study aims to evaluate many weather conditions simultaneously and identify the most important weather variables in predicting excess death counts associated with hot weather by evaluating their prediction performance. This analysis takes advantage of a recent advance in statistical learning methods—the random forests approach, and accounts for exposures to multiple weather conditions in a data-driven way. This approach reduces substantial scientific meteorological-related judgments while taking many weather conditions into consideration. It is important to note that this paper is not to demonstrate that random forests are an alternate method to GAM or GLM in heat-related epidemiological studies.

2. Methods

2.1. Data sources

This study uses daily mortality data and weather observations from four U.S. cities (Chicago, IL; Detroit, MI; Philadelphia, PA; and Phoenix, AZ) from 1998 to 2006. Death records were obtained from the National Center for Health Statistics. To prepare the data for analysis, we created daily counts of deaths, first for all-cause mortality and then for cause-specific mortality. International Classification of Diseases tenth revision (ICD-10) codes were in use for the period 1998–2006. Daily total mortality excluded injuries and external causes (ICD-10 beginning with S through Z). Mortality counts were further classified as cardiovascular diseases (CVD; ICD-10 codes I01–I52), stroke (ICD-10 codes I60–I69), myocardial infarction (MI; ICD-10 codes I21–I22), congestive heart failure (CHF; ICD-10 codes I50), pneumonia (ICD-10 codes J12–J18), chronic obstructive pulmonary disease (COPD; ICD-10 codes J40–J44 and J47) and respiratory disease (ICD-10 codes J00–J99).

We aimed to evaluate whether hot weather conditions would be associated with increased levels of daily mortality counts, compared to the expected levels for any given day, based on a long-term average. To define the generally expected level of daily mortality counts, we modeled mortality counts as a smooth function (a cubic spline) of day of the year (degrees of freedom=5) while adjusting for day of week and year over the time period of our study (1998–2006). Day of the year indicates a seasonal trend, which has been assumed to be the same each year and has thus been coded as 1 to 365/366. The indicator variable for year enables control of long-term trends, if present. From this smooth function, we created a single smooth function that represented the annual 'expected' pattern of daily mortality averaged over the entire 9 years of data. A smooth function was created for all-cause mortality as well as for the cause-specific mortality. Then, using the daily deaths predicted by this smooth function for a given calendar date (e.g., July 10), we calculated the difference between the observed daily and the 'expected' for various categories of mortality. This variable can take on negative or positive values and we refer to it as deviation from typical daily mortality counts. We used this concept in our previous work to evaluate our proposed hybrid clustering method to identify potentially 'dangerous' hot days (Zhang et al., 2012).

Weather measurements from four cities were obtained from the National Climatic Data Center (NCDC, 2010). From this data, we created variables of daily minimum, mean and maximum temperature, dew point, apparent temperature, barometric pressure and absolute humidity. Each variable was calculated on the same day as, one day before, and two days before the deaths occurred. Besides these weather variables, calendar month as an additional variable was used to account for timing in season as a potential indicator of early season heat waves in the data analysis. Apparent temperature was derived using the equation from Zanobetti and Schwartz (2008). The description of all variables is shown in Table 1.

2.2. Approach

We applied a machine learning method called random forests to select the most important variables among all available variables in predicting deviation from typical daily mortality counts. Random forests are an extension of regression tree methods. Before discussing the specifics of the analysis, we next provide an overview of these statistical methods.

A regression tree is a non-parametric statistical learning technique described by a tree-structured algorithm (Faraway, 2006). Using this method, a dataset is partitioned in a recursive manner. This algorithm evaluates every possible division point of every predictor of the variable of interest to make a split in the data at each step, and the choice of a predictor variable and its value are determined by minimizing variance in predictions (Faraway, 2006). For example, our objective in this paper was to use weather variables as inputs to predict deviation from typical daily mortality counts. The basic idea is to partition the space of weather variables recursively into two smaller regions. At each step, the algorithm chose one of the weather variables and the value to split it on which better predicted deviation from typical daily mortality counts compared to other variables and values. In other words, the algorithm chose the most "dangerous" weather condition during each split. Each leaf or terminal node represents a partition region, characterized by a set of weather conditions associated with a deviation from typical expected mortality. Importantly, these conditions include potentially high order interactions among the predictors. (We present an example to illustrate the regression tree structure with terminal nodes in Supplementary material, S1). Regression tree methods are relatively straightforward to understand and implement, and can be used to find interaction effects among predictor variables, but its results are sensitive to small changes in the data, especially outliers (Faraway, 2006). The recursive nature of the regression tree method derives from the fact that it is performed on the most important predictors selected from the previous step.

Random forests are a collection of classification and regression trees that can be used to predict values or categories of target variables (Breiman, 2001). Each individual tree in the forest represents results from a specific regression tree (Breiman, 2001). Each tree is constructed based on a bootstrap sample of a dataset and a random subset of predictors. A final classification decision is a majority vote or the weighted average of all individual trees. Random forests have shown better prediction performances compared to other classification and regression tree methods, and can deal with missing values and a combination of binary and continuous variables automatically (Breiman, 2001). The importance of each predictor can also be quantified by assessing averaged prediction error across all random trees. Random forests can allow for complicated interactions among highly correlated predictors, and can decrease prediction errors compared to traditional regression tree methods (Breiman, 2001) because results are averaged among all trees.

In this paper, various weather variables and metrics were assessed in predicting deviation from typical daily mortality counts using random forests: daily minimum/maximum temperature, dew point, barometric pressure and absolute humidity on the same day as, one day before, and two days before the deaths occurred. The most important weather variables were determined by the importance scores derived from random forests, which are quantified as the average percent increase in mean squared error. Note that the outputs of random forests (e.g., importance scores here) are different from GAM and GLM in heat-related epidemiological studies which provide estimates of relative risk (e.g., percent change in mortality risk). In this analysis, the random forests' approach took 20,000 bootstrap samples of summertime (May 1st to September 30th) weather and mortality data from each one of the four cities, and each sample resulted in a tree. For each bootstrap sample, prediction error was derived by predicting the data not included in this bootstrap sample commonly called out-of-bag data, and the importance score of an independent variable was calculated by comparing the prediction errors from the permuted sample of that variable in the out-of-bag data to those from the unpermuted sample of that variable. A concrete example of the permutation approach is as follows: when we used a bootstrap sample to construct a regression tree using weather variables and heat-related mortality in the study period, we