STATS506 Final Report

Yu Lin, Shukun Liu, Yizhou Zhang, Yupeng Yang

12/7/2022

1. Description

This project is aiming at investigating whether the the health (mortality rate and diseases) of the aged is correlated to hot weathers, heat waves, and *heat island effects*. Linear regression is our primary method to investigate the relationship.

About the scope of this project, we focused on the data in California, including weather satellite data data, death profile, census data, and the geographical data of CA. The time period we investigated is basically from 2011 to 2015.

The challenges we met are mostly the realization of algorithms, e.g. transforming the longitude and latitude information into the desired format. We did not meet really challenging structural or theoretical problems, as we made good organization about the flow of this project, which is presented by sequence below. The assumptions, technical challenges and solutions are mixed in the following description.

2. Data

2.1 Data Collected

- Death data in csv files was directly downloaded from California State: Death Profiles by County by Month, 1970-2020 (https://data.chhs.ca.gov/dataset/death-profiles-by-county). This website provides monthly mortality data of all counties in California. We picked the monthly data from 2011 to 2015, based on which our analysis is county-wise.
- Census data of each county in CA was downloaded from the R package "tidyverse". The relative process can be check in the file "get_census_data.R" in the src folder.
- NASA MODIS Terra Land Surface 8-Day Average Temperature Emissivity Satellite data was acquired from NASA Earthdata website (https://search.earthdata.nasa.gov/search). We only picked the data for regions covering California from 07/20/2013 to 07/27/2013. The satellite collects data by regions in the shapes of curved parallelograms, and then projects the data to a square region through averaging. The square region is a 1200*1200 matrix, with each point corresponding to the emissivity of a 1km*1km grid after the compression (from parallelogram to square). We acquire the longitude and latitude information of the vertexes of the square regions covering California, and approximately find the real longitude and latitude information for each point. The average value and variance of the emissivity in a county will help to measure the heat-island effect of that county.
- Daily California station-wise temperature and humidity data was crawled from Western Regional Climate Center (WRCC) (https://raws.dri.edu/). Each county in California has multiple stations, but the WRCC website does not label which county each station belong to. The codes for crawlers are written in Python and we utilized the online computational platform Colab to run multiprocessing jobs.

- The geographical information (latitude and longitude) of each station is then crawled from the same WRCC website and utilized to map the stations onto counties.
- The latitude and longitude information of the border of each county was retrieved from California Open Data Portal (https://data.ca.gov/dataset/california-counties).

2.2 Data Collection (Crawler)

From the WRCC website, We crawled daily weather data of over 500 stations from all over California. The data we collected features daily maximum air temperature and relative humidity data of all stations. The website sends the data for a specific station to the explorer only after the users fill out a request form, indicating the target year and month. We found that the data request form can be accessed through the URLs in the form "https://wrcc.dri.edu/cgi-bin/wea_monsum.pl?ca" + station code. Therefore we crawled all the station codes first, and then enumerate the code list to collect the longitude & latitude information and weather data of each station, which is stored in csv files. The above process can be check in files "crawl_all_station_address_code.ipynb" and "crawl_all_station_weather.ipynb" in the src folder.

3. Feature Engineering & Data processing

3.1 Heat Wave Definition

The heat waves in this project are identified by a temperature-humidity threshold: one day or longer periods when the daily maximum Wet Bulb Globe Temperature (WBGTmax) > 28 °C. This WBGTmax threshold follows the International Standards Organization (ISO) criteria for recognizing the risk of occupational heat related illnesses.

3.2 Feature Engineering

3.2.1 Calculate HImax and WBGTmax

Methodology To convert the huge daily station-wise weather data for all counties into some features or indicators that are more applicable to data analysis (regression), we decided to use the temperature and humidity data to calculate the WBGT index (a numerical measure of extremely hot weather). Apart from counting the heat waves defined above (WBGTmax > 28 °C), we also took into account the intensity of heat waves. To compare days with heat waves to days without extreme temperature, we reused the heat intensity formula for heat waves to quantify the intensity for days without extreme temperature. To calculate this heat intensity, we used daily minimum relative humidity and maximum temperature for each county we collected. This function can be checked in the file "HI_WBGT_cal.R" in the src folder.

$$HI_{max} = \frac{(0.5 \times (T_{max} + 61.0 + (T_{max} - 68.0) \times 1.2) + (0.094 RH_{min})) + T_{max}}{2}$$
$$WBGT_{max} = -0.0034 HI_{max}^2 + 0.96 HI_{max} - 34$$

3.2.2 Feature Engineering & Aggregation

We processed the maximum Air Temperature and minimum Relative Humidity information to generate the daily max heat index (HImax) and the daily maximum Wet Bulb Globe Temperature (WBGTmax). It should be noticed that a small fraction of weather data collected from the WRCC website is theoretically wrong. E.g., -99 Fahrenheit degrees is impossible but existing in the data collected. After checking the

source of data, we recognized that it is the problem of the data providers. As our project focuses on the hot situation, and the input variables for the regression model later are averaged monthly and county-wise features, we chose to drop the daily station data in which the temperature is lower than 30 Fahrenheit degrees. We also dropped the data with relative humidity out of the range from 0 to 100, since a relative humidity is rigorously defined in that range. After the data being processed, we stored all the daily weather data as a big csv file (44 MB) and wrote it into our RMDBS.

To parallel all the data into the same level, as our death profile data is monthly. we merged the daily HImax and WBGTmax data into the monthly data for each station by averaging, and generate 4 monthly features. This process can be concluded in the function of "month_summary.R" in the src folder.

Finally we got 4 features for each station: (1) average WGBT_max in each month, (2) duration of heat wave (WGBT_max >28) in each month, (3) the maximum WGBT_max in each month, and (4) the average WGBT_max when heat wave (WGBT_max > 28) is happening in each month. Then we stored the data as a csv file and pushed the data into our RMDBS too.

3.2.3 Map Station Data onto Counties

To map stations onto their corresponding counties, we first transformed the string longitude and latitude information we collected into decimal numbers. To map the monitoring station spots with county border polygons, the coordinates of the 2 sets of data were unified. An overview map Fig.1 was generated for reference.

We then averaged the 4 station-wise features of all stations in the same county, and the outcome averaged values were used as the 4 features for the county. At this moment, we possess the monthly county-wise weather features. The above process was realized by the file "station_to_county.R" in the src folder.

3.3 Death Profile Data & Census data

As regards the death data, we accessed the death data of each county per month from the CA government website. We used the data of total population, total death number and the elderly's death number of each county from 2011 to 2015 to generate our response variables. The data set also includes the causes of deaths.

The Californian census data was obtained from the R package "tidycensus" in county level. We acquired the total population of each county in California and the population of elderly (Age ≥ 65).

3.4 Measure Heat Island Effect in each county

A heat island means that a place that has a abnormally higher temperature than its surrounding areas. Here we employed the NASA 8-day average temperature emissivity satellite data covering California State from 7/20/2013 to 7/27/2013 to measure the heat island effects. The raw data, including the emissivity values and longitude & latitude information, was downloaded manually, processed and stored as 1200*1200 matrices in txt files. This process can be check in the file "NASA.ipynb" in the src folder. Then we mapped each data point onto its corresponding county, which is similar to 3.2.3. The processing of the data can be directly realized in "main.R" in the src folder.

The final step to handle this satellite data was to aggregate point-wise data into county-wise data and conduct feature engineering. We calculated the mean value and the variance of the emissivity data in each county, since we assumed the the variance can reflect the tensity of heat island effect. The above processing steps of the data can be directly realized in "main.R" in the src folder.

Then the 2 features, mean and variance, were also used to do the linear regression. This regression is in a smaller time scale, since the data points are just the average value in 8 days.

4. Model Fitting & Results

4.1 Method

We used linear regression models to find the correlation between features and the death data. By features, we mean (1) monthly average WGBT_max in each station, (2) monthly average duration of heat wave in each station(WGBT_max >28), (3) monthly maximum WGBT_max in each station, and (4) monthly average WGBT_max in each station for heat wave (WGBT_max > 28). Meanwhile, the response variables (death data) included the total death rate, the elder's death rate, and the deaths rates caused by specific reasons. To make the result more accurate, we ignored the counties with very small populations (<100000). The executions are also presented in "main.py" directly.

4.2 Results & Discussion

With the weather monitoring station data, we made linear regressions using all 4 features as predictors, with elderly death rates by different causes as response variables separately. We discovered that both the max_WBGT_max_monthly value and the avg_WBGT_max_monthly value have significant influence on all the death rates by different causes. The avg_WBGT_max_monthly value may decrease all the death rates, whereas he max_WBGT_max_monthly value may increase those rates. The p-value results are shown in Fig.2.

The other 2 predictors,heatwave duration and avg_WBGT during heat waves, do not have obvious influence on all the response variables. However, both predictors seem to be positively correlated with the death rate caused by malignant neoplasms.

We may conclude that normal hot weathers may be helpful to the older people's health, while extremely hot weathers are harmful. The malignant neoplasm is especially positively correlated with hot weathers.

In the analysis of NASA satellite data, the response variable of the elderly's proportion was also added to make regressions. We found that as the average and variance values of the temperature emissivity increase, the elderly proportion tends to decrease. In other words, the higher the both features are, the less older people living in that region. This implies that the elderly tends not to live in a high-heat-island-effect area, as we assumed higher variance of temperature emissivity means stronger heat island effects.

As a result, it is possible that the heat island has negative influence on the health of the elderly, which urges the older people to move to somewhere (some counties) cooler. It can also be that heat island effects come with some other conditions, for example noisy streets or high living costs, resulting in the above situation. In fact, the average and variance values of the temperature emissivity are not significant predictors in the regressions with other response variables.

The above operations and some more explorations can be found in "main.py" in the src folder. Some sample exploration outcome are presented as Fig.3 and Fig.4, showing the signs of the correlations between predictors and different response variables. The former used data from all counties, and the latter only consider counties that had at least 1 heat waves in each target month.

5. References

California Counties. California Open Data. Retrieved 12/13/2022 from https://data.ca.gov/dataset/california-counties

Death Profiles by County. California Health and Human Services Open Data Portal. Retrieved 12/13/2022 from https://data.chhs.ca.gov/dataset/death-profiles-by-county

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software, 4*(43), 1686. doi:10.21105/joss.01686.

Western Regional Climate Center. *RAWS USA Climate Archive*. Website of Western Regional Climate Center. Retrieved 12/13/2022 from https://raws.dri.edu/

Zhengming Wan. *MOD11A2 v061 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1*. NASA Earth Observation Data. Retrieved 12/13/2022 from https://lpdaac.usgs.gov/products/mod11a2v061/

6. Appendix

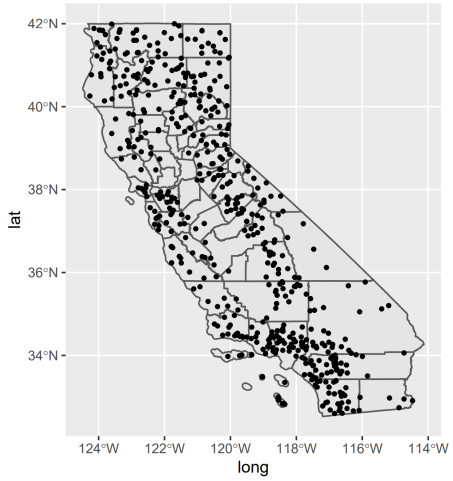

Figure 1: The weather monitoring stations are mapped onto their counties

Response Variable \ Predictor	max_WBGT_max_monthly	avg_WBGT_max_monthly
All causes (total)	1.79E-02	6.36E-36
Alzheimer's disease	7.51E-06	1.11E-10
Malignant neoplasms	1.11E-03	3.23E-12
Chronic lower respiratory diseases	2.46E-11	3.79E-39
Diabetes mellitus	3.66E-09	1.58E-14
Assault (homicide)	9.27E-03	3.42E-02
Diseases of heart	9.58E-06	2.86E-35
Essential hypertension and hypertensive renal disease	5.62E-07	5.10E-09
Accidents (unintentional injuries)	5.62E-05	4.90E-07
Chronic liver disease and cirrhosis	6.95E-13	5.37E-09
Nephritis, nephrotic syndrome and nephrosis	5.00E-11	5.98E-05
Parkinson's disease	1.48E-06	1.46E-02
Influenza and pneumonia	1.44E-03	2.44E-15
Cerebrovascular diseases	3.70E-10	2.45E-17
Intentional self-harm (suicide)	1.89E-09	2.29E-03

Figure 2: Correlation beween response variables and 2 features. Blue = negative correlation, Red = positive correlation.

	avg_WBGT	max_WBGT
death_rate	-	-
elderly_death_rate		

Figure 3: Correlation analysis considering all counties.

	avg_WBGT	max_WBGT	duration	avg_heatwave_WBGT
death_rate	-	+		-
elderly_death_rate	-	+		-
elderly_proportion			-	
total_population	-	+	-	-

Figure 4: Correlation analysis considering only counties with heatwaves in each targte month