STATS 506 Project Report Draft

Zicong Xiao, Yulin Gao, Qianang Chen, Xiaoyang Sheng

December 13, 2022

1 Introduction

An urban heat island (UHI) is an urban area which becomes islands of higher temperatures relative to the surrounding rural areas. This phenomenon is caused by human activity, materials of infrastructures, climate change and etc. It will cause residents some severe diseases like asthma and stroke, especailly to population who are over 65 years old. Yet the government can adopt some measures, for example, increase the insurance rates and medical expenditure to mitigate the harmful effects. We looked into data from mainly two sources: First we collected the climate data from earth engine, with covariates like surface temperature and percipitation, which are significant factors in our model. On the other hand we extracted the health data from CDC and chose the variables like mortality of different age groups, insurance rates and prevalence of asthma. Then we'll use different models (linear regression, Bayesian hierarchical model, random forest) to make vulnerability assessment towards the health of older adults. Next we'll do some model checking and validation to evaluate the model fitness. Finally based on the results we obtained, we'll do some analysis on negative impact of UHI on public health, along with the effect of each covariate.

2 Data

Urban Heat Island(UHI) Intensity:[1] The data includes the UHI intensity of summer day time/night time and winter day time/night time, for the selected counties that are influenced by the UHI, and the time unit is year, for 2003 - 2018. Here is some example of the data table:

id	county	year	su_daytime	su_nighttime	win_daytime	win_nighttime
1	21093	2003	1.1965928	0.1929673	0.2432011	0.0344972

Climate:[2] For data concerned with climate, the atmospheric reanalysis dataset ERA5 of global climate is collected from Earth Engine database. Reanalysis combines model data with observations across the world into a globally complete and consistent dataset. It records various climate features, like mean temperature, and provides aggregated values monthly throughout the world, which are available from 1979 to 2020.

county	yyyymm	mean_temperature	dewpoint_temperature	total_precipitation	surface_pressure	u_wind	v_wind
21093	201101	272.4396	267.1925	0.0579825	99304.19	1.0082280	0.1254208

Elderly Death:[3] The data is downloaded from CDC Wunder, consists of the death of different age groups (55-64,65-74,75-84,85+) in selected counties from 2010 to 2020, the time unit is month.

id	county	county_code	age	month	death
46	Baldwin County, AL	1003	55-64	2010/01	16

Elderly population:[4] The data is downloaded from Census API, consists of the population of several elderly age groups (55-64,65-74,75-84,85+) of the selected counties in 2010, with time unit of one year.

id	county_fips	population	age	county	year
1	48441	13649	55-64	Taylor County, Texas	2010

Insurance Rate:[5] The data of insurance rate consists of the uninsured population, total population and the calculated uninsured rate of the counties that are selected to be greatly influenced by UHI, and the time unit is also year, from 2006 - 2019. When we apply analysis, we only calculate and use the insurance rate, i.e. 1-uninsured rate. Here is the head of the data:

id	county_fips	county	state	year	name	population	uninsured_population	uninsured_rate
1	1003	Baldwin	AL	2008	Baldwin County, AL	143,932	23,631	16.4

Stroke Indicator:[6] The data from CDC inclue the stroke indicator of two age group 45-64 and greater than 64 (we pick the latter), of the counties every three years. Here is the view of the data:

id	county_fips	county	state	start_year	end_year	stroke_indicator	age
1	1003	Baldwin	Alabama	2012	2014	296.3	¿= 65

Asthma data:[7] The asthma dataset include the asthma indicator of counties in the unit of year, but only 2018 and 2019. Here is the view of the data:

id	county_fips	county	state	year	asthma_indicator
1	1003	Baldwin	Alabama	2018	9.6

3 Assumption

- To unify the primary time unit, we take the mean of climate features in both summer and winter. Here we assume these features in summer months (Jun, Jul, Aug) each year follow a same distribution. The same assumption is also made for data in winter months (Dec, Jan, Feb). Therefore, the mean of data in three months can represent the feature this year.
- We assume that the impact of urban heat islands can be explained by the UHI intensities, which allows us to perform quantitative analysis. Also, we assume that only a negligible portion of elderly people don't stay within one urban heat island area and thus we regard the relationship between all the elderly people within every single urban heat island area and their health as an ideal model to study the UHI effects. As a result, the statistics on the death toll and death rate of elderly people are a good representative of their general health condition.
- Since the insurance is closely related to the health investment and condition of citizens, the insurance rate could be used as an indicator to reflect the influence of UHI towards the citizen's health.
- The stroke indicator is a direct indicator that reflects the stroke situation in one region. Since the data of stroke is in the period of 3 years, we calculate the average of the UHI intensity in that 3 years to match the stroke indiactor, considering three years as a whole so that in this amount of time the change is not very significant.

4 Primary Analysis: UHI Intensity v.s. Climate

Regressing summertime, and wintertime, daytime, and nighttime UHI intensities respectively on climate features, we obtain, for both summertime and wintertime, surface pressure has large correlation with daytime UHI intensities while total precipitation has large correlation with nighttime UHI intensities according to their significant p-values. In addition, values of wind have great impact on both UHI intensities. Due to their correlations, all four climate features will be selected in the analysis of next stage.

UHI	total_precipitation	surface_pressure	u_component_of_wind_10m	v_component_of_wind_10m
su_daytime	0.0126	<2e-16	0.3848	2.95e-14
su_nighttime	<2e-16	0.37143	0.00172	7.41e-14
win_daytime	0.0188	<2e-16	<2e-16	0.1167
win_nighttime	<2e-16	7.39e-13	<2e-16	0.798

Table 1: Hypothesis testing

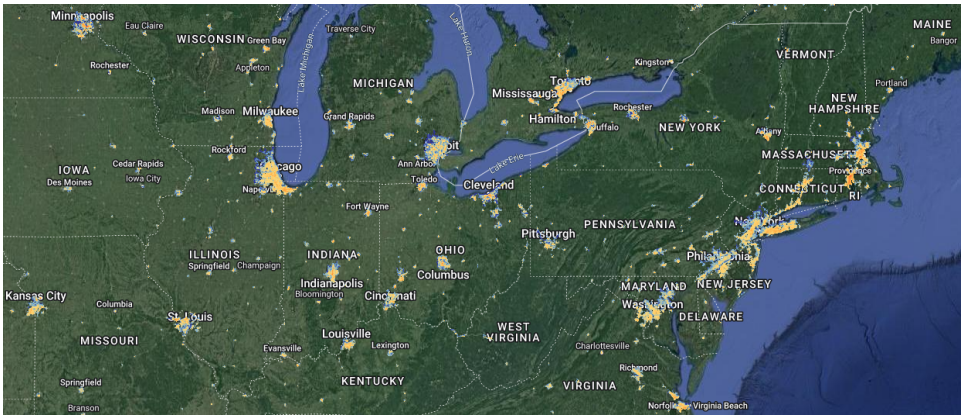

Figure 1: UHI map

5 Main Analysis

5.1 Elderly Death v.s. UHI Intensity

In this part, we investigate the impact of urban heat islands and heat waves on elderly people. To quantify the analysis, we analyze the relationship between the UHI intensity and the death toll/rate.

First, we analyze the impact of UHI in a long-term way. We perform a linear regression of the elderly death toll from 2010 to 2020 on the average UHI data from 2003 to 2018. The significant values of the UHI predictors is shown in the table.

	(Intercept)	summer_day	summer_night	winter_day	winter_night
P-value	0.00523	0.09608	0.88034	0.36418	0.18696
Significance	**	.			

From the table above, we can infer that the total elderly death toll is slightly correlated with the UHI intensity in summer daytime. However, the correlation isn't strong enough for us to conclude the relationship between UHI and the elderly' health. Taking into account that the population in each county is not the same, and even differs a lot, we calculate the elderly's death rate based on the population statistics in 2010 as an estimation. The significant values are as follows.

	(Intercept)	summer_day	summer_night	winter_day	winter_night
P-value	< 2e-16	0.00629	0.26069	0.98382	0.70151
Significance	***	**			

Therefore, we can conclude that the elderly's death rate is linearly correlated with the UHI intensity in summer daytime with a 99% confidence level.

Second, we investigate the short-term effect of UHI intensity on the elderly's health. We perform various methods of regression of the elderly's death ratio in 2010 on the UHI intensities in 2010. The significance P-value of predictors in linear regression is shown below.

	(Intercept)	summer_day	summer_night	winter_day	winter_night
P-value	< 2e-16	5.22e-6	0.6019	0.2630	0.0418
Significance	***	***			*

From the table above, the P-value of the predictor summer_day is very low, and we can conclude that the urban heat islands and heat waves may have adverse effects on elderly people.

Also, to identify the importance of each UHI predictor, we perform a random forest regression on the elderly's death rate, and the importance plot is shown as follows, from which we can infer that the UHI intensity in summer daytime is the major significant predictor.

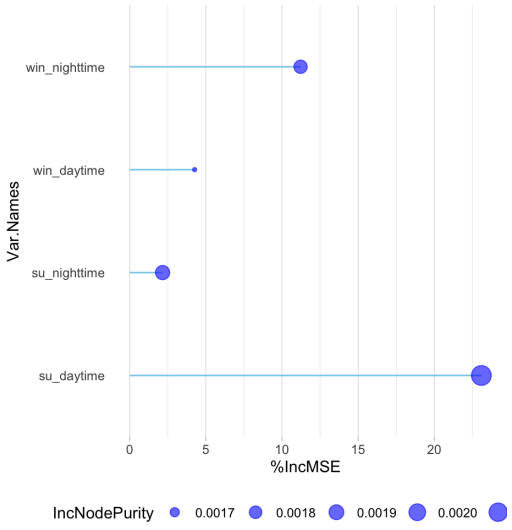

Figure 2: Importance Plot of the Predictors

5.2 Insurance Rate v.s. UHI Intensity

In this section, we try to study the relationship between Urban Heat Island(UHI) Intensity and the Insurance cover rate in those counties influenced by UHI in the years of 2006-2018.

We study the relationship between them using linear regression and random forest model. The p-value of the predictor in linear regression and the importance in random forest could show the influence of the UHI on the response, in this way the Insurance rate.

5.2.1 Random Forest

Then we apply the random forest to the dataset and obtain its importance plot to see the potential relation there. We set the number of the trees as 1000. The result shows that the the percent of Var explained is 18.39. The importance plot is:

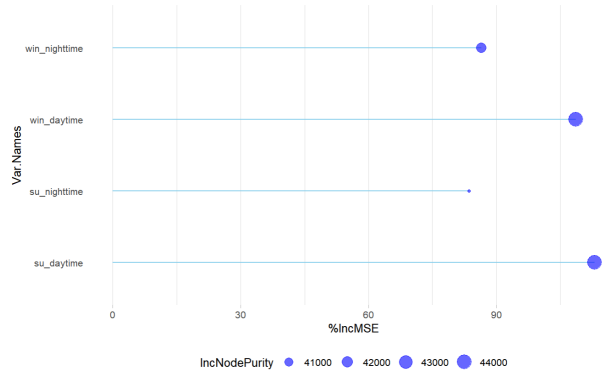

Figure 3: The importance plot of the random forest on the UHI vs insurance rate.

We can see that it matches the result of linear regression. UHI in daytime could have an influence on the insurance rate.

5.2.2 Linear Regression

The summary of the linear regression:

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	85.727049	0.160714	533.412	<2e-16
su_daytime	1.193068	0.101203	11.789	< 2e-16
su_nighttime	0.003799	0.227925	0.017	0.987
win_daytime	-3.069843	0.195545	-15.699	< 2e-16
win_nighttime	-0.989565	0.238195	-4.154	3.31e-05

The p-value of each regressor could be considered as the significance towards the influence to the Insurance rate. The p-value of the su_daytime, win_daytime is significantly small, which shows there is evidence that they have close relation to the insurance rate. We may conclude that the UHI intensity of the day time of a year could have influence on the insurance rate, and since the estimate is positive, which means that the UHI intensity for daytime summer is higher, the insurance rate tends to increase, which means will negatively impact the health of citizens.

5.3 Stroke v.s. UHI Intensity

It is similar to the analysis of the insurance rate, we apply both random forest and linear regression methods.

5.3.1 Random Forest

Apply the random forest, we get percentage of Var explained is 31.95, and the importance plot:

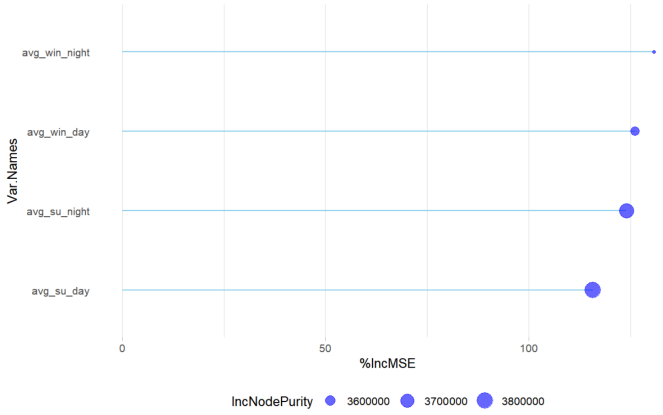

Figure 4: The importance plot of the random forest on the UHI vs stroke indicator.

The importance plot shows that avg_su_day is the least important in the model. The other factors could have a influence on the stroke indicator.

5.3.2 Linear Regression

Again, apply the linear regression on them and see the result:

	Estimate	Std. Error	t value	Pr(¿—t—)
(Intercept)	267.5877	1.5180	176.272	<2e-16
avg_su_day	-0.6048	1.0525	-0.575	0.566
avg_su_night	12.3947	2.3019	5.385	7.56e-08
avg_win_day	13.1866	2.1715	6.072	1.34e-09
avg_win_night	-19.9064	2.6647	-7.470	9.23e-14

The Multiple R-squared is 0.02659, Adjusted R-squared is 0.02589. Among the four regressors, except avg_su_day, all other three shows evidence of being significant in the influence to the stroke indicator. The estimate of winter tend to cancel each other in day and night. Also, the estimate of the avg_su_night is 12.3947, which means with a higher UHI intensity in summer night would probably increase the risk of stroke.

5.4 Elderly Asthma v.s. UHI Intensity

5.4.1 Model Selection

In this section we build the Bayesian linear model with respect to the prevalence of stroke and UHI/climate effects, and the insurance rate is also taken into account. Let $y_i (i = 1, 2, ..., N)$ be the mean of people who had got stroke within a county in three years(2017-2019), $N = 427$ is the number of counties we choose. We define x_1 as the UHI index in summer during the day, x_2 as the UHI index in summer at night, x_3 as the mean air temperature and x_4 as the insurance rate within a county. The likelihood distribution is defined as follows:

$$y_i \overset{\text{iid}}{\sim} N(\mu_i, \sigma_2) \tag{1}$$

$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{2}$$

Then we define the noninformative priors of parameters as follows:

$$\begin{aligned} \beta_i &\sim U(-1e6, 1e6), i = 0, 1, 2, 3, 4 \\ \sigma^{-2} &\sim \mathrm{G}(-1e6, 1e6) \end{aligned} \tag{3}$$

5.4.2 Fitting and Results

We use flat priors, 5 chains, 100K iterations and 50K burn-in periods in our MCMC simulation. The posteriors of each parameter of population over 65 and under 65 are as follows:

Parameter	mean(over 65)	95% CI(over 65)	mean(under 65)	95% CI(under 65)
β_0	-261.758	(-803.703,173.850)	-168.118	(-219.075,-99.355)
β_1	5.135	(-0.028,10.241)	1.258	(0.595,1.920)
β_2	1.956	(-10.971,15.151)	1.010	(-0.630,2.643)
β_3	1.960	(0.772,3.502)	0.703	(0.517,0.840)
β_4	-0.557	(-1.899,0.884)	-0.174	(-0.360,0.001)

5.4.3 Analysis and Discussion

From the tables above we find that the effect of UHI index is more significant than mean air temperature and the prevalence of stroke in all age groups will grow with the increase of UHI index and mean air temperature.

We also find that although β_2 is positive, but the 95% credible intervals both contain 0. Use R, the probabilities that $\beta_2 > 0$ from each group are 0.6136 and 0.8854, so we can conclude that the UHI index in summer at night has weaker effect to public health than the other two covariates. The UHI effect is stronger at night based on scientific literatures, which means the index tend to be higher. But the temperature at night is also lower, so people aren't as vulnerable to heat-related illness, which causes $\beta_2 < \beta_1$.

The mean value of β_4 is negative in both tables. If the rate of insurance coverage within a county is higher, which depends on government funding and people's awareness of health, then the residents will be less vulnerable to all kinds of diseases.

The first three parameters in table 1 are all greater than the corresponding parameters in table 2, so the variables of mean air temperatures and UHI index have a greater effect to older adults who are 65 years old. Since their immunities are going down, they are more susceptible to external factors than young adults. The parameter β_4 in table 1 is less than that in table 2, which implies that the insurance coverage guarantee more interests to older populations, and their diseases may be more treatable.

We can improve the fitness of model by increasing the number of chains, burn-in periods and thinning rates, or considering the interaction between variables.

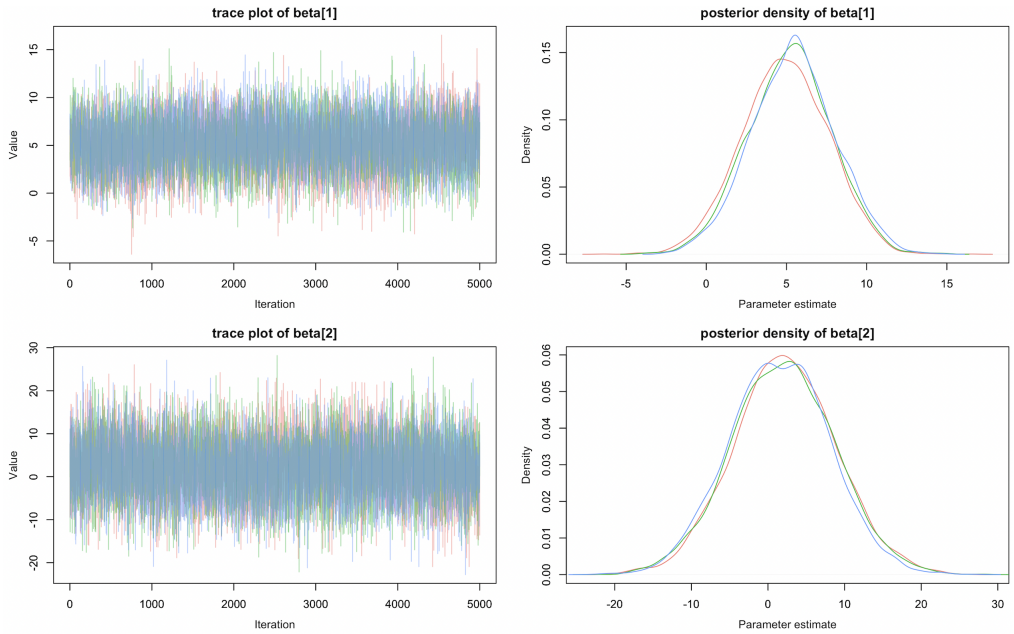

Figure 5: The trace plot and posterior pdf of β_1 and β_2 from the MCMC output of population over 65

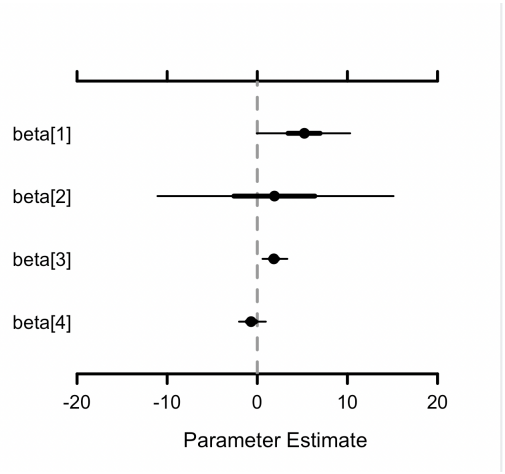

Figure 6: The caterpillar plot of β_i from the output of population over 65

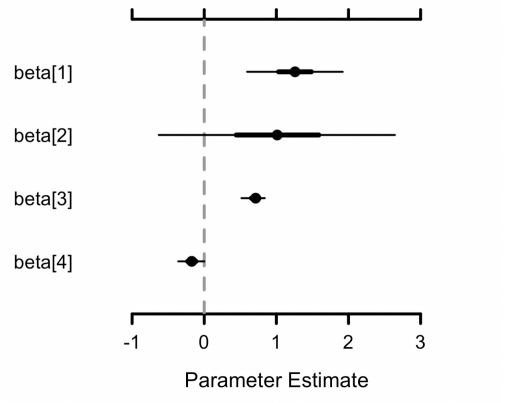

Figure 7: The caterpillar plot of β_i from the output of population under 65

6 Conclusion and Discussion

We first analyze the UHI intensity and the climate data from Google Earth Engine, and find that there is large correlation between summer/winter UHI intensity and surface pressure, nighttime UHI intensity and total precipitation. We confirm that the UHI intensity data is valid since it is closely related to the climate data, and further we may could set up the model to predict or calculate the UHI Intensity in some other regions in the future.

Then we study the relation between the elderly death rate and the UHI intensity. After take some assumption and the transformation, there is evidence that UHI intensity (especailly for summer daytime) could affect the elderly death rate, therefore threat the elderly health to a large extent.

Apart from the death rate, we study the insurance covered rate to analyze the possible health effect from another aspect. After applying the linear regression and random forest, almost all predictors of UHI intensity show a importance in the relation, and the increase of UHI intensity will possibly lead to the insurance rate, which indicates that the UHI would likely influence the health of citizens.

Besides, we also picked some typical illnesses data that could be caused by extreme heat or climate change, including stroke and asthma. For the stroke, we apply the same method as the insurance rate, and get the similar result. There is evidence that the high UHI intensity could lead to high stroke rate(indicator).

For the asthma, we apply the Bayesian linear model, together with the insurance rate also taken into consideration. The result shows that the air temperatures and UHI index have a greater effect to older adults who are 65 years old, and the insurance coverage guarantee more interests to older populations.

However, we did not come up with a complete model that could predict the elderly health data by the UHI intensity due to the different data structures and limitation of the methods we have applied. In the future, we could improve the tasks by finding better data with uniform structure or find better models that could accomplish the prediction work.

In conclusion, we carry out the vulnerability assessment towards the health of elderly people by the UHI from many aspects including death rate, insurance rate, stroke and asthma. The results all show that there is strong evidence that the UHI could affect the health of elderly people, and high UHI intensity could possess greater threats to their health by causing diseases and higher death and insurance rate. Therefore, it is necessary and urgent to take precautions to help elderly people from UHI, especially in large cities. Methods like take better urban planning and health care could be taken in the future.

7 Code Work(Gitlab Repository)

All of our code works have been uploaded to the gitlab, the link is `https://gitlab.umich.edu/zicongx/stats_506_term_project`, the permission to the professor and GSI should be added.

8 Reference

[1] UHI Intensity Data, Earth Engine[online], https://developers.google.com/earth-engine/datasets/catalog/YALE_YCEO_UHI_UHI_all_averaged_v4
[2] Climate Data, Earth Engine[online], https://developers.google.com/earth-engine/datasets/catalog/ECMWF_ERA5_MONTHLY
[3] Elderly Death Data, CDC Wunder[online], https://wonder.cdc.gov/controller/datarequest/D76
[4] Elderly Population Data, US Census[online], https://www.census.gov/data/developers/data-sets/decennial-census.2010.html#list-tab-99P2A1SGILQAEXII31
[5] Insurance Rate Data, US Census[online], http://data.ctdata.org/dataset/health-insurance-coverage
[6] Stroke Indicator Data, CDC[online], https://ephtracking.cdc.gov/DataExplorer/?query=51ED8370-BE00-4813-A4F8-AE641EF61672&fips=26161&G5=9999
[7] Asthma Data, CDC[online], https://ephtracking.cdc.gov/DataExplorer/?query=51ED8370-BE00-4813-A4F8-AE641EF61672&fips=26161&G5=9999