Report

Vishnu Rendla, Christian Ortmann, and Gillian McGinnis

Introduction

Duncan Lee of the university of Glasgow has developed a method within the Bayesian setting for analyzing spatio-temporal patterns of disease risk, using MCMC simulations. As an illustration of this method, he produced a worked application on pneumonia data across England between 2002 and 2017 to derive the posterior distribution of mortality risk. The results of his work are impressive, with very narrow credible intervals for the posterior medians. We were inspired by his work, and chose to adapt his workflow and methodology to analyze spatio-temporal trends of skin cancer risk across the U.S.. Skin cancer has been one of the most common types of cancer in the U.S. over the past few decades, and unlike other common types, such as breast cancer, has been suggested to be associated heavily with a few covariates. Namely, sunlight exposure and an individual's skin color. [Source: IDPH] We wanted to test if these associations are true, and if they are, how strong the link is, and how cancer rates have changed spatially and temporally across the U.S. over a specific period of interest. For this, we used the following datasets: bridged-race population estimates across the U.S., skin cancer rates by race, and average daily sunlight received by U.S. states. Each of these datasets has been obtained from different sources, but they've been wrangled and combined into one cohesive dataset. Although we wished to study trends from 2000-2020, each dataset only covers different periods and some have missing values across regions. Because our model cannot handle missing values, we had to take the overlap of all the datasets, and ended up with data for all states of the continental U.S. from 2003-2011.

Exploratory Data Analysis (EDA)

As a first step, we decided to conduct EDA on the dataset and visualize the basic structure of the data. In this step, we also generate the necessary shapefiles for the map of the U.S., by state, which will facilitate analysis in the next steps.

Plotting spatially averaged boxplots of the 3 variables, we make the following observation. Daily sunlight remains virtually the same throughout the years, which is expected due to the constancy of the climate during this short period. Overall cancer rates have also remained relatively unchanged throughout this period, and the overall fraction of population identifying

as white has also remained relatively stable throughout this period. However, there are outliers across all years, which may be very impactful in a spatio-temporal context, which is what the method we've adopted attempts to capture.

Question 1

Impact of sunlight and diversity on skin cancer risk across the U.S.

The first question we are trying to answer is if cancer rates among populations across regions of the U.S. are dependent on the amount of sunlight received by those regions and the diversity of the resident populations of the regions. We are interested in this question because sunlight exposure and a person's skin color are commonly linked to risk cancer. It is often suggested that frequent exposure to sunlight, which carries ultraviolet radiation, increases a person's risk of skin cancer. Also, melanin in the skin is known to be a protective factor against damage induced by ultraviolet radiation. So, a person of a lighter skin tone, with lower amounts of melanin in the skin, is expected to be at a higher risk of skin cancer.

Methods

Model

In order to first ensure if a spatially autocorrelated model is appropriate for our analysis for our chosen covariates, we first fitted a naive Poisson log-linear regression model incorporating spatial structure, computed residuals, produced a spatial adjacency matrix (which will be used as a parameter in the next steps) and tested for spatial autocorrelation using Moran's I test. We obtained a Moran's I value of 0.13095, indicating a small spatial autocorrelation.

The general form of the model used in our analysis is:

$$Y_{kt} \sim Poisson(\theta_{kt})$$

$$ln(\theta_{kt}) = \beta_0 + \beta 1 S_k + \beta 2 W_{kt} + \psi_{kt}$$

where k and t are spatial and temporal indices, respectively, Y_{kt} are cancer rates, S_k is the average sunlight, W_{kt} is the percentage of population identifying as white, β_0, β_1 are regression coefficients and ψ_{kt} is the spatio-temporal correlation term. Because we are trying to find the spatial distribution of disease risk over time, we model ψ_{kt} as a spatially autocorrelated first-order autoregressive process, given by:

$$\psi_t = \rho_T \psi_{t-1} + \epsilon_t$$

where $\psi_t = (\psi_{1t}, ..., \psi_{Kt})$ is the vector of random effects for all spatial units at time t, and $\epsilon_t = (\epsilon_{1t}, ..., \epsilon_{Kt})$ is the vector of errors accounting for spatial autocorrelation. Temporal autocorrelation is accounted for by the term $\rho_T \psi_{t-1}$. Please see [2 and 3] for the exact form of ϵ_t.

Priors

The regression parameters β_0, β_1 are assigned independent and weakly informative normal priors given by $N(0, 100000)$. This is a hierarchical model, so we further need to specify the priors, $\epsilon_{kt}|\epsilon_{-kt}, W$. They are assumed to follow a Normal distribution (please see [2] for their exact form) whose mean and variances are functions of the hyperpriors $\rho S, \rho T$, and τ^2, which are assumed to be weakly informative.

MCMC parameters

We ran 3 MCMC chains, each with the following specifications: 2200000 samples, while discarding 200000 samples (burnin) and a thinning step of 1000 iterations.

Convergence diagnostics

We checked convergence diagnostics by both examining traceplots and producing Gelman-Rubin statistics.

Credible intervals

Please see the code file.

Posterior predictive checks

Please see the code file.

Discussion

Looking at the trace plots, we can see that we have roughly consistent variation and no trend from the first to final samples in each plot, suggesting that the chains have likely converged. Additionally, based on the Gelman-Rubin statistics, we see our two covariates "proportion white" and "sunlight daily average" have point estimates and upper CI of less than 1.1, further supporting convergence.

To determine if sunlight_daily_avg or proportion_white significantly affect skin cancer, we examine their credible intervals (2.5%-97.5%). Both intervals include zero, suggesting no conclusive evidence that these variables have a significant influence on skin cancer. rho.S is moderate (~0.41), indicating moderate spatial correlation, while rho.T is close to one, suggesting strong temporal correlation. These findings are consistent across all three chains, as indicated by effective sample sizes and convergence diagnostics.

Combining posteriors of the regression coefficients from all chains, scaling by their respective standard deviations and looking at their credible intervals, we can see that a 1 standard

deviation increase in the proportion of white people (0.112) is associated with a 14.9% increase in the cancer count, and this effect is statistically significant as the credible interval does not include 1. However, the effect of daily sunlight average is inconclusive due to numerical instability caused by its very large standard deviation (1630).

Question 2

Spatio-temporal trend of skin cancer risk across the U.S. from 2003 - 2011

In this part, we wish to answer 2 complementary questions: 1. What is the temporal trend of skin cancer risk across the U.S., and 2. What is the spatial distribution of skin cancer risk over the same time period? Together, these will help us understand the overall spatio-temporal trend of skin cancer risk across the U.S. from 2003 - 2011.

We are interested in this question for two reasons. We are curious if states that receive more sunlight such as the Southwest U.S. or states with a relatively higher percentage of white population such as the Mountain states, have higher risk of skin cancer. Furthermore, we want to know if rise in skin cancer awareness over the years, and people increasingly taking precautions such as sunscreen and umbrella usage, has reduced skin cancer rates over time.

Methods

In order to answer these questions, we use the same dataset as in the previous part, and adopt the same model.

Discussion

Temporal trends

Looking at the posterior distribution of average risk per year along with the 95% CI intervals, we find 2 periods of interest: 2003-2004 and 2004-2011. There is a sharp decline in average risk during the first period, and remaining relatively stable throughout the second period. However, the credible intervals are quite wide throughout, indicating a some uncertainty in the risk estimates.

Spatial trends

We interpret the spatial distribution of cancer risk in terms of the posterior exceedance probability (PEP). This is defined as the posterior probability that disease risk is greater than one given the data Y, where the baseline risk of 1 is the average risk across U.S. from 2003-2011. A higher PEP indicates higher risk and a lower PEP indicates a lower risk. We have also plotted the median of the direct risk rates. Both plots indicate that certain states, such as New Mexico, Nevada and Wyoming have a significantly elevated risk. We also observe that the Pacific states have a higher risk.

References

[1] https://dph.illinois.gov/topics-services/diseases-and-conditions/cancer/type/skin.html

[2] https://www.sciencedirect.com/science/article/abs/pii/S1877584520300319

[3] https://link.springer.com/chapter/10.1007/978-1-4612-1284-3_4