Distribution Agreement

Signature:

________________________________ _______________

Chenhao Lu Date

Assessment of Clustering of Cutaneous T-Cell Lymphoma in Metropolitan Atlanta

By

Chenhao Lu

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

____

Jeffrey M. Switchenko, PhD

Committee Chair

____

Yuan Liu, PhD

Committee Member

Assessment of Clustering of Cutaneous T-Cell Lymphoma in Metropolitan Atlanta

By

Chenhao Lu

B.S./B.A.

University of California, San Diego

2018

Thesis Committee Chair: Jeffrey M. Switchenko, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2020

Abstract

Assessment of Clustering of Cutaneous T-Cell Lymphoma in Metropolitan Atlanta

By Chenhao Lu

Background: Cutaneous T-cell lymphoma (CTCL) is a rare type of non-Hodgkin lymphoma (NHL). In the United States, the overall CTCL incidence increased from the year 1973 to 2004 and just started stabilizing in recent decades. The cause of previous increase in incidence remains unknown but was believed to be related with environmental factors. Recent analysis found significant geographic clustering of CTCL at county level in Georgia. There was also correlation between the clustering and an increased concentration of benzene and TCE exposure at county level.

Objective: (1) Identify the geographic clustering of CTCL in metropolitan Atlanta, including five counties: Fulton, DeKalb, Gwinnett, Cobb, and Clayton, at the census tract level. (2) Analyze the correlation of CTCL incidence with ambient benzene and TCE levels.

Methods: The CTCL patients data and chemical toxins data were collected from Georgia Cancer Registry (GCR) and EPA's National Air Toxics Assessment (NATA) respectively. Standardized incidence ratio (SIR) was estimated for each census tract to assess the incidence of CTCL. Global and local Moran's I Statistics were used to assess the geographic clustering. Non-spatial generalized linear models were fitted to study the associations between CTCL incidence and chemical toxins exposures. Spatial generalized linear mixed model within a Bayesian setting was also performed because of the small sample size.

Results: Clusters of census tracts with high CTCL incidence and high chemical toxins exposure were found in metropolitan Atlanta. Expected positive correlation between the exposures to benzene and TCE and the distribution of CTCL incidence at the census tract level were not identified. Majority of the results from non-spatial and spatial models were not statistically significant, except those in logistic regression model, where benzene exposures and concentration showed negative correlation with the log odds of CTCL cases.

Conclusions: The results failed to support the previous finding at finer level, but the Bayesian setting model provided a solid theoretical base for study at small geographic unit level. Further CTCL incidence study should be based on larger patients sample size, more detailed patients information, and more environmental toxins.

Assessment of Clustering of Cutaneous T-Cell Lymphoma in Metropolitan Atlanta

By

Chenhao Lu

B.S./B.A.

University of California, San Diego

2018

Thesis Committee Chair: Jeffrey M. Switchenko, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Department of Biostatistics and Bioinformatics

2020

Table of Contents

1. **Introduction**……………….……………………………………………………………1

2. **Methods**…………………….…..……………………………………………………….3

3. **Results**……………………………………………………………………………….13

4. **Discussion**………………….……………………………………………………..…16

5. **Reference**………………….………….……………………………………………..…19

6. **Tables and Figures**…….……………………………………..…………………..…23

Introduction

Cutaneous T-cell lymphoma (CTCL) is a rare and heterogeneous type of non-Hodgkin lymphoma (NHL) which refers to a group of lymphoproliferative disorders characterized by infiltration of the neoplastic T lymphocytes to the skin at presentation. Mycosis fungicides (MF) and Sézary syndrome (SS) are the two most common forms, accounting for the majority of CTCLs.[1,2] In the United States, overall annual incidence of CTCL from 1973 to 2009 was 7.5 per million people and overall CTCL survival was 78.3%. However, the overall incidence increased from 2.8 per million people (1973-1979) to 10.5 per million people (2000-2004) and just started stabilizing in recent decades.[3,4] There are age, racial and gender differences found in incidence of CTCL. Studies have shown that incidence was higher among older people, males and African Americans, who were specifically diagnosed at younger ages and had worse survival than Caucasians.[5-7]

The cause of the previous increase in incidence remains unknown and one of the obstacles is that the molecular pathogenesis of CTCL is only partially acknowledged.[7] Investigation demonstrated that the increase in NHL can be attributed to increased diagnosis and misdiagnosis of NHL as Hodgkin's in the past, but this does not completely explain the trends in CTCL. Some hypotheses were made, including family factors, medical conditions, radiation, occupation, and environmental exposure.[8] Geographic clustering of CTCL has been recognized across the world, including Vasternorrland county, Sweden[9], Canada[10,11], Texas[7,13], Pittsburgh, Pennsylvania[12], and Georgia[1] therefore implying possible environmental and chemical exposures elicit CTCL. Economic development and industrial construction since last century, as well as association of occupational chemicals with hematologic malignancies could help to justify the trends of rise in CTCL.[14,15]

Benzene and trichloroethylene (TCE), among numerous occupational carcinogenic chemicals, are well-known carcinogens and associate closely with NHL.[16,17] Benzene (C_6H_6) is produced from gasoline, coal, and cigarette smoke that are all commonly used in daily life. In the chemical industry, benzene and its chemical compound are widely applied as well.[18] The carcinogenic property and other toxicity of benzene has been studied extensively and many government agencies such as United States Environmental Protection Agency (EPA)[19] have set regulation for benzene. TCE(C_2HCl_3) is often used as a solvent in metal degreasing and an additive to the dry-cleaning products. The EPA has characterized TCE as carcinogenic to humans.[20]

Recent analysis found significant geographic clustering of CTCL on the level of counties in Georgia, particularly the four most populous counties around Atlanta with incidence rates of CTCL that were between 1.2 and 1.9 times higher than the state average. There was also correlation between the clustering and an increased concentration of benzene and TCE exposure.[1]

Based on the finding from previous studies, this thesis focused on the geographic clustering of CTCL in metropolitan Atlanta, including five counties: Fulton, DeKalb, Gwinnett, Cobb, and Clayton, at the census tract level instead of county level. The census tract level of study enables us to identify the geographic difference in CTCL risk and chemical exposure more comprehensively. The investigation used geocoded incidences of CTCL from the Georgia Cancer Registry and demographic data from the US Census Bureau. The aim is to evaluate clustering and to analyze the correlation of CTCL incidence with ambient benzene and TCE levels, obtained through the EPA's National Air Toxics Assessment (NATA) database in each census tract from 1999 to 2005. Then, Bayesian spatial modeling was applied to estimate the correlation, because of the small sample size encountered in this thesis.

Method

CTCL Patients Data

The CTCL patients data used in the study were collected from the Georgia Cancer Registry and the previous study[1]. We included patients age ≥ 15 years old who had a new diagnosis of CTCL between 1999 and 2015 in metropolitan Atlanta. Patient demographic and disease characteristics were also collected. Patients with missing demographic information (race, age, and sex) were excluded from the data. In total, 566 CTCL patients complying with standards were identified.

We also identified the count of CTCL cases in each census tract. Census tract is an areal unit similar to a neighborhood established by the Census Bureau. It is a small and relatively permanent statistical subdivision of a county. The population in census tracts is approximately 4000 on average. Some census tracts were updated over the years, because of the change in population. For example, census tracts with a population over 8000 would be split into two or more tracts in new Census. Small boundary corrections were also encountered in some cases. Due to the time length of our study, our data contained census tracts from 2000, and 2010 Census. For consistency, we used a linkage file to assign a 2000 census tract to its most likely 2010 census tract. According to 2010 US Census data[21], there were totally 632 census tracts in Fulton, DeKalb, Gwinnett, Cobb, and Clayton. Within those census tracts, 347 of them had at least one case of CTCL from 1999 to 2015. Four census tracts with zero population were excluded from the analysis.

Standardized Incidence Ratio (SIR)

SIR is frequently used in epidemiology to indicate whether the occurrence of certain disease in a population is high or low. We estimated the SIR for each census tract in metropolitan Atlanta. An SIR > 1 showed that the number of observed CTCL cases of that census tract is higher than expected and indicated a greater increase in the risk of CTCL.

The SIR of each census tract could be calculated by the following formula:

$$\mathrm{SIR}_i = \frac{\text{Observed number of CTCL cases from 1999-2015 in census tract}_i}{\text{Expected number of CTCL cases from 1999-2015 in census tract}_i},$$

where the expected cases per year in census tract *i* equal to:

$$\sum_{j=1}^{n} \text{Population in subgroup } j \text{ in census tract } i \times \text{National CTCL rate for subgroup } j,$$

and *n* = the number of subgroups.

The expected number of CTCL cases per year were estimated by multiplying the national CTCL incidence rates for each subgroup by the number of individuals in the corresponding subgroup living in each census tract. Numbers of individuals within each subgroup at the census tract level were obtained from the 2010 US Census Data[21]. Then we multiplied the expected CTCL cases in each census tract for the year of 2010 by the length of study period (17 years) to get the total expected number of CTCL cases from 1999 to 2015 in each census tract. The national CTCL incidence rates were obtained from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program[22]. For the calculation, we constructed eight subgroups with combinations of age (ages 15-59, ≥ 60 years old), race (white, non-white), and gender (male, female).

Benzene and TCE Data

We obtained benzene and TCE exposure level data from the EPA's National Air Toxics Assessment (NATA) database[23], which provides assessment of outdoor air toxics and estimates the cancer and other health issue risks for common air pollutants. NATA calculated the concentration and risks for air toxics using different dispersion models at both county and census tract level since 1999. Based on our interest in census tract level data, we compiled 1999, 2002, and 2005 NATA for Benzene/TCE in Georgia, created a subset for census tracts in metropolitan Atlanta, and averaged across time the concentration and exposure data for our analysis. The exposure concentration of toxic is estimated by the Hazardous Air Pollutant Exposure Model (HAPEM), based on ambient air concentration data, indoor/outdoor environment concentration relationships, population data, and human activity pattern data[24]. The exposure level of toxic is estimated by the Assessment System for Population Exposure Nationwide (ASPEN), based on the EPA's Industrial Source Complex Long Term model (ISCLT) which simulates the behavior of the pollutants after they are emitted into the atmosphere[25]. Both HAPEM and ASPEN report the estimates by microgram per cubic meter ($\mu g/m^3$). All census tracts in the data from 1999 to 2005 were based on the 2000 US Census Data. Again, we used the linkage method to assign each 2000 census tract to its most likely 2010 census tract for the consistency of our analysis.

Spatial Analysis

We performed spatial analysis of SIRs and chemical toxins (benzene/TCE) at the census tract level using several R packages, including *sf*, *RColorBrewer*, *classInt*, *map*, *ape*, and *spdep*. First, we obtained the shape files of census tracts in metropolitan Atlanta from US Census Bureau's 2010 TIGER/Line files[26]. All shape files were read into R by package *sf* with function *read_sf*. Then, we used packages *RColorBrewer* and *classInt* to link variables SIRs, benzene and TCE with five color categories that showed on the maps. Color categories was defined using Jenks optimization method, which is a data clustering method created to determine the best classification of values into difference classes and to minimize class's deviation from the mean. Next, we plot those variables overlaid with shape files to create maps at the census tract level using package *maps*. Finally, we evaluated the geographic clustering of SIRs, benzene, and TCE by both global and local spatial analysis to create Moran's *I* and local Moran's *I* statistics.

For global spatial analysis, we first calculated the distances among all the census tracts using their latitude and longitude centroid coordinates and stored the results in a matrix. We inverted the values in this matrix and inserted zeros to all diagonals to obtain a matrix of inverse distance weights, which we applied in the function *Moran.I* to achieve global Moran's *I* statistic as well as pseudo *P* value. Global Moran's *I* statistic, which is often called as Moran's *I* statistic, is always between -1 to 1 and it measured multi-dimensional spatial autocorrelation and indicated clustering (statistic > 0), randomness (= 0), and dispersion (< 0) of the studied variables. The *P* value was based on the significance test with null hypothesis that our interested variables were randomly distributed. The detailed calculation of Moran's I statistic is shown here:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2},$$

where z_i = deviation of an attribute for feature *i* from its mean $(x_i - \overline{X})$, $w_{i,j}$ = spatial weight between feature *i* and *j*, n = total number of census tracts, and

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \,.^{27}$$

For local spatial analysis, we used package *spdep* to calculate the Local Moran's *I* of SIRs at census tract level and to create the cluster map. Local Moran's *I* was one kind of local indicators of spatial association (LISA), which allowed us to access the influence of each census tract and to identify clusters and outliers at each census tract. Local Moran's I of each census tract was calculated by the function *localmoran*. We standardized the SIRs as observed values and calculated the spatial lagged values of each census tract using function *lag.listw*. LISA value was calculated by multiplying those two values. Based on LISA value, we divided the census tracts which had significant local Moran's I statistic into four quadrants: "High-High" (HH), "Low-Low"(LL), "High-Low"(HL), and "Low-High"(LH). HH and LL quadrants indicated spatial clustering of similar LISA value in the census tract. LH and HL quadrants indicated spatial outliers and spatial association of dissimilar LISA value[28]. More general speaking, a HH census tract meant this census tract had high SIRs and was surrounded by census tracts with high SIRs. A LL census tract was interpreted similarly, except it had low SIRs and was surrounded by low SIRs census tracts. HH and LL census tracts could also been seen as "hot spots" and "cold spots" respectively.

Statistical Analysis

To study the associations between CTCL SIRs and chemicals, we used two approaches: (1) Non-spatial Generalized Linear Models (GLM) (Logistic and Poisson regression models); (2) Spatial generalized linear mixed models (GLMM) for areal unit data within a Bayesian setting and conditional autoregressive priors (CAR).

Non-spatial Generalized Linear Model (GLM)

Poisson regression is often used to model count data and rate data, for example, the count and SIRs of CTCL cases in our study. We directly modeled the SIRs of CTCL at census tract level by adding an offset variable to normalize the fitted cell means per space interval. Benzene / TCE concentration or exposure were the only predictors in the model and they were again fitted separately in four models. The basic Poisson regression model for SIRs is shown here:

$$log(\mu) = \beta_0 + \beta_1 * X + log(t),$$

where $\mu\ = E(Y_i)$, Y_i = number of observed CTCL cases at the census tract i, X = the chosen benzene / TCE concentration or exposures variables, and t = number of expected CTCL cases the census tract. $log(t)$ was referred as the offset variable. The model could be rearranged to:

$$log(\mu) - log(t) = \beta_0 + \beta_1 * X$$

$$log(\mu/t) = \beta_0 + \beta_1 * X,$$

where μ/t = the estimated SIR of CTCL cases at the census tract. Based on this regression model we could check and compare how the chemical concentrations and exposures associate and affect the SIRs of CTCL at the census tract.

Logistic regression, a part of GLM is used when the dependent or target variable is binary. In order to meet that, we created two indicator variables for CTCL cases (cases > 0 / = 0 , cases > =2 / < 2) as the target variables. Then, we modeled on the odds of CTCL cases instead of SIR. In order to control for tract-level covariate direction, we also introduced three census tract-level demographic variables: the median age of population, male percentage in population, and African American percentage in population given we did not include an expected count offset. These estimates were obtained from American Fact Finder[29] for the 2010 US Census and 2006-2010 American Community Survey by US Census Bureau. They were independent and acted as the predictor variables in the model along with the benzene / TCE concentration and exposures, which were fitted separately in four models due to the close association among them. The logistic regression model is shown here:

$$logit(p) = \beta_0 + \beta_1 * X + \beta_2 * \text{Median Age} + \beta_3 * \text{Male Percentage} + \beta_4 * \text{African American Percentage}$$

where p = probability of number of CTCL cases based on the chosen indicator variable and X = the chosen benzene / TCE concentration or exposures variables. We fitted two indicator variables each with four chemical variables. By this model setting, we could check and compare how the chemicals and selected demographic variable affect the odds of CTCL. All non-spatial GLMs were fitted in R with *glm* fucntion.

Spatial Generalized Linear Mixed Model (GLMM)

GLM models such as logistic and Poisson regression models usually operate under an independence assumption, but the data from diverse areas could be spatially correlated. Often, observations located closer tend to be more similar than those located far away from each other. For example, in our study, census tracts located closer were more likely to have similar concentration and exposures of chemicals. Thus, ignoring the spatial dependence in the data when fitting GLM models could violate the assumption and result in bias and incorrect statistical inference. Also, the standard models might not handle the small numbers of CTCL cases we faced in the study. We expected large number of census tracts having zero CTCL cases during the study period. The estimation from standard models particularly, the Poisson regression model, could be unstable and biased.

Therefore, we used spatial GLMMs for areal unit data within a Bayesian setting and conditional autoregressive priors (CAR). Areal unit data is one type of spatial data, where observations were obtained from regions with well-defined boundaries, such as the census tracts in our study. We assumed CAR with the random effects which allowed the estimation of each census tract to borrow strength from its neighbors. So, the effect of "zero case" in census tracts could be lightened and the estimates could be more stable and accurate. The spatial structure underlying the CTCL cases and chemicals at census tract level was summarized through an adjacency matrix W, which stores the neighborhood structure of each areal unit. The matrix connecting units i and j is as following:

$$W_{ij} = \begin{cases} 0 & \text{if } i = j \\ 0 & \text{if } i \text{ and } j \text{ are not neighbors} \\ c_{ij} & \text{if } i \text{ and } j \text{ are neighbors} \end{cases},$$

where c_{ij} represents the strength of neighborhood between areal units i and j.[27]

Spatial GLMM is a kind of hierarchical spatial model where normal maximum likelihood estimation could be difficult to apply. Instead, we used a Bayesian setting approach by computing a posterior distribution of unknowns and making inference in terms of probability statements conditional on the observed data. The Bayesian approach provided some advantages over the frequentist statistical approach. The Bayesian method allowed us to include spatial correlation among the random effects and to acknowledge other uncertainties in our model caused by the low number of CTCL cases in our study. To compute and estimate the Bayesian posterior distribution, we applied the Markov Chain Monte Carlo (MCMC) simulation, which is a numeric algorithm sampling from high dimensional probability distributions.[30] The basic Poisson GLMM took the following form:

$log(\mu_i) = \beta_0 + \beta_1 * X + log(t_i) + \theta_i$,

$Y_i \mid \mu_i \sim$ Poisson(μ_i),

$\theta_i \sim$ CAR(σ^2),

where $\mu_i = E(Y_i)$, Y_i = number of observed CTCL cases at the census tract i, X = the chosen benzene / TCE concentration or exposures variables, t_i = number of expected

CTCL cases the census tract i, and θ_i is the vector of random effects assumed to have conditional autoregressive priors. R package *CARBayes* was used for fitting the spatial GLMM.

Results

Descriptive Statistics

In total, 566 new cases of CTCL in the time period of 1999 to 2015 were included in our study. The median and maximum SIR for all census tracts was 1.053 and 12.456, with mean equal to 1.374.

The ASPEN concentration and HAPEM exposure levels of benzene (in $\mu g/m^3$) per census tract, averaged from 1999 through 2005, were both 2.11. The ASPEN concentration and HAPEM exposure levels of TCE were 0.090 and 0.077.

Spatial Analysis

Figure 1 and Figure 2 are maps of SIRs, benzene, and TCE in metropolitan Atlanta at census tract level. Areas of high SIRs were distributed randomly and did not have a straight look of geographic clustering pattern. Areas of high concentration and exposures of benzene and TCE were concentrated in the center of metropolitan Atlanta. The Moran's I statistic and p-values for SIRs, benzene concentration and exposure levels, and TCE concentration and exposure levels are 0.003 (0.046), 0.166 ($p < 0.001$), 0.182 ($p < 0.001$), 0.140 ($p < 0.001$), and 0.144 ($p < 0.001$), respectively. For SIRs, the Moran's I statistic is significant but close to zero, which indicated spatial randomness and almost no geographic clustering existing. For other variables, clustering exists in metropolitan Atlanta.

The Local Moran's I map for the SIRs (Figure 3) identified hot spots and cold spots through out the metropolitan Atlanta. Majority of "High-High" census tracts were located in Fulton and DeKalb Counties. Those census tracts also had high estimates of concentration and exposures of benzene and TCE. "High-High" census tracts in Clayton and Cobb counties contained lower estimates of benzene and TCE. "Low-Low" census tracts of SIRs were located in the downtown Atlanta. However, those census tracts had high estimates of benzene and TCE.

Statistical Analysis

Non-spatial Poisson Regression Model

The resulting parameter estimates are presented in Table 1. The β estimates and P value for SIR and benzene concentration (in $\mu g/m^3$) was 0.080 (P = 0.299), for SIR and benzene exposure (in $\mu g/m^3$) was 0.096 (P = 0.237), for SIR and TCE concentration (in $\mu g/m^3$) was 0.902 (P = 0.466), and for SIR and TCE exposure (in $\mu g/m^3$) was 1.255 (P = 0.438). All estimates were not statistically significant (P > 0.05). Therefore, concentration and exposure of both chemical toxins were not associated with SIR at census tract level with the data we had.

Non-spatial Logistic Regression Model

We had two indicator variables corresponding to two different p for the logistic regression models. Figure 4 illustrates the map of census tract based on these indicator variables. First, we fitted the indicator variable (CTCL cases = 0 / > 0) with chemical toxins and other three demographic factors we chose (median age, male population percentage, and African American population percentage). The results are presented in Table 2. The signs of parameter estimates for Benzene and TCE were all negative, which means that those chemical toxins influenced the number of CTCL cases negatively. For example, the β estimate and P value for Benzene concentration (in $\mu g/m^3$) in logistic regression model was -0.228 (P = 0.466). It indicates that, for a 1-unit increase in $\mu g/m^3$, there is a decreased in CTCL log odds. But, again none of the estimates for predictors and intercepts were statistically significant. The results of logistic regression models based on the second indicator variable (CTCL cases < 2 / > =2) are shown in Table 3. This time, β estimates concentration and exposures of chemical toxins indicated greater negative impacts on the log odds of CTCL cases. Also, the β estimates for benzene concentration and exposures were statistically significant. However, all other β estimates in those models were still non-significant.

Spatial Poisson regression Model with Bayesian setting

Table 4 illustrates the resulting posterior parameter estimates of spatial Poisson regression model within a Bayesian setting. We performed the model based on 5,000 post burn-in and thinned MCMC situation samples, which were obtained following a burn-in period of 50,000 and thinning remained samples by 10 to reduce their autocorrelation.

The estimates of median represented the β estimate in standard Poisson model. Bayesian model smoothed SIR map is illustrated in Figure 5. The estimated median for benzene concentration (in $\mu g/m^3$) was 0.076, for benzene exposure (in $\mu g/m^3$) was 0.092, for TCE concentration (in $\mu g/m^3$) was 0.808, and for TCE exposure (in $\mu g/m^3$) was 1.088. The results were close to those from non-spatial Poisson regression model. Deviation information criterion (DIC) for each model was 1505.94, 1505.59, 1506.32, and 1506.39 respectively. Typically, smaller DIC value indicates better fit of the model.

Discussion

In this thesis, we performed spatial and statistical analysis for the CTCL incidence and the environmental risk factors including benzene and TCE in five counties in metropolitan Atlanta at the census tract level. We developed non-spatial Logistic and Poisson regression models and spatial Poisson regression model within a Bayesian setting. Clusters of census tracts with high CTCL incidence and high chemical toxins exposure were identified, but we failed to find an expected positive correlation between the exposures to benzene and TCE and the distribution of CTCL incidence at the census tract level.

The geographic clustering of CTCL has been identified in Georgia and several other regions nationally and internationally[1,7-12]. More specifically, a correlation between increased exposures to benzene and TCE and increased incidence of CTCL was observed in Georgia at county level[1]. Our study at the census tract level did not support this previous analysis. In addition, the majority of the prior studies included over 2000 patients, with the exception of a smaller study analyzing 274 patients with CTCL in Pittsburgh which also failed to find a correlation between environmental toxins and CTCL incidence because of the biased in patient selection[12]. Also, most studies were based on county or city level. The Canadian analysis of 6685 patients from 1992 to 2010 was the only study that apply analysis by postal code and street level. Deeper analysis showed that high incidence regions were linked to industrial sites.

CTCL is a rare cancer. In our study, we only identified 566 patients from 1999 to 2015. The sample size was small compared to the 632 census tracts in metropolitan Atlanta (2010 U.S. Census). 285 (49%) of the census tracts in our study did not have cases in 16 years of period. The challenge of small sample size brought bias and underpowered our analysis. For small geographic units, like census tracts in our study, they were more easily to be affected by unstable rates. A small change in the number of CTCL cases could cause a large impact on the SIR estimation of the census tract, especially for those with low population sizes. Therefore, it is crucial to make sure the count of CTCL cases is accurate. But it was possible that not all the CTCL cases was captured, because CTCL could be misclassified as other peripheral T-cell lymphoma or T-cell lymphoma NOS[1]. Another limitation is the absence of confounders in our patients data, such as income, education level, job type, and so on. Patients could possibly work at an area with high chemical toxins exposure, but live in the suburbs, where the case was identified. In addition, the molecular pathogenesis of CTCL still remained unknown. We do not know the exact chemicals or risk factors that caused CTCL. Except benzene and TCE, other risk factors and chemical toxins should be included in future studies.

Overall, the thesis demonstrated that there is small sign of geographic clustering of CTCL in the metropolitan Atlanta at census tract level. We identified census tracts with high CTCL incidence and surrounded by other census tracts with high CTCL incidence. But both standard models and spatial model within a Bayesian setting showed that the CTCL incidence was not correlated or negatively correlated with the benzene and TCE exposures. For further study, there should be larger number of CTCL cases. The study could be focused on areas with more CTCL cases or extend the study period of time. More detailed information on patients could be collected, for example, the income, education level, job type, and working location of the patients. Location of factories, industrial facilities, gas station, and other faculties that generated chemical toxins could be included in the spatial analysis. Meanwhile, more environmental toxins should be included in the study. Both standard Poisson regression and logistic regression models showed their advantages of fitting count and binary data, but they could not completely justify the spatial dependence and small sample size issue we faced in the analysis. Spatial Poisson regression model within a Bayesian setting showed its advantages in this circumstance. Even though the Bayesian setting model did not provide statistically significant results in our study, it still provided a solid theoretical base for future study at small geographic unit level. Similar methods should be applied in other areas to help identify the specific etiologic triggers for CTCL.

Reference

1. Clough, L., Bayakly, A. R., Ward, K. C., Khan, M. K., Chen, S. C., Lechowicz, M. J., . . . Switchenko, J. M. (2020). Clustering of cutaneous T-cell lymphoma is associated with increased levels of the environmental toxins benzene and trichloroethylene in the state of Georgia. *Cancer*. doi:10.1002/cncr.32665
2. Bagherani, N., & Smoller, B. R. (2016). An overview of cutaneous T cell lymphomas. *F1000Res, 5*. doi:10.12688/f1000research.8829.1
3. Korgavkar, K., Xiong, M., & Weinstock, M. (2013). Changing incidence trends of cutaneous T-cell lymphoma. *JAMA Dermatol, 149*(11),1295-1299. doi:10.1001jamadermatol.2013.5526
4. VD, C., & Weinstock, M. (2007). Incidence of cutaneous T-cell lymphoma in the United States, 1973-2002. *Arch Dermatol, 143*, 854-859.
5. Wilson, L. D., Hinds, G. A., & Yu, J. B. (2012). Age, race, sex, stage, and incidence of cutaneous lymphoma. *Clin Lymphoma Myeloma Leuk, 12*(5), 291-296. doi:10.1016/j.clml.2012.06.010
6. Imam, M. H., Shenoy, P. J., Flowers, C. R., Phillips, A., & Lechowicz, M. J. (2013). Incidence and survival patterns of cutaneous T-cell lymphomas in the United States. *Leuk Lymphoma, 54*(4), 752-759. doi:10.3109/10428194.2012.729831
7. Litvinov, I. V., Tetzlaff, M. T., Rahme, E., Jennings, M. A., Risser, D. R., Gangar, P., . . . Duvic, M. (2015). Demographic patterns of cutaneous T-cell lymphoma incidence in Texas based on two different cancer registries. *Cancer Med, 4*(9), 1440-1447. doi:10.1002/cam4.472

8. Hartge P, Devesa SS. Quantification of the impact of known risk factors on time trends in non-Hodgkin's lymphoma incidence. *Cancer Res*. 1992;52(19 suppl):5566s-5569s.

9. Gip L, Nilsson E. [Clustering of mycosis fungoides in the County of Vasternorrland]. *Lakartidningen*. 1977;74:1174-1176.

10. Ghazawi FM, Netchiporouk E, Rahme E, et al. Distribution and clustering of cutaneous T-cell lymphoma (CTCL) cases in Canada during 1992 to 2010. *J Cutan Med Surg*. 2018;22:154-165.

11. Ghazawi FM, Netchiporouk E, Rahme E, et al. Comprehensive analysis of cutaneous T-cell lymphoma (CTCL) incidence and mortality in Canada reveals changing trends and geographic clustering for this malignancy. *Cancer*. 2017;123:3550-3567.

12. Moreau JF, Buchanich JM, Geskin JZ, Akilov OE, Geskin LJ. Non- random geographic distribution of patients with cutaneous T-cell lymphoma in the Greater Pittsburgh Area. *Dermatol Online J*. 2014;20: pii: 13030/qt4nw7592w.

13. Litvinov IV, Tetzlaff MT, Rahme E, et al. Identification of geographic clustering and regions spared by the cutaneous T-cell lymphoma in Texas using 2 distinct cancer registries. *Cancer*. 2015;121:1993-2003.

14. Ghazawi FM, Alghazawi N, Le M, et al. Environmental and other extrinsic risk factors contributing to the pathogenesis of cutaneous T cell lymphoma (CTCL). *Front Oncol*. 2019;9:300.

15. Smith MT. Advances in understanding benzene health effects and susceptibility. *Annu Rev Public Health*. 2010;31:133-148.

16. Switchenko JM, Bulka C, Ward K, et al. Resolving uncertainty in the spatial relationships between passive benzene exposure and risk of non-Hodgkin lymphoma. *Cancer Epidemiol*. 2016;41:139-151.

17. Bassig BA, Zhang L, Vermeulen R, et al. Comparison of hematological alterations and markers of B-cell activation in workers exposed to benzene, formaldehyde and trichloroethylene. *Carcinogenesis*. 2016;37:692-700.

18. Wilbur S, Wohlers D, Paikoff S, Keith LS, Faroon O. ATSDR evaluation of potential for human exposure to benzene. *Toxicol Ind Health*. 2008;24(5-6):399-442.

19. Control of Hazardous Air Pollutants From Mobile Sources, March 29, 2006, Volume 71, Number 60, Page 15853-15902, wais.access.gpo.gov.

20. US Environmental Protection Agency. Fact Sheet on Tricholoroethylene (TCE). Accessed March 1, 2019. https://www.epa.gov/assessing-and- managing-chemicals-under-tsca/fact-sheet-trichloroethylene-tce.

21. US Census Bureau. Decennial Census of Population and Housing. Accessed November 1, 2018. https://www.census.gov/programs-surveys/decennialcensus/data/datasets.2010.html.

22. Surveillance, Epidemiology, and End Results Program, National Cancer Institute, https://seer.cancer.gov/.

23. US Environmental Protection Agency. National Air Toxics Assessment. Accessed March 1, 2019. https://www.epa.gov/national-air-toxics-assessment.

24. US Environmental Protection Agency. Human Exposure Modeling - Hazardous Air Pollutant Exposure Model (HAPEM). https://www.epa.gov/fera/human-exposure-modeling-hazardous-air-pollutant-exposure-model-hapem.

25. US Environmental Protection Agency. The ASPEN Model. https://archive.epa.gov/airtoxics/nata/web/html/aspen.html.

26. US Census Bureau. TIGER/Line Shapefiles and TIGER/Line Files. Accessed November 1, 2018. https://www.census.gov/geo/maps-data/ data/tiger-line.html.

27. P. A. P. MORAN, NOTES ON CONTINUOUS STOCHASTIC PHENOMENA, *Biometrika*, Volume 37, Issue 1-2, June 1950, Pages 17–23.

28. Anselin, L. (1995), Local Indicators of Spatial Association —LISA, *Geographical Analysis*, 1995; 27(2):93-115.

29. American Factfinder. https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t.

30. Duncan L.,CARBayes version 5.1.2: An R Package for Spatial Areal Unit Modeling with Conditional Autoregressive Priors.

Tables and Figures

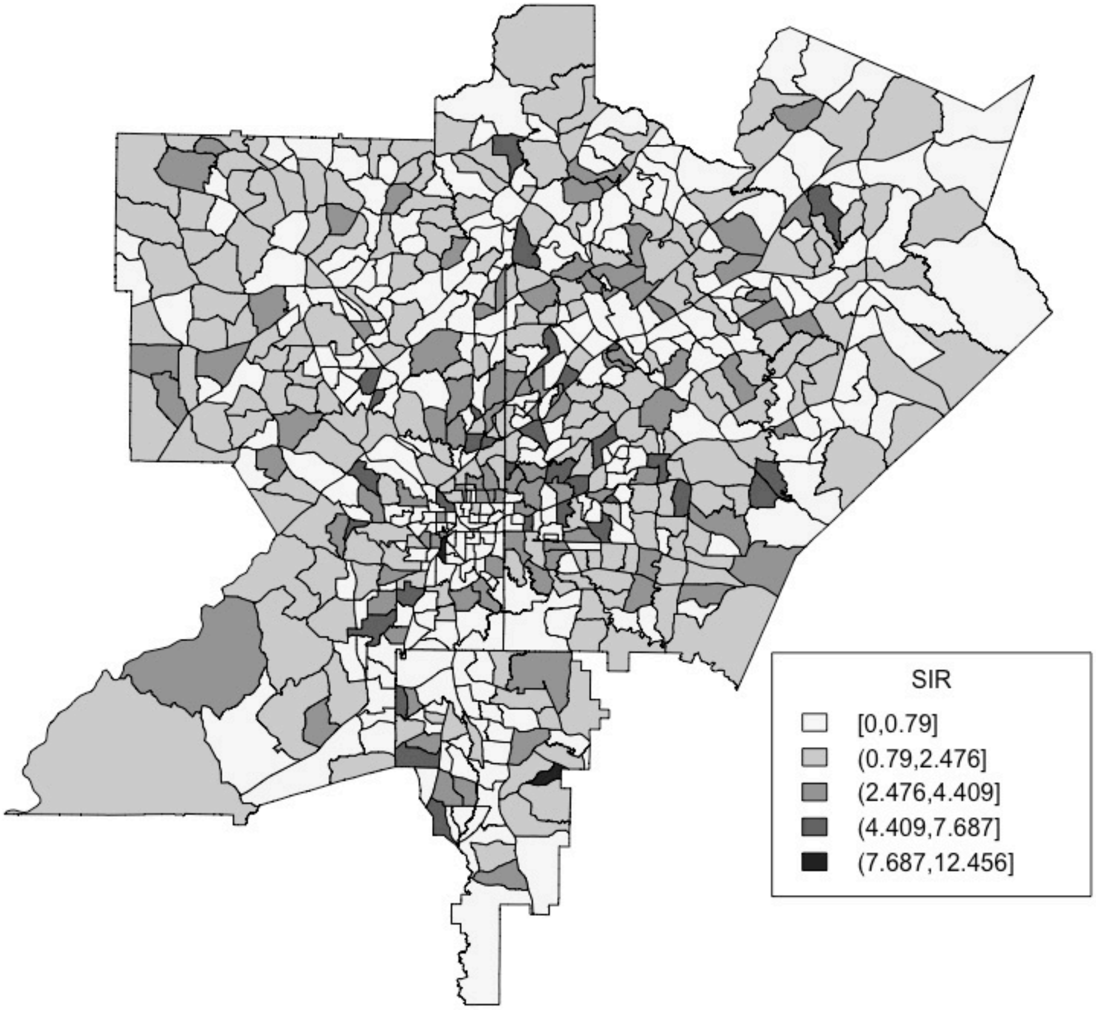

Figure 1. Standardized incidence ratios (SIR) of cutaneous T-cell lymphoma are illustrated for each census tract in metropolitan Atlanta, aggregating cases from 1999 through 2015.

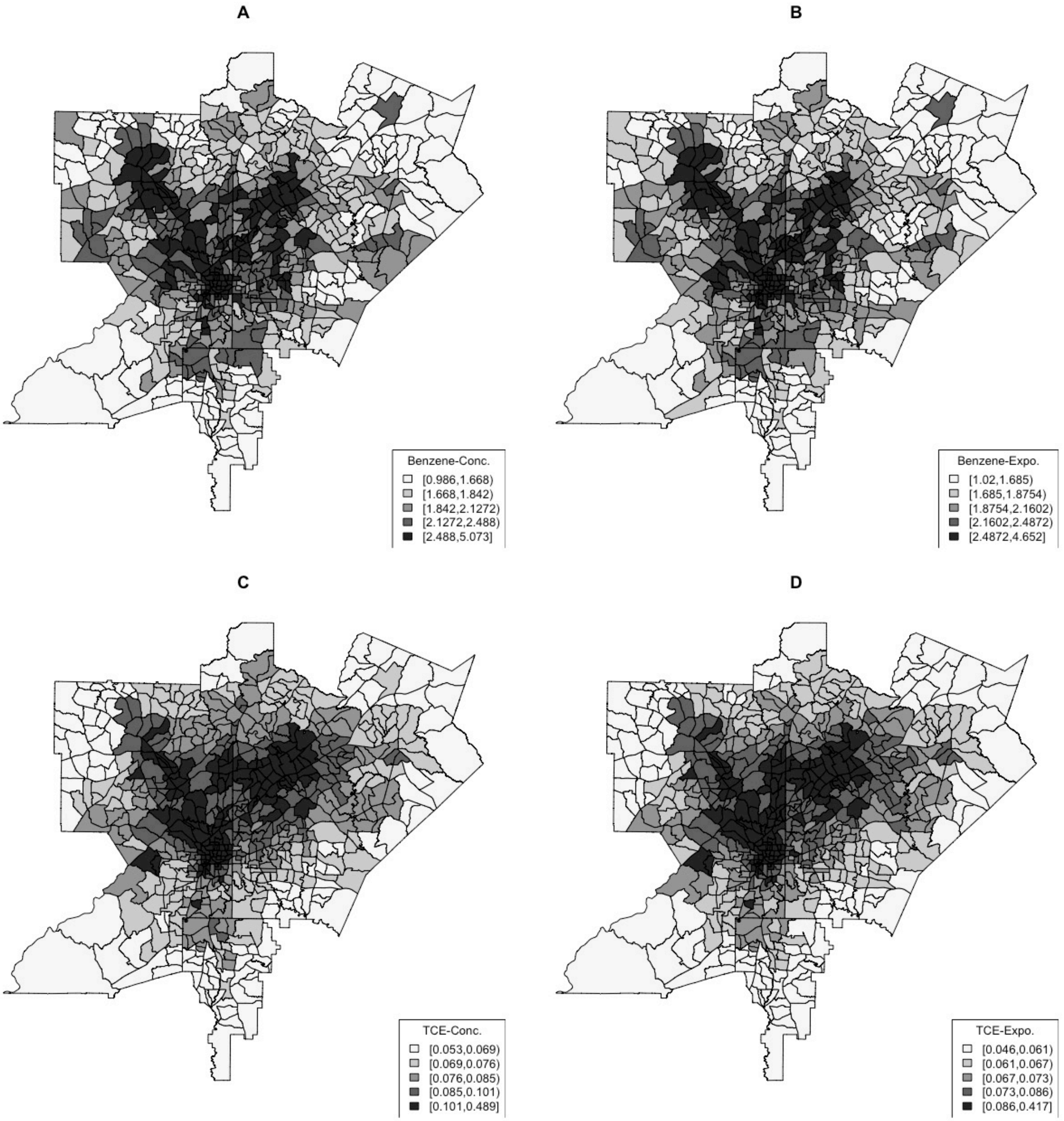

Figure 2. Image illustration (A) Assessment System for Population Exposure Nationwide (ASPEN) concentrations (in $\mu g/m^3$) and (B) Hazards Air Pollutant Exposure Model (HAPEM) exposure (in $\mu g/m^3$) for benzene, and (C) ASPEN concentrations (in $\mu g/m^3$) and (D) HAPEM exposure (in $\mu g/m^3$)for trichloroethylene (TCE) for each census tract in metropolitan Atlanta.

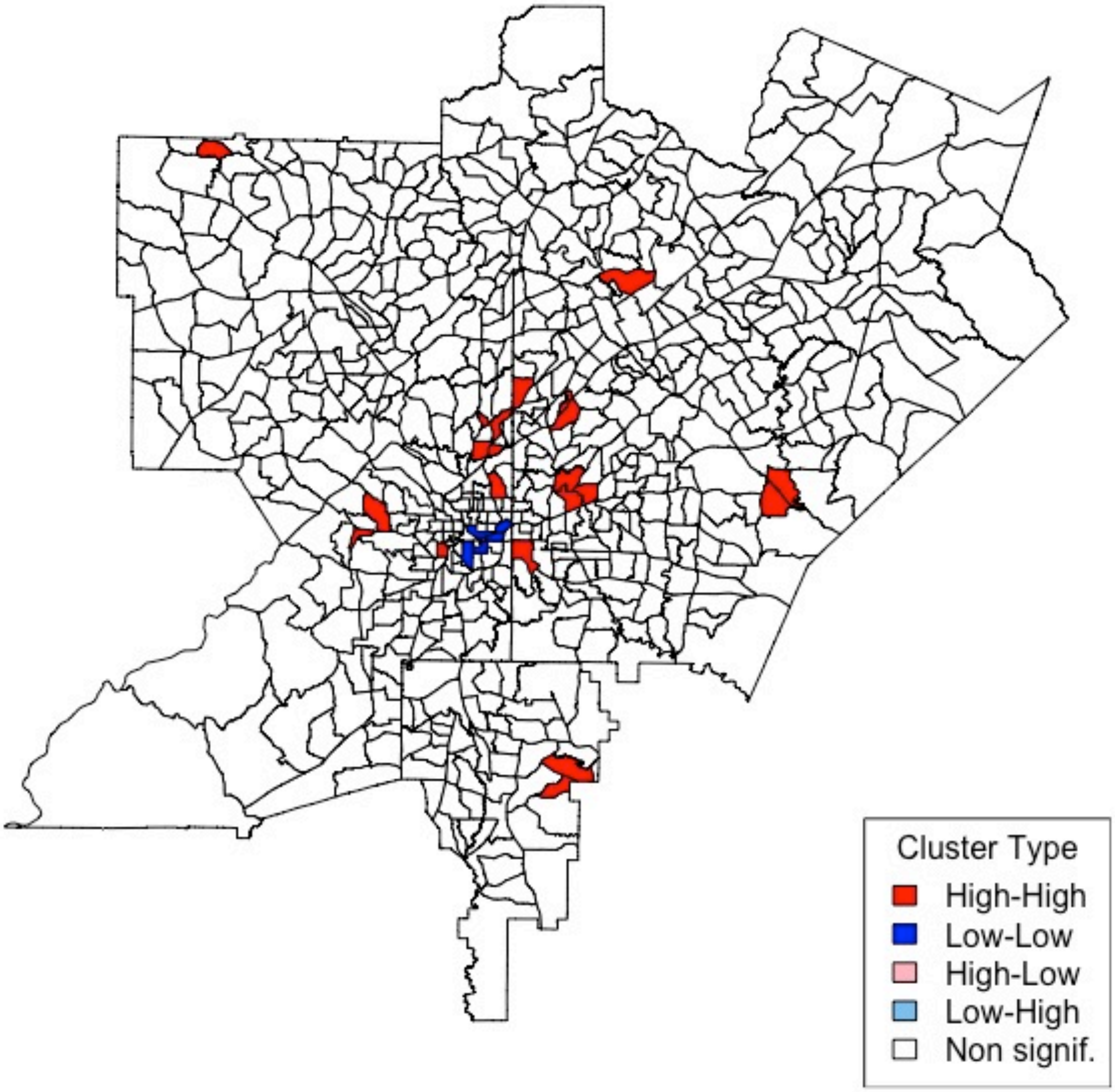

Figure 3. Local Moran's I statistic results are illustrated by census tract in metropolitan Atlanta. Dark red indicates "high-high" area and dark blue indicates "low-low" area.

Table 1. Parameter Estimates Summaries of Poisson Regression Models

Variable	β Estimate	Standard Error	Z	P
Model 1				
Intercept	0.172	0.160	1.074	0.283
Benzene Concentration (ASPEN)	0.080	0.076	1.039	0.299
Model 2				
Intercept	0.138	0.170	0.809	0.418
Benzene Exposure (HAPEM)	0.096	0.081	1.182	0.237
Model 3				
Intercept	0.255	0.114	2.240	0.025
TCE Concentration (ASPEN)	0.902	1.236	0.729	0.466
Model 4				
Intercept	0.240	0.126	1.903	0.057
TCE Exposure (HAPEM)	1.255	1.619	0.775	0.438

Table 2. Parameter Estimates Summaries of Logistic Regression Models (cases = 0 / > 0)

Variable	β Estimate	Standard Error	Z	P
Model 1				
Intercept	0.555	0.991	0.569	0.576
Benzene Concentration (ASPEN)	-0.228	0.149	-1.537	0.124
Median Age	-0.150	0.262	-0.574	0.566
Male Population %	-0.780	1.465	-0.532	0.594
African American Population %	0.017	0.015	1.111	0.266
Model 2				
Intercept	0.582	0.992	0.587	0.557
Benzene Exposure (HAPEM)	-0.257	0.158	-1.624	0.104
Median Age	-0.136	0.261	-0.521	0.602
Male Population %	-0.735	1.467	-0.501	0.616
African American Population %	0.017	0.015	1.128	0.259
Model 3				
Intercept	0.374	0.983	0.380	0.704
TCE Concentration (ASPEN)	-2.227	2.292	-0.972	0.331
Median Age	-0.166	0.263	-0.629	0.529
Male Population %	-1.068	1.45	-0.736	0.462
African American Population %	0.018	0.015	1.199	0.230
Model 4				
Intercept	0.410	0.987	0.415	0.678
TCE Exposure (HAPEM)	-3.171	3.024	-1.049	0.294
Median Age	-0.167	0.263	-0.633	0.527
Male Population %	-1.049	1.451	-0.723	0.470
African American Population %	0.018	0.015	1.197	0.231

Table 3. Parameter Estimates Summaries of Logistic Regression Models (cases < 2 / >= 2)

Variable	β Estimate	Standard Error	Z	P
Model 1				
Intercept	0.449	1.326	0.339	0.735
Benzene Concentration (ASPEN)	-0.410	0.298	-2.072	0.038
Median Age	0.111	0.313	0.355	0.723
Male Population %	-2.860	2.084	-1.372	0.170
African American Population %	0.014	0.018	0.794	0.427
Model 2				
Intercept	0.448	1.324	0.339	0.735
Benzene Exposure (HAPEM)	-0.432	0.208	-2.078	0.038
Median Age	0.137	0.313	0.439	0.661
Male Population %	-2.826	2.077	-1.360	0.174
African American Population %	0.015	0.018	0.834	0.404
Model 3				
Intercept	0.400	1.347	0.297	0.767
TCE Concentration (ASPEN)	-7.140	4.055	-1.761	0.078
Median Age	0.028	0.318	0.089	0.929
Male Population %	-3.103	2.138	-1.452	0.147
African American Population %	0.013	0.018	0.756	0.450
Model 4				
Intercept	0.508	1.359	0.374	0.708
TCE Exposure (HAPEM)	-10.000	5.331	-1.876	0.061
Median Age	0.029	0.317	0.091	0.928
Male Population %	-3.077	2.139	-1.439	0.150
African American Population %	0.014	0.018	0.760	0.447

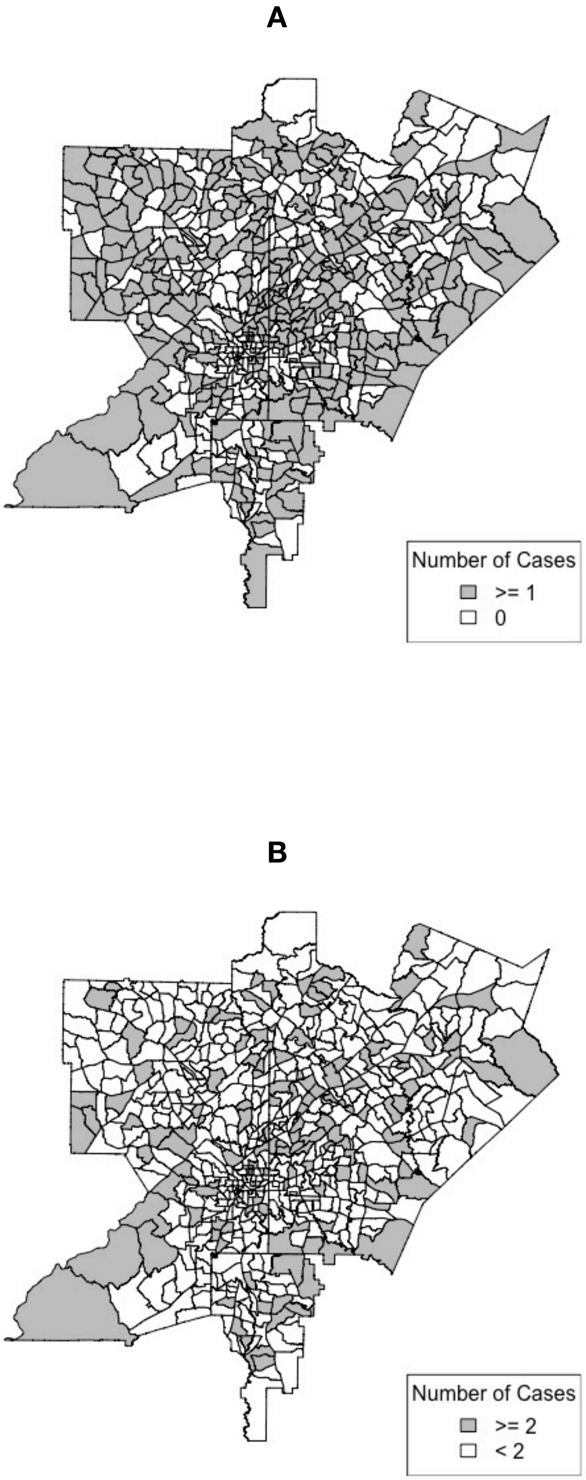

Figure 4. Image Illustration (A) census tracts with >= 1 or = 0 CTCL cases and (B) census tracts with >= 2 or < 2 CTCL cases in metropolitan Atlanta.

Table 4. Posterior Parameter Estimates Summaries of Spatial Poisson Generalized Linear Mixed Models

Variable	β Estimate	95% Credible Interval
Model 1		
Intercept	0.203	(-0.049, 0.447)
Benzene Concentration (ASPEN)	0.076	(-0.076, 0.222)
σ^2	0.005	(0.002, 0.034)
τ	0.456	(0.028, 0.939)
DIC = 1505.94		
Model 2		
Intercept	0.143	(-0.191, 0.475)
Benzene Exposure (HAPEM)	0.092	(-0.070, 0.252)
σ^2	0.007	(0.002, 0.034)
τ	0.448	(0.026, 0.936)
DIC = 1505.59		
Model 3		
Intercept	0.259	(0.052, 0.493)
TCE Concentration (ASPEN)	0.808	(-1.950, 3.003)
σ^2	0.007	(0.002, 0.0976)
τ	0.451	(0.030, 0.935)
DIC = 1506.32		
Model 4		
Intercept	0.247	(0.004, 0.512)
TCE Exposure (HAPEM)	1.088	(-2.501, 4.098)
σ^2	0.009	(0.002, 0.092)
τ	0.369	(0.014, 0.923)
DIC = 1506.39		

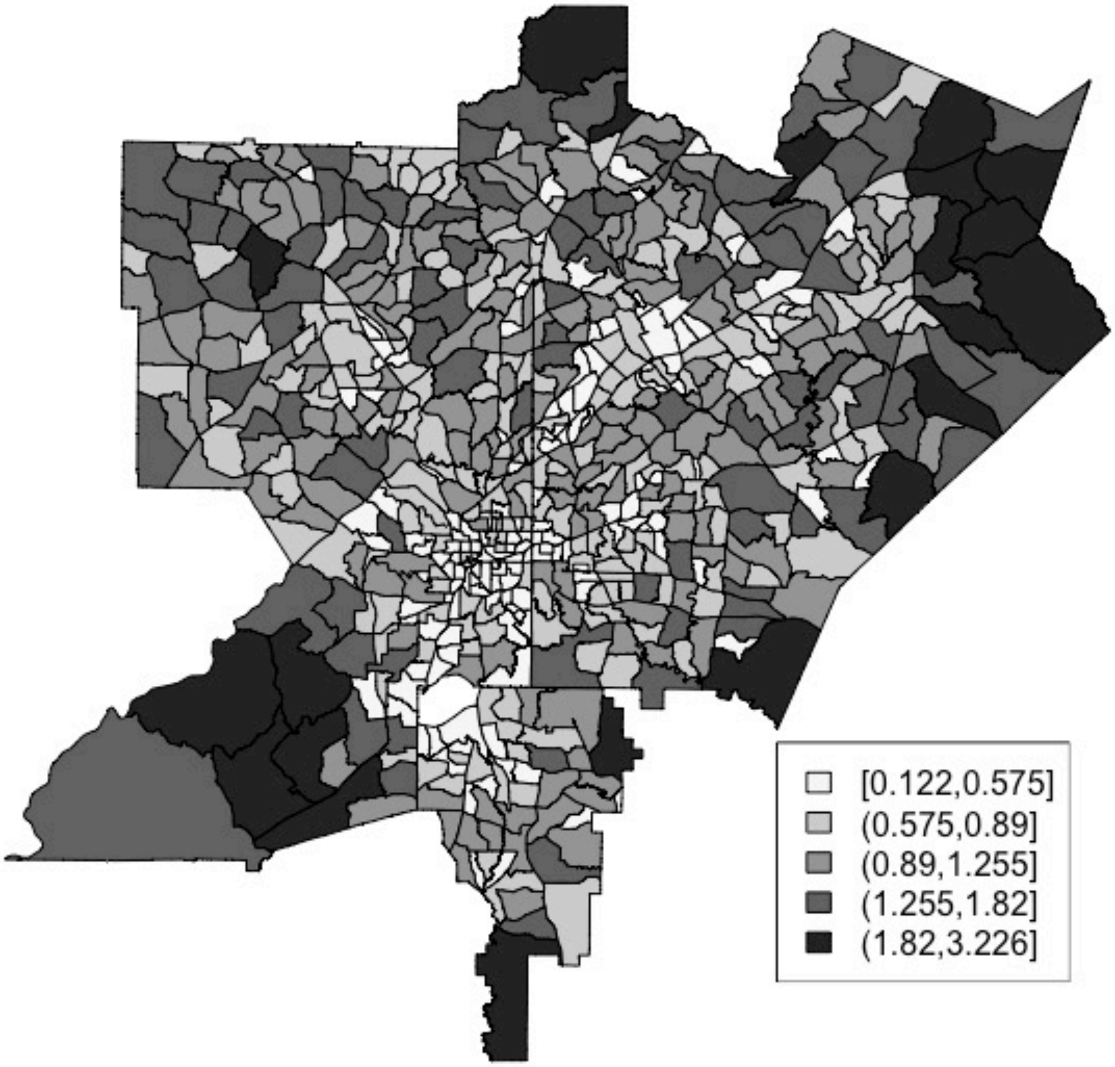

Figure 5. Bayesian smoothed SIR of cutaneous T-cell lymphoma generated from Spatial Poisson Model within a Bayesian setting.