Spatial Regression-Based Environmental Analysis in Infectious Disease Informatics

Daniel D. Zeng\(^1,2\), Ping Yan\(^1\), and Su Li\(^3\)

\(^1\) Department of Management Information Systems, the University of Arizona
\{zeng,pyan\}@email.arizona.edu
\(^2\) Institute of Automation, Chinese Academy of Sciences
\(^3\) Beijing Technology and Business University
lis@lreis.ac.cn

Abstract. Studying relationships between environmental factors and infectious diseases is an important topic in public health research. The existing studies have been focused on temporal correlations among environmental risks and infectious disease outbreaks. In this paper, we advocate the importance of spatial data analysis in infectious disease-related environmental analysis. Using data from the Beijing CDC, we have conducted spatial regression analysis to study correlation between Measles occurrences and the following environmental factors: population density and proximities to railways, roads, and water systems. We report some preliminary findings concerning significant spatial autocorrelation identified from our analysis.

Keywords: Environmental analysis, infectious disease informatics, spatial regression.

1 Introduction

The public health community has accumulated significant knowledge about how various infectious diseases emerge and spread. However, studies about the impact of environmental factors on infectious disease emergence and spreading remain sketchy. It has been well-argued that environmental factors such as temperature, humidity, proximity to water body, may have compounded impact on infectious disease transmission. But detailed models studying such kind of impact are yet to developed and evaluated due to a number of reasons such as the complex nature of the interaction between the environment and disease transmission, lack of comprehensive unbiased datasets, and lack of appropriate analysis tools or informatics environments.

The existing studies have started to focus on temporal correlations among environmental risks and infectious disease outbreaks. We argue that it is critical to add the spatial dimension in these studies. In this paper, we report a research effort aiming to analyze in a spatial-temporal context correlation between several environmental factors with Measles outbreaks in Beijing. This analysis is based on spatial regression, which has been a popular statistical tool in infectious disease informatics practice but has not yet been widely applied to study the impact of environment on infectious
diseases. As powerful software packages such as ArcGIS and GeoDA [1] are being increasingly adopted and environmental data samples are being collected more easily with GPS-enabled devices, we expect to see increasing interest in this type of spatial and spatial-temporal analysis framework. In Section 2, we briefly survey related work on environmental health risk analysis. The spatial regression analysis using the Beijing Measles dataset is presented in Section 3. We conclude in Section 4 with a summary and a brief discussion of future research.

2 Related Work

A few previous studies have aimed to identify correlation or association between one or a set of infectious diseases and certain environmental factors. We sample below four research topics. 1) Many respiratory diseases can be significantly impacted by climate, bearing obvious seasonal characteristics [2-5]. 2) Highly intensive human movement, typically associated with dense railway and road networks, has been shown to have significant impact on disease spreading [5]. 3) With an increasing population, the chances of stable transmission cycles between infected and susceptible persons are higher [6, 7]. 4) Vegetation coverage can reflect an area's environmental conditions, such as air quality [8].

Most existing work has focused on detecting temporal correlations between environmental factors and infectious disease cases. For instance, Wavelet coherency analysis and least squares regression analysis were used to identify statistical correlations between disease occurrences and climatic indices [2, 3]. However, these methods lack the ability to identify environmental factors potentially correlated with certain infectious diseases in a spatial context.

3 Measles and Environmental Factors: A Spatial Regression Analysis

While spatial data analysis has received increasing attention in many fields, including epidemiological studies, it remains underutilized in environmental analysis in the context of infectious disease informatics. One key reason lies with the difficulty of accessing environmental data and quantifying certain environmental factors. With accelerated adoption of technologies such as GIS, GPS, and remote sensing, environmental data are becoming available in a finer geographical granularity and it is very likely that environmental analysis in the public health context will become routine and lead to real-time actionable findings.

Our reported study is focused on a spatial regression analysis. In general, spatial regression first quantifies the spatial pattern through a pre-specified neighborhood structure and then examines relations between the attributes of interest and potential explanatory variables that can account for the observed spatial pattern. Spatial autocorrelation is automatically captured by this kind of analysis. In our work, we apply the spatial lag model to study the relationship between five selected spatial-geographical environmental factors and the Measles incidence rate in Beijing. In a spatial lag model, spatial autocorrelation is modeled by a linear relation between the response vector (y) and the associated