GIS-BASED, DATA-DRIVEN TECHNIQUES FOR SPATIAL ANALYSIS OF INFECTIOUS DISEASES AT THE REGIONAL, STATE, AND NATIONAL LEVELS

By

ABOLFAZL MOLLALO

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2019
To my family
ACKNOWLEDGMENTS

I want to thank my parents for their unconditional love and immeasurable supports in my life. Words cannot express how grateful I am for all you have done for me. I am extending my heartfelt thanks to my brother and lovely sister for their love, encouragement, and help in my life. You are the most important people in my life, and I dedicate this dissertation to you.

This work wouldn’t have been possible without the support and advice of a number of individuals. First and foremost, I would like to express my deepest and sincere gratitude to my mentor professor Gregory Glass for believing in me, giving me the freedom to bridge my interests in data science and medical geography and providing feedback over the years. I also would like to express my appreciation to my other committee members, Dr. Liang Mao, Dr. Jason Blackburn, and Dr. Parisa Rashidi for their insight, collaboration, feedback, and ideas throughout the entire PhD program.

I would like to give a special thank you to my friends especially those who live in Gainesville for being there for me and are and will be like a family to me.

Finally, many thanks go to the UF Geography department including my teachers, staff and all other people who helped me, directly or indirectly. Thanks for working hard to create a positive learning environment. It was a great pleasure and honor for me to be a part of this program.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>4</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>7</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>8</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>10</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>12</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>1  BACKGROUND</td>
<td>14</td>
</tr>
<tr>
<td>Application of New Technologies and Tools in Public Health</td>
<td>16</td>
</tr>
<tr>
<td>Disease Mapping</td>
<td>19</td>
</tr>
<tr>
<td>Global Clustering Techniques</td>
<td>21</td>
</tr>
<tr>
<td>Local Cluster Detection Techniques</td>
<td>23</td>
</tr>
<tr>
<td>Space-Time Clustering Techniques</td>
<td>26</td>
</tr>
<tr>
<td>Environment and Infectious Diseases</td>
<td>27</td>
</tr>
<tr>
<td>Disease Modeling</td>
<td>30</td>
</tr>
<tr>
<td>Knowledge-Driven Models</td>
<td>30</td>
</tr>
<tr>
<td>Data-Driven Models</td>
<td>31</td>
</tr>
<tr>
<td>Research Questions and Hypotheses</td>
<td>34</td>
</tr>
<tr>
<td>2  A 24-YEAR EXPLORATORY SPATIAL DATA ANALYSIS OF LYME DISEASE INCIDENCE RATE IN CONNECTICUT</td>
<td>41</td>
</tr>
<tr>
<td>Materials and Methods</td>
<td>43</td>
</tr>
<tr>
<td>Data Collection and Preparation</td>
<td>43</td>
</tr>
<tr>
<td>Global Clustering</td>
<td>44</td>
</tr>
<tr>
<td>Local Clustering</td>
<td>45</td>
</tr>
<tr>
<td>Spatial smoothing</td>
<td>45</td>
</tr>
<tr>
<td>Local moran’s</td>
<td>45</td>
</tr>
<tr>
<td>Spatial scan statistics</td>
<td>46</td>
</tr>
<tr>
<td>Accuracy Assessments</td>
<td>47</td>
</tr>
<tr>
<td>Results</td>
<td>48</td>
</tr>
<tr>
<td>Discussion</td>
<td>50</td>
</tr>
<tr>
<td>3  MACHINE LEARNING APPROACHES IN GIS-BASED ECOLOGICAL MODELING OF THE SAND FLY PHLEBOTOMUS PAPATASI, A VECTOR OF ZOONOTIC CUTANEOUS LEISHMANIASIS</td>
<td>60</td>
</tr>
<tr>
<td>Material and Methods</td>
<td>62</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1-1</td>
<td>Application of GIS in the study of infectious diseases</td>
</tr>
<tr>
<td>1-2</td>
<td>Application of RS in the study of infectious diseases</td>
</tr>
<tr>
<td>1-3</td>
<td>Application of GPS in the study of infectious diseases</td>
</tr>
<tr>
<td>1-4</td>
<td>Application of GIS, RS, Spatial statistics and machine learning in the study of infectious diseases (Lyme disease, leishmaniasis, tuberculosis)</td>
</tr>
<tr>
<td>2-1</td>
<td>Comparison of LISA and SaTScan clusters by confusion matrix and its derivatives statistics</td>
</tr>
<tr>
<td>2-2</td>
<td>Results of the global Moran statistic of LD incidence rate, Connecticut, 1991-2014</td>
</tr>
<tr>
<td>2-3</td>
<td>Characteristics of LD spatial clusters detected by SSS with 5% of the population at risk throughout Connecticut, USA for the period 1991-2014</td>
</tr>
<tr>
<td>3-1</td>
<td>Bioclimate variables used in this study</td>
</tr>
<tr>
<td>3-2</td>
<td>Pearson correlation coefficients between selected variables</td>
</tr>
<tr>
<td>4-1</td>
<td>Top 10 states with the largest number of hotspot counties ($p &lt; 0.10$) of smoothed tuberculosis (TB) incidence rate (STIR) in the continental US, 2006–2010</td>
</tr>
<tr>
<td>4-2</td>
<td>Pearson correlation analysis between selected variables for modelling STIR, continental US.</td>
</tr>
<tr>
<td>4-3</td>
<td>Results of linear regression (LR) model for modeling log (STIR), continental US.</td>
</tr>
<tr>
<td>4-4</td>
<td>Effects of environment and socio-economic factors on the log (STIR) using LR model.</td>
</tr>
<tr>
<td>4-5</td>
<td>Comparison of multi-layer perceptron (MLP; one and two hidden layers), and LR model’ performance for predicting log (STIR) in the continental US.</td>
</tr>
<tr>
<td>A-1</td>
<td>Statistical test for comparing AUCs of SVM, LR and RF classifiers:</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1-1</td>
<td>Contingency table of Knox method</td>
</tr>
<tr>
<td>1-2</td>
<td>Principle of linearly separable SVM using maximum margin</td>
</tr>
<tr>
<td>2-1</td>
<td>Geographic location of the Connecticut, its towns and approximate populations. The names on the map show the location of the towns which are mentioned in this paper.</td>
</tr>
<tr>
<td>2-2</td>
<td>Temporal trend of LD incidence throughout Connecticut from 1991 to 2014. Black dots show the incidence rate for each year and the blue line represents a scatter with smooth lines.</td>
</tr>
<tr>
<td>2-3</td>
<td>Locations of spatial clusters of LD incidence in Connecticut, USA based on the true-cluster definition of with LISA and SSS methods targeting 5% of the population at risk.</td>
</tr>
<tr>
<td>2-4</td>
<td>Comparison of LISA and spatial scan statistics with 5% and 50% of the population at risk of LD in Connecticut, USA for each period from 1991 to 2014. Sensitivity, specificity and overall accuracy expressed as per cent (%).</td>
</tr>
<tr>
<td>3-1</td>
<td>Geographic location of the study area and its counties, NE Iran. The map inset (B) illustrates the location of presence/absence sampling of <em>Ph. papatasi</em>.</td>
</tr>
<tr>
<td>3-2</td>
<td>Methodology flowchart used in this study</td>
</tr>
<tr>
<td>3-3</td>
<td>The ROC curves and AUCs for SVM, RF and LR classifiers</td>
</tr>
<tr>
<td>3-4</td>
<td>Comparison of the SVM, LR and RF classifiers. Overall accuracy, AUC, Kappa index, sensitivity, specificity, PPV and NPV expressed as %.</td>
</tr>
<tr>
<td>3-5</td>
<td>Habitat suitability map of <em>Ph. papatasi</em> in Golestan province, northeast Iran based on support vector machine (note: this map is based on 1km units).</td>
</tr>
<tr>
<td>4-1</td>
<td>Topological architecture of multi-layer perceptron neural network (MLPNN) used in this study.</td>
</tr>
<tr>
<td>4-2</td>
<td>Spatial distribution of training, cross-validation, and test data used for modeling log (STIR).</td>
</tr>
<tr>
<td>4-3</td>
<td>The frequency of TB cases (left) and the cumulative TB incidence rate (right) across the continental US (2006–2010).</td>
</tr>
</tbody>
</table>
4-4 Hotspot map for the STIR in the continental US identified by hotspot analysis (Getis–Ord Gi*) technique, 2006–2010. ........................................................... 104
4-5 The Normal P-P Plot of LR model. ................................................................. 105
4-6 Scatter plot of observed and predicted log (STIR) (by single hidden layer MLP model) for test data in the continental US. ................................................. 106
4-7 The contribution of input features on predicting log (STIR) according to sensitivity analysis of single hidden layer MLP. RMSE: Root mean square error. ................................................................................................................. 106
B-1 Habitat suitability map of *Ph. Papatasi* based on logistic regression classifier. 113
B-2 Habitat suitability map of *Ph. Papatasi* based on random forest classifier...... 114
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIDS</td>
<td>Acquired Immune Deficiency Syndrome</td>
</tr>
<tr>
<td>AMOEBA</td>
<td>A Multidirectional Optimal Ecotope-based Algorithm</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>CDC</td>
<td>Center for Disease Control and Prevention</td>
</tr>
<tr>
<td>CSR</td>
<td>Complete Spatial Randomness</td>
</tr>
<tr>
<td>CT</td>
<td>Connecticut</td>
</tr>
<tr>
<td>CTDPH</td>
<td>Connecticut Department of Public Health</td>
</tr>
<tr>
<td>DEM</td>
<td>Digital Elevation Model</td>
</tr>
<tr>
<td>DF</td>
<td>Dengue Fever</td>
</tr>
<tr>
<td>EBS</td>
<td>Empirical Bayes Smoothing</td>
</tr>
<tr>
<td>ENFA</td>
<td>Ecological Niche Factor Analysis</td>
</tr>
<tr>
<td>ENM</td>
<td>Ecological Niche Models</td>
</tr>
<tr>
<td>ESDA</td>
<td>Exploratory Spatial Data Analysis</td>
</tr>
<tr>
<td>EWS</td>
<td>Early Warning Systems</td>
</tr>
<tr>
<td>GAM</td>
<td>Geographic analysis machine</td>
</tr>
<tr>
<td>GARP</td>
<td>Genetic Algorithm for Rule-set Production</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GLM</td>
<td>General Linear Model</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>HIV</td>
<td>Human Immunodeficiency Virus</td>
</tr>
<tr>
<td>KDE</td>
<td>Kernel Density Estimation</td>
</tr>
<tr>
<td>LASSO</td>
<td>Least Absolute Shrinkage and Selection Operator</td>
</tr>
<tr>
<td>LD</td>
<td>Lyme Disease</td>
</tr>
<tr>
<td>LISA</td>
<td>Local Indicator Spatial Autocorrelation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>LR</td>
<td>Logistic Regression (in Chapter 3)</td>
</tr>
<tr>
<td>LR</td>
<td>Linear Regression (in Chapter 4)</td>
</tr>
<tr>
<td>LST</td>
<td>Land Surface Temperature</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>Maximum Entropy</td>
</tr>
<tr>
<td>MCDA</td>
<td>Multi Criteria Decision Analysis</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi-Layer Perceptron</td>
</tr>
<tr>
<td>MLT</td>
<td>Machine Learning Technique</td>
</tr>
<tr>
<td>MODIS</td>
<td>Moderate-resolution Imaging Spectroradiometer</td>
</tr>
<tr>
<td>NDVI</td>
<td>Normalized Difference Vegetation Index</td>
</tr>
<tr>
<td>NPV</td>
<td>Negative Predictive Value</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive Predictive Value</td>
</tr>
<tr>
<td>RF</td>
<td>Random Forest</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>RS</td>
<td>Remote Sensing</td>
</tr>
<tr>
<td>SA</td>
<td>Spatial Autocorrelation</td>
</tr>
<tr>
<td>SDM</td>
<td>Species Distribution Model</td>
</tr>
<tr>
<td>SSS</td>
<td>Spatial Scan Statistics</td>
</tr>
<tr>
<td>STIR</td>
<td>Smoothed Tuberculosis Incidence Rate</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TB</td>
<td>Tuberculosis</td>
</tr>
<tr>
<td>USGS</td>
<td>United States Geological Survey</td>
</tr>
<tr>
<td>VIF</td>
<td>Variance Inflation Factor</td>
</tr>
<tr>
<td>VL</td>
<td>Visceral Leishmaniasis</td>
</tr>
</tbody>
</table>
Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

GIS-BASED, DATA-DRIVEN TECHNIQUES FOR SPATIAL ANALYSIS OF INFECTIOUS DISEASES AT THE REGIONAL, STATE, AND NATIONAL LEVELS

By

Abolfazl Mollalo

August 2019

Chair: Gregory E. Glass
Major: Geography

The present research emphasized the spatial epidemiological aspects of three different infectious diseases: Lyme disease, zoonotic cutaneous leishmaniasis, and tuberculosis at the state, regional, and national levels, respectively. We combined relatively innovative methods/tools, for example, remote sensing, exploratory spatial data analyses, GIS and data science techniques. The findings of these studies can provide useful insight to health authorities on prioritizing resource allocation to risk-prone areas.

We examined changes in the spatial distribution of significant spatial clusters of Lyme disease incidence rates at the town level from 1991 to 2014 as an approach for targeted interventions. Local clustering was measured using a local indicator of spatial autocorrelation (LISA). Elliptic spatial scan statistics (SSS) in different shapes and directions were also performed in SaTScan. The accuracy of these two cluster detection methods was assessed and compared for sensitivity, specificity, and overall accuracy.

In another case study, we compared several approaches to model the spatial distribution of *Phlebotomus papatasi*, the primary vector of zoonotic cutaneous leishmaniasis, in an endemic region of the disease in Golestan province, northeast of
Iran. We gathered and prepared data on related environmental factors including topography, weather variables, distance to main rivers and remotely sensed data such as normalized difference vegetation cover and land surface temperature (LST) in a GIS framework. Applicability of three classifiers: (vanilla) logistic regression, random forest and support vector machine (SVM) were compared for predicting presence/absence of the vector. Predictive performances were compared using an independent dataset to generate area under the ROC curve (AUC) and Kappa statistics.

Despite the usefulness of artificial neural networks (ANNs) in the study of various complex problems, ANNs have not been applied for modeling the geographic distribution of tuberculosis (TB) in the US. Likewise, ecological level researches on TB incidence rate at the national level are inadequate for epidemiologic inferences. We collected 278 exploratory variables including environmental and a broad range of socio-economic features for modeling the disease across the continental US. We investigated the applicability of multilayer perceptron (MLP) ANN for predicting the disease incidence.
"Person", "place", and "time" are known as three key components in every descriptive epidemiologic research. Application of place (geography) in the study of diseases backs to more than 2000 years ago. Hippocrates (460-377 BC) in his essay entitled “On Airs, Water, and Places” proposed a theory suggesting that environmental factors can have a significant influence on the occurrence of diseases. In the late seventeenth century, Filippo Arrieta (1694) utilized maps to investigate the outbreak of plague in Bari, Italy. At the end of the eighteenth century, Valentine Seaman (1795) mapped fatal cases of yellow fever in New York city to detect a possible association between the location of yellow fever cases and waste sites. Perhaps the most famous application of place (map) in the study of disease distribution is the work of John Snow in mapping cholera death cases in the 1850s in London. Using the distribution map of (cholera) deaths, he could trace and detect the source of the outbreak to a contaminated public water pump.

Before 1950, there were very few studies that incorporated spatial components in health researches. During the 1950s, Jacques May (1958) proposed a new concept titled “Disease-ecology” or “geographic pathology” which surprisingly is still valid in epidemiological studies. “Disease-ecology” investigates the interaction of people with the environment. His viewpoint was "process-oriented" and applied natural scientific explanations: "Once the person, disease, and place are known, we may be able to understand why someone is afflicted and someone else is not". The principal objective of his approach was to better comprehend the dynamics of disease which varies per weather condition, mineral particles in water, vegetation cover and other influencing...
factors. May divided an environmental health problem into three main categories: 1) inorganic environment such as weather/climate variables which can reflect many features of traditional environmental philosophy, 2) organic environment which can reflect disease pattern because of animals/plants activities, and 3) socio-economic factors which can explain the disease associated with the behavior and culture of human.

By the end of the 1960s, remarkable changes in the concepts of May's approach began to emerge within the literature. The variations were partially associated with the raised attention of geographers to the disease causation through visual interpretation. This was in contrasts with May's approach that studies disease occurrence as understanding the processes of disease-ecology. In this decade, the World Atlas of Disease was issued under the supervision of May. The atlas represented a significant milestone in medical geography in this century. It contained both small and relatively large scale maps of disease morbidity which were shaded by colors or plain symbols. Mapping continued as one of the essential fields where geographer contributed substantially and is widely used in spatial epidemiological researches as the exploratory tool for generating hypotheses useful in health care planning.

Before 2000, medical geography mainly focused on clustering and cluster detection analysis. Along with the advances in technology, the field of medical geography has progressively evolved. Several sophisticated spatial tools such as geographic information system (GIS), remote sensing (RS), and novel spatial statistical analysis such as machine learning algorithms played an important role.
Although historical researches in medical geography rarely considered geographical aspects, recent (decade) studies increasingly include "spatial" component in epidemiological inferences \(^{14}\). This increment might be due to the broader availability of spatial data and quality of spatial health data than in the past because of the advances in technology \(^{15}\). Moreover, increased availability and sharing more accurate environmental, socio-economic, cultural, and biological data has fueled researches in medical geography. The availability of user-friendly software packages designed to expedite statistical analysis is another explanation, particularly for non-expert users. Therefore, medical geographers now have more powerful tools and high-quality and update data at hands, and thus, spatial epidemiological studies are expected to continue to play a significant role in the epidemiology of diseases. In the later sections of Chapter 1 we will briefly demonstrate the application of a wide range of new tools and technologies applied in medical geography, such as GIS, RS, GPS, and spatial statistics and models.

**Application of New Technologies and Tools in Public Health**

Over the last two decades, mapping techniques, spatial analyses of disease patterns, and risk modeling of diseases have substantially improved with the advent of GIS \(^ {16}\). GIS has provided unprecedented opportunities to collect, organize and integrate geospatial data from various sources and in different formats. It has enabled conducting very complex analysis in public health in conjunction with population and environment that in earlier researches were very hard or impossible to be studied. This tool has equipped medical geographers to find answers for overly complicated questions more efficiently \(^ {17}\). The application of GIS in recent decades have been progressively increased. In Table 1-1, we have summarized some of the widely-used applications of
GIS in the studies of infectious diseases. An overview of GIS and its application in the studies of infectious diseases has been published by Nykiforuk et al. (2011)\(^{18}\).

According to Campbell and Wynne (2011) remote sensing refers to obtain some characteristics of an object without making physical contact with it (i.e., through an aerial sensor)\(^{19}\). It allows collecting a vast amount of data in a short time and high accuracy which facilitates disease surveillance and control. An overview of RS and its application in public health has been reviewed by Hay et al. (1997)\(^{20}\). High-resolution satellite images can help medical geographers to investigate geographic variations of diseases at a small-area scale. Moreover, with the help of RS, some products such as normalized difference vegetation cover (NDVI) which reflects vegetation cover on the ground, land surface temperature (LST), soil property, fog, and land cover can be derived and incorporated as candidate factors in modeling. Also, several biophysical features can be quantified by this technology including, biomass, chlorophyll absorption, surface texture, and moisture content\(^{21}\). In public health, RS data can be used for generating habitat suitability maps for species, predicting vector or host population or presence/absence and identifying vector or host habitat. Table 1-2 summarizes some applications of RS in the study of infectious diseases.

Another technology that has made a significant role in collecting geospatial data is the global positioning system (GPS). This technology has made collecting massive amounts of spatial and attributes data in surveillance of infectious disease faster, easier and with an affordable cost\(^{22}\). Table 1-3 summarizes some applications of GPS in the study of diseases.
Numerous researches have integrated the above technologies in various types of infectious diseases researches. These studies include mapping disease prevalence, predicting habitat suitability map of vectors, identifying risk-prone areas of infections, etc. As an early study that integrated the above techniques, Glass et al. (1990) used LANDSAT TM satellite images to derive land cover/land use in developing an environmental geodatabase. These remotely sensed products and GIS were combined to identify risk factors and consequently high-risk areas of Lyme disease in Maryland. Mollalo et al. (2018) captured the location of collected sandflies of leishmaniasis in an endemic province in Iran using handheld GPS. They then used several environmental factors including NDVI in a GIS framework. They could predict the presence/absence of the sandfly with an accuracy of 90% for a hold-out dataset.

Spatial statistics in epidemiology are statistical tools that are used to mainly map, explain, and predict the spatial distribution of health problems. In general, the four main types of spatial statistics frequently used in the studies of infectious diseases are: 1) mapping (visualization) disease count or rate 2) clustering and cluster detection analysis (exploratory spatial data analysis) 3) Correlation analysis and 4) modeling (explain or predict disease rates/counts). The very first step in spatial statistics is linking disease counts/rates data to their corresponding locations and visualize them as maps. The mapping step, however, may reveal some useful information, can also conceal the actual affected areas. Therefore, some statistical analyses are required to evaluate the observed pattern of disease occurrences statistically. Spatial clustering and cluster detection techniques can consider locations and attributes and can be used to address this issue by identifying hotspot(s) and cold spot(s) of infectious diseases. Identifying
hotspots are very helpful in generating hypothesis and can provide useful information for further analysis. Finally, using a proper model the possible relationship between disease frequency/rate and several explanatory variables (such as environmental and socio-economic factors) can be explained. A properly developed model can also be utilized to predict (spatially/temporally) disease distribution. Detailed information about each step of the application of spatial statistics in the study of infectious diseases is provided as follows.

**Disease Mapping**

The topic of disease mapping has a very long history and is an ongoing topic among health researchers and organizations to produce more accurate atlas/maps of morbidity or mortality rates of various infectious diseases. An early example of this topic is mapping the spatial distribution of cancer mortality in England and Whales by Stocks (1936). Maps of infectious diseases can illustrate a summary of the complicated status of spatial distributions of infectious diseases. The disease map can disclose spatial patterns in the data that are not readily detectable in a tabular format. Some applications of disease mapping are basic descriptions of the spatial disease distribution, hypothesis generation of possible relations between factors and health outcome and risk-prone areas representations useful for budget and resource allocations.

Choropleth maps also known as shaded or thematic maps are widely used to illustrate mortality and morbidity rates of infectious diseases. These maps are particularly useful to identify health disparities, generating hypothesis and representing variations of infectious disease counts/rates over time. However, several issues regarding choropleth maps exist. This representation can be non-informative or to some
extent misleading regarding the choice of coloring; classification techniques and the choice of cut-off values. Moreover, shaded mapping for areas with a small population can lead to high variations of estimated risks compared to the large populations.

In the situation of sparse or highly clustered data, Bayesian models can be a suitable alternative. It provides a more robust solution as it can avoid unstable estimates of rates/risks. These models are increasingly applied in small area estimations and disease mapping by spatial epidemiologists. They are especially useful for rare mapping diseases. For instance, empirical Bayesian estimation, a type of Bayesian models, can either shrink unstable risks/rates toward the local mean by obtaining information from neighboring areas or weights the unstable risks/rates toward global average from all areas. It is evident that the areas with higher rates are less smoothed than areas with lower rates. Thus, smoothing produces more stable estimations of disease risks/rates. However, it should be noted that global mean smoothing can produce over-smoothed rates/risks which can mask informative local variations. An exciting feature of Bayesian methods is that they can help to remove random components resulting from correlated and unmeasured factors from the maps. Another benefit of this method is that it accounts for uncertainty measures associated with the relative risks with confidence intervals. It is evident that including uncertainty as confidence interval are more reliable for decision makers. Sharmin et al. (2016) used a Bayesian generalized linear model to adjust for underreporting cases of dengue by incorporating climate factors in Dhaka, Bangladesh. Randremanana et al. (2010) used the integration of a Bayesian approach and a generalized linear mixed model to analyze the geospatial distribution of TB in an endemic area, Madagascar. They
generated (smoothed) heat maps of TB to compare the association of TB rate with the nationwide TB indicators. The more detailed information regarding Bayesian approach can be found in Lawson (2013).

**Global Clustering Techniques**

As already mentioned, although disease mapping can help obtain some useful information about the spatial pattern of infectious diseases, they can be misleading and unable to evaluate the significance of the pattern statistically. Spatial dependence or spatial autocorrelation (SA) is a fundamental concept that is applied to evaluate the pattern from the statistical viewpoint. SA indicates the association of a feature with itself located nearby. SA proposed by Tobler (1979) as the first law of geography: "everything is related to everything else but near things are more alike than the distant things." SA can be positive (i.e., similar attributes are closer to each other), negative (i.e., different variables are closer to each other), or none (i.e., random distribution). SA is an essential concept in medical geography and should be investigated in analyzing spatial data. There are several spatial statistic tools to evaluate the presence of SA (overall pattern) known as global clustering techniques. The null hypothesis significance (i.e., random distribution) is tested against the alternative hypothesis (i.e., clustered distribution).

One of the most straightforward measures of spatial autocorrelation, when the variable of interest is categorical, is join counts statistics. The null hypothesis of join counts states that the distribution is random. In this technique, a binary attribute is classified into black and white colors, and a join (or connections between the zones) is named as either WW (0,0), BB (1,1), or BW (1,0). By counting the number of WW, BB and BW, type of SA can be determined. The overall distribution is clustered (i.e.,
positive SA) if the number of BW joins is significantly lower than expected by chance \(^{44}\). The distribution is dispersed (i.e., negative SA) if the number of BW is significantly higher than expected by chance. The distribution is random if the number of BW joins approximately the same as what would be expected by chance. The test of significance evaluating the BW statistic as a standard deviate is as follows \(^{45}\):

\[
Z(BW) = \frac{BW - E(BW)}{\sqrt{\sigma_{BW}^2}}
\]  

(1-1)

In this formula, \(Z(BW)\) indicates the magnitude of SA. Join counts statistics have been applied in the disease studies. Gilbert et al. (1994) used join counts statistics to evaluate the spatial distribution of canker disease of trees in Panama. They counted the number of cankered-cankerded tree joins (BB), healthy-healthy tree joins (WW), and cankered -healthy tree joins (BW) and compared them with the expected value \(^{46}\).

Global Moran’s statistic is the most widely used measure of SA in continuous variables which relies on both the variable's location and attribute \(^{47}\). This index ranges from -1 to +1. In general, the values close to the value of -1 indicate a dispersion, the values close to the value of +1 shows a clustering, while the values close to the value of 0 indicate a random pattern \(^{48}\). The null hypothesis states that the distribution is random. This index is calculated as \(^{47}\):

\[
I = \frac{N \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \sum_{i=1}^{N} (x_i - \bar{x})^2}
\]  

(1-2)

Where \(N\) is the count of spatial units \(i, j\); \(x\) is the variable of interest (e.g., malaria incidence rate); \(w_{ij}\) is the connection between units \(i\) and \(j\). The spatial weight matrix can be constructed in two main ways \(^{49}\): 1) contiguity-based neighbors: sharing a border (rook) or sharing a border or point (queen) 2) distance-based neighbors: k-nearest
neighbors or buffer (threshold distance). Many studies utilized this index to investigate the overall pattern of infectious diseases. For instance, Naish et al. (2014) used global Moran’s statistic to assess the presence of SA in the study of dengue incidence rates in Queensland, Canada and found a positive SA.

Geary’s C is another common method for measuring SA which has similar calculations to Moran’s I. The value of this statistic ranges from 0 to 2 where the values close to 0 shows a positive SA, the value close to +2 shows a negative SA, while the index value close to +1 shows a random distribution. Therefore, this index is inversely related to the global Moran’s statistic. The null hypothesis expresses that there is no SA (i.e., C=1). Using the same notation as for global Moran’s I, the Geary’s C statistic is computed as follow:

\[
C = \frac{(N - 1) \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - x_j)^2}{2 \left( \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \right) \sum_{i=1}^{N} (x_i - \bar{x})^2}
\]

Compared to Moran’s I, which is a global indicator of SA, Geary’s C is highly sensitive to the difference in neighbors. Simões et al. (2004) applied Geary’s C to investigate the distribution of paracoccidioidomycosis, a fungal infection, in southern Brazil. Their finding rejected the null hypothesis of SA and showed a significant spatial autocorrelation. Other global clustering techniques are widely used in assessing overall pattern are average nearest neighbors, K-function (used in point pattern analysis), general C and Cuzick-Edwards.

Local Cluster Detection Techniques

In global clustering techniques, only a single SA value is assigned to the entire pattern. Thus, global techniques are unable to identify the location of clusters. While, in local measures of clustering (i.e., cluster detection techniques), an SA is calculated for
each areal unit and its significance is evaluated with statistical tests. Also, the techniques can identify hotspot(s), coldspot(s) or outlier(s) in a pattern.

Perhaps one of the simplest ways to visualize the location of clusters is by using density function such as kernel density estimation (KDE). In the KDE method, a weighting kernel function is fitted over each point or line. The weight decreases as we move away from the point/line feature. This method is criticized for the subjective choice of search radius (bandwidth) and not providing statistical evaluation of the results. As an example, Aikembayev et al. (2010) used KDE in a GIS environment to identify the areas of outbreak concentration of Bacillus anthracis by livestock species, Kazakhstan.

The Geographic analysis machine (GAM) is an automated cluster detection technique developed by Openshaw et al. (1987) for point data. The method developed to identify spatial clusters of childhood leukemia incidence. In this method, a two-dimensional grid is superimposed over the study area. Then, a series of different circles with various sizes is generated and scans the whole study area. The observed intensity of events within each circle is compared with a threshold based on Monte Carlo simulation. If the observed intensity exceeds the threshold, it draws a circle on the map. The final output of GAM is the map of significant circles. A significant disadvantage of this technique is that no conclusion can be drawn about the significant level of clusters.

Local indicator spatial autocorrelation (LISA) is perhaps the most widely used cluster detection technique for identifying clusters of various infectious diseases. The statistic is based on the decomposition of global Moran’s I for each areal unit. LISA
coefficients for each unit $i$ can be calculated as the deviation of values from the mean of neighbors $63$:

$$L_i = \frac{(y_i - \bar{y}) \sum_{j=1, j \neq i}^{n} w_{ij} (y_j - \bar{y})}{s_i^2} \quad (1-4)$$

$$s_i^2 = \frac{\sum_{j=1, j \neq i}^{n} (y_i - \bar{y})^2}{n - 1} \quad (1-5)$$

One of the advantages of this statistic compared to some other cluster detection techniques is its ability to identify outliers (i.e., areas with high values of an attribute are surrounded with low values and vice versa). Using LISA, Hassarangsee et al. (2015) detected hotspots of tuberculosis incidence in several districts in Thailand $64$. Szonyi et al. (2015) identified several outliers of Lyme disease in western Texas by applying this technique $65$.

Getis-Ord $G_i^*$ statistic $66$ is one of the most popular hotspot detection methods. It is a polygon-based method; however, point events can be aggregated by superimposing a grid. The statistic is calculated as follows $66$:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\left[ n \sum_{j=1}^{n} w_{ij}^2 - \left( \sum_{j=1}^{n} w_{ij} \right)^2 \right] / (n - 1)}} \quad (1-6)$$

$$S = \sqrt{\frac{\sum_{j=1, j \neq i}^{n} (x_j - \bar{x})^2}{n - 1} - \bar{x}^2} \quad (1-7)$$

The statistic can identify both hotspots and cold spots. The $G_i^*$ is usually standardized to z-scores (of normal distribution) so that a large (positive) Z-score is corresponded to a significant hotspot (i.e. $Z$-score $>1.96$ standard deviation); a large negative value of z-score is associated with a location as coolspot ($Z$-score $<-1.96$ standard deviation); and
value close to zero shows a location that is neither hotspots nor coolspot. Sun et al. (2017) utilized the Getis-Ord $G_i^*$ statistic to identify hotspots of the dengue fever epidemic in Sri Lanka.

One of the novel tools to determine statistically significant elevated disease rate is spatial scan statistic (SSS). The tool developed by Kulldorff (1997) for point and polygon units in SaTScan software. The SSS can also find significant clusters of individual health events (i.e., in case-control studies) using the Bernoulli model. In this method, circular windows with varying radius sizes or ellipsoidal windows in different shapes and directions are drawn around each point or centroid of polygon data. The size of the window is changed from 0 to the specified cluster size. For each window, a maximum likelihood test is applied to compare the risk within the window with the outside. Monte Carlo simulation is used to test for a significance level of clusters. The cluster with maximum likelihood is considered as primary clusters and the rest of the cluster(s) as a secondary cluster(s). A significant advantage of this method over the most cluster detection techniques is that it can take confounding into account by adjusting for variables.

**Space-Time Clustering Techniques**

Space-time analysis of disease clusters is another crucial topic in epidemiological studies which is usually overlooked. Space-time analysis is an ongoing research topic which provides useful insight for public health decision-makers into how an infectious disease propagates over the landscape. It can help to find direction and periodic patterns helpful for predictions. Some space-time clustering techniques have been proposed. In the following section, we have provided a general description of two widely used space-time cluster detection techniques.
Knox method (1964) is a primary method that is used for examining space-time interactions (clusters) than expected by chance. Space-time interactions exist when many of cases that are close in time are also close in space. Therefore, in this method for each pair of cases, the interval (i.e., time and space interval) is examined. A 2*2 contingency table is constructed with cross-comparison proximity in space and time (Figure 1-1). The number of pairs for each situation is counted, and the actual number of pairs that fall into each category is compared with the expected number (i.e., cross products of rows and column totals). Then a final Chi-square test with \((n-1)\times(r-1)\) degree of freedom measures a significant level of difference. The CrimeStat software (https://www.icpsr.umich.edu/CrimeStat/) allows users to apply Knox test. This approach has been criticized due to the subjective cut-off value for closeness in space and time.

Similar to the pure spatial scan statistic, in space-time approach a window scans the whole study region, and the risk within the window is compared with the risk outside of the window. However, instead of a 2-dimension circle/ellipse window, a cylindrical window with the base of a circle/ellipse is used. The base of the cylinder depicts space while the height corresponds to time. Also, p-value for each cylinder is computed using Monte Carlo simulation. Compared to the Knox test, this method accounts for possible geographic/temporal population heterogeneity (i.e., different population growth in different regions).

**Environment and Infectious Diseases**

The occurrence of infectious diseases represents a serious public health challenge. Infectious diseases annually kill over 15 million people (>25% of the total death), worldwide. The number is progressively growing, and the geographic
distribution is expanding to non-endemic areas. One strategy to fight against the infectious diseases is to monitor and control the causative agents such as a virus, bacteria of the disease. In this regard, identifying geographic patterns of infection and underlying risk factors such as environmental and socio-economic factors is helpful. Layers of environmental data can be obtained from various sources such as landcover data from remote sensing, weather data from weather stations or WorldClim, land use data from government records, soil maps from the department of agriculture, field investigation, etc. The data are provided in different formats like excel, grid, shapefile, text, etc., then using spatial analysis tools like GIS they are integrated and combined for further analysis. Another strategy for control and monitor infectious diseases is the surveillance of risk factors for diseased patients or deaths. An explanation for this approach is because of a strong correlation between underlying risk factors and the distribution of human cases (rates). For instance, to control Lyme disease in human, monitoring and controlling *Ixodes scapularis* population is applied as a proxy for disease morbidity.

The diseases (or parasites/vectors) highly dependent on the local and global environment and consequently can influence disease prevalence in a community. Environmental determinants (such as weather/climate, vegetation) can provide favorable conditions for breeding, feeding, resting sites of certain vector-borne diseases. However, it should be noted that socio-economic factors such as race, gender, poverty, lifestyle (such as drinking alcohols or smoking) can play an essential role in the occurrence of infectious disease such as Tuberculosis. Climate change is an essential determining factor in explaining variations of most infectious diseases.
especially vector-borne diseases \(^7\). It can lead to migration of pathogens (causative agents) and animals to new areas and consequently expansion of the diseases to uninfected areas \(^7\). The relationship between climate condition and infectious diseases have levels of complexity. For instance, weather and climate factors can have different influences on the vector or pathogen abundance of vector-borne diseases. The differences are mainly because of differences in the life cycle or life stage of vectors. For instance, four stages of tick-borne diseases' life cycle (i.e., eggs, larvae, nymph, adult) last two years of being completed. While the life stage of mosquitos (i.e., eggs, multiple larvae, pupae, adults) only take a few weeks to a few months to be completed. This difference indicates that mosquito's abundance responds to short-duration variations in weather or climate, while tick's population reacts to longer-term changes in weather condition and with little inter-variations \(^8\). Also, extreme high/low temperature prevents both mosquitoes and ticks host for blood meal seeking \(^8\). Under other climate conditions, meteorological factors can have relatively complex effects on mortality rates. In harsh weather conditions, all the four stages of tick's life remain secured (because ticks spend most of their life time under layers of soil), while the only dipteran can seek refuges. Thus, the direct consequences of temperature have fewer impacts on the survival of the tick's population compared with mosquitoes' abundance. As almost all development stages of mosquitoes are affected by the presence of stagnant water, the reproduction rates depend on the amount of rainfall \(^8\). However, heavy rainfall can flush away larval and eggs. Therefore, reproduction rates of ticks cannot be influenced by weather changes apart from long-lasting impacts on host densities \(^8\).
Disease Modeling

Although correlation analysis can indicate associations between factors and disease morbidity/mortality, the associations do not indicate causality. Disease modeling is an advanced level of spatial analysis in medical geography. It includes testing generated hypothesis obtained from cluster detection or observations. Disease modeling can involve the integration of previously mentioned tools (i.e., GIS, RS, and GPS) with statistical and epidemiological analysis. Modeling disease occurrence helps medical geographers to describe morbidity/mortality rates and identify underlying factors. It can also be used for predictions: classification (such as presence/absence of vectors) or regression (such as predicting disease rate/count). The prediction can be spatial (i.e., for locations with unknown mortality/morbidity) or spatiotemporal (i.e., future (near/far) status of disease) or for other areas (projection). Space-time predictions can be used for early detection of infection or outbreak. Forecasting disease occurrence can provide valuable guidelines for public health decision makers for cost-effective planning and targeted interventions of future outbreaks.

Knowledge-Driven Models

In general, disease modeling techniques can be classified into two categories: data-driven and knowledge-driven models. According to Pfeiffer, data-driven models are mainly based on statistical analysis to quantify the relationship between disease morbidity/mortality and underlying factors. While knowledge-driven models use knowledge of experts as evidence. These models depend heavily on environmental layers and are widely used in spatial epidemiology for mapping suitable areas for disease/vector transmission. Some techniques such as multi-criteria decision analysis (MCDA) can use experts' knowledge to provide a habitat suitability map by converting...
knowledge to decision rules. Analytical hierarchy process (AHP) proposed by Saaty (1990) is one of the most common ways to define the weights of factors. By overlying (combining) factors based on their corresponding weights, suitability for each pixel in the study area can be computed. Moreover, uncertainty associated with variables, the relationship between independent factors and the dependent variable, and the degree of risk can be modeled by incorporating fuzzy logic. One of the significant limitations of the knowledge-driven methods is that incorporating risk factors, weights and type of membership functions of fuzzy rules are subjective. Knowledge-driven models have been applied to modeling several infectious diseases. Clements et al. (2006) used them in modeling rift-valley fever in Africa, Mollalo and Khodabandehloo (2016) applied them in the study of zoonotic cutaneous leishmaniasis in an endemic area in Iran; and Rakotomanana et al. (2007) applied them in modeling malaria vector control in Madagascar.

**Data-Driven Models**

Data-driven models can generally be classified into two main categories: 1) presence-absence models 2) presence-only models. Here, we first describe the characteristics of several presence-absence models.

Classification and regression tree (CART) is a decision-tree based method that can be applied for solving both classification and regression problems in spatial epidemiology. CART is particularly useful in working with noisy data (such as datasets with missing values or highly skewed data) and extensive and complex datasets. CART is flexible in modeling a variety of response variables such as categorical, survival and continuous data. Another advantage of CART is that the results obtained from the model are easily interpretable. However, it should be acknowledged that the model
does not have a statistical test for the significance level of variables and is subject to overfitting\textsuperscript{93}. Other features of CART have been discussed in detail in Franklin (2009)\textsuperscript{94}.

One of the significant problems of CART is attributed to the low predictive ability and high variations in results. To address this drawback, bagging, boosting and random forest can be useful. These algorithms generate numerous decision trees by random selection of variables and random selection of samples with replacement (bootstrap sampling). In bagging, in each repetition of sampling, one-third of data is used for testing the model performance (i.e., out-of-bag samples). Boosting is like bagging: in bagging, samples have equal weights, while in boosting samples are weighted. Random forest (RF) is a type of bagging\textsuperscript{95-97}.

Support vector machine (SVM) is another data-driven model that is used for binary classification problems. In SVM, a dataset of high dimensional points viewed as vectors \( \{x_i \in R^d: i=1,\ldots,n\} \), \( d>1 \), where each point belongs to one of two classes defined by \( \{y_i \in \{0,1\}: i=1,\ldots,n\} \). Here \( y_i \) corresponds to the presence/absence of points. If we assume these points to be linearly separable (i.e., can be separated via a linear boundary), the goal of SVM is to find the d-dimensional hyperplane maximizing the margin (i.e., the distance between the closest points or support vectors) as illustrated in Figure 1-2. More detailed explanations can be found in Noble (2006)\textsuperscript{98}.

Artificial neural networks (ANNs) is another data-driven model that works with presence-absence data. More detailed information of ANN is provided in Chapter 4 of this dissertation.
Presence-only model in data-driven techniques is a relatively new concept. Traditional statistical models such as the generalized linear model (GLM), however, have a well-established algorithm, require presence and absence of the diseases (or vector) data. However, collecting absence data can be very expensive and may not cover most of the study region. Species distribution model (SDMs) can be used for modeling presence-only data and a vast area. They are highly applied in conservation and ecology for modeling geographic distribution of species (i.e., disease macro-organism). Another advantage of the SDM model is that they can be used for projection (i.e., extrapolation beyond the study region) with decent accuracy.

Ecological niche models (ENMs) use background or pseudo-absence data instead of real absence data to predict niche (i.e., conditions that enables species to maintain population without immigration). These models often are developed at coarse spatial scales. Most important ENMs are ecological niche factor analysis (ENFA), genetic algorithm for rule-set production (GARP) and maximum entropy (MaxEnt).

ENFA assumes that species occurrence is not random and is clustered in a small section of the study region. ENFA extracts marginality and specialization. Marginality would be close to 1 when variables contribute to the life of species in a specific region, and the value would be close to 0 when species could be found everywhere. Specialization shows species tolerance. Thus, it can summarize all the variables into a few factors. As an example of the application of ENFA in infectious diseases, Ayala et al. (2009) used ENFA to provide habitat suitability map of malaria species in Cameroon.
GARP uses four rules to select or reject, evaluate, and test rules from the final model. The set of rules with the highest predictive performances are selected and used for habitat suitability. GARP can obtain high accuracy even with a small sample size. It can also be used for projection or extrapolation. More information about GARP has been provided in Stockwell and Noble (1992)\textsuperscript{103} and Stockwell and Peterson (1999)\textsuperscript{104}.

Maximum Entropy (MaxEnt) is another presence-only ecological niche model (ENM) that is widely used in ecology and geography. In this model, the contribution of each explanatory variable to the final model is computed by removing each variable and examining the AUC of the model. The MaxEnt can be used even with minimal sample size (n<100) and is easily interpretable\textsuperscript{105}. MaxEnt has been successfully used in predicting the distribution of several infectious diseases (vectors) such as leishmaniasis in France\textsuperscript{106} and the West Nile virus in Iowa\textsuperscript{107}. More detailed information presented in Phillips et al. (2006)\textsuperscript{108}.

Although, many published pieces of research have examined the relationships between environmental conditions and infectious diseases, novel techniques such as machine learning techniques have been underutilized in spatial epidemiology. In this dissertation, we examine three different infectious diseases (i.e., Lyme disease, zoonotic cutaneous leishmaniasis, tuberculosis) from a spatial perspective and at three different levels. Table 1-4 summarizes some applications of GIS, RS, and machine learning techniques in the study of the mentioned infectious diseases.

**Research Questions and Hypotheses**

The main research questions and hypotheses explored in this dissertation are as follows:
In Chapter 2, we examine changes in the spatial distribution of significant spatial clusters of Lyme disease (LD) incidence rates in Connecticut at the town level from 1991 to 2014 as an approach for targeted interventions. Spatial distribution of LD in CT has been rarely investigated. Thus, we intend to help policymakers for targeted interventions by prioritizing towns with high incidence rate. The primary objective of the first paper is to integrate GIS and ESDA to better describe changes in the spatial pattern of LD, supposing the passive cases of LD serve as a spatially random sample of the infection in the state. Our specific research questions in Chapter 2 include:

- What is the spatial distribution (random/ dispersed/ clustered) of LD incidence over the past 24 years?
- If not-random, where are the locations of the clusters/hotspots in CT?
- Can we avoid or minimize the effects of edges in the spatial scan statistic technique?
- Which cluster detection technique (LISA/ spatial scan statistic) is more sensitive, specific, and accurate?

In Chapter 3, we compare three popular machine learning classifiers (i.e. SVM, LR, and RF) in predicting spatial distribution of the Phlebotomus Papatasi, one of the major vectors of zoonotic cutaneous leishmaniasis (ZCL). Support vector machine and random forest are two-widely used classifier in environmental studies but have been rarely applied for studying the species geographic distribution. We suppose that potential sampling and measurement errors during collecting sand flies do not have much influence on the accuracy of the classifiers. Also, we assume that ecological factors used in this study are sufficient predictors of Ph.papatasi. The main research questions in Chapter 1 are:

- Which machine learning technique (i.e., SVM, LR, and RF) is more sensitive, specific, and accurate in predicting Ph.papatasi in the study area?
In Chapter 4, we integrated geographic information system (GIS), spatial statistics, and artificial neural networks (ANNs) in analyzing TB distribution in the US to assist national TB control programs. ANN is another data science technique which is a favorite modeling tool in other scientific domain but has never been investigated for geographic modeling of TB. We examine the spatial distribution of the disease and applicability of machine learning techniques in TB modeling with the following assumptions 1) all the reported county-level TB incidence rates represent the status of TB in the US and 2) ecological and socio-economic factors influence the TB infection. Our specific research questions in Chapter 4 include:

- What are the most important ecological and socio-economic contributing factors in predicting TB incidence rate in the US?
- Which modeling technique (i.e., multilayer perceptron (MLP) or linear regression) is more accurate in predicting TB incidence rate?
<table>
<thead>
<tr>
<th>Basic GIS Function</th>
<th>Applications in Infectious Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizing data from multiple sources and in different formats</td>
<td>Sofizadeh et al. (2016) developed a geodatabase of sandflies and explanatory variables to study the distribution of <em>Phlebotomus Paptasi</em>. Sandflies were collected during the field operation. The explanatory data were climatic conditions and the elevation layer. These data were obtained from the WorldClim database. Normalized differentiated vegetation index (NDVI) was used to reflect vegetation cover and obtained from MODIS satellite products.</td>
</tr>
<tr>
<td>Computing slope, and aspect of study region as a candidate factor</td>
<td>Hasyim et al. (2016) derived topographic features (slope and aspect) of the resampled digital elevation model (DEM) of the study area as two explanatory variables in modeling malaria cases with ordinary least square and geographically weighted regression models.</td>
</tr>
<tr>
<td>Calculating Euclidean distance to water bodies</td>
<td>Franke et al. (2015) used Spatial Analyst Tool, to compute Euclidean distance to inland and wetland waters for malaria risk modeling.</td>
</tr>
<tr>
<td>Buffer analysis</td>
<td>Nakahapakor and Tripathi (2005) applied 500 m and 1000 m buffering operation to identify the geographic environment conditions surrounding village affected by Dengue fever.</td>
</tr>
<tr>
<td>Overlay of environmental layers</td>
<td>Palaniyandi et al. (2014) used overlay analysis of various environmental factors to map potential breeding areas of vector-borne diseases in an endemic area in India.</td>
</tr>
<tr>
<td>Identifying spatial pattern and clusters</td>
<td>Mollalo et al. (2017) used global and local Moran's I to identify geospatial pattern and location of clusters of Lyme Disease incidence rate in Connecticut in using Spatial Statistics tool in a GIS environment. Li et al. (2013) established a decision support system, as an accessible and inexpensive approach, for the response to infectious disease surveillance based on WebGIS and intelligent mobile services.</td>
</tr>
</tbody>
</table>
| WebGIS | }
<table>
<thead>
<tr>
<th>RS Function</th>
<th>Applications in Infectious Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI and LST</td>
<td>Mollalo et al. (2018)(^{24}) derived NDVI and LST from MODIS satellite images and used it for prediction habitat suitability of <em>Phlebotomus Papatasi</em>, in an endemic area in Iran.</td>
</tr>
<tr>
<td>Sea surface temperature and Sea surface height</td>
<td>Lobitz et al. (2000)(^{116}) linked public remotely sensed data (Sea surface temperature and height) with cholera cases in Bangladesh.</td>
</tr>
<tr>
<td>Crop disease management</td>
<td>Franke and Menz (2007)(^{117}) used high-resolution multi-spectral data to detect in-field heterogeneities of crop diseases over time. Hugh-Jones et al. (1992)(^{118}) used a land cover map derived from a Landsat TM image to distinguish grazing areas with several levels of animal’s tick infestation.</td>
</tr>
<tr>
<td>Image Classification</td>
<td></td>
</tr>
<tr>
<td>Tick habitat suitability map</td>
<td>Using Landsat TM satellite images, Glass et al. (1995)(^{23}) indicated that many proper tick habitats coincide with residential properties in proximity to the forested areas in Baltimore County, Maryland.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPS function</th>
<th>Applications in Infectious Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping</td>
<td>Coburn and Blower (2013)(^{119}) used handheld GPS to establish geographic coordinates at each sampling sites to map HIV epidemics in sub Saharan Africa.</td>
</tr>
<tr>
<td>Dynamic mobility network</td>
<td>Paz-Soldan et al. (2010)(^{120}) used GPS device to quantify human mobility among tracked individuals to study dengue virus transmission in Peru.</td>
</tr>
<tr>
<td>Outbreak investigation</td>
<td>Masthi et al. (2015)(^{121}) used GPS along with google earth to investigate outbreak of cholera in a village in India to accurately pinpoint the location of household of cases and follow up them.</td>
</tr>
</tbody>
</table>
Table 1-4. Application of GIS, RS, Spatial statistics and machine learning in the study of infectious diseases (Lyme disease, leishmaniasis, tuberculosis)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Infectious disease</th>
<th>Aim</th>
<th>Study Area</th>
<th>Techniques/tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garcia-Marti et al. (2017)</td>
<td>Lyme disease</td>
<td>Mapping and modeling tick dynamics using volunteered data</td>
<td>Netherlands</td>
<td>Random forest; Remote sensing (NDVI); Volunteer geographic information</td>
</tr>
<tr>
<td>Ostfeld et al. (2006)</td>
<td>Lyme disease</td>
<td>To investigate the effect of variations of temperature, humidity, and deer and mice on the risk of Lyme disease</td>
<td>New York</td>
<td>GPS (field data); GIS; Spatial statistics (density mapping)</td>
</tr>
<tr>
<td>Pepin et al. (2012)</td>
<td>Lyme disease</td>
<td>To investigate relations between human Lyme disease incidence and density of nymphs</td>
<td>36 eastern states</td>
<td>GPS (field data); Spatial statistic (zonal statistics, density mapping of nymphs; spatial autocorrelation); Modeling (negative binomial model)</td>
</tr>
<tr>
<td>Ramezankhani et al. (2018)</td>
<td>Leishmaniasis</td>
<td>Predicting cutaneous leishmaniasis incidence based on environmental factors</td>
<td>Isfahan, Iran</td>
<td>Decision trees; Remote sensing (NDVI)</td>
</tr>
<tr>
<td>Nieto et al. (2006)</td>
<td>Leishmaniasis</td>
<td>To predict spatial distribution and potential risk of visceral leishmaniasis</td>
<td>Bahia, Brazil</td>
<td>Ecological niche model (GARP); GIS</td>
</tr>
<tr>
<td>Paixao Seva et al. (2017)</td>
<td>Leishmaniasis</td>
<td>To predict future status of visceral leishmaniasis and identify underlying risk factors To explain spatial pattern of hospitalization due to tuberculosis, and to identify spatial and space-time clusters of TB</td>
<td>Sao Paulo, Brazil</td>
<td>Spatio-temporal Bayesian model; GIS; GIS; Spatial statistics (spatial scan statistics)</td>
</tr>
<tr>
<td>Yamamura et al. (2016)</td>
<td>Tuberculosis</td>
<td>To investigate social behavior of individuals who developed TB</td>
<td>Riberiaro, Brazil</td>
<td>GIS, GPS (locations of houses)</td>
</tr>
<tr>
<td>Patterson et al. (2017)</td>
<td>Tuberculosis</td>
<td>To investigate social behavior of individuals who developed TB</td>
<td>A town in South Africa</td>
<td>GIS, GPS (locations of houses)</td>
</tr>
</tbody>
</table>
### Figures

<table>
<thead>
<tr>
<th>Distance</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close</td>
</tr>
<tr>
<td>Close</td>
<td></td>
</tr>
<tr>
<td>Not Close</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1-1. Contingency table of Knox method

![Contingency table of Knox method](image1.png)

Figure 1-2. Principle of linearly separable SVM using maximum margin

![Principle of linearly separable SVM using maximum margin](image2.png)
Lyme disease (LD), a tick-borne, bacterial, zoonotic infection, remains a serious challenge for public health. The disease is distributed globally, predominantly in temperate portions of the Northern Hemisphere such as Europe, Canada and USA. In the United States (US), the geographical distribution of LD is primarily confined to the north-eastern and mid-western areas. Past studies have shown that in these areas, LD is caused by *Borrelia burgdorferi sensu stricto*. The pathogen is mainly transmitted to humans during blood meals by the bite of infected blacklegged ticks (*Ixodes scapularis*) with white footed mice (*Peromyscus leucopus*) serving as the primary reservoir for this bacterium. The disease is the most common vector-borne disease in U.S. with an estimated average number of 30,000 new cases every year; however, the genuine number is likely much higher. The mean incidence rate of LD in the top 13 U.S. states with the highest incidence rate during 2005-2009 progressively rose from 29.6±10.6 per 100,000 in 2005 to 49.6±15.5 per 100,000 in 2009. At the same time, in 11 states with the lowest incidence rate, the mean incidence developed from 1.3±0.7 to 2.3±1.7 per 100,000 individuals. Although this common zoonotic disease rarely leads to death, it can cause severe symptoms related to skin, joints and heart in addition to anxiety and depression if untreated. LD can also be a socioeconomic burden to society.

In recent years, exploratory spatial data analyses (ESDA) to describe spatial patterns of LD has increased significantly as a strategy to improve our understanding of disease transmission and risk. Several recent studies from different parts of the U.S. have examined the spatial pattern of LD using ESDA. For example, Kugeler and colleagues (2015) applied circular scan statistics to detect high-risk counties of LD in the U.S. from 1993 to 2012. They showed that the number of counties with high incidence of LD successively increased from 69 (1993-1997) to 130 (1998-2002) and further to 197 (2003-2007) and 260 (2008-2012) counties, respectively 134. In Texas, which is a nonendemic LD area, Szonya et al. (2015) applied global and local Moran’s I-tests to determine the distribution and location of possible clusters, respectively, with respect to the spatial distribution of LD at the county level (2000-2011). They observed a clustered distribution with a high incidence cluster in central parts of the state, mainly in a cross-timbers eco-region 65. In Virginia, Li et al. (2014) utilized the Empirical Bayes smoothing (EBS) method on census tract LD cases with the aim of lessening random variations, especially in censuses with small populations. Then, they applied space-time scan statistic and found a primary cluster in northern Virginia which had experienced population growth and urban-sub-urban improvements between 2008 and 2011 135.

The town of Lyme in Connecticut was the first spot that LD was recognized in the U.S. The initial cluster in 1976 was observed in children 136. Since then, in spite of all endeavors conducted by the Connecticut Department of Public Health (CTDPH) to control the disease, it remains endemic with substantial morbidity rates. Although LD is a well-investigated epidemiological subject in Connecticut, historical changes in patterns of disease have been minimally studied 137. Additionally, even though geographical
information systems (GIS) is a useful tool to study infectious diseases \(^1\), powerful GIS-based studies of LD from this region are insufficient for prioritizing counties for intervention. Thus, the main objective of this study is to use the combination of GIS and ESDA to better describe changes in the spatial pattern of LD, supposing that the reported passive cases of LD represent a spatially random subsample of the disease in the state. Our specific research questions included: 1) what is the spatial distribution (random/ dispersed/ clustered) of LD incidence over the past 24 years?; 2) If clustered, where have the clusters/hotspots occurred? 3) Can we avoid or minimize the effects of edges in the spatial scan statistic technique?; and 4) Which cluster detection technique (LISA/spatial scan statistic) is more sensitive, specific and accurate?

**Materials and Methods**

**Data Collection and Preparation**

We used passively reported indigenous LD cases over a period of 24 years from 1991 to 2014 throughout the state of Connecticut. We retrieved data from the CTDPH containing yearly counts and rates of LD at the town level. The CTDPH has a well-established LD surveillance system operating since 1987 \(^{138}\). Reports of cases were based on the National Surveillance Case Definition from the Centers for Disease Control and Prevention (CDC) for LD \(^{139-141}\). Data were geocoded and grouped into four equal intervals (each period included six years: 1991–1996, 1997–2002, 2003–2008 and 2009–2014) to further explore clustering, possible clusters and how hotspots had changed.

Administrative boundaries of towns were obtained from the Map and Geographic Information Center of Connecticut GIS data using the shapefile format (http://magic.lib.uconn.edu/). Similarly, annual population statistics were downloaded
from CTDPH (http://www.ct.gov/dph/site/default.asp). The study area and the names of the towns mentioned in this paper are shown in Figure 2-1.

**Global Clustering**

We applied global clustering techniques to statistically evaluate whether the existing pattern of LD incidence was random, clustered, or dispersed. We used the global Moran’s I statistic \(^{142}\) to measure spatial autocorrelation using GeoDa software version 1.6.7 \(^{143}\). The null hypothesis assumes that there is no spatial pattern among the incidence of LD in different towns (i.e. complete spatial randomness) \(^{144}\). This statistic employs a covariance term between each town and its neighbours as follows \(^{145}\):

\[
I = \frac{N}{S_0} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} w_{ij} (x_i - \bar{x})(x_j - \bar{x}) \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}} \tag{2-1}
\]

\[
S_0 = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \tag{2-2}
\]

where \(x_i\) and \(x_j\) are incidences of LD in the \(i\)th and \(j\)th towns, respectively; \(N\) the aggregate number of towns; and \(wij\) the spatial neighborhood weight for towns \(i\) and \(j\) generated based on the first order Queen’s contiguity which shares all common points including boundaries and vertices. The generated spatial weight is used as a criterion for recognizing neighbors of each town. The weight is defined taking into account adjacent neighbors and written as:

\[
w_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ are adjacent neighbours} \\
0 & \text{otherwise} 
\end{cases} \tag{2-3}
\]

The Moran’s I index varies between -1 and +1, with 0 showing spatially random distribution, while negative values indicate dispersed distributions and positive values
for clustered distributions. We assessed significance of the index using both the Z-score and P-value.

**Local Clustering**

**Spatial smoothing**

The global clustering techniques provide information about the overall distribution of LD (random, clustered or dispersed), but we were also interested in identifying local clusters. First, we applied the EBS routine to account for variation in town sizes and populations. Contrasts in population size among the spatial areal units (i.e. towns of Connecticut) may lead to variance instability and spurious outliers. This is due to the observed raw rate in spatial areal units with small population being profoundly affected by small changes of adding or removing few cases. Thus, crude rates might not reflect underlying risk compared with other areal units with large populations. EBS provides a solution to avoid this type of possible bias as it adjusts the estimated risk toward the global mean to reduce variance instability; areas with low population are adjusted more than areas with larger populations. Since there was a considerable difference in areas of some towns (e.g., Derby and New London are approximately 5 mi² while Woodstock and New Milford cover more than 60 mi²) and also population size of towns (e.g., Union and Canaan have about 1,000 people, whereas New Haven and Bridgeport have more than 120,000 individuals) applying the EBS is justifiable. We calculated spatial weights for each time interval using the first-order Queen’s contiguity. EBS smoothed rates were employed in local cluster detection analyses.

**Local moran’s**

To detect local clusters of the LD rates smoothed by EBS, we applied Anselin’s local indicator of spatial autocorrelation (LISA) statistics. LISA identifies hotspots
(towns with a high incidence surrounded by high incidences); coldspots (towns with a low incidence surrounded by low incidences) and outliers (towns with a high incidence surrounded by low incidences, or towns with a low incidence surrounding by high incidences). We used GeoDa (https://geodacenter.github.io/) for LISA analyses. We set the number of permutation tests to 999 and 95% significance level (p<0.05). We mapped significant clusters using ArcGIS 10.2 (ESRI, Redlands, CA, USA).

**Spatial scan statistics**

We were interested in comparing LD clusters identified by LISA and the Spatial Scan Statistic (SSS), for which we used an ellipsoidal, moving window situated on the centroid of each town so that at any point the window incorporated different sets of neighbors. At each position, the radii of ellipse was set to vary continuously from 0 to a maximum that never included more than half of the total population at risk. If the window contained the centroid of the neighboring towns, then that whole town was included. This procedure produces a very large number of ellipsoidal windows and each one can be a possible cluster of LD with different set of neighbors. The ratio of the length of longest to the shortest axis of the ellipse was 1.5, 2, 3, 4 or 5. For each shape, a different number of angles (i.e. the angle between the horizontal east-west line and the longest axis of the ellipse) of the ellipse were also tested. For each ellipse, a likelihood ratio statistic was computed based on the number of observed and expected cases within and outside the ellipse. The null hypothesis (i.e. LD incidence is equal inside and outside of the window) was tested against the alternative hypothesis that the risk was elevated within the ellipse. The likelihood ratio is reported by P values calculated based on the Monte Carlo simulation approach which finds the maximum likelihood ratio over the entire study region. The ellipse with the maximum likelihood was signified the most
likely (primary) cluster. This approach also detected secondary clusters which were additional ranked clusters that had high likelihood ratios but did not overlap the primary cluster.\textsuperscript{147}

Three sets of data were built for the analysis based on discrete Poisson probability model in SaTScan software version 9.4.2.\textsuperscript{147} The datasets were: Case file representing the annual number of cases of LD for each town (n=169) from 1991 to 2014; Coordinate file was the two-dimensional Cartesian coordinates of centroid of each town; and Population file was the population size of each town. We used the mean population and mean number of cases per time interval. To scan the study area, we applied two criteria: no geographic overlap between clusters and 5\% of the population as the maximum to search for hotspots with the aim of comparing the results with LISA’s high-high clusters as the LISA analysis only considered the neighboring towns. Also in another run, we used no geographic overlap between clusters and 50\% population at risk to investigate whether the results depended on the primary settings. To ensure statistical power, the number of Monte Carlo replications was set to 999 and only clusters at 99\% confidence interval were considered (p<0.01).

Accuracy Assessments

Table 2-1 shows a 2x2 confusion matrix used to compare the performance of LISA and SSS (for 5\% and 50\% of populations at risk) to detect hotspots, sensitivity, specificity and overall accuracy.\textsuperscript{148} In this case, sensitivity measures how well the cluster detection techniques correctly detected the presence of LD hotspots, whereas specificity provides a measure of how well the techniques correctly identified the absence of LD hotspots. Overall accuracy shows the ability of cluster detection techniques to identify true positive and true negative LD hotspots. Locations of the
detected clusters with both LISA and SSS techniques were compared with the location of true clusters as defined by Birnbaum et al. (1996) that “true clusters explain fewer than 5% of all reported clusters” \(^{149}\). Therefore, the top 5% towns with regard to high LD incidence were considered as True Hotspots. We calculated these statistics for each technique in each period, separately.

**Results**

There were 54,478 reported human LD cases from 1991 to 2014, out of which 10,328 cases (19.0%) occurred in first period (1991-1996), 20,234 cases (37.1%) in second period (1997-2002), 11,210 cases (20.6%) in third period (2003-2008) and 12,706 cases (23.3%) in the last period (2009-2014). The annual incidence of LD ranged between 84.7 (in 1993) and 305.6 (in 2002) cases per 100,000 individuals with a mean of 144.8 cases per 100,000 people (Figure 2-2). The P values for the global Moran’s I statistic were close to zero which reject the null hypothesis of complete spatial randomness (CSR) for all time periods (Table 2-2). The index values ranged from 0.55 to 0.71, which indicates significant clustering. Although, we detected clustering with both raw and smoothed methods, and in all four periods, the variation in rates required EBS before running LISA. Based on the results of EB smoothing technique, LISA found High-High clusters (hotspots), which varied for each study period. Results of LISA showed that in the first and last period, the hotspots were completely restricted to the towns in the East. In other words, for the first period, 24 towns and for the last period 30 towns were identified as the hotspot in eastern Connecticut. The towns in the West were more influenced in the second and third periods. The number of towns in the West identified as hotspots in the second period was 8, while 7 towns were hotspot in the East of Connecticut. In addition, in the third period, 10 towns in the West and 6 towns in the
East were identified as hotspots. It should be noted that applying EBS before running LISA, reduces the likelihood of detecting a false cluster in low-population areas where the cases were detected. But comparison of the detected clusters by raw and EBS methods in LISA showed small differences with regard to the location of the detected clusters.

Spatial scan statistics with 5% and 50% of the populations at risk with no geographic cluster overlaps identified clusters with high incidence rates for each period. Except for the second period, The SSS results revealed that the most likely cluster predominantly occurred in the eastern region of the state, while the secondary cluster occurred in towns in the West (Figure 2-3). Primary and secondary hotspots were observed in different locations when 5% of the population at risk was investigated using SSS for comparison with LISA. For the first period, the primary cluster occurred in the eastern parts of the state and included 22 towns and 792 cases. The risk of LD incidence within the primary cluster was 7.67 times greater than outside the cluster. The secondary cluster occurred in the western region and contained 5 towns and 152 cases. During the second period, the primary cluster occurred only in the eastern parts of the state and included 23 towns and 1,208 cases. The risk of LD incidence within this cluster was 3.40 times higher than outside. During the 2003-2008 period, the disease largely affected the eastern region with 22 towns and 738 cases as compared to the western parts with 10 towns and 145 cases. The relative risks of primary and secondary clusters were 3.33 and 9.18, respectively. In the last period, only eastern parts of the state showed statically significant clustering (no secondary cluster). The primary cluster
contained 24 towns and 942 cases. The risk of LD incidence in this cluster was 3.69 times more than in other areas (Table 2-3).

Comparison of the results of accuracy assessments of the spatial scan statistics at 5% and 50% of the populations at risk, shows that sensitivity of this method increases with an increment of the population defined to be at risk. However, increasing the population at risk also leads to a decrease in the specificity of the results. In addition, LISA tended to have higher specificity and had the highest overall accuracy (Figure 2-4).

**Discussion**

This retrospective study examined the spatial structure of LD incidence distribution in Connecticut based on 24 years of reported data with the aim of describing the spatial distribution and the changes that have occurred with regard to the disease. It differs from most other studies in the region by not focusing on local studies of the pathogen, reservoir or vector and their associations with the environment. Instead, it focuses on the changing patterns of documented human disease occurrences. We assumed that the number of affected human cases corresponded to the density of *Ixodes scapularis* in Connecticut 150-152. In addition, it should be noted that ticks have limited capabilities to move to new areas because of their small size 153. Therefore, one of the means to fight against the risk of the disease in the study area would be targeted intervention in the areas (towns in this case) that were constantly affected with a very high morbidity rate. Targeted control, planning and management of the disease can assist with resource allocation to the towns with persistent high incidence rates resulting in time and costs savings.
Results of this study confirm and extend the findings of Xue et al. (2015) in Connecticut\(^{154}\). They analyzed yearly clusters of LD and showed that the distribution was clustered and that this clustering occurred in western and eastern Connecticut with few cases in the central region supposing the yearly incidence weighted-geographic mean analysis represents the clusters of the case distributions. According to the paper, the epidemic of LD reached equilibrium in 2007 in the western parts of Connecticut, while this happened in 2009 in the East. Comparison of the results shows that periods 1-3 included epidemic conditions while period 4 contained equilibrium condition. Therefore, actual/sudden shifts in the locations of clusters occurred roughly before equilibrium (period 1-3); and it may be that epidemic conditions affected both western and eastern parts of the state; while equilibrium conditions only occurred in the eastern regions.

This study identified towns with high LD incidence rates using the LISA test and spatial scan statistics approaches. By intersecting the locations of the clusters identified by LISA some towns, including Chaplin, Windham and Scotland, were persistently detected to have high-rate clusters. Likewise, neighbors of these towns including Andover, Columbia and Lebanon, were recognized as having high-rate clusters in 3 out of the 4 periods of study (Figure 2-1). This indicates that these areas were almost persistently affected by a high incidence of LD during the 24 years of study and therefore they deserve closer consideration.

In this study, we used an elliptical window in the scan rather than highly applied circular shape used in other studies for several reasons. First, according to previously published papers, the elliptic shape had better power and precision compared with
circular one and also follows more accurately certain geographic features with varying shapes and directions. Another reason was to avoid edge effects at the borders of the study area. When circles are used for scanning the study area, particularly when the circle should center on the centroid of border towns, considerable parts of it inevitably covers the neighboring state (i.e. Rhode Island in the East, Massachusetts in the North, Long Island Sound to the South and New York in West, where states are also LD-endemic areas where data were not accessible) and the true number of cases would be underestimated. The use of ellipses with different shapes and directions helps to reduce this problem to some degree.

There were similarities in the locations of the spatial hotspots identified by SaTScan and LISA even though these methods apply different methodologies to detect hotspots. It should be noted that we smoothed the incidence rate for the LISA cluster detection method, while this is not available for spatial scan statistics. The results were also in agreement with the findings of Barro et al. (2015), who compared different cluster detection techniques including Getis-Ord Gi* statistic, a multidirectional optimal ecotope-based algorithm (AMOEBA) and the spatial scan statistic for identifying hotspots of human cutaneous anthrax in Georgia for point data. They also found that SSS was more sensitive (due to augmentation in the quantity of true positives) but less specific (by increment of the population at risk because of a declining number of true negative clusters). Here, the results were highly dependent on defining the weight matrix in LISA or the percentage of the population at risk in spatial scan statistics.

The most important limitation of this study is attributed to the data reported that were used in this study. As indicated by the CDC, surveillance data are subject to
under-reporting and misclassification in highly endemic areas such as Connecticut; however, this problem would not be very severe in light of well-designed LD surveillance system in this state. Additionally, as long as there is simply a spatially random thinning of reported cases, the overall spatial analysis should not be affected. However, if there is persistent over- or under-reporting in selected regions, it may be difficult to recognize the impact of such biases. Another factor influencing the results is that the definition of LD has changed several times: from using a two-test approach for laboratory affirmation to the addition of probable and suspect categories with less strict criteria and subtle changes in the specification of confirmed cases. However, empirically (Figure 2-3), these changes do not seem to substantially have altered the areas of high-risk clusters. Therefore, although the analyses conducted were based on data reported by CTDPH, the findings of this study should be interpreted with some caution.

The findings of this study can be regarded as a basis for generating hypotheses about underlying risk components. For instance, visual comparison of the locations of towns that were never identified as cluster areas (Figure 2-3) and the digital elevation model (DEM) of the study area suggest that low-risk towns are located in low-lying areas; thus it seems that people who live at moderate or higher altitudes in Connecticut are at a higher risk. This would be consistent with the earliest epidemiologic investigations that identified increased risk in areas away from the seashore. Thus, altitude as a proxy of other environmental factors can uncover the reasons of the spatial distribution of the disease in the study area. Moreover, except for Windham, all the other towns identified as persistent clusters had populations lower than 8,000 people showing that persistent towns as clusters occurred in less populated areas. Therefore,
further studies should incorporate other environmental factors that have significant influence on the spatial and temporal patterns of the disease.
## Tables

### Table 2-1. Comparison of LISA and SaTScan clusters by confusion matrix and its derivatives statistics

<table>
<thead>
<tr>
<th>Identified hotspots (LISA or SaTScan)</th>
<th>Presence</th>
<th>Absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual hotspots (top 5% incidence rates)</td>
<td>Presence</td>
<td>True Positive (TP)</td>
</tr>
<tr>
<td></td>
<td>Absence</td>
<td>False Positive (FP)</td>
</tr>
</tbody>
</table>

Sensitivity = (TP)/(TP+FN); Specificity = (TN)/(TN+FP); Overall accuracy = (TP+TN)/(TP+FN+FP+TN)

### Table 2-2. Results of the global Moran statistic of LD incidence rate, Connecticut, 1991-2014

<table>
<thead>
<tr>
<th>Year</th>
<th>Index</th>
<th>Z-score</th>
<th>P-value</th>
<th>Type of distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991-1996</td>
<td>0.71</td>
<td>16.05</td>
<td>≈ 0</td>
<td>Clustered</td>
</tr>
<tr>
<td>1997-2002</td>
<td>0.59</td>
<td>13.54</td>
<td>≈ 0</td>
<td>Clustered</td>
</tr>
<tr>
<td>2003-2008</td>
<td>0.55</td>
<td>12.73</td>
<td>≈ 0</td>
<td>Clustered</td>
</tr>
<tr>
<td>2009-2014</td>
<td>0.59</td>
<td>13.28</td>
<td>≈ 0</td>
<td>Clustered</td>
</tr>
</tbody>
</table>

### Table 2-3. Characteristics of LD spatial clusters detected by SSS with 5% of the population at risk throughout Connecticut, USA for the period 1991-2014

<table>
<thead>
<tr>
<th>Cluster (C)</th>
<th>Year</th>
<th>Observed Cases (N)</th>
<th>Expected Cases</th>
<th>Ratio (Obs/Exp)</th>
<th>P-value</th>
<th>Relative Risk (RR)</th>
<th>Log likelihood ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Likely (East)</td>
<td>1991-1996</td>
<td>792</td>
<td>171.26</td>
<td>4.62</td>
<td>&lt;0.0001</td>
<td>7.67</td>
<td>736.02</td>
</tr>
<tr>
<td>Secondary (West)</td>
<td>1991-1996</td>
<td>152</td>
<td>32.26</td>
<td>4.71</td>
<td>&lt;0.0001</td>
<td>5.07</td>
<td>120.17</td>
</tr>
<tr>
<td>Most Likely (East)</td>
<td>1997-2002</td>
<td>1,208</td>
<td>474.52</td>
<td>2.55</td>
<td>&lt;0.0001</td>
<td>3.40</td>
<td>496.50</td>
</tr>
<tr>
<td></td>
<td>2003-2008</td>
<td>738</td>
<td>305.46</td>
<td>2.42</td>
<td>&lt;0.0001</td>
<td>3.33</td>
<td>284.31</td>
</tr>
<tr>
<td>Most Likely (East)</td>
<td>2009-2014</td>
<td>942</td>
<td>376.77</td>
<td>2.50</td>
<td>&lt;0.0001</td>
<td>3.69</td>
<td>400.80</td>
</tr>
</tbody>
</table>

P-values were obtained from the Monte Carlo hypothesis test
Figure 2-1. Geographic location of the Connecticut, its towns and approximate populations. The names on the map show the location of the towns which are mentioned in this paper.
Figure 2-2. Temporal trend of LD incidence throughout Connecticut from 1991 to 2014. Black dots show the incidence rate for each year and the blue line represents a scatter with smooth lines.
Figure 2-3. Locations of spatial clusters of LD incidence in Connecticut, USA based on the true-cluster definition of with LISA and SSS methods targeting 5% of the population at risk (A), for the period 1991-1996; (B), for the period 1997-2002; (C), for the period 2003-2008; (D), for the period 2009-2014.
Figure 2-4. Comparison of LISA and spatial scan statistics with 5% and 50% of the population at risk of LD in Connecticut, USA for each period from 1991 to 2014. Sensitivity, specificity and overall accuracy expressed as per cent (%).
CHAPTER 3
MACHINE LEARNING APPROACHES IN GIS-BASED ECOLOGICAL MODELING OF
THE SAND FLY PHLEBOTOMUS PAPATASI, A VECTOR OF ZOONOTIC
CUTANEOUS LEISHMANIASIS

Zoonotic cutaneous leishmaniasis (ZCL) is an important neglected tropical
disease due to insufficient considerations for controlling and ignored remarkable public
health impacts in endemic countries *158. It poses unpleasant skin lesions to the patients
which may take a long time to heal and injects serious socio-economic burdens to the
society 159,160. This vector-borne disease is highly prevalent in developing countries
including Iran 161. Previous studies in Iran have demonstrated that Leishmania major is
the causative agent of ZCL and Ph. papatasi which is the target species in this study
serves as the major vector of the disease. It is reported that the disease is still endemic
in many rural areas of 17 of 31 provinces of Iran and almost 70% of
cutaneous leishmaniasis cases in Iran are ZCL 162,91. A large number of factors impact
the distribution of Ph. papatasi and/or its pathogen which have made it a
multidimensional and complex health problem particularly when the underlying
relationship is ambiguous 163.

In recent years, novel machine learning techniques such as random forest (RF)
and support vector machine (SVM) have become popular research tools for ecologists
and medical researchers due to improved predictive performances of the techniques,
particularly in solving binary classification tasks 164. The rapid application of these
techniques might be due to their abilities to approximate almost any complex functional

* Reprinted with permission from Mollalo, A., Sadeghian, A., Israel, G. D., Rashidi, P., Sofizadeh, A., &
Phlebotomus papatasi, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. Acta
tropica, 188, 187-194.
These techniques are mathematical models that can map the relationship between input data and the target output by receiving several examples (training data). A trained model subsequently can predict a variable of interest for a new independent (test) dataset.

To date, there are few studies using machine learning methods that model the spatial distribution of leishmaniasis. In South America, Peterson and Shaw (2003) used the genetic algorithm for rule-set prediction (GARP) to model ecological niche of three sand fly vector species: Lutzomyia whitmani, Lutzomyia intermedia, and Lutzomyia migonei. They projected the geographic distribution of vectors for the year 2050 under global climate change scenarios and predicted that Lutzomyia whitmani will dramatically increase in southern Brazil. Samy et al. (2016) used Maximum Entropy (MaxEnt), a presence-only ecological niche model (ENM), to map the potential distribution of ZCL across Libya. Their model successfully predicted the risk areas and distribution of most species associated with ZCL including Ph. Papatasi. This sand fly was present mostly in lowland areas and the risk areas of ZCL were almost confined to the northern coast of the country. In a study in northwestern Iran, Rajabi et al. (2014) used radial basis functional link nets (RBFLN), a neural network modeling technique, to map potential areas of visceral leishmaniasis (VL), or kala-azar, a fatal form of leishmaniasis. They produced a susceptibility map of VL, with the overall accuracy of 92% in endemic villages and concluded that riverside and nomadic villages without health centers were high-risk areas. RF and SVM have been applied in the classification of other vector-borne diseases such as malaria and dengue fever. In an endemic district in South Africa, Kapwata and Gebreslasie (2016) successfully applied a
random forest approach to select important features to create predictive maps of malaria. They found that the most determining features were altitude, normalized difference vegetation index (NDVI), rainfall and temperature. Gomes et al. (2010) trained an SVM model for a small sample size (n = 28 patients) to classify dengue fever (DF) and dengue hemorrhagic fever (DHF) patients based on gene expression data. They achieved 85% accuracy for the validation dataset and identified 7 out of 12 genes that can differentiate DF patients from DHF patients.

To the best of our knowledge, no effort has been made to model *Ph. papatasi* or other vectors of leishmaniasis distribution using SVM and RF, which are two effective supervised techniques in ecological studies. This study analyses the applicability of these techniques in GIS-based modeling of the species in Golestan province, northeast of Iran. These techniques are used to classify this endemic area of ZCL into two distinctive classes: ‘presence’ and ‘absence’ of *Ph. papatasi*. The accuracy of these classifiers was compared with the frequently applied (vanilla) logistic regression. This study is a starting point for further studies and application in control measures such as “automatic early warning systems”. Findings of this study can help decision makers target surveillance, and to control and mitigate species propagation into new foci.

**Material and Methods**

**Study Area**

Surveillance was conducted in Golestan province, a notorious endemic area of leishmaniasis. The province has over 1,700,000 inhabitants and covers an area of 20,368 square kilometers in the northeast of Iran. This area lies within latitudes 36°30’ and 38°8’ N and longitudes 53°57’ and 56°22’ E (Figure 3-1). Geographically, this province is bounded from north to Turkmenistan and from west to the Caspian Sea and
Mazandaran province. The altitude of the study area ranges from −32 to 3,945 m above mean sea level. The climate of province varies from cold to moderate weather in the south because of Alborz mountain ranges and arid and semi-arid with hot summers in northern areas. As one moves from southern to northern areas, generally, the amount of rainfall and percentage of humidity decreases and air temperature increases.

Data Collection and Preparation

Several categories of data in different formats including vector and raster data models and excel spreadsheets were incorporated. The data were obtained from the Worldclim, the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite images, the United States Geological Survey (USGS) and GIScenter of Golestan province. These data were used as candidate explanatory variables to model spatial distribution of vector presence/absence in the study region. It was essential to determine candidate factors to model species distribution. Thus, in the first step, a ZCL related geo-database was developed in GIS and stored in a common projection system.

Sand fly collection

We conducted sampling during the primary activity period of sand flies from early April 2014 to late October 2014 (adult sand flies are absent in other months). In each of the 16 counties of the province, 5 to 10 villages were selected based on their topographic conditions, spatial distribution of sampling sites (almost distributed uniformly across the study area), and area of each county. For each sampling site, sticky paper traps smeared with castor oil with the size of 15*20 cm used and placed indoors (bedrooms, stables,…) and outdoors (animal shelters, rodents burrows,…). In each site, 30 indoor and 30 outdoor sticky traps were monthly installed. We installed traps at sunset and collected them the next day before sunrise. Also, no unprecedented
changes in climate condition or ZCL incidence was observed for the study period.

Collected sand flies were then washed to remove oil and other debris. We preserved the sand flies in 70% ethanol and transferred them to an entomology laboratory for species identification. Identification of collected material was based on morphological identification keys of dissected sand flies. In 44 out of 142 sampling sites, *Ph. papatasi* was present (Figure 3-1B). Although other species of sand flies were also found in the study area, we did not consider them in habitat modelling, as *Ph. papatasi* is the main vector of ZCL in the study area. Detailed information on collected sand flies has been presented in Sofizadeh et al. (2016) 109.

**Exploratory data**

To investigate the effects of temperature and precipitation on habitat suitability of *Ph. papatasi* in this area, we used the WorldClim database. Nineteen climate variables (bio1 through bio19) were obtained from the database (http://www.worldclim.org) and clipped with the boundaries of the study area (Table 3-1). These variables are at 1 km (30 arcsec) spatial resolution. In addition, to investigate the possible effects of proximity to water on presence/absence of the sand fly, we calculated the Euclidean distance from each sampling site to the nearest river in the GIS environment. The river coverage obtained from GIS and Environment Center of Golestan province.

Two widely used remotely sensed products including land surface temperature (LST) and NDVI were used to measure land temperature and vegetation cover, respectively. These products were downloaded from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite imagery from March to October 2014 (corresponding to the same time of the sand fly collection). Both LST and NDVI are at
1 km spatial resolution; the temporal resolution for the LST dataset is 8 days which aggregated to a monthly resolution. LST downloaded for day and night, separately. The NDVI dataset is at a monthly temporal resolution. They were pre-processed in ENVI 5.4 image analysis software. All remotely sensed satellite images are freely available at the MODIS website: http://www.modis.gsfc.nasa.gov.lp.hscl.ufl.edu.

We used NDVI to reflect the density of vegetation cover on the ground. This index, which has been highly correlated with the plant species presence, is calculated via the red (R) and near-infrared (NIR) spectral bands from the following formula:

\[
\text{NDVI} = \frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}}
\]  

(3-1)

The value of NDVI differs from −1 to +1 so that a greater value of NDVI represents a higher amount of vegetation cover.\textsuperscript{158, 160, 174}

The 30-meter shuttle radar topography mission (SRTM) elevation data was obtained from the United States Geological Survey (USGS) website (https://www-usgs.gov.lp.hscl.ufl.edu/). Also, a slope map was generated from the digital elevation model (DEM) of the study area. To calculate slope, the average maximum algorithm was used in ArcGIS 10.2 software (ESRI, Redlands, CA, USA).

In summary, we developed a set of 31 candidate variables representing different weather conditions, topographies, vegetation regimes, land temperature and proximity to rivers that might plausibly impact sand fly occurrence (Supplementary Table). The selection of these environmental variables was based on previously published literature that found direct and/or indirect impacts on the distribution of the sand fly populations. The values of each raster variable were extracted for each of the 142 sampling sites.
through "Extract values to points" function in ArcGIS 10.2. The presence and absence of *Ph. papatasi* are coded with 1 and 0, respectively.

**Classifiers Used in This Study**

Logistic regression (LR) is a generalized linear regression model that is appropriate for situations where the dependent variable is dichotomous. The output of LR is a likelihood of species occurrence as a function of several exploratory variables.

The random forest (RF) classifier, an ensemble tree-based algorithm, generates a large number of binary decision trees (for instance 1000 trees) and then combines them to provide a final classification for the presence/absence of the outcome. This algorithm injects randomness through random selection of observations using bootstrap sampling (i.e. sampling with replacement) and random selection of exploratory variables. This process learns binary trees that are shallow (i.e. have a small number of nodes) but highly specialized on a subset of the dataset/variables. Ensembling the decisions of this diverse set of trees (forest) leads to an improvement in the performance of the method. A final decision is achieved based on the majority of labels (i.e. votes or outputs of trees) obtained from individual trees. In this study, the number of trees was optimally selected from the set of {10, 30, 50, 100, 500} by cross-validation. In addition, the optimal maximum depth (i.e., number of layers or number of edges from the root to the node) of the trees was chosen from the set of {2, 3, 4}. The maximum depth is selected via the cross-validation done on the training data.

SVM, originally proposed by Vapnik (1999), can extract patterns from complex non-linear data. We used this SVM's ability to extract spatial patterns from the collected data and features. SVM has several appealing characteristics compared to
most machine learning techniques. It guarantees global optima and leads to a unique solution when solved using iterative methods, rather than stochastic randomness. We used a grid search to find the optimum values for the SVM parameters. Detailed explanations of SVM can be found in Noble.

We trained all above-mentioned models (i.e. LR, RF, and SVM) using sklearn (http://scikit-learn.org/stable/) in Python programming language to predict Ph. *papatasi* occurrences.

**Pre-processing of the Data**

In the first step, to avoid/reduce the existence of the multicollinearity and redundancy of the variables, we applied Pearson’s correlation analysis in SPSSver. 23 software. We investigated Pearson correlations among all pairs of selected variables. We omitted highly correlated factors (i.e. correlation coefficient >0.7) before modeling. All of the models are sensitive to the number of exploratory variables because redundant variables add extra noise which reduces model accuracy and generalizability. A proper selection of the most effective input variables is crucial to improving predictions. To address this, feature selection via L1 regularization was utilized before developing the models to reduce the effects of possible noise. More detailed information of L1 regularization method can be found in Park and Hastie (2007).

After selecting the variables, we standardized (i.e., scaled) the input dataset before incorporating the variables into the models because the data have different ranges and units. This important pre-processing step can increase the performance of models by converging faster, particularly in SVM and RF which are solved with iteration-based optimization techniques. There are several standardization formulas.
presented in literature and, here, we use the following equation which re-scales input data to z-scores:

\[ Z_n = \frac{Z_i - \mu}{\sigma} \quad (3-2) \]

Where \( Z_i \) is the initial (actual) value; \( \mu \) and \( \sigma \) are the average and standard deviation of the initial value, respectively and \( Z_n \) is the standardized value. Moreover, after developing the models, the output of the model is returned to the original form through the following equation:

\[ Z_i = (Z_n \ast \sigma) + \mu \quad (3-3) \]

**Overview of Training Models**

To develop models, we used 5-fold cross-validation where in each fold, the dataset is randomly partitioned into two different categories including training set, 75% of total data (\( n_{\text{train}} = 106 \)), and a remaining 25% (\( n_{\text{test}} = 36 \)) as a hold-out test set used to assess the accuracy and predictive power of the models after training process. We saved and used the same training and test datasets for all models for further comparisons. Also, the data are randomly selected, thus, on average 75% of the training set is the same from one fold to the other.

To fine-tune each model’s hyper-parameters (e.g. parameters in SVM), using the earlier mentioned grid search method, we performed a 3-fold cross validation within the training set, breaking it into 66.6% and 33.3% subsets for each fold. As a final step, LR, SVM and RF classifiers with optimal parameters were trained and then used to distinguish presence from the absence of *Ph. papatasi*. All the reported results are the average from the 5-fold cross-validation.
Models Validation

A confusion matrix or error matrix is used to assess the generalization capabilities of the classifiers. Sensitivity (or true positive rate) and specificity (true negative rate) were used to evaluate the ability of the model to correctly identify presence and absence of *Ph. papatasi*, respectively. Also, two important statistics have been calculated to measure the performance of classifiers, namely positive predictive value (PPV) and negative predictive value (NPV). PPV is the probability that the sampling sites identified by the model with the presence of *Ph. papatasi* that actually have that species, while NPV is the probability that the sampling sites identified by the model as the absence of *Ph. papatasi* that actually don’t have that species. The models’ accuracies were assessed using overall accuracy and area under receiver operating characteristics (ROC) curve. In this study, we also used Cohen’s Kappa index to exclude the probability of chance agreement between predictions and ground truth.

Due to black-box nature of SVM, we are unaware of the weights/contribution of each variable in predicting presence/absence of *Ph. papatasi*. We applied sensitivity analysis to understand the contribution of each factor on the output variable by excluding each variable from the model one by one and comparing the performance of the model in terms of Cohen’s Kappa index. The absence of a variable that reduces the Kappa index of the model the most is considered as the most important factor and so forth. Finally, variables were imported into the most accurate model to generate habitat suitability map for the entire study area. The obtained map was classified into 5 different categories: very low, low, moderate, high and very high probabilities. Natural break
classification (a.k.a. jenks) was utilized to classify the obtained suitability map. The overall flowchart of the procedures is shown in Figure 3-2.

**Results**

After conducting variable selection, 7 variables selected as inputs of the models. Table 3-2 shows the list of selected variables and the correlation coefficients among them.

As stated in the Methods section, the parameters C and $\gamma$ in the SVM model were determined experimentally by systematically increasing these values from 0.5 to 20 and 0.005 to 0.1, respectively. The best accuracy of SVM was obtained with $C = 17.5$ and $\gamma = 0.105$ and the radial basis function (RBF) kernel.

Results of this study using the test dataset (25% of the sampled areas not used in developing the models) demonstrated that all three classifiers have good generalization capabilities. The SVM model generated the highest overall accuracy (91%) followed by the LR (88%) and the RF (85%). The ROC curves (prediction rates) and AUCs for each method (Figure 3-3) showed all the classifiers generated high AUCs (more than 90%); however, there were slight differences between AUCs of the classifiers. The AUCs were 0.974, 0.965 and 0.935 for SVM, LR, and RF classifiers, respectively.

Kappa indices were 0.79, 0.73 and 0.65 for the SVM, LR, and RF models, respectively, suggesting substantial agreements between models outputs and ground truth. Similarly, the lowest accuracy expected better than chance was obtained by the RF method.

The highest positive predictive value (PPV) was achieved by the SVM model (85%), indicating that the probability that the sampling sites identified by the SVM
classifier with the presence of *Ph. papatasi* that actually have that species was 85%. It was followed by the LR model (82%) and the RF model (79.0%). Similarly, the SVM had the highest negative predictive value (NPV = 93.2%) suggesting that the probability that the sampling sites identified by the SVM as the absence of *Ph. papatasi* that actually don’t have that species was 93.2% and followed by the LR (90.7%), and RF models (87.0%).

The SVM model was the most sensitive model, with 84.9% of the correctly classified sampling sites, followed by LR model with 80.0% of the correctly classified sampling sites whereas the RF model was the least sensitive classifier with 73.4%. The SVM model also had the most specific model (93.2%), implying that over 90% of absence locations are correctly classified by this method. It was followed by the LR model (91.9%) and RF models (90.4%). A graphical comparison of evaluation metrics of classifiers (i.e. LR, SVM and RF) is shown in Figure 3-4.

Based on the above results, the best accuracy was obtained by SVM (with the RBF kernel) compared to the other classifiers. Next, we assessed the contribution of each variable for this best classifier using sensitivity analysis. Results of this analysis indicated that the lowest Cohen’s Kappa index obtained when the slope was removed from the SVM model. This suggested that this factor is the most influential factor in predicting *Ph. papatasi* in the study region. The most influential factors in order of contributions (from higher to lower) as predictors for *Ph. papatasi* were: slope, nighttime LST in October, bio8 (mean temperature of wettest quarter), bio4 (temperature seasonality), bio18 (precipitation of warmest quarter), bio5 (max temperature of
warmest month), bio15 (precipitation seasonality), distance to major rivers and bio13 (precipitation of wettest month), respectively.

According to the produced habitat suitability map (Figure 3-5) from the SVM method, considerable parts of the study area are under very high probabilities of the species occurrence. Results of the natural break classification of the map show that 28.8%, 5.0% and 7.7% of the total study area were predicted in the very low (0–0.18), low (0.18–0.48) and moderate (0.48–0.73) probabilities of existence classes, while areas covering high (0.73–0.90) and very high (>90) probabilities of *Ph. papatasi* occurrences represent 16.5% and 42% of the total area, respectively. Very high probabilities of the species occurrence happened mostly in central and north (eastern) parts, while no and low occurrence of the species was restricted to the southern half of the study area.

**Discussion**

This study focused on geospatial features that can predict the presence or absence of *Ph. papatasi* in an endemic province of ZCL in Iran. We combined relatively novel techniques/tools: remote sensing (to extract NDVI and LST), GIS (mainly for data preparation and mapping) and machine learning (for model development). Our results successfully demonstrated the utility of machine learning techniques in predicting the spatial distribution of *Ph. papatasi* across Golestan province, Iran, from a surveillance sampling strategy. Among the compared models, the SVM was an efficient classifier with the highest prediction capability for producing habitat suitability map of *Ph. papatasi* occurrence. Phlebotomine sand flies have limited flight range of about 300 m \(^{185}\) and a few have been known to travel 1.5–2 km \(^{186}\). Regarding 1 km
spatial resolution of LST and bioclim variables, it seems that it is safe to assume that this scale is operative for intervention or allocation of resources.

Regardless of the SVM benefits in *Ph. papatasi* habitat modeling, our results did not indicate that the SVM classifier necessarily outperforms other binary classifiers for species modeling or that RF is not a very accurate classifier. For instance, Ding et al. (2018) found SVM inferior to the other models including random forest and gradient boosting machine in modeling the global distribution of *Aedes aegypti* and *Aedes albopictus* \(^{187}\). Also, Williams et al. (2009) identified RF as the best model in predicting appropriate habitats of six rare plant species in Creek Terrane, northern California. The weaker results of RF compared to the other classifiers might be due to overfitting (memorization). But generally, it is recommended to take SVM into consideration for binary classification problems when having a dataset with a small sample size and a relatively large number of exploratory variables \(^{188}\).

The findings from this study can be compared with previous work conducted in the same area by Sofizadeh et al. (2016). They utilized presence-only ecological niche models (i.e. MaxEnt and GARP) for *Ph. papatasi* habitat modeling. The SVM model had a 27% improvement in accuracy over the MaxEnt model \(^{109}\). All three presence-absence models developed in this study were more accurate than those developed using ecological niche models. However, niche models generate pseudo-absence data from the background, AUCs derived from pseudo-absence models are difficult to interpret and thus a spurious pseudo-model might be selected. In this study, the lowest accuracy for test data obtained by RF was 85%, while the overall accuracy of MaxEnt, which outperformed GARP in the study of Sofizadeh et al. (2016), was no better than 64%.
This suggests that presence-absence models perform better for habitat suitability models of the sand flies. Our results also confirm and extend the findings of Carvalho et al. (2015) in predicting the vector of *Leishmania amazonensis* in South America. In their study, random forest, as a presence-absence model, had the best performance compared to several models including ENMs. They concluded that incorporating species absence data significantly improved model performance. However, it should be acknowledged that we incorporated some new variables such as monthly NDVI (as opposed to yearly NDVI) and monthly LST (for day and night, separately) and distance to main rivers together with considering absence points in modeling. Among these new variables, nighttime LST and distance to main rivers were found to be important determinants.

Sensitivity analysis indicated that slope was the most influential factor which is in agreement with the results of previous ecological niche model. Topography and its derivatives such as slope can have influence on biological and physical factors. For instance, an increase in altitude leads to decreases in temperature and rainfall, and these changes cause different plant types to grow. Also, Mollalo et al. (2014) found altitude (and slope) a determining factor in the incidence of human ZCL. This suggests that there is a strong correlation between human cases of ZCL and occurrence of *Ph. papatasi*. Nighttime LST found the most important factor after slope which rarely is incorporated in modeling *Ph. papatasi*. Land temperature can be associated with soil moisture, evapotranspiration and vegetation water stress, which can be determining factors in larval development and egg survival. Boussaa et al. (2016) also found LST a significant factor in describing the distribution of *Ph. papatasi*. Similarly, no significant
association of daytime LST was observed which might be attributed to the nighttime activities of sand flies. Temperature and precipitation variables (bio8, bio4, bio 13, bio 15, bio18) were found as the next important contributing predictors. Cross et al. (1996) and Medlock et al. (2014) found that temperature and precipitation play important role in development and survival of different life stages of *Ph. papatasi*. These two factors have different effects on sand fly populations. For instance, lack of rainfall sometimes doesn’t reduce larval frequency while moderate rainfall facilitates transmission and intense rainfall reduces transmission because it flushes away suitable resting sites for adult sand flies and destroys immature stages of the life cycle. In regards to temperature, tropical climate provides favorable conditions to facilitate transmission, while, the sand fly is not able to survive in temperature less than 15 °C under laboratory conditions.

It should be noted that the accuracy of the classifiers used in this study depends entirely on the data. Additionally, there are potential sources of errors (random errors or systematic errors) in the dataset such as sampling bias or errors during vector collection which can influence the accuracy of the classifiers. Therefore, the results of this study should be used with some caution. We also recommend testing the performances of the models on other datasets (e.g., regions) with comparable environmental characteristics to select the best unbiased estimator for predictions. As a next step, a new sampling validation in the same area of study is suggested to be performed in absence and presence sites not sampled previously. Future studies should also consider actual counts of the *Ph. papatasi* for each sampling site. Depending on the frequency distribution of sand flies, proper modeling techniques such as Poisson regression,
negative binomial models, and artificial neural networks can be used. However, higher counts of vector imply a greater risk of ZCL, the main purpose of this study was to compare predictive capabilities of three different binary classifiers and modeling for count data was not the purpose of this study.

In this paper, the use of novel machine learning techniques to predict the presence/absence of an arthropod vector, *Ph. papatasi*, was investigated. The findings showed that the SVM classifier was a reliable tool to solve specific problems in medical geography and ecological modeling. The results of this study can be extended to develop standalone early warning services (EWS) to actively monitor sand flies' presence in endemic areas. Due to sophisticated predictive capabilities of SVM compared to other classification methods, this classifier can be considered/incorporated into the design of early warning systems (i.e. predictions of vector presence greater than a threshold as a warning). This is applicable even for other vector-borne diseases such as tick- and mosquito-borne diseases. However, the implementation of EWS is more feasible in countries with well-established climate stations that systematically record data.
Table 3-1. Bioclimate variables used in this study

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>slope</td>
<td></td>
</tr>
<tr>
<td>bio13</td>
<td></td>
</tr>
<tr>
<td>bio15</td>
<td></td>
</tr>
<tr>
<td>bio5</td>
<td></td>
</tr>
<tr>
<td>bio8</td>
<td></td>
</tr>
<tr>
<td>LST_OctN</td>
<td></td>
</tr>
<tr>
<td>Dist_major</td>
<td></td>
</tr>
<tr>
<td>slope</td>
<td>1</td>
</tr>
<tr>
<td>bio13</td>
<td>-.567**</td>
</tr>
<tr>
<td>bio15</td>
<td>.376**</td>
</tr>
<tr>
<td>bio5</td>
<td>-.610**</td>
</tr>
<tr>
<td>bio8</td>
<td>.031</td>
</tr>
<tr>
<td>LST_OctN</td>
<td>-.356**</td>
</tr>
<tr>
<td>Dist_major</td>
<td>.457**</td>
</tr>
</tbody>
</table>

Table 3-2. Pearson correlation coefficients between selected variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bio1</td>
<td>Annual Mean Temperature</td>
</tr>
<tr>
<td>Bio2</td>
<td>Mean Diurnal Range (Mean of monthly (max temp – min temp))</td>
</tr>
<tr>
<td>Bio3</td>
<td>Isothermality (BIO2/BIO7) (*100)</td>
</tr>
<tr>
<td>Bio4</td>
<td>Temperature Seasonality (standard deviation *100)</td>
</tr>
<tr>
<td>Bio5</td>
<td>Max Temperature of Warmest Month</td>
</tr>
<tr>
<td>Bio6</td>
<td>Min Temperature of Coldest Month</td>
</tr>
<tr>
<td>Bio7</td>
<td>Temperature Annual Range (BIO5-BIO6)</td>
</tr>
<tr>
<td>Bio8</td>
<td>Mean Temperature of Wettest Quarter</td>
</tr>
<tr>
<td>Bio9</td>
<td>Mean Temperature of Driest Quarter</td>
</tr>
<tr>
<td>Bio10</td>
<td>Mean Temperature of Warmest Quarter</td>
</tr>
<tr>
<td>Bio11</td>
<td>Mean Temperature of Coldest Quarter</td>
</tr>
<tr>
<td>Bio12</td>
<td>Annual Precipitation</td>
</tr>
<tr>
<td>Bio13</td>
<td>Precipitation of Wettest Month</td>
</tr>
<tr>
<td>Bio14</td>
<td>Precipitation of Driest Month</td>
</tr>
<tr>
<td>Bio15</td>
<td>Precipitation Seasonality (Coefficient of Variation)</td>
</tr>
<tr>
<td>Bio16</td>
<td>Precipitation of Wettest Quarter</td>
</tr>
<tr>
<td>Bio17</td>
<td>Precipitation of Driest Quarter</td>
</tr>
<tr>
<td>Bio18</td>
<td>Precipitation of Warmest Quarter</td>
</tr>
<tr>
<td>Bio19</td>
<td>Precipitation of Coldest Quarter</td>
</tr>
</tbody>
</table>

Correlation is significant at the 0.01 level (2-tailed).
Figure 3-1. Geographic location of the study area and its counties, NE Iran. The map inset (B) illustrates the location of presence/absence sampling of *Ph. papatasi*. 
Figure 3-2. Methodology flowchart used in this study

Figure 3-3. The ROC curves and AUCs for SVM, RF and LR classifiers
Figure 3-4. Comparison of the SVM, LR and RF classifiers. Overall accuracy, AUC, Kappa index, sensitivity, specificity, PPV and NPV expressed as %.
Figure 3-5. Habitat suitability map of *Ph. papatasi* in Golestan province, northeast Iran based on support vector machine (note: this map is based on 1km units).
Tuberculosis (TB) is a contagious disease caused by *Mycobacterium tuberculosis*. The disease is primarily transmitted through the respiratory route by coughing or sneezing. The disease mostly attacks the lungs but can also affect other organs such as kidney and brain. It can promote the course of human immunodeficiency virus (HIV) infection into acquired immune deficiency syndrome (AIDS). According to the World Health Organization (WHO) global TB report, it is estimated that 10.4 million incident cases in 2016 developed the disease, of which almost 1.7 million patients died. This agency has ranked TB as the leading cause of death among HIV patients, the most common killer from a single infectious agent, and the 9th leading cause of death, worldwide.

According to the statistics by the WHO, most TB cases (>90%) are reported in developing countries; however, it can also occur in developed countries. Despite efforts to eradicate TB in the US, the disease remains a major public health challenge. In 2016, more than 9200 TB cases were reported in various parts of the US, which placed the disease among the top notifiable infectious diseases in the country. Although the frequency of TB has decreased in recent years, it is not expected that the US will achieve the goal of TB elimination in this century.

There are many factors that influence the spatial distribution of TB, which has made the disease a multidimensional and complex public health problem.

---

Previous researches from different parts of the world have demonstrated that TB transmission is related with various individual factors, for example, age, gender, education level, race, migration, drinking alcohol, and presence of diseases (such as HIV and diabetes) \(^{211-213}\). Moreover, at the ecological level, factors such as climate, altitude, air pollution, economic level, unemployment rate, and poverty have found significant on TB occurrence \(^{214,215}\). One of the major drawbacks of the highly applied traditional statistical models in the study of TB is that these models are often based on several hard-to-meet assumptions \(^{191,91}\). This can bias the estimations of TB frequency/incidence rate \(^{216}\). For instance, some assumptions of the linear regression (LR) model are normality of all variables, the linear relationship between inputs and output, constant variance of errors, and little or no multicollinearity. They also often need a complete and/or long-term recorded dataset to achieve unbiased estimations \(^{217}\). On the other hand, machine learning techniques (MLTs) may lead to appropriate estimations even with noise-contaminated and incomplete data \(^{218,219}\). As advanced tools, MLTs have been successfully used in analyzing and modeling various complex environmental disciplines, including in ecology, geography, biomedicine, and epidemiology \(^{24,159,220,221,166}\). The growing popularity of MLTs can be attributed to their abilities to approximate almost any complex non-linear functional relationship \(^{222-224}\). Despite their capabilities in working with noisy and incomplete data as in most epidemiological studies, they have been underused in spatial epidemiology \(^{169,225}\).

Inspired by human neural processing, artificial neural networks (ANNs) are among the most popular MLTs used in recent years in environmental studies. Artificial neural networks have a large number of highly interconnected processing elements...
(neurons) working in unison to solve specific problems. Compared to the traditional statistical models, ANNs are independent of the statistical distribution of data and do not require a priori knowledge about the data for deriving patterns. ANNs are simplified mathematical models that can map the relationship between input and output layers by receiving several examples (training data). A properly trained network can further be used to predict outcome(s) from new data (test data).

There have been few published ANN architectures in spatial modeling of infectious diseases, worldwide. Aburas et al. applied an ANN model with a back-propagation algorithm to predict the frequency of confirmed dengue cases using Singaporean National Environment Agency data. Results of their model showed a correlation coefficient of 0.91 between actual and predicted values. They also identified influential environmental factors predicting the number of dengue cases including mean temperature, mean relative humidity, and total rainfall. Laureano-Rosario et al. used ANN to predict dengue fever (DF) occurrence in a region in Puerto Rico, and in several coastal municipalities in Mexico. The developed ANN models were trained with 19 years of DF data for Puerto Rico and six years’ data for Mexico. Sea surface temperature, precipitation, air temperature, humidity, previous DF cases, and population size were used as explanatory variables. Their results showed that the ANN successfully modeled DF outbreak occurrences with the overall power of 70% in both areas. The variables with the most influence on predicting DF outbreak were population size, previous DF cases, air temperature, and date. In the US, Xue et al. developed the least square (LS) regression analysis and a neural network trained by the genetic algorithm to evaluate influenza activity in 10 geographic regions. They compared the
models using three evaluation metrics: mean square error, mean absolute percentage error, and relative mean square error. All three evaluation indices used in their study were lower than the corresponding metrics for the LS regression model showing the superiority of the genetic algorithm-based neural network to the LS regression model.

To date, TB control efforts have relied on the empirically developed WHO DOTS (directly observed treatment, short-course) control strategy which focuses on “case-finding” rather than “place-finding” 230. Previous studies from different parts of the US and at different levels have shown the association of TB with socio-economic status. Mullins et al. used purely spatial scan statistics to identify spatial clusters of census tracts with high TB prevalence rates in Connecticut. They found six clusters of TB containing 126 census tracts 231. Persons in these clusters were more likely to be black non-Hispanic and less likely to be Asian. Bennett et al. used multivariate logistic regression to assess the association between demographic and clinical characteristics and latent TB infection in refugees in San Diego, California 232. They found that the highest prevalence rate was among refugees from sub-Saharan Africa and those with less education. In a study in Harris County (in Texas), Feske et al. showed a positive association between the percent of individuals using public transportation in census tract and location of clusters detected by Getis–Ord’s Gi* hotspot analysis 233.

Ecological level researches on TB incidence rate at the national level are inadequate for epidemiologic inferences especially in the US. Therefore, it is crucial to perform an ecological study across the continental US to identify the location of statistically significant hotspots and to determine the relationship between environmental and socio-economic factors and TB incidence to provide useful insight to
policymakers in planning for TB control at a larger scale. To our knowledge, no study has utilized ANNs in modeling the geographic distribution of TB incidence rate in the US. Integration of the GIS and ANN can improve policymakers’ insight in identifying potential TB high-risk areas and risk factors useful for future mitigation efforts. We examined the spatial distribution of the disease and applicability of MLTs in TB modeling with the following assumptions (1) all reported county-level TB incidence rates represent the status of TB in the continental US and (2) the TB incidence rate is influenced by environmental and socioeconomic factors.

**Material and Methods**

**Tuberculosis Data**

Data on all reported TB cases in the continental US between 2006 and 2010 were obtained from the paper of Scales et al. in the American Journal of Preventive Medicine. All data are at the county level (n = 3109) and are publicly available. Latent TB cases were not reported and included in this study. To alleviate variations of TB incidences, particularly in counties with a small population size such as Loving and King counties in Texas (n < 300 populations), the cumulative incidence was calculated (2006–2010). For this purpose, five-year corresponding population estimates from the American Community Survey (ASC) were used. Tuberculosis incidence rates were imported into ArcGIS 10.5 (ESRI, Redlands, CA, US) and geocoded at the county level.

**Explanatory Data**

In this study, 278 environmental and socio-economic factors were collected from various sources and considered as explanatory variables based on previous studies and domain knowledge. From the Center for Disease Control and Prevention (CDC) wonder database, climate data were derived including daily maximum and minimum
air temperature (°F), and daily maximum heat index (°F). Also, the average number of
diabetes cases during the study period were obtained from this database.
Topographic data including minimum, maximum and mean of altitude and slope were
obtained from the national map website. The county-level values of these data were
calculated using zonal statistics in ArcGIS Spatial Analyst extension. Socioeconomic
data were acquired from the US Census Bureau. A broad range of socio-economic
factors including age group, agriculture, immigration, education level, employment rate,
health, Hispanic or Latino population, income, poverty, and race were obtained across
the nation from this database. All population-based predictors were normalized to the
county’s population size. The full dataset description of exploratory variables is in
Supplementary Table. All data used in this study are downloadable from the above
sources.

Global and Local Clustering

After mapping TB incidence rates at the county level, the empirical Bayes
smoothing method was implemented in the GeoDa software to adjust the crude
incidence rates toward the global mean. This helps to reduce the variance instability
associated with counties with a small population size. Thus, the response variable
changed to (logged) smoothed TB incidence rate (STIR) rather than the crude incidence
rate with more than 900 counties with 0 value which makes modeling difficult. All these
counties have non-zero STIR values. Next, the spatial pattern was statistically
evaluated using the global Moran’s I and Getis–Ord General G.

Global Moran’s I measures the similarities between the TB incidence rates of
neighboring counties as follows:
\[ I = \frac{n \sum_{i=1}^{n} \sum_{j=1,j \neq i}^{n} w_{ij} z_{i} z_{j}}{s_{0} \sum_{i=1}^{n} z_{i}^{2}} \]  

(4-1)

\[ s_{0} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \]  

(4-2)

where \(z_{i}\) and \(z_{j}\) are the deviations of STIR for counties \(i\) and \(j\) from average incidences (i.e., \((x_{i} - \bar{x}), (x_{j} - \bar{x})\)), respectively; \(w_{ij}\) is the spatial weight based on Rook’s contiguity (i.e., common borders between counties \(i\) and \(j\)); and \(s_{0}\) is the aggregation of all spatial weights.

Moreover, the Getis–Ord General \(G\) statistics, developed by Getis and Ord was used as a measure of clustering of the high or low value of STIR \(^{57}\). A positive or negative Z-score for \(G\) indicates spatial clustering of high (hotspot) or low (coldspot) values, respectively. The formula for the general \(G\) of spatial association is:

\[ G = \frac{\sum_{i=1}^{n} \sum_{j=1,j \neq i}^{n} w_{ij} x_{i} x_{j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \forall j \neq i \]  

(4-3)

Getis–Ord \(Gi^{*}\) statistics \(^{66}\) was applied on the smoothed rates to identify the locations of statistically significant hotspots of the STIR (\(p<0.05\)). Using the same notation as in Equations (1) and (2), this statistic is computed as follows \(^{233,114}\):

\[ Gi^{*}_{i} = \frac{\sum_{j=1}^{n} w_{ij} x_{j} - \bar{X} \sum_{j=1}^{n} w_{ij}}{S \sqrt{\frac{[n \sum_{j=1}^{n} w_{ij}^{2} - (\sum_{j=1}^{n} w_{ij})^{2}]}{n - 1}}} \]  

(4-4)

\[ S = \sqrt{\frac{\sum_{j=1,j \neq i}^{n} (x_{j} - \bar{x})^{2}}{n - 1} - \bar{x}^{2}} \]  

(4-5)
Artificial Neural Networks

An ANN is a computational model, which consists of several simple processing elements called neurons. The neurons are usually structured in layers: the input layer, the hidden layer(s) and the output layer. In this study, we used ANNs with one and two hidden layers, however, theoretical research has shown that almost any complex and non-linear function can be estimated by an ANN with a single hidden layer. Additionally, having more hidden layers increases the number of parameters, which may lead to over-fitting (i.e., memorizing the data while training). The aim of the hidden layer is to find a multi-dimensional expansion of the input layer, which can be better transformed to the pattern in the output layer. The neurons in the input layer are connected to all neurons in the hidden layer. Similarly, all the neurons in the hidden layer are connected to every neuron in the output layer (Figure 4-1). In this system, selected explanatory variables are fed to the input layer and passed through the hidden layer that processes them using simple mathematical operations. The relationship between input and output layers of ANNs is trained by observing a series of known examples from the training dataset and adjusting the weights accordingly. Once the training phase is completed, the network is usually able to generalize what it has learned to the test data with similar attributes of input. We developed a multi-layer perceptron (MLP) neural network with one and two hidden layers to approximate the dependency of log (STIR) in the continental US to environmental and socio-economic factors.

Multi-layer perceptron is the most commonly used ANN structure in environmental modeling. During the training phase of MLP, each input feature is multiplied by its corresponding weight. The results are then summed and passed
through a smooth non-linear activation function to produce the output. The Logistic function and hyperbolic tangent are among the widely used activation functions. In a supervised learning setup, the difference between the (MLP) output and actual output/target (i.e., the error or cost function) can be calculated as in Equation (5): 

\[
\text{error} = \frac{1}{2} \sum_{i=1}^{n} (t_i - o_i)^2
\]  

Equation (4-6)

where \( o \) and \( t \) are model output and target respectively, and \( n \) is the training sample size. To minimize the cost function, a back-propagation algorithm based on stochastic gradient descent is used to adjust each weight in the MLP model. The updated value for each weight is calculated as in Equation (6):

\[
w^{(k+1)} = w^{(k)} - \alpha \frac{\partial \text{error}}{\partial w^{(k)}}
\]  

Equation (4-7)

where \( \alpha \) or learning rate controls how much the coefficients can change on the \( k^{th} \) update. More detailed information about back-propagation MLP is presented in.

**Model Pre-Processing**

To develop the models, the entire dataset was randomly divided into three different partitions: (1) training data: to learn and update the weights and biases in network (60% of the total data) (2) cross-validation data: to avoid overfitting problem by tuning models' parameters during training phase (15% of the total data) (3) test data: to evaluate accuracy and predictive power of the network after the training process (25% of the total data). The spatial distribution of training, cross-validation, and test data as shown in Figure 4-2. For the purpose of comparison, we used the same training, cross-validation and test data for all developed models.
The next step was to standardize the input data before using them in the ANN models. Input data have different ranges and units, this pre-processing step can enhance the performance of models by faster convergence\textsuperscript{24,252}. There are several standardization formulas presented in the literature. Here, we use the following equation which transforms the input data to the range of 0 to 1 as in Equation (8):

$$X_s = \frac{X_i - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$  \hspace{1cm} (4-8)

where $X_i$ is the initial (actual) value; $X_{\text{min}}$ and $X_{\text{max}}$ are the minimum and maximum of the initial values and $X_s$ is the respective standardized value. Moreover, after training the network, the output of the model is returned to the original form through the Equation (9):

$$X_i = X_s \times (X_{\text{max}} - X_{\text{min}}) + X_{\text{min}}$$  \hspace{1cm} (4-9)

Multi-layer perceptron and linear regression (LR) models are sensitive to redundant explanatory variables because noise reduces model accuracy and generalizability. Thus, a ‘proper’ selection of independent variables is crucial for better performance. We applied L1-regularization or least absolute shrinkage and selection operator (LASSO) on the training and cross-validation dataset before developing the models. This process produces a sparse solution (i.e., few non-zero coefficients) which reduces overfitting and enhances interpretability of the results\textsuperscript{253}. Detailed information of L1 regularization technique can be found in Park and Hastie\textsuperscript{179}.

In the MLP model, 8 and 1 neurons were used in the input and output layers, respectively. These numbers correspond to the 8 explanatory variables selected during L1 regularization and one response variable (i.e. log (STIR)). There is no deterministic
rule to determine the number of neurons in the hidden layer. The grid search was used to tune hyper-parameters in MLP with one and two hidden layer(s). This approach systematically evaluates the developed model for each combination of model parameters. For MLP, the tangent hyperbolic activation function \( \phi(x) = \frac{1-e^{-x}}{1+e^{-x}} \), a non-linear and symmetric function which maps any real value to \([-1,1]\), was used in the hidden layer(s) \(^{254}\). In addition, a linear identity function \( f(x) = x \) was applied as an activation function in the output layer. All computer codes were developed in the Python programming language.

**Model Evaluation**

We computed three types of evaluation metrics to assess and compare the generalization capability of MLP with LR in predicting \( \log(\text{STIR}) \) in the continental US. These statistics include root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (R) between model output and ground-truth.

Due to the fully connected architecture of MLP, it is difficult to define the explicit relationship between input and output variables by coefficients \(^{255}\). Sensitivity analysis is a common way to address this problem. In this analysis, each factor was excluded from the model, individually, and the RMSE of resulting models were compared. The most influential factor, among the selected features, is the one that its absence increases the RMSE of the model the most. Using sensitivity analysis, we identified the most influential factors in predictions of the log (STIR). Finally, we ranked them according to their decreasing importance.
Results

Between 2006 and 2010, 64,496 TB cases were reported across the continental US. The number of TB cases showed a consistent declining trend from 14,119 to 11,284 annual cases (Figure 4-3). Among the states, the highest average TB incidence rates were identified in Louisiana (5.30 cases per 100,000), Arizona (5.05 per 100,000) and Georgia (4.7 per 100,000).

The global Moran’s *I* and general *G* indicated significant spatial clustering of STIR in the continental US for the study period (Moran’s *I* = 0.13, Z-score = 32.13, p<0.005; General *G* = 0.002, Z-score = 15.3, p<0.005). The hotspot analysis (Getis–Ord Gi*) identified that about 7% of the continental US counties (*n* = 216) were part of hotspots. The hotspots of STIR were distributed unequally, almost restricted to the southern half of the country and particularly in the southern and southeastern counties of the US (Figure 4-4). Table 4-1 summarizes the top 10 states with the largest number of counties detected as part of STIR hotspots by the Getis–Ord Gi* technique.

Based on the results of L1 regularization, the environmental and several socio-economic factors were selected. Out of 278 explanatory variables, only 8 factors were incorporated as input variables for the models: (1) RHI820: resident population: not Hispanic, white alone (July 1-estimate) (proportion of county population); (2) LFE330: employed persons by industry (NAICS)-agriculture, forestry, fishing and hunting, and mining (proportion of county population); (3) minimum temperature; (4) POP778: year of entry by citizenship status in the United States entered 2000 or later-foreign-born (proportion of county population) (5) IPE110: people of all ages in poverty (proportion of county population); (6) SPR440: social security-benefit recipients (proportion of county population); (7) HIS305: Hispanic or Latino persons, educational attainment, 25 years
and over, male (proportion of county population); (8) POP730: population one year and
over by residence-moved from different county within same state 2005–2009
(proportion of county population).

All Pearson correlation values among the selected variables were under 0.5, thus
we considered the selected factors as relatively uncorrelated (Table 4-2). The log
(STIR) was positively correlated with all variables except for household income
(p<0.05).

Preliminary results of the developed LR model showed that the selected
predictors generated R = 0.666, R² = 0.443, and F = 184.246 (Sig. F Change < 0.001).
The R value which represents the simple correlation between predictions and reality
indicated an acceptable degree of correlation. The R² value indicates that 44.3% of total
variations in the log (STIR) can be explained by the predictors. The F-test was
significant showing the developed LR model as a whole has statistically significant
predictive capability. Durbin–Watson test was close to 2 which verifies independency of
errors assumption (Durbin–Watson statistic = 2.04) (Table 4-3).

The t-test showed that all selected variables are statistically significant at a 99%
confidence interval. Based on the standardized coefficients, among the variables, in
order of strength, “RHI820”, “LFE330”, “HIS305”, “SPR440”, and “POP730” have
negative impacts on the log (STIR) while the variables “POP778”, “Min Temp”, and
“IPE110” have positive impacts on the dependent variable, respectively. Tolerance and
the variance inflation factor (VIF) were used as two collinearity diagnostic tests to
assess multicollinearity level for all variables. As the values of VIF didn’t exceed 3 and
tolerance statistic were above 0.1, it seems that there is no cause for concern about collinearity (Table 4-4).

The normal probability-probability plot (P-P Plot) of LR residuals showed that data points were closely aligned with the diagonal line suggesting the distribution of the residuals was almost normal. This indicated that the assumption of normality of errors was almost met. Nevertheless, there were a few samples which departed from the diagonal line (Figure 4-5).

Table 4-5 presents the performance of MLP (1-hidden layer), MLP (2-hidden layer), and LR models, in terms of MAE, RMSE, and R, for training, cross-validation and test datasets, respectively. The correlation coefficient for a single hidden layer MLP, with 20 nodes in the hidden layer, was larger than double hidden layers MLP, and LR (Table 4-5), which showed a better agreement between the predicted and the ground-truth. In addition, RMSE in MLP (single layer) was lower than the other models. However, MAE in MLP with a two-hidden layer (20 nodes in first and 10 nodes in second hidden layers) had the same test errors as single layer MLP. Results suggest that the single layer MLP model outperformed the rest of models with higher generalizability for predicting the log (STIR) in the continental US. Similarly, double hidden layer MLP outperformed the LR model.

The scatter plot between the output of the MLP model and the corresponding observed log (STIR) (i.e., ground-truth values) for test data (Figure 4-6) showed that the model was able to predict the average variations of log (STIR), while it was unable to predict some counties with exceptional rates.
The best model accuracy was achieved by the single layer MLP compared with the other models. We examined the contribution/relative importance of each input feature on the log (STIR) using sensitivity analysis. The results revealed that the highest RMSE occurred when “resident population of American Indian and Alaska native” was removed from the model, which implies that this factor has the maximum contribution, among the selected factors, in predicting log (STIR) at the county level in the continental US. The most influential factors in order of contributions were: “RHI820: resident population: not Hispanic, White alone (July 1-estimate) (proportion)”, “LFE330: employed persons by industry (NAICS)-agriculture, forestry, fishing and hunting, and mining (proportion)”, “Minimum Temperature”, “POP778: year of entry by citizenship status in the United States entered 2000 or later-foreign-born (proportion)”, “IPE110: people of all ages in poverty (proportion)”, “SPR440: supplemental security income-average monthly payments per recipient”, “HIS305: Hispanic or Latino persons, educational attainment, 25 years and over, male (proportion)” and “POP730: population one year and over by residence-moved from different county within same state 2005-2009 (proportion)”. (Figure 4-7).

Discussion

According to the Institute of Medicine (2000), eliminating TB will require the development of new tools to identify risk factors and high-risk areas. As expressed by Feske et al., effective TB elimination in the US would require geographic elucidation of high-risk areas and systematic surveillance of location-based risk factors. In this study, we combined more advanced tools to examine the relationship between environmental and socio-economic factors, and the log (STIR) across the continental US. Integration of GIS, spatial statistics, and ANNs resulted in an efficient multi-
disciplinary approach, which can provide helpful guidelines for health decision makers. The benefits obtained from this approach can enhance mitigation efforts such as budget allocation, educating people who live in high-risk areas and drug distribution. Due to limited research on the spatial modeling of TB at the national level, our study can be regarded as a basis for future nationwide TB program researches.

In this study, we examined the spatial pattern and hotspots of the STIR in the continental US. Moran’s I and General G statistic were used to investigate the presence of spatial autocorrelation of local STIR. Our results showed that the distribution of STIR at the county levels is clustered at the county level (p<0.05). We then conducted the hotspot analysis to identify the counties with statistically significant STIR using the hotspot analysis (Getis–Ord Gi*) approach. Our findings showed that STIR hotspots were concentrated in the southern and western states (Figure 4-4); however, the South, Southeast, and Southwest counties of the country were more severely affected. The concentration of hotspots suggests that there were more cases observed in these areas that would be expected if everyone were equally at risk.

Visual comparison of the location of identified hotspots in some states which had a very high proportion of counties falling into hotspots (such as Georgia and Florida) with recent surveillance reports of Georgia\textsuperscript{257} and Florida\textsuperscript{258} showed pieces of evidence of similarities with some differences. This suggests the stability of location of counties falling into hotspots which require close attention. Since after detecting the statistically significant hotspots of STIR in the region, associated risk factors were not known, we used ANN to model the relationship between environmental and socio-economic factors and the log (STIR).
Based on sensitivity analysis, the environmental factors identified the contribution of the selected variables to the log (STIR) at the county level. We found that the average daily minimum temperature (with a positive effect) was an important climate factor in the log (STIR). This is probably due to the adverse effects of minimum air temperature on the respiratory system of patients, and more close lifestyle of people in cold weather which increases the risk of exposure to infectious agents. Similar results were reported in other studies. In a time-series analysis study in Fukuoka (Japan), Onozuka and Hagihara found a positive significant relationship between extreme cold temperature and TB incident cases. Our results are also consistent with the findings of Mourtzoukou et al. and Khalid et al. in Pakistan.

Sensitivity analysis showed that economic factors are important for log (STIR) in the continental US. One of the most important economic factors was “proportion of population county of all ages in poverty” which suggests that underserved segments of the population are at higher-risk of STIR in the US. Conversely, counties with a higher proportion of the population employed by industry or higher security income had lower STIR. These factors potentially describe the impact of poverty/deprivation on lifestyle choice. These findings agree with the individual-level studies of McKenna et al., Ho, Weis et al., and Unemployment rate has been found to be an important factor in TB transmission in the studies of Munch et al. who indicated the strongest correlation with TB caseload, and the studies of Dos Santos and Jackubowiak; while Sun et al. did not find it significant in China. In a cohort study in Georgia, Djibuti et al. showed that TB patients with lower-income households are at higher risk of poor TB treatment. Bamrah et al. analyzed the genotyping of homeless persons in the US,
1994–2010. Their results showed that homeless TB patients had an approximately 10-fold increase in TB incidence and had more than twice the odds of not completing treatments.

Results of this study also showed the importance of race distribution and STIR as there was a strong negative relationship between the proportion of county population who are white and STIR. Visual comparison of the locations of the hotspots (Figure 4-4) with the distribution map of race and ethnicity provided by US Census Bureau, confirms that, in general, counties with more than 80% of the white population didn’t fall into the hotspots. This agrees with the findings of Cantwell et al. and CDC report.

Artificial neural networks and GIS were effective in modeling log (STIR), but several limitations exist. First, the data used in this study were collected from multiple online sources. It should be noted that TB reported data are subject to spatial differences in case detection, thus, a standard passive case-finding approach needs to be considered. In addition, methods of data collection and preparation are different which may result in biased estimations. In this study, we only included active reported TB cases. It should be noted that there are many people with latent TB infection who have not developed TB disease, and therefore, are not detected or counted as cases. This is more prevalent (exists in greater proportions) among those from other countries. The last limitation stems from the study design. Our study at the county level should be considered to characterize population rather than individual-level characteristics of risk (ecological fallacy). Thus, the findings of this study should be applied to population-level targeting rather than considering individual treatment.
This study showed machine learning techniques in spatial modeling can be applied to TB incidence rate across the continental US. For future works, we recommend conducting researches at multiple scales, particularly at finer scales such as at the census tracts or block group level for more focal interventions. However, as most policy decisions on TB control are performed at the state and federal government levels, this represents a meaningful first step. To optimize the structure of ANNs, choice of parameters (e.g., learning rate, activation functions) and training the weights in ANN, and heuristic algorithms such as genetic algorithm may be useful leading to global optima, because it can help to escape from local optima\textsuperscript{272}. Also, hyper-parameter tuning coupled with more recent ANN approaches in terms of activation function or optimization (e.g., Adam or other adaptive methods) is highly recommended. We also recommend incorporating genetics data into the model or examine the genetic characteristics of individual patients in counties with high modeling errors. In this regard, national TB genotyping surveillance coverage which refers to the proportion of TB cases with culture-positive with at least one genotyped isolate in the US can be useful. Also, other statistical techniques such as Poisson model (through including the size of local population at risk as an offset in the model) or the logistic regression model (through binomial distribution) can be investigated in modeling TB incidence rate. Coupled with GIS, the ANNs techniques successfully identified some determinant factors of log (STIR), ranked them, and had better prediction ability than the traditional linear regression analysis. The findings of this study can provide useful insight to health authorities on prioritizing resource allocation to risk-prone areas.
Tables

Table 4-1. Top 10 states with the largest number of hotspot counties \((p < 0.10)\) of smoothed tuberculosis (TB) incidence rate (STIR) in the continental US, 2006–2010.

<table>
<thead>
<tr>
<th>Rank</th>
<th>State</th>
<th>No. Hotspot Counties</th>
<th>Percentage (#hotspots/#counties)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Georgia</td>
<td>57</td>
<td>35.8%</td>
</tr>
<tr>
<td>2</td>
<td>Texas</td>
<td>30</td>
<td>11.8%</td>
</tr>
<tr>
<td>3</td>
<td>North Carolina</td>
<td>23</td>
<td>23.0%</td>
</tr>
<tr>
<td>4</td>
<td>Louisiana</td>
<td>22</td>
<td>34.3%</td>
</tr>
<tr>
<td>5</td>
<td>Florida</td>
<td>20</td>
<td>29.9%</td>
</tr>
<tr>
<td>6</td>
<td>California</td>
<td>17</td>
<td>22.7%</td>
</tr>
<tr>
<td>7</td>
<td>South Carolina</td>
<td>17</td>
<td>37%</td>
</tr>
<tr>
<td>8</td>
<td>Arkansas</td>
<td>12</td>
<td>16.0%</td>
</tr>
<tr>
<td>9</td>
<td>Mississippi</td>
<td>12</td>
<td>14.6%</td>
</tr>
<tr>
<td>10</td>
<td>Alabama</td>
<td>10</td>
<td>14.9%</td>
</tr>
</tbody>
</table>

Table 4-2. Pearson correlation analysis between selected variables for modelling STIR, continental US.

<table>
<thead>
<tr>
<th></th>
<th>POP730</th>
<th>LFE330</th>
<th>IPE110</th>
<th>POP778</th>
<th>Min Temp</th>
<th>SPR440</th>
<th>HIS305</th>
<th>RHI820</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP730</td>
<td>1.000</td>
<td>0.051</td>
<td>0.041</td>
<td>0.064</td>
<td>-0.124</td>
<td>-0.078</td>
<td>-0.138</td>
<td>-0.024</td>
</tr>
<tr>
<td>LFE330</td>
<td>0.051</td>
<td>1.000</td>
<td>0.018</td>
<td>0.057</td>
<td>0.136</td>
<td>-0.499</td>
<td>-0.186</td>
<td>-0.040</td>
</tr>
<tr>
<td>IPE110</td>
<td>0.041</td>
<td>0.018</td>
<td>1.000</td>
<td>0.266</td>
<td>-0.231</td>
<td>-0.108</td>
<td>0.066</td>
<td>0.384</td>
</tr>
<tr>
<td>POP778</td>
<td>0.064</td>
<td>0.057</td>
<td>0.266</td>
<td>1.000</td>
<td>-0.005</td>
<td>0.091</td>
<td>-0.390</td>
<td>0.248</td>
</tr>
<tr>
<td>Min Temp</td>
<td>-0.124</td>
<td>0.136</td>
<td>-0.231</td>
<td>-0.005</td>
<td>1.000</td>
<td>0.066</td>
<td>-0.032</td>
<td>0.308</td>
</tr>
<tr>
<td>SPR440</td>
<td>-0.078</td>
<td>-0.499</td>
<td>-0.108</td>
<td>0.091</td>
<td>0.066</td>
<td>1.000</td>
<td>-0.015</td>
<td>0.003</td>
</tr>
<tr>
<td>HIS305</td>
<td>-0.138</td>
<td>-0.186</td>
<td>0.066</td>
<td>-0.390</td>
<td>-0.032</td>
<td>-0.015</td>
<td>1.000</td>
<td>0.403</td>
</tr>
<tr>
<td>RHI820</td>
<td>-0.024</td>
<td>-0.040</td>
<td>0.384</td>
<td>0.248</td>
<td>0.308</td>
<td>0.003</td>
<td>0.403</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 4-3. Results of linear regression (LR) model for modeling log (STIR), continental US.

<table>
<thead>
<tr>
<th>Model</th>
<th>R Square</th>
<th>Adj. R Square</th>
<th>R Square Change</th>
<th>F</th>
<th>df1</th>
<th>df2</th>
<th>Sig.</th>
<th>Durbin–Watson</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR</td>
<td>0.666 a</td>
<td>0.443</td>
<td>0.440</td>
<td>184.246</td>
<td>8</td>
<td>1854</td>
<td>0.000</td>
<td>2.041</td>
</tr>
</tbody>
</table>

a. Predictors: (Constant), POP73, LFE330, IPE110, POP778, Min Temp, SPR440, HIS305, RHI820.
Dependent Variable: log (STIR).
### Table 4-4. Effects of environment and socio-economic factors on the log (STIR) using LR model.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Unstandardized Coefficient</th>
<th>Standardized Coefficient</th>
<th>t</th>
<th>Sig.</th>
<th>95.0% Confidence Interval for B</th>
<th>Collinearity Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>0.001</td>
<td></td>
<td>0.009</td>
<td>0.993</td>
<td>-0.198 to 0.200</td>
<td></td>
</tr>
<tr>
<td>RHI820</td>
<td>-0.007</td>
<td>-0.294</td>
<td>-11.11</td>
<td>0.000</td>
<td>-0.009 to -0.006</td>
<td>0.429</td>
</tr>
<tr>
<td>LFE330</td>
<td>-0.023</td>
<td>-0.166</td>
<td>-7.929</td>
<td>0.000</td>
<td>-0.029 to -0.017</td>
<td>0.683</td>
</tr>
<tr>
<td>Min Temp</td>
<td>0.013</td>
<td>0.210</td>
<td>9.809</td>
<td>0.000</td>
<td>0.010 to 0.016</td>
<td>0.653</td>
</tr>
<tr>
<td>POP778</td>
<td>0.083</td>
<td>0.282</td>
<td>12.621</td>
<td>0.000</td>
<td>0.070 to 0.095</td>
<td>0.602</td>
</tr>
<tr>
<td>IPE110</td>
<td>0.012</td>
<td>0.140</td>
<td>6.662</td>
<td>0.000</td>
<td>0.008 to 0.015</td>
<td>0.677</td>
</tr>
<tr>
<td>SPR440</td>
<td>-0.009</td>
<td>-0.097</td>
<td>-4.703</td>
<td>0.000</td>
<td>-0.013 to -0.005</td>
<td>0.701</td>
</tr>
<tr>
<td>HIS305</td>
<td>-0.019</td>
<td>-0.145</td>
<td>-5.976</td>
<td>0.000</td>
<td>-0.026 to -0.013</td>
<td>0.508</td>
</tr>
<tr>
<td>POP730</td>
<td>-0.015</td>
<td>-0.080</td>
<td>-4.489</td>
<td>0.000</td>
<td>-0.021 to -0.008</td>
<td>0.950</td>
</tr>
</tbody>
</table>

VIF: Variance inflation Factor.

### Table 4-5. Comparison of multi-layer perceptron (MLP; one and two hidden layers), and LR model’ performance for predicting log (STIR) in the continental US.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Cross-Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>RMSE</td>
<td>R</td>
</tr>
<tr>
<td>LR</td>
<td>0.27</td>
<td>0.35</td>
<td>0.66</td>
</tr>
<tr>
<td>MLP (1 hidden layer)</td>
<td>0.25</td>
<td>0.33</td>
<td>0.70</td>
</tr>
<tr>
<td>MLP (2 hidden layers)</td>
<td>0.26</td>
<td>0.34</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Figure 4-1. Topological architecture of multi-layer perceptron neural network (MLPNN) used in this study. 

Figure 4-2. Spatial distribution of training, cross-validation, and test data used for modeling log (STIR).
Figure 4-3. The frequency of TB cases (left) and the cumulative TB incidence rate (right) across the continental US (2006–2010).

Figure 4-4. Hotspot map for the STIR in the continental US identified by hotspot analysis (Getis–Ord Gi*) technique, 2006–2010.
Figure 4-5. The Normal P-P Plot of LR model.
Figure 4-6. Scatter plot of observed and predicted log (STIR) (by single hidden layer MLP model) for test data in the continental US.

Figure 4-7. The contribution of input features on predicting log (STIR) according to sensitivity analysis of single hidden layer MLP. RMSE: Root mean square error.
CHAPTER 5
CONCLUSION

Summary

Despite a significant growth in spatial analysis of various types of infectious diseases, most studies remain non-spatial. Our primary focus of this research was to study spatial aspects of (three different) infectious diseases. Our studies confirmed that “geography” is an essential part of the infectious diseases. In other words, we showed that LD hotspots in CT occurred in specific geographic towns. Also, using geographic factors we could predict presence/absence of one of main vector of ZCL and TB incidence with a decent accuracy. Thus, it is necessary to incorporate it in future control programs and predictions to fight against the diseases.

Due to interdisciplinary nature of the most infectious diseases, we underscored innovative techniques such as RS, machine learning and GPS in a GIS platform and as cost-effective and time saving approaches. Our main intent was to make contribution in predicting presence/absence of the vector(s) that cause the disease and predicting incidence of the diseases with higher accuracy than highly-applied traditional techniques in literature. We examined a range of machine learning techniques that are widely-used in many disciplines but underutilized in spatial epidemiology.

In Chapter 2, using ESDA in a GIS framework, we analyzed resulting pattern of LD distribution in CT. We identified the towns where controlling LD was most required and could provide the highest benefits in allocating resources. In Chapter 3, we promoted machine learning classifiers such as random forest and support vector machine as powerful algorithms in binary classifications tasks. However, there are many other machine learning classifiers that have not been included in the research such as
k-means and Naïve Bayes classifier. Most of the binary classifiers are transferable to medical geography and are likely to have better prediction abilities than traditional statistical techniques. We therefore recommend medical geographers to develop and evaluate novel models in infectious disease research to examine the applicability of techniques that better predict the presence/absence of other vectors or disease morbidity/mortality rates.

In Chapter 4, the results of ANN showed an improvement of prediction accuracy of continuous health outcome (TB rate) compared to the classic linear regression. ANN coupled with GIS provided insight about the heterogeneity of TB in the US. We applied the data science technique in the context of national level control strategy which are rarely implemented in nationwide control programs of infectious diseases. However, less insight is obtained in county level studies compared to individual level studies. Our search of ANN in spatial epidemiology indicated that very few studies have applied the technique. We recommend examining other robust and novel ANN techniques. Depending on the nature and complexity of the infectious disease, size of the study area, spatial distribution of the vectors, various types of the neural networks could be applied, such as radial basis function, regulatory feedback, and recurrent neural networks.

**Limitations**

The success of classifiers and models used in this research highly depend on the quality of collected data and the variables used for predictions. Final spatial data that represent a feature in the models undergo several transformations such as measurement error or analysis errors (such as converting addresses to locations (geocoding)). Also, there are several obstacles and sources of uncertainty that need to
be considered. First, neglecting the role of spatial autocorrelation especially in sparse data may result in biased estimation of variables’ importance. SA is one of the most important concepts that are usually ignored in spatial epidemiology. Misclassification due to population migration or movements is another major source of uncertainty in spatial and temporal pattern that can influence the validity of the results of our LD and TB studies. Thus, misclassification can enhance or reduce the strength of associations. One of the strengths of aggregated data compared to individual level studies is that they are less influenced by misclassification. Selection bias can be another important limitation. Inaccurate or incomplete distribution of collected health data (such as collected sand flies) may lead to improper representation of the population of interest. In our ZCL study, selection bias during sand fly collection can distort associations between prediction of sand fly occurrence and descriptive variables.

Another problem is attributed to the administrative boundaries. All LD and TB data and associated variables which were collected as a single value that represent administrative boundaries (because of data restrictions, we used aggregated data) rather than physical boundaries. This might be problematic for spatial and particularly space-time analysis when comparing changes of pattern over time. The boundaries may be subject to change (split or merge of some areas) because of governmental policies, however in our study very few changes in the administrative boundaries occurred. Spatial scale was another major concern in our LD and TB researches. In these two papers, data were obtained at the town and county levels, respectively. The values within each division is uniform but there might be sharp contrasts between neighboring polygons. Thus, selection of geographic resolution for conducting analysis
is important. Because aggregation at various spatial scales can lead to different results. Also, different aggregation techniques (median, mode, sum, etc.) can lead to different results. In our studies, the choice of spatial unit was dictated by the available data.

These biases can cause spurious associations. By taking advantage of technologic advances such as GIS as well as data collection techniques such as GPS and RS, it is likely to improve reliability of results. Spatial data with higher spatial and temporal resolutions can provide more accurate estimations of risk and predictions in researches in medical geography. However, even with having accurate data spurious associations are inevitable.

**Outlook and Future Challenges**

Future researches in GIS and spatial analysis need to find optimal solutions for challenges such as scale and edge effect. Data collection techniques need to make accurate and reliable spatial data available. Organizations that provide demographic data need to make the same data available at multiple scales such as census tracts, black group and county level. Analysis has to be (re)conducted at all the possible scales, to address Modifiable areal unit problem (MAUP). Due to exponential growth of availability and use of spatial data which is likely to continue, it is anticipated that GIS software packages become more available. The software should include advanced and robust techniques such as data science techniques and other approaches from other disciplines. Also, spatial autocorrelation and uncertainty inherent in spatial data need to be incorporated in the packages. These two areas deserve closer attention of medical geographers. Analyzing big-data particularly in national and global health in near future can be a new era of investigations in spatial epidemiology. In this regard, deep-learning
neural network is a machine learning technique that can be useful when working with noisy and very large datasets.\textsuperscript{274}
### APPENDIX A
STATISTICAL COMPARISON OF CLASSIFIERS

Table A-1. Statistical test for comparing AUCs of SVM, LR and RF classifiers:

<table>
<thead>
<tr>
<th>Test Statistic</th>
<th>AUC 1</th>
<th>AUC 2</th>
<th>AUC 1 - AUC 2</th>
<th>Lower 95% CL</th>
<th>Upper 95% CL</th>
<th>Test Statistic Value</th>
<th>Prob Level</th>
<th>Reject H₀ at α = 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Z</td>
<td>0.974 (SVM)</td>
<td>0.965 (LR)</td>
<td>0.0090</td>
<td>-0.0704</td>
<td>0.0884</td>
<td>0.222</td>
<td>0.8243</td>
<td>No</td>
</tr>
<tr>
<td>Wald Z</td>
<td>0.974 (SVM)</td>
<td>0.935 (RF)</td>
<td>0.0309</td>
<td>-0.0569</td>
<td>0.1349</td>
<td>0.794</td>
<td>0.4272</td>
<td>No</td>
</tr>
<tr>
<td>Wald Z</td>
<td>0.965 (LR)</td>
<td>0.935 (RF)</td>
<td>0.0300</td>
<td>-0.0704</td>
<td>0.1304</td>
<td>0.584</td>
<td>0.5592</td>
<td>No</td>
</tr>
</tbody>
</table>

H₀: AUC ₁ = AUC ₂  
H₁: AUC ₁ ≠ AUC ₂  
Sample Size = nₜᵉˢᵗ = 36  
Therefore, the P-values do not indicate a significant difference between the AUCs.
APPENDIX B
HABITAT SUITABILITY MAPS

Figure B-1. Habitat suitability map of *Ph. Papatasi* based on logistic regression classifier.
Figure B-2. Habitat suitability map of *Ph. Papatasi* based on random forest classifier
LIST OF REFERENCES


115


80. Ogden, N. H., & Lindsay, L. R. (2016). Effects of climate and climate change on vectors and vector-borne diseases: ticks are different. *Trends in parasitology, 32*(8), 646-656.


BIOGRAPHICAL SKETCH

Abolfazl Mollalo started his PhD program in the Department of Geography, University of Florida in August 2015. He received his Ph.D. from this university in August 2019. His areas of expertise include medical geography, spatial epidemiology, GIS and applied machine learning in public health. He has obtained his master degree in GIS and bachelor degree in surveying engineering from national universities in Iran.