Novel Applications of Geospatial Analysis in the Modeling of Infectious Diseases

Pyrros Alexander Telionis

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics and Computational Biology

Bryan L. Lewis, Co-Chair
Stephen G. Eubank, Co-Chair
Kaja M. Abbas
Korine N. Kolivras

April 1st, 2019
Blacksburg, Virginia

Keywords: Autocovariate, Boosted Regression Trees, Computational Epidemiology, Ebola, Forecast, GIS, Gravity Model, Hepatitis C, Incidence, Infectious Disease, Maximum Entropy, Melioidosis, Metapopulation-Patch, Mobility, Social Transmission Niche, Spatial Autocorrelation, Spatial Epidemiology, Travel Network.

(ABSTRACT)

At the intersection of geography and public health, the field of spatial epidemiology seeks to use the tools of geospatial analysis to answer questions about disease. In this work we explore two areas: the use of geostatistical modeling as an extension of niche modeling, and the use of mobility metrics to augment modeling for epidemic responses.

Niche modeling refers to the practice of using statistical methods to relate the underlying spatially distributed environmental variables to an outcome, typically presence or absence of a species. Such work is common in disease ecology, and often focuses on exploring the range of a disease vector or pathogen. The technique also allows one to explore the importance of each underlying regressor, and the effect it has on the outcome. We demonstrate that this concept can be extended, through geostatistical modeling, to explore non-logistic phenomena such as incidence. When combined with weather forecasts, such efforts can even predict incidence of an upcoming season, allowing us to estimate the total number of expected cases, and where we would expect to find them. We demonstrate this in **Chapter 2**, by forecasting the incidence of melioidosis in Australia given weather forecasts a year prior. We also evaluate the efficacy of this technique and explore the impact of environmental variables such as elevation on melioidosis.

But these techniques are not limited to free-living and vector-borne pathogens. We theorize that they can also be applied to diseases that spread exclusively by person-to-person contact. Exploring this allows us to find areas of underreporting, as well as areas with unusual local forcing which might merit further investigation by the health department. We also explore this in **Chapter 4**, by relating the incidence of hepatitis C in rural Virginia to demographic data.

The West African Ebola Outbreak of 2014 demonstrated the need to include mobility in predictive disease modeling. One can no longer assume that neglected tropical diseases will remain contained and immobile, and the assumption of random mixing across large areas is unwise. Our efforts with modeling mobility are twofold. In **Chapter 3**, we demonstrate the creation of mobility metrics from open source road and river network data. We then demonstrate the usefulness of such data in a meta-population patch model meant to forecast the spread of Ebola in the Democratic Republic of Congo. In **Chapter 4**, we also demonstrate that mobility data can be used to strengthen outbreak detection via hotspot analysis, and to augment incidence models by factoring in the incidence rates of neighboring areas. These efforts will allow health departments to more accurately forecast incidence, and more readily identify disease hotspots of atypical size and shape.

(GENERAL AUDIENCE ABSTRACT)

The focus of this work is called “spatial epidemiology”, which combines geography with public health, to answer the *where*, and *why*, of disease. This is a growing field, and you’ve likely seen it in the news and media. Have you ever seen a map of the United States turning red in some virus disaster movie? The real thing looks a lot like that. After the Ebola outbreak of 2014, public health agencies wanted to know where the next one might hit. Now that there is another outbreak, we need to ask where and how will it spread? What areas are hardest hit, and how bad is it going to get? We can answer all these questions with spatial epidemiology. Our work adds to two aspects of spatial epidemiology: niche modeling, and mobility.

We use niche modeling to determine where we could find certain diseases, usually those that are spread by insects or animals. Consider Lyme disease, you get it from the bite of a tick, and the tick gets it from a white-footed mouse. But both the mice and ticks only live in certain parts of the country. With niche modeling we can determine where those are, and we can also guess at what makes those areas attractive to the mice and ticks. Is it winter harshness, summer temperatures, rainfall, and/or elevation? Is it something else?

In **Chapter 2**, we show that you can extend this idea. Instead of just looking at where the disease is, what if we could guess how many people will get infected? What if we could do so, a year in advance? We show that this can be done, but we need a good idea of what the weather will be like next year.

In **Chapter 4**, we show that you can do the same thing with hepatitis C. Instead of Lyme’s ticks and mice, hepatitis C depends on drug-use, unregulated tattooing, and unsafe sex. And like with Lyme, these things are only found in certain places. Instead of temperature or rainfall, we now need to find areas with drug-problems and poverty. But we can get an idea of this from the Census Bureau, and we can make a map of hepatitis C as easily as we did for Lyme. But hepatitis C spreads person-to-person. So, we need some idea of how people move around the area. This is where mobility comes in.

Mobility is important for most public health work, from detecting outbreaks to estimating where the disease will spread next. In **Chapter 3**, we show how one could create a mobility model for a rural area where few maps exist. We also show how to use that model to guess where the next cases of Ebola will show up. In **Chapter 4**, we show how you could use mobility to improve outbreak and hotspot detection. We also show how it’s used to help estimate the number of cases in an area. Because that number depends on how many cases are imported from the surrounding areas. And the only way to estimate that is with mobility.

Dedication

To my wife and father, for believing when I didn’t... And to ο Μεγάλος and ο Πρόεδρος, to Zettaki and Lulaki, and to my mom Vasso, who should have been here… μας λείπεις.

Acknowledgements

I could not have done this work without the support, understanding, and prudence of my advisers and committee: Bryan Lewis, Stephen Eubank, Kaja Abbas, and Korine Kolivras. I owe them each a great debt which I can only hope to pay forward. The same can be said for the departments which so graciously hosted me, Population Health Sciences, Geography, and GBCB, as well as the Biocomplexity Institute. I'm especially thankful to Population Health Sciences, which started me on this journey on a sunny winter day in Kerry Redican's office.

I also cannot thank enough Dennie Munson, the mother of our department, who is both protective and demanding of her sometimes-disobedient flock. And to my friend Paige Bordwine, who has the best stories of anyone I've met, and who kept me connected to the real world of public health. I must also thank Dr. Wu and our friends at the Defense Threat Reduction Agency, who gave me the most exciting and meaningful work of my entire career.

I cannot forget the support and comradery I've had with my academic brothers and sisters. Gloria, James, and Dan, the original four "health nerds", plus Gabs, Meghana, Arinjoy, Ashley, Liz, Danielle, Logan, Shreejana, and Megan. Not to mention the BI facilities and IT people who saved me so many times, Robert, Andrew, and Amy.

Finally, I'd like to acknowledge President Charlie Steger, whose vision made the Biocomplexity Institute a reality, and Governor Gilbert C. Walker, who on March 19th, 1872, signed the bill creating our beloved alma mater.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BRT	“Boosted Regression Trees” (a machine-learning algorithm)
CNIMS	“Comprehensive National Incident Management System”
DCW	“Digital Chart of the World”
DRC	“Democratic Republic of Congo”
DTRA	“Defense Threat Reduction Agency”
ENM	“Environmental Niche Model”
Esri	“Environmental Systems Research Institute”
GARP	“Genetic Algorithm for Rule-Set Prediction” (a machine learning algorithm)
GIS	“Geographic Information Systems”
HCV	“Hepatitis C Virus”
HDX	“Humanitarian Data Exchange”
HWSD	“Harmonized World Soil Database”
IPCC	“Intergovernmental Panel on Climate Change”
MaxEnt	“Maximum Entropy” (a machine-learning algorithm)
MIDAS	“Models of Infectious Disease Agent Study”
ML	“Machine Learning”
MSE	“Mean Squared Error”
NASA	“National Aeronautics and Space Administration”
NCEP	“National Centers for Environmental Prediction”
NIGMS	“National Institute of General Medical Sciences”
NIH	“National Institutes of Health”
NOAA	“National Oceanic and Atmospheric Administration”
OD	“Origin-Destination”
OSM	“OpenStreetMap”
SEIR	“Susceptible, Exposed, Infected, Recovered”
SRTM	“Shuttle Radar Topography Mission”
SVM	“Support Vector Machines” (a machine-learning algorithm)
VGIN	“Virginia Geographic Information Network”
WHO	“World Health Organization”

Forward

It perplexes me. It seems so incredibly self-evident – that the marriage of geography and public health enriches both – that I cannot understand why the interface between the two worlds remains so desolate. Certainly, there are a few titans with one foot in each, but the multitudes of both sides remain largely unaware or uninterested in their counterparts.

Consider the Master of Public Health (MPH) degree, the mainstay of its field. MPH programs are accredited by the Council on Education for Public Health (CEPH). At of the time of this writing, there were 1,387 accredited MPH programs and concentrations, offered at more than 186 different universities. Of these, only one offered a specialization in geographic analysis: The University of Southern California's "MPH in GeoHealth". A search on Sophas.org, the web portal through which most aspiring public health students apply for graduate matriculation, yields similar results: an unaccredited "Masters of Spatial Analysis for Public Health" at Johns Hopkins University, and a graduate certificate in "Public Health Geographic Information Systems" at the University of North Texas. The world of Geography is no more inclusive. Of the 581 geography departments in the US and Canada cataloged by the American Association of Geographers, a single department, at East Central University, is part of its parent University's College of Health Sciences. Internet searches for the geography degrees with the terms "public health" or "epidemiology" included in the search terms yielded only the aforementioned program at Johns Hopkins, and an online program at the London School of Hygiene and Tropical Medicine.

Certainly, many universities offer dual-degree programs, and many students can forge their own way with minors or dual majors. I have been one of the lucky ones to do so. Increasingly we have also seen programs on both sides eager to reach across the divide and collaborate. But formal training at the program level remains limited and somewhat novel.

I suspect that this scarcity of public health GIS modeling may be the result of skepticism on the part of the mainstream public health community regarding the usefulness of such work. I do not refer to researchers here, but the bulk of the professional public health community, the ones in the trenches so to speak. Indeed, in personal correspondence, many have questioned the usefulness of such efforts at all, especially when the endemicity is known. Moreover, many in the public health community may feel that these tools are exclusively useful for zoonotic diseases and environmental contaminants, when in nations like the United States chronic diseases are far more significant. Yet others are likely unaware of the power of geospatial analysis and consider the typical heatmap the extent of the field's offerings.

But I am confident that geospatial analysis can answer questions that are much more relevant to the daily operation of public health systems. I hope that once these techniques are demonstrated, and once researchers on both sides take notice of the complementary nature of the two fields, both sides may be more willing to engage the other. Certainly, the work of one student is not going to change much, but every small jaunt across the divide helps bring the two fields together. For my part, I hope I can contribute a small part to this cause by continuing my cross-disciplinary work and engaging in pedagogical outreach.

Chapter

1. Introduction

1.1 Spatial Epidemiology

The field of spatial epidemiology is informed by the triangle of human disease ecology, which posits that human health depends on a confluence of population, habitat and behavioral factors, each of which influences the others (1). In our case, we are primarily investigating habitat and its effects on health; the *where* and *why* of disease. Spatial epidemiology is a vital tool in public health, offering new insights and allowing epidemiologists to satisfy many of Hill's Criteria of Causation (2) during their work.

The field is older than one might imagine. Historically, one can trace the origin to Hippocrates, who noticed a spatiotemporal variation in malaria cases, noting a strong seasonality and localization, which was later echoed by both Celsus and Galen (3). The field may be even older than that, with mentions of disease environment associations appearing in the classical Chinese text, the Huangdi Neijing (4). The modern origins of the field are less ambiguous. Dr. John Snow's Cholera investigation, which famously identified the Broad Street pump as the source of the 1855 outbreak, is inarguably the most well-known example of the field, and a staple of modern public health and medical geography classrooms (5). Other eminent examples include the work of Dr. William Gorgas to reduce the impact of yellow fever during the construction of the Panama Canal (6), and the work of the Tennessee Valley Authority to eliminate malaria in the 1930s. Both cases relied on spatial efforts to map out and eliminate mosquito breeding sites.

The digital revolution revolutionized the field in the 1960s, when geographic information systems allowed computerized mapping and analyses. The first such system was created by Roger Tomlinson, on behalf of the Canadian government which sought to manage natural resources. The era also saw the rise of advanced statistical analyses, such as multivariate regression of spatial data (7), investigations into spatial autocorrelation (8), and network driven location-allocation problems (9). These techniques were quickly applied to disease modeling, and today GIS tools allow quick analysis of spatial patterns within public health data.

1.2 Common Applications of GIS in Epidemiology

Perhaps the most obvious application to the casual readers is applied cartography, the creation of maps and visualization of data. Cartography is arguably the origin of the entire field of geography, and GIS certainly modernizes the task. There are numerous applications of data visualization in epidemiology, but modern cartography is only a small fraction of what GIS has to offer an epidemiologist.

Aside from cartography, the reader is most likely aware of several common GIS tools, such as interpolation, topology analyses, proximity analyses, overlays (10). Drawing buffers around point-sources of disease risks, such as tire dumps which amplify malaria, is a common example of proximity analysis. Overlays are even more simple, combining maps of different data from different sources to augment surveillance and investigations. Such efforts can be supplemented by georeferencing, which allows one to import unreferenced cartographic data into their GIS by lining up visible landmarks with known coordinates. An example of this is seen in **Chapter 2** of

this work. Interpolation allows us to readily estimate the values between data sampling points, and can be used to create a trend surface of anything from ground water quality (10) to tuberculosis transmission rates (10).

Exploratory data analysis is another strength of GIS, allowing one to look at the distribution and sampling density of their data, to plot semivariograms to estimate relatedness and autocorrelation, and to investigate the distribution of model residuals (10). Perhaps the most well-known of all GIS tools, are clustering, heatmap, and density functions, made famous by John Snow's original work. Examples of this can be found in every Hollywood film depicting a biological disaster. These tools still have relevance today, and have recently been used to map a variety of modern ailments from the incidence and mortality of pneumonia (10) to the distribution of human Brucellosis in China (11). Hotspot detection is a related application by which epidemiologists can detect outbreaks. Unlike a heatmap, where the absolute values of each point are paramount, hotspot analyses are more concerned with relative values. Specifically, hotspot analyses detect if the high and low values clustered are together in a way that defies random chance (12). Doing so allows us to find clusters of unusually high incidence rates (13). Such hotspot techniques are often combined with space and time scanning which passes a scanning window over the study area comparing case counts to an expected distribution, finding clusters and generating a map of relative risk (14,15).

Another common application is in epidemiological surveillance. Beyond simple cartographic visualizations, geospatial analysis can greatly augment traditional surveillance methods. By combining clustering analyses and regression testing with reported disease data, tools like BioSense can act as early warning systems for epidemics (16,17). And in the age of social media, mining services like Twitter for geocoded data can be just as effective as mining electronic-medical records when it comes to detecting and pinpointing outbreaks (18), and more timely.

In our work, we will focus on two more facets of geospatial analysis which have applications in public health and disease modeling; namely, geostatistical modeling derived from environmental niche modeling, and mobility estimation.

1.3 Geostatistical Modeling

Environmental niche modeling is a form of species distribution modeling, allowing one to map out the potential range of a species and investigate the underlying conditions that give rise to this distribution. By relating the presence or absence of the species in question at certain points, with the environmental conditions at each point, one can estimate the causal relationship between those variables and the outcome (19). Originally these relationships were explored by simple logistic regression but advances in computer science and machine learning have given us far more powerful tools today. These include most notably Maximum Entropy Modeling (MaxEnt), a form of multinomial logistic regression which seeks to maximize informational entropy (20–22), and Boosted Regression Trees (BRT) a decision tree generating algorithm similar to Random Forests but with gradient boosting (23–25). A variety of other applications exist, including genetic algorithm for rule-set prediction (GARP) (26), the minimum-volume ellipsoid (MVE) and Marble

algorithms (27), and broadly applicable classifier Support Vector Machines (28). But at the time of this writing, MaxEnt and BRT dominate the field.

Though more commonly found in ecology, niche modeling has a significant role to play in epidemiology. It can be used to identify the range of the vectors and reservoirs (29,30), free-living pathogens (31), and to model the range in which the potential of zoonotic spillover may occur (19,32). The technique also allows one to gauge the effect that each variable has on the outcome, by mapping out the response curve, and estimating variable importance (33). When modeling zoonotic spillover, it is ideal to include some measure of the density of relevant vectors and reservoirs, but in their absence, a "black box" model built exclusively on abiotic factors, such as elevation and rainfall, is often perfectly serviceable (19). The idea being that all the biotic factors of the area, the presence and density of each species in question, are bound by their own niches, which are ultimately dependent on abiotic factors anyway. As such, a model informed exclusively by abiotic factors, can still approximate presence or absence of each species in question and ultimately the disease in question. Certainly, there will be localized variations, caused by specific predator-prey interactions. An extra fox or two in a neighborhood may depress local Lyme incidence by reducing the population of white-footed mice. But at a wider scale, the Eltonian noise hypothesis suggests that variations disappear and do not affect the overall niche (19). We must note here that some studies have cast doubt on this theory (34), but it remains a foundational idea of the field.

Niche models are sufficiently generalizable to be projected onto new areas in both space and time. Successful attempts have been made to project niche models from one continent to another (19,35,36), and plausible estimates of future ranges have been projected decades into the future when combined with climatological estimates from the Intergovernmental Panel on Climate Change (IPCC) (37,38).

But for all their strengths, niche models are limited by a few key criteria. Most notably, their output is, by definition, logistic. They estimate the probability of presence or absence of a species. They are also typically restricted to the modeling of free-living pathogens, and zoonotic diseases. We feel there is room to expand this application to cover both human-to-human pathogens as well as to estimate a continuous response, rather than a logistic one. As niche modeling is specifically restricted to exploring presence and absence, our work is more appropriately referred to as geostatistical modeling (39).

In **Chapter 2**, we will explore the use of geostatistical modeling techniques to estimate the incidence of melioidosis in Australia. We then project our trained model onto climate predictions to estimate incidence for the coming season. Though this is similar to the efforts to project a model onto IPCC climatological estimates, it is the only example we are aware of that incorporates real weather forecasts rather than generalized climatic trends. In **Chapter 4**, we demonstrate the feasibility of doing the same with a disease that spreads exclusively from person-to-person: hepatitis C. We also introduce the concept of the "Social Transmission Niche", which is an extension of the "Zoonotic Transmission Niche" from infectious disease ecology (32). We suggest that the endeavor of modeling a pathogen like the hepatitis C virus is fundamentally the same as modeling any zoonotic disease. Instead of reservoirs and vectors, we need some understanding

of risky behaviors that drive the disease. But just like reservoirs and vectors, these behaviors are not randomly distributed. Rather they depend heavily on underlying conditions, in this case poverty, disenfranchisement, and a lack of educational opportunities. And the distributions of these phenomena are just as easily mapped as temperature or elevation.

1.4 Mobility Modeling

The other area we wish to explore is in the modeling of human mobility as it applies to epidemiology. Most spatial analyses require some conceptualization of the spatial relationships between the points or areas under scrutiny. If you'd like to run a hotspot analysis to detect unusual clustering of high incidence rates, the algorithm will compare the values of each areal unit to those of its neighbors (40). But what constitutes a "neighbor" is another question entirely. One could make assumptions given proximity (41,42), but in truth, human mobility is often at odds with pure distance metrics (43,44). If the disease in question is transmitted from person to person, the disease is as mobile as the local population. Ignoring such mobility will therefore drastically confound analyses. The same can be said for geostatistical models of person-to-person diseases. In the end, we are not so concerned about proximal neighbors after all. Rather, we must account for interactions with highly connected areas, regardless of their proximity. If the disease is only spread by contact with infected individuals, it is self-evident that the density of cases in these highly connected areas will be important to include in modeling efforts.

Mobility is just as vital to models aimed at predicting the growth and spread of diseases. Consider a compartmental model, which aims to approximate the epidemic curve of an outbreak by binning the population into categories such as susceptible, infected, and recovered. Such models assume random mixing of the population, which is certainly unrealistic across wide-scale study areas. It is possible to break this study area up into smaller regions, "metapopulation patches", in which random mixing is more reasonable an assumption. But after doing so, one must estimate the interactions between each (45). And for that, we need some understanding of mobility. We can also use mobility to augment more complex agent-based models. Such efforts make use of a synthetic population, modeling each individual as an *in silico* agent, and considering each contact that agent has with another, as a possible transmission event (46,47). For this, we need an adequate understanding of the social network and contacts each agent makes on a regular basis. But we also need to incorporate some approximation of long-range travel outside the normal scope of daily life. Which leads us again to a need to understand local mobility.

Taking mobility into account is also necessary when considering facility placement through the use of location-allocation analyses (48,49). Such work generally considers travel network and modalities, optimizing the placement of a given facility, by the use of some network clustering algorithm such as k-means or k-medians (50,51). Such techniques have been used in a variety of public health applications, from the placement of rural healthcare facilities in India (52) to this author's own work concerning the placement of Ebola treatment centers in Liberia. Mobility metrics also pair well with phylogeography, allowing researchers to follow the genomic changes

in a pathogen as it moves through the environment, and estimate the basis for these changes (53,54).

In **Chapter 3**, we will explore the creation of a mobility model to inform a meta-population patch model for Ebola in Democratic Republic of Congo (DRC). Given the extreme dearth of properly curated travel networks in rural regions of developing nations such as the DRC, we also explore the creation of such travel networks from open source data. In **Chapter 4**, we explore the creation of a mobility model given an existing travel network from Here/NavTeq, and the calibration of this model using commuter flow data from the US Census Bureau's American Community Survey. We then explore the application of incorporating mobility data into hotspot analyses, and geostatistical models to improve predictive power and reduce the autocorrelation of model residuals.

Chapter

2. Methods for Forecasting Spatial Incidence Patterns with Geostatistical Models

Attribution

The following chapter is based on work done with partners at the Defense Threat Reduction Agency. In addition to assisting with literature review, they provided us with climatic data and did significant work reformatting these data for our use.

2.1 Abstract

In this work we seek to demonstrate that traditional niche modeling methods can be combined with weather forecasting to predict incidence rates for a future disease season. As a case study, we will focus on melioidosis, a lethal pneumonia-like disease caused by the saprophytic bacterium *B. pseudomallei*. Predominantly limited to Australia and Southeast Asia, infection is typically the result of exposure to wet soil, and often linked to heavy flooding or monsoon rains. Accordingly, the incidence of melioidosis is strongly associated with predictable climatological factors. As such we feel it is an appropriate target for weather-based forecasting.

In September of 2016, we trained a geostatistical model on historical incidence data and weather patterns from 2000-2014. In conjunction with weather forecasts from the National Centers for Environmental Prediction (NCEP), we used this model to forecast incidence of melioidosis in Australia for the then upcoming rainy season (November 2016 to May 2017). After the conclusion of said rainy season, we compared our forecast to real case counts.

Our initial forecast underestimated the disease incidence. However, the NCEP weather forecast significantly underestimated rainfall and humidity, which strongly drives melioidosis. To account for weather prediction errors, we then reran our analysis using the real-world weather data and found that our second disease forecast was much more accurate. We conclude that it is possible to forecast disease incidence, but the accuracy of weather predictions may limit long-term disease forecasts.

2.2 Introduction

A mainstay of ecology, environmental niche modeling allows one to mathematically define the confluence of environmental conditions associated with the presence or absence of an observed species. This allows one to map the range of the species, and to forecast changes to this range given plausible scenarios of environmental modification. In the field of public health, niche modeling is particularly useful to disease ecologists as it allows them to identify the ranges of the vectors and reservoirs of a zoonotic disease or the range of a free-living pathogen. These models can also explicate the relationship between disease occurrence and underlying environmental variables, and can be projected onto future conditions or novel landscapes (45). The primary shortcoming of niche modeling is that the output is necessarily logistic, and mapping applications are primarily limited to determining the probability of presence or absence.

More sophisticated modeling techniques, such as model-based geostatistics, extend the concept of niche modeling beyond the considerations of occurrence alone (39). Typically involving Bayesian inference and specific considerations for spatial autocorrelation, these techniques have found use in mapping the intensity of parasitic infections (55,56), disease prevalence (57,58), and malarial transmission intensity (59).

Despite the efficacy and ease of use of modern statistical and machine learning tools, disease mapping remains a rarity in the public health community. A recent systematic review found that of 355 identified clinically relevant diseases with significant spatial variation, fifty percent had never

been formally mapped whatsoever (39). The same study found that rigorous explorations of the disease transmission niche existed for a mere twenty percent of these diseases, while less than three percent had been subjected to thorough geostatistical modeling. Such efforts also rarely attempt to make prospective forecasts and are almost exclusively focused on zoonotic diseases. In our literature review, we found a single paper which attempted to predict the spatial distribution of disease prevalence using machine learning (ML) techniques (60), and the authors only did so with historical data rather than making a forecast.

2.1.1 Forecasting Incidence

It is not uncommon to encounter linear regression-based forecasts, but these efforts must make do without the advantages of ML: automated variable selection, the ability to fit to non-linear jagged responses, and the ability to make use of pseudo-absence points (25). Many efforts have been made to forecast case counts using historical environmental or climatic data (61,62), but typically they do not include a spatial component. Combining climate and demographic projections with ML-based disease modeling efforts to forecast the geospatial distribution of incidence has, to the best of our knowledge, never been attempted. Additionally, prospective forecasting, in which the author makes a testable prediction of future case counts, is very rare.

We intend to demonstrate this methodology by forecasting melioidosis incidence in Australia. The disease was selected for its strong association with predictable climatic conditions. Additionally, melioidosis is caused by a free-living pathogen, and as such, our analyses are not confounded by the interaction of vectors, reservoirs, and dilution hosts. The disease is not contagious, which eliminates the need to consider person-to-person transmission. Finally, Australia was chosen as a study area due to the higher quality surveillance data and case reporting when compared to developing nations in the melioidosis range.

2.1.2 Case Study: Melioidosis

Melioidosis is an often-fatal disease caused by the Gram-negative coccobacillary bacterium *Burkholderia pseudomallei*. Closely related to *Burkholderia mallei*, the etiological agent of the agricultural pest Glanders, both pathogens are considered select agents by the CDC. Both pose a severe threat to human health and could potentially be weaponized by malicious actors. Accordingly, both are of particular interest to infectious disease modelers.

Typically found in Southeast Asia and Oceana, the *B. pseudomallei microbe* is a saprophyte and lives freely in moist soils. The primary exposure route is contact with contaminated soil or water (63). Melioidosis is not contagious, but does have a high case-fatality rate of 20-50% depending on the quality and promptness of treatment (63,64). The disease has a mean incubation period of nine days (65), and often follows flood and extreme rain events with a roughly two week lag (62). Whether inhaled or introduced by percutaneous event, the bacteria settle in the lungs and cause a severe pneumonia-like illness, killing through suffocation or overwhelming sepsis (66). In Australia, the incidence is estimated to be 20-50 per 100,000 person-years (67), albeit with a significant underreporting rate. Its range is mostly limited to the northern coastal regions of the country between Darwin in the Northern Territory (NT) and Townsville in Queensland (QLD) (31). Risk factors include immunodeficiencies caused by alcoholism or diabetes, as well as lifestyle risks such as recreational gardening and rice-paddy farming (63,68,69). Ethnicity may also play

a role as Aboriginal Australians and Torres Strait Islanders experience fourteen times higher incidence rates than Caucasians (70), though this remains controversial and may be explained by sociological collinearity with other risk factors such as diabetes (63,71,72). Most significantly for our purposes, melioidosis is strongly associated with climate, in particular heavy rainfall (62,67). Other factors that influence melioidosis are cloud cover, dew point, humidity, temperature, and wind speed (62,73,74), the latter the result of the propensity for *B. pseudomallei* to be aerosolized by high winds (74).

Efforts to model the environmental niche of melioidosis are well established (31). Furthermore, an effort to forecast the fortnightly number of cases at a single hospital in Darwin, based on historical weather and case data, and a 14-day weather prediction from Australian Bureau of Meteorology, has already proved successful (62). As such we seek to test whether it is possible to combine such efforts and forecast incidence of melioidosis, provided we have access to climatological projections.

2.2.3 Study Objective

The objective of this study is to demonstrate the feasibility of combining weather predictions with disease modeling to forecast incidence, and to evaluate the plausibility of doing so accurately. In the end, we hope these efforts play a role in establishing spatiotemporal disease forecasting as a science which can inform public health resource allocation and improve intervention efficacy. We also hope to create a methodology by which existing prospective agent-based disease models can account for the changing environmental risk to each agent as they move through space and time.

2.2.4 Significance

If we can establish the feasibility of accurately forecasting disease incidence, this could have wide-sweeping implications for public health. It could inform public health agencies in both their response and surveillance efforts. For example, an agency which expects an increase of a particular disease in a specific area, may choose to do educational outreach in the area, may step up surveillance efforts, may communicate the increased risk to local physicians, and may order additional resources to combat the epidemic. Such efforts could also be useful to the public if it were easily accessible. Imagine if you will, that the weather app on your smart phone could also give you an estimate of the local Lyme disease risk for the day in question; that would surely be relevant if you had plans to go hiking or have a picnic.

A better understanding of how each disease responds to specific climatic variables will also allow for improved estimate of the impact of climate change. At the moment, forecasting the change in the distribution of specific diseases or vectors for given Intergovernmental Panel on Climate Change (IPCC) climate change scenarios, is quite common. But as with all niche models, these efforts have a logistic output, and produce merely presence or absence. But substantial climate change will likely change the burden of such diseases, and the density of their vectors as well. These variations cannot be detected by traditional projected niche modeling but can be clarified with geostatistical methods.

2.3 Methods and Materials

2.3.1 Data Acquisition

Case records were sourced from the open dataset attached to the Limmathurotsakul (31) publication. These data include 1,072 cases between 2000 and 2014, geocoded to the hospital of diagnosis. Though this spatial limitation greatly hinders our investigation of fine-scale environmental conditions, such as elevation or slope, the cases are temporally resolved to the year of diagnosis. The population density of Australia is highly autocorrelated, clustered and mostly limited to coastal population centers. Accordingly, we assume that the cases in question came from somewhere within a 30-mile radius of each hospital. Using historical LandScan data from Oak Ridge National Lab (75), we calculated the annual melioidosis incidence within each of these cells. We also created 10,720 pseudo-absence points placed randomly in both space and time, to represent the background conditions of the study area. These pseudo-absence points were limited to areas which LandScan showed as populated, a process which eliminated large swaths of the country's interior as seen in **Figure 2.1**. We chose the number 10,720 to achieve a 10:1 ratio with the presence points, as well as reach the required point density prescribed by literature (76).

2.3.2 Initial Investigation

As we are basing our efforts on the case data collected by another team, our first effort involved approximating the Limmathurotsakul (31) study to ensure that we achieve similar results. Following the prescribed methodology, we created a niche model relating the presence and pseudo-absence data to the BioClim19 climatic variables from WorldClim.org (77), and soil unit type from Harmonized World Soil Database v 1.2 (HWSD) (78). This was done using Dismo (25), a Boosted Regression Trees (BRT) niche modeling package for the R statistical programming language (79). BioClim19 is a gridded raster dataset which includes 19 variables representing temperature or precipitation at certain temporal ranges, each averaged over the years 1950-2000. We selected the same six variables as the original study, namely minimum, mean, and maximum grids for precipitation and temperature respectively. We did not include Enhanced Vegetation Index data and other soil quality metrics such as salinity and percent gravel found in the original study as Limmathurotsakul (31) showed these to be insignificant contributors to melioidosis incidence.

For the sake of comparison we repeated the analysis using Maximum Entropy modeling (MaxEnt) (20,22,80), as it is the primary competitor to BRT in the field of niche modeling. We then compared the outputs of both of our models to the original work done by Limmathurotsakul (31). Finally, to investigate the impact of ethnicity, we repeated the analysis adding the percentage of the population which identifies as Aboriginal Australian or Torres Strait Islander as an explanatory variable. These data were sourced from the 2011 Australian Census as reported by the Australian Bureau of Statistics. The data were in a vector format at the "SA2" level which appears to be roughly the equivalent of counties in the United States. As with most demographic data, the polygonal density is directly proportional to population density, and sparsely populated areas have limited detail.

2.3.3 Endeavors in Forecasting

Though the BioClim19 data were appropriate for a niche model, they do not suffice for a model meant for forecasting incidence. Such a model must be trained by relating historical incidence to climate data from the corresponding years. As such we turned to the National Centers for Environmental Prediction (NCEP), a division of the NOAA. Critically, NCEP offers both historical climate data and weather forecasts of nine months in a gridded raster format (81). Though the NASA Global Land Data Assimilation Dataset offers higher spatial and temporal resolution for historical data, it is not a source of weather forecasting. Furthermore, using the same source for both historical data and forecasts reduces the likelihood of unit errors, alignment errors, and problems with differing methodologies. One may question the accuracy of a nine-month forecast, and indeed it is unlikely that the NCEP can accurately predict the weather of a specific day. However, for our purposes, some generalized measure of how the coming year will compare to prior years will suffice.

Using NCEP data, we calculated the mean, minimum, maximum, and standard-deviation of monthly heat flux, humidity, precipitation rate, soil moisture, temperature, and wind speed, for the melioidosis season, which we defined as August 1st to July 31st of the following year. We also included elevation, slope, and landform curvature values calculated from Shuttle Radar Topography Mission (SRTM) data, as well as soil type from the HWSD. We then found the mean values of each variable, aside from soil, within the 20-km radius cells around the presence points and pseudo-absence points. For soil type we calculated majority value, as there is no way to average the categorical data. Using LandScan data for the appropriate year we calculated the population of each circular cell, then using the dated case data, we calculated the annual incidence for that season.

Finally, we fit a Gaussian BRT model to these data, relating incidence to historical weather conditions. BRT was chosen for its ability to do variable selection and its ability to fit to jagged responses (25,82), as well as its familiarity to the disease mapping community (31,39) which could encourage replication. Though we expect significant multicollinearity in our climate data, BRT is highly resistant to confounding from collinearity and in fact is commonly used for the purpose of variable selection (25,62,83). As such we chose not to pursue the typical path of reducing dimensionality via Principal Component Analysis or selecting variables by more common Variance Inflation Factor metrics.

For fine-tuning the BRT model, we started with the default settings as recommended by Elith (23), including a bag fraction of 0.75, a learning rate of 0.01, and a tree complexity of 5. Varying these factors did not substantially improve the model's predictive power as measured by cross-validation, and as such we retained the original values for all further work.

After the first iteration we found that a subset of ten variables explained the majority of the variation in our model. As shown in **Table 2.1**, the top ten variables accounted for 90% of the model contribution. Accordingly, the remaining variables were discarded for the sake of parsimony. Though elevation and landform curvature did contribute to predictive power of the model, they were also removed after further consideration. While melioidosis cases are seemingly associated with both, we cannot simply average the values of these variables within a 20-km

radius of each hospital and assume that we have an adequate understanding of the effects of terrain on melioidosis. It is unlikely that a model trained on these mean values could be reasonably projected onto fine scale data.

The parsimonious model, built upon the remaining seven variables whose contributions are shown in **Table 2.2**, was then projected onto NCEP climatic predictions for the coming melioidosis season. The output of this was a raster image where the value of each cell represents the forecast incidence. As the NCEP forecast data were of a very low spatial resolution, 2.5 x 2.5 decimal degrees, we also attempted to resample them to match the resolution of the LandScan population data using cubic convolution interpolation. When the model was applied to these resampled data, we noticed Moiré banding, an issue more commonly seen in image manipulation. As the NCEP historical data were of higher resolution, 1x1 degrees, an alternative to the resampling was to project the model on historical data which most closely matched the 2017 forecast. We did so by minimizing the mean squared error (MSE) between the 2017 forecast and historical grid on 1000 randomly dropped points. We also took note of the correlation coefficient between the grids to ensure that the historical grid was a suitable representation of the variable's spatial distribution between the test points. In most cases, the year with the smallest MSE also had the highest raster correlation coefficient. An example of this is seen in **Table 2.3,** which shows that the NCEP forecasts suggests that 2017 should most closely resemble 2003 in terms of mean humidity. After determining which historical year was most representative of the 2017 forecast for each variable, we projected the forecasting model onto these grids as well.

Finally, using the 2017 LandScan gridded population data, we also estimated the total number of cases in each state by simply multiplying the two grids to estimate total cases per cell, then summarizing using Zonal Statistics.

2.3.4 Exploring the Effects of Fine-Scale Terrain

The fact that our case data are geocoded to the diagnostic hospital, rather than the site of infection, is a significant hindrance to our work. It forces us to average climatic conditions across each 20-km cell, and more importantly prevents us from examining the effects of fine-scale variations in landscape. There is no way for us to account for this at a national level without geocoded case data which are not available. But we can examine the effect it has on melioidosis in a smaller area, where individual cases can be resolved.

To do so, we made use of data from Corkeron (84) which included a risk map showing points representing the home addresses of melioidosis cases in Townsville. These data represent the only geocoded cases available to us. We georeferenced the data against landmarks in aerial photography and used them to create an environmental niche model of melioidosis in the area, this time using MaxEnt. A BRT-based geostatistical model does not make sense in this case, as the study area is small, and we have a single year of data available to us. Furthermore, MaxEnt allows for jackknifing to augment our understanding of variable importance. In this case, jackknifing refers to the process of cross-validating while excluding specific variables from the final model. By comparing the model's predictive power, with and without a specific variable, we can estimate the importance of that variable to model fitness (80).

In addition to the aforementioned climatic data, we included SRTM derived elevation, slope derived from said elevation, soil type from the HWSD, and calculated Euclidean distance to open water, which in combination with elevation should provide a reasonable estimate of flooding and inundation risk. Inland water data were sourced from the Digital Chart of the World (85) and available from diva-gis.org. The variable importance figures from this study will allow us to evaluate the significance of these landscape variables. MaxEnt was run with a 5-fold random seed cross-validation with a maximum of 1500 iterations towards convergence.

2.3.5 Evaluating the Forecasts

It is difficult to evaluate the accuracy of our incidence forecasts without individual case data. However, we have estimated the total number of cases in each state for 2017. These figures can easily be compared to ground truth after the conclusion of the season. The difficulty is in determining whether the error is the result of the disease model, or the climate forecasts upon which the disease model was based. To investigate this, we compared the true climatic conditions, again sourced from NCEP, with the 2017 weather forecasts made the year before. We then projected our original BRT model on the true climatic conditions, producing an incidence map entirely free from weather prediction error. To avoid confusion, we will refer to this as the re-analysis model as opposed to the forecast model made earlier. Using the same procedures, we also estimated the total number of cases of melioidosis by state and compared these values to the true figures.

2.4 Results

2.4.1 Niche Modeling Output

The output of the initial BRT and MaxEnt niche models are shown in **Figure 2.2**. Though they were not identical, they were reasonably similar, with a Pearson correlation coefficient of 0.74662. In both cases, the models predicted melioidosis would be found predominantly in the northern coastal regions of Australia, in general agreement with the model made by Limmathurotsakul (31). A common metric of the predictive power of a niche model is the area under the curve (AUC) of the Receiver Operating Characteristic (80). In this case, both models had very large AUC values exceeding 0.95, though they may be overfitting due to the significant spatial sparsity of the data. Comparing the variable contributions, which can be seen in **Table 2.4**, we note that both algorithms considered rainfall in the wettest month to be the most significant explanatory variable. Both algorithms produced very similar response curves for rainfall as well. As seen in **Figure 2.3**, the expected ideal rainfall amount for the presence of melioidosis cases, or more specifically the survival of *B. pseudomallei* in the soil, is estimated to be around 350-450 mm per month during the rainy season. Adding to our model the percent of the local population that identifies as Aboriginal or Torres Strait Islander was nearly inconsequential, contributing just 0.3% of the variable contributions to the final model, and not improving the AUC in any meaningful way. The response curve for this variable, seen in **Figure 2.4** was also nonsensical, with an abnormal plateau between 30 and 90% and strange jagged features. This suggests that these data are completely superfluous to our efforts.

2.4.2 Forecast for the 2017-18 Season

The final model for our forecast included just seven variables, with mean specific humidity being the most significant as seen in **Table 2.2**. With a final total of 4250 decision trees, the model's training data correlation was 0.721, with a cross-validation correlation of 0.608. The variable response curves shown in **Figure 2.5** were as expected, with melioidosis cases being strongly associated with high humidity and temperature. Rainfall seasonality and wind speed were also significant.

The outputs of our incidence model projected onto the composite historical grids, unaltered NCEP forecasts, and resampled NCEP forecasts are seen in **Figures 2.6**, **2.7**, and **2.8** respectively, while the number of expected cases is shown in **Table 2.5**. As expected, we predicted that the northern coasts of Australia would be most affected by the disease, with Darwin bearing the brunt of the cases. Meanwhile, **Figures 2.9** and **2.10** show the expected distribution of melioidosis cases in the 2017 season. We estimated a total of 36 cases for 2017, with 14 in NT and 14 in QLD. In comparison to seasons with significant cases such as 2009 and 2014, both shown in **Figure 2.11**, we expected the 2017 to be quite mild.

2.4.3 Terrain Effects

The localized niche models we created using georeferenced data in Townsville had an average AUC of 0.880, with a standard deviation of 0.043. As seen in **Table 2.6**, elevation was the most substantial contributor to this model, followed by soil type and precipitation in the wettest month. When considering the results of the AUC jackknifing, we get a slightly different picture, with soil type being far less significant to the model AUC than model contribution suggested. On the other hand, removing minimum temperature of the coldest month from the permutation substantially lowered the AUC. The jackknifing did corroborate the importance of elevation, which, as seen in **Figure 2.12**, was the most significant variable when it came to AUC. Elevation was so critical in fact, that a model built upon elevation alone would still have an AUC greater than 0.8.

Selected response curves for these models are shown in **Figure 2.13** but are mostly unremarkable. Melioidosis cases are found at lower elevations and case density falls off exponentially as elevation increases. This may be the result of increasing population density, or the result of an increased propensity to flooding. Cases are also associated with higher rainfall values and milder winters. Soil type gives us minimal information. While the variable is significant in the model, the only significant response is an apparent aversion to solonchak, a soil found in arid areas with poor drainage characteristics. Slope was a modest contributor, fourth most significant according to the model contribution, but only accounting for 5.7% of the AUC variation. Though the response curve for Euclidean distance to water does support the expectation that melioidosis cases are associated with proximity to water, the variable contributed almost nothing to the final model, and removing it did not affect the AUC.

2.4.4 Forecast Accuracy

Our model predicted a relatively mild melioidosis season, with fewer than 40 cases, to accompany a reasonably dry year with few flooding events. This proved not to be the case, as most of Australia's northern coast experienced historic flooding, in particular after landfall of Cyclone Debbie (86). **Figure 2.14** shows the difference between forecast and actual climatic conditions

for the top six variables in our model. Most significantly, our two top contributors, mean specific humidity and maximum humidity, saw substantially higher values than expected along the melioidosis endemic northern coastal regions. Together, these two variables account for sixty-seven percent of the model contributions. Given the positive association between melioidosis and increasing humidity rates, this resulted in far more cases than expected.

As seen in **Table 2.5**, the model projected on NCEP weather forecasts, predicted a total of 14 cases in both NT and QLD. In reality, health officials in NT diagnosed 53 cases, and we estimated that QLD experienced 37 cases. This estimate is based on a report which showed 28 cases by April 6th, and the epidemic curve from Stewart (87) which shows that approximately 25% of melioidosis cases per season occur after this date. The model projected onto the re-analysis data collected after the season had ended yielded far more accurate results, predicting 49 cases in NT and 40 in QLD. The output of this model is seen in **Figure 2.15**. We note unusual results in Western Australia where the forecast model predicted four total cases, while the re-analysis model estimated 24. In actuality, the region experienced eight recorded cases.

2.5 Discussion

Despite our omission of the vegetation moisture index variable used in the Limmathurotsakul (31), our initial attempts at approximating the melioidosis niche was reasonably successful. The minor disparity between the BRT and MaxEnt models is not unusual, and we note that most of the areas in which the results differed are found deep in the interior of Australia with minimal population density. We suspect that the difference between the two results from over or under-fitting to training data, issues common to BRT and MaxEnt respectively. This could be solved by a larger number of well distributed presence points which unfortunately are unavailable to us. Furthermore, we expect that the disparity between our models and those of Limmathurotsakul (31) result from differences in pseudo-absence point generation. The exact methodology is not specified in the 2016 paper, nor is the method for distributing presence points geocoded to hospitals to their surrounding areas.

2.5.1 The Involvement of Race

The question of whether to include percent indigenous in our initial niche models was thoroughly debated. Though both Aboriginal Australians and Torres Strait Islanders experience a substantially higher than national average incidence of melioidosis (71), the cause remains ambiguous. Some have suggested that it may be a result of increased type II diabetes and alcohol-consumption, both of which are strongly associated with melioidosis as a result of immunosuppression (66). It may be simply a case that Indigenous people are more likely to encounter *B. pseudomallei* as a result of cultural beliefs and practices that are strongly linked with the environment (88) or differences in employment type. Either would invalidate our use of these racial demographics, as using race as a proxy for health-behavior and employment data is certainly inappropriate. Still there is evidence that some other factor may be at play. Currie (71) included a multitude of health factors, including alcohol consumption and diabetes, in a logistic regression to estimate adjusted relative risk. In that study, Aboriginal Australians are still shown

to have a higher than average risk even after accounting for the effect of other risk factors (71). In this case, percent indigenous may be a relevant cofactor, and its inclusion may in fact play a role in combatting the disease by justifying additional resources be diverted to the Aboriginal Australian and Torres Strait communities.

In our case, the addition of percent indigenous peoples had no effect on the model whatsoever. It may be the case that the data gathered from the Australian Bureau of Statistics simply lacked the resolution to separate signal from noise. If there is an association between race and melioidosis, we failed to detect it.

2.5.2 Forecasting Models

At first glance, one may be surprised to see the primary contributor to our forecasting model is mean specific humidity, followed by maximum specific humidity. Typically, melioidosis is associated with rainfall or flooding, but an association between humidity metrics and melioidosis cases is biologically plausible. The disease manifests after an individual is exposed to free-living *B. pseudomallai* found in the soil. But the bacteria require wet soil to survive and are quickly killed by desiccation. Accordingly, areas with very dry air are unlikely to experience melioidosis cases, while areas with perpetually moist soil are likely to have both a high incidence of melioidosis and high humidity values. At first glance, it is tempting to assume that multicollinearity is confounding these results, that rainfall is the primary driver of melioidosis, and that humidity is itself dependent on rainfall and therefore associated with melioidosis. This does not appear to be the case as within our data, humidity metrics and rainfall metrics exhibited a mean annual correlation of 0.851. Though they are strongly related, they are sufficiently different that we would expect BRT to have focused on rainfall if it were truly the better predictor of melioidosis. Note that, strangely, the correlation between mean rainfall and humidity in forecast data was a mere 0.307, a substantial outlier given the high correlation between these data for other years. Perhaps this is indicative of a flaw in the forecasting methodology, and had our model favored rainfall it may have been more accurate.

2.5.3 Fine-Scale Analysis

The results of our fine-scale Townsville analysis reveal a limitation in our national analyses. Elevation, slope, and soil type all played a substantial role in the trained model, suggesting that these variables would greatly improve the predictive power if included in the nationwide models. Unfortunately, we simply lacked the data to implement this at a national level. The HWSD lists 23 soil types for Australia, but only four are found in the Townsville area. A model trained on data from Townsville would have no way to estimate the response of melioidosis to any of the other soil types. The same applies to elevation and slope, as Townsville does not exhibit extremes of either. This issue could be greatly mitigated if we were able to obtain national case data resolved to home address rather than diagnosing hospital.

Note that a fraction of the Townsville data had no associated soil type as a result of the resolution and clipping of the HWSD, which left a few coastal regions without data. This could be mitigated by moving to a polygonal soil database such as the Digital Atlas of Australian Soils.

2.5.4 Limitations

The primary limitations of this study stem from the lack of spatial resolution. Over 80% of the cases in our presence data originated from Darwin or Townsville, with 691 of the 1072 cases geocoded to the parking lot of the Royal Darwin Hospital. In our study, we assumed that most of these cases were from within a 20-km radius of the admitting hospital in question. There is ample reason to suspect this, given the incredible disparity of population density between coastal regions of Australia and its interior. Moreover, melioidosis is so common in Darwin that a colloquial term for the disease is “Night Cliff Gardener’s Disease”, after the eponymous neighborhood located just 10 minutes by car from the Royal Darwin Hospital. Nevertheless, we cannot be certain that a fraction of patients are not coming from hundreds of kilometers away. After all, Royal Darwin Hospital is the largest hospital in the entire Northern Territory. The same can be said for the Townsville Hospital, which is the only tertiary care facility in the area.

The high concentration of cases in Darwin and Townsville also establishes a significant spatial sampling bias with regards to the data and brings into question the use of pseudo-absence points spread across the continent. Perhaps it is more reasonable to limit the study area, and accordingly, the pseudo-absence points to areas near the two cities. Unfortunately, the extreme concentration of population into these three cities means that case data from the interjacent areas will be very rare. Even with case data geocoded to place of residence, we would expect a substantial spatial bias in sampling.

Finally, we must reevaluate whether it is appropriate to train a model on cells with a 20-km radius, then project the same model onto rectangular cells of roughly a kilometer-squared. It may be more appropriate to use a single uniform grid for both training and projection purposes, in which case we would suggest a hexagonal pattern (89) as it is more realistically representing hospital catchment area, which likely isn’t square.

2.5.5 Uncertainty and Sensitivity Analysis

The response curves and variable contributions for our models give us a strong understanding of the sensitivity the predicted outcome has to variations in each regressor. We can confidently say that minor fluctuations in the mean humidity will have a far more serious effect on the predicted incidence than similar fluctuations in temperature. With this knowledge we could explore uncertainty by accounting for confidence intervals in our weather forecast data. Ideally, one could use these data to generate confidence intervals for the number of predicted melioidosis cases by state, as well as maps showing the low and high end of our incidence estimates. Unfortunately, these data were not available at the time of our original study.

2.5.6 Conclusion and the Future of Forecasting

The results of this study are encouraging. Despite the limitations, with accurate weather forecasting we find that a trained model can adequately forecast total cases up to nine months ahead of time. The most glaring issue with this is that accurate weather forecasts are far from guaranteed. In fact, the forecasts used in this study were specious, hindering the predictive power of our models. Given the speed at which *B. pseudomallei* reacts to inundation, the possibility that a single flood may contribute the majority of the cases for an entire year, and the challenge in forecasting specific weather events months in advance, it seems unlikely that incidence

forecasting of diseases like melioidosis will become viable in the near future. Still we feel that the methodology presented here can, with minor modification, be put to use with other infectious diseases. These methods seem particularly well suited to vector-borne illnesses that are more heavily modulated by past climatic conditions such as Lyme disease.

Superficially, Lyme disease may seem like a very peculiar choice for a weather-based forecasting effort. Lyme depends on a complex web of interactions between the black-legged tick vector, white-footed mouse reservoir, dilution hosts, reproductive hosts, predators of the white-footed mice, and predators of said predators (90). But Lyme disease is strongly associated with temperature, rainfall, and humidity on one and two year lags (91–93), forest fragmentation and density of edge-habitats (94–98), and land cover (99,100). Lyme is also strongly associated with acorn mast (101), which itself is associated with temperature and rainfall on a one year lag (102).

It is critical to note that all the aforementioned factors are either generally static or on a substantial lag. Though a recent study showed that Lyme emergence depends on spring rainfall and temperature (103), future incidence rates for an area, as well as their geospatial distribution, could be predicted almost entirely with historical climate data and abiotic factors. This eliminates the seemingly insurmountable challenge of accurately forecasting weather nine months into the future, while allowing the public health infrastructure to optimize their surveillance and response. Though mosquitoes are not as cooperative in this endeavor, generally depending on a monthly rather than yearly climate lag (104), the same methodology used for Lyme should be applicable to other tick-borne illnesses such as Rocky Mountain spotted fever and Powassan virus. As such, we are confident these methods will find use in the future.

2.6 Funding and Acknowledgements

This study was supported by the Defense Threat Reduction Agency (DTRA) Comprehensive National Incident Management System (CNIMS) Contract HDTRA1-17-0118; and the National Institutes of Health (NIH) and National Institute of General Medical Sciences (NIGMS) Models of Infectious Disease Agent Study (MIDAS) Cooperative Agreement U01GM070694.

We thank our partners and collaborators at DTRA, Aiguo Wu, Akeisha Owens, Dana Kuan, Dana Meranus, Jerry Mothershead, Jonathan Butler, Kierstyn Schwartz, Scott Runyon, Terence Hill, Yaitza Luna-Cruz and Zyg Dembek. We also thank Direk Limmathurotsakul and his collaborators for releasing their melioidosis dataset to the public.

Figure 2.1: The presence points sourced from the Limmathurotsakul (31) study and randomly generated pseudo-absence points used for modeling melioidosis. Note that these data are distributed in time as well as space.

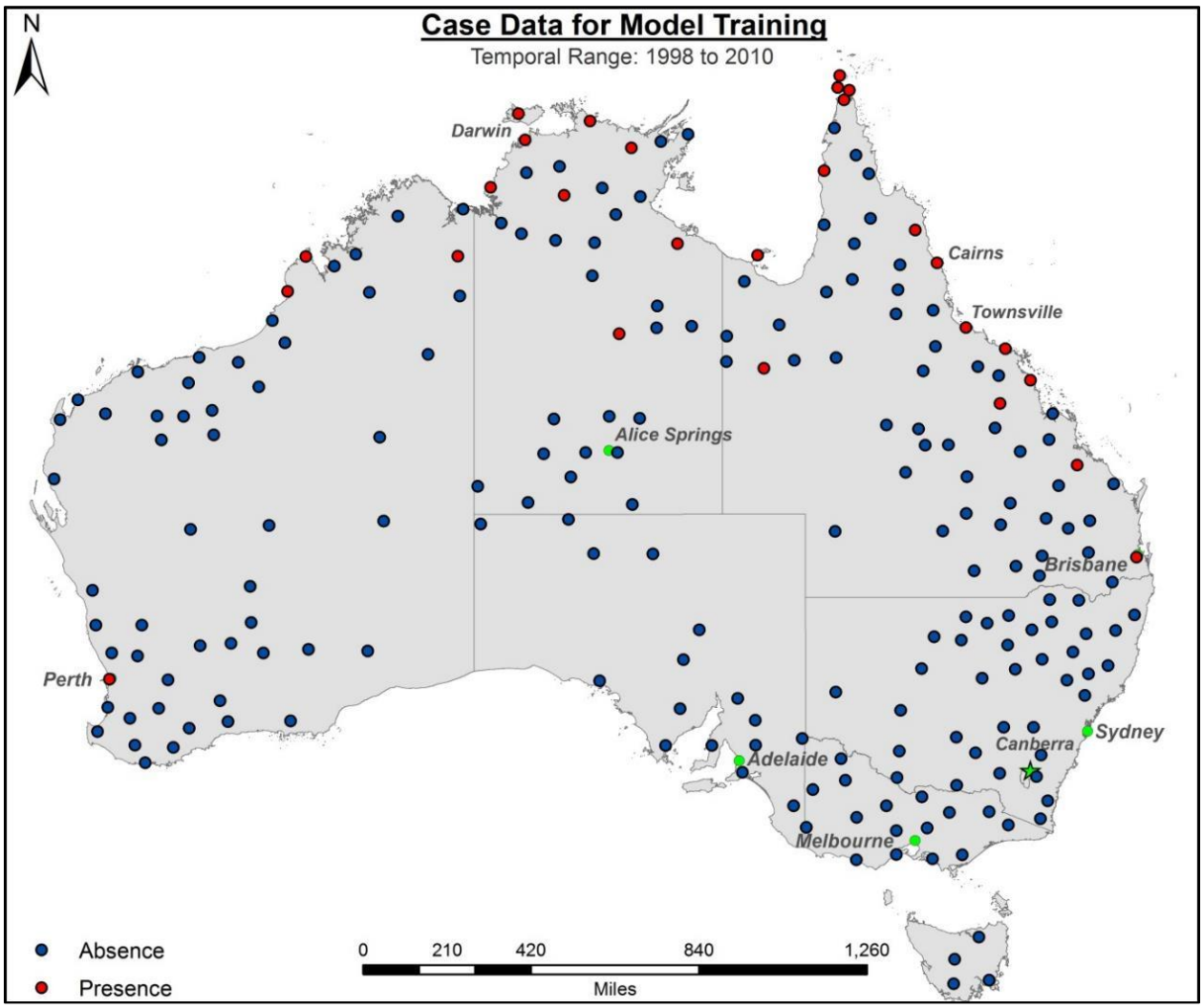


Figure 2.2: The output of the transcontinental niche models forecasting the presence/absence of melioidosis in Australia. The MaxEnt model (upper) has an AUC of 0.952, while the BRT model (lower) has an AUC of 0.995. The Pearson Correlation coefficient between the two is 0.74662. As expected, both models predicted cases in the northern coastal regions, which matches the general predictions of Limmathurotsakul (31).

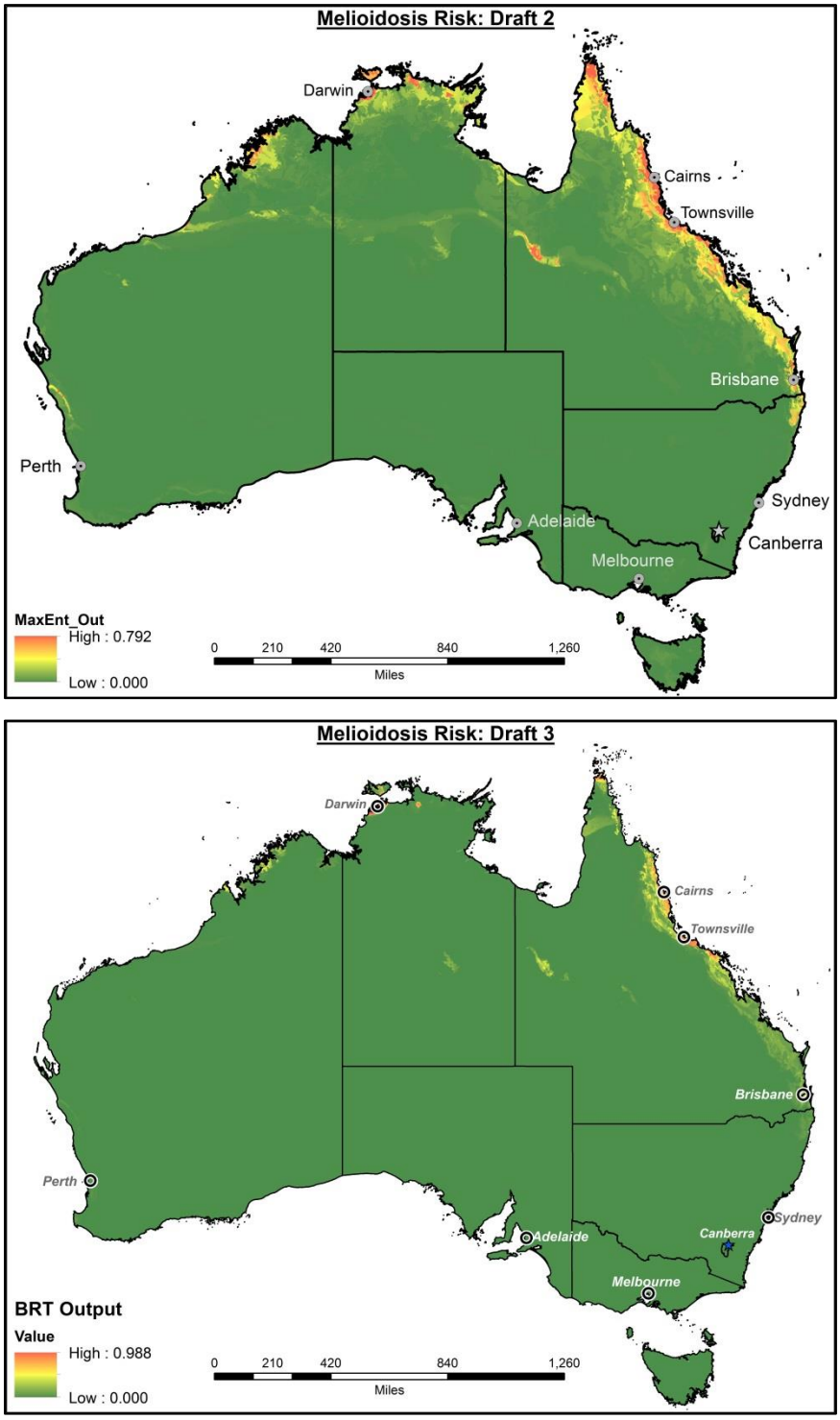


Figure 2.3: Here we see the response curve relating melioidosis cases to rainfall in the wettest month for both the MaxEnt (upper) and BRT (lower) generated transcontinental niche models. In both cases, the algorithms identified this variable as the most consequential contributor to model fitness. Both produced very similar curves, with a peak between 350-450 mm of rainfall per month.

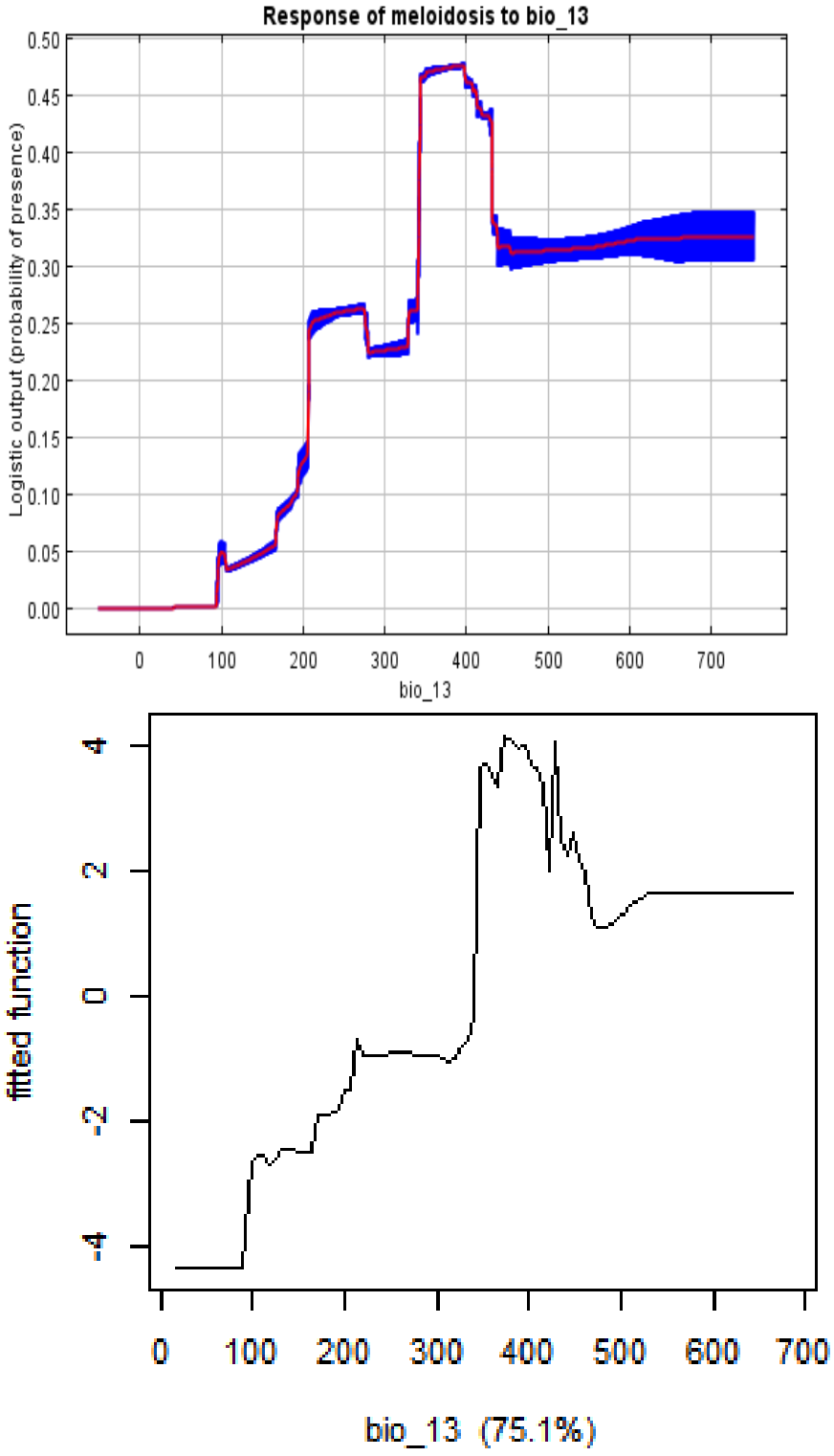


Figure 2.4: The response curve for the percent indigenous population variable used in our MaxEnt derived transcontinental niche model. The mostly nonsensical curve, in addition to the extremely small variable contribution of 0.2%, strongly suggests that the variable is either inconsequential to melioidosis incidence, or that its effect on melioidosis cannot be resolved without finer resolution spatial data. Note that the scale in this case represents the response of melioidosis to the variable, rather than contribution which is fixed at 0.2% for this variable.

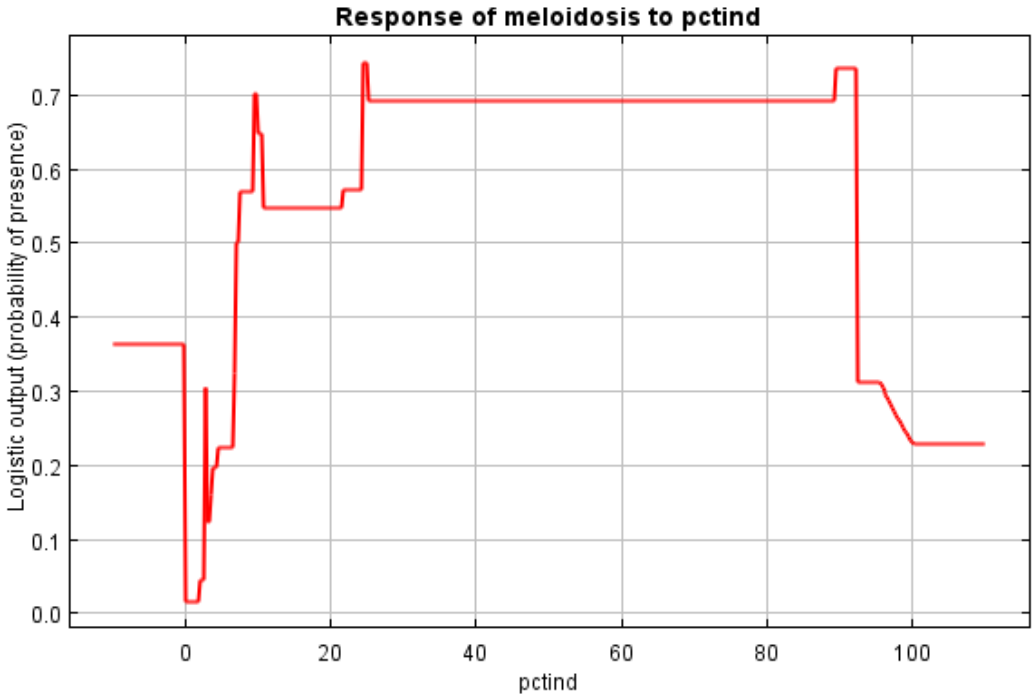


Figure 2.5: The BRT generated response curves for the incidence forecasting model.

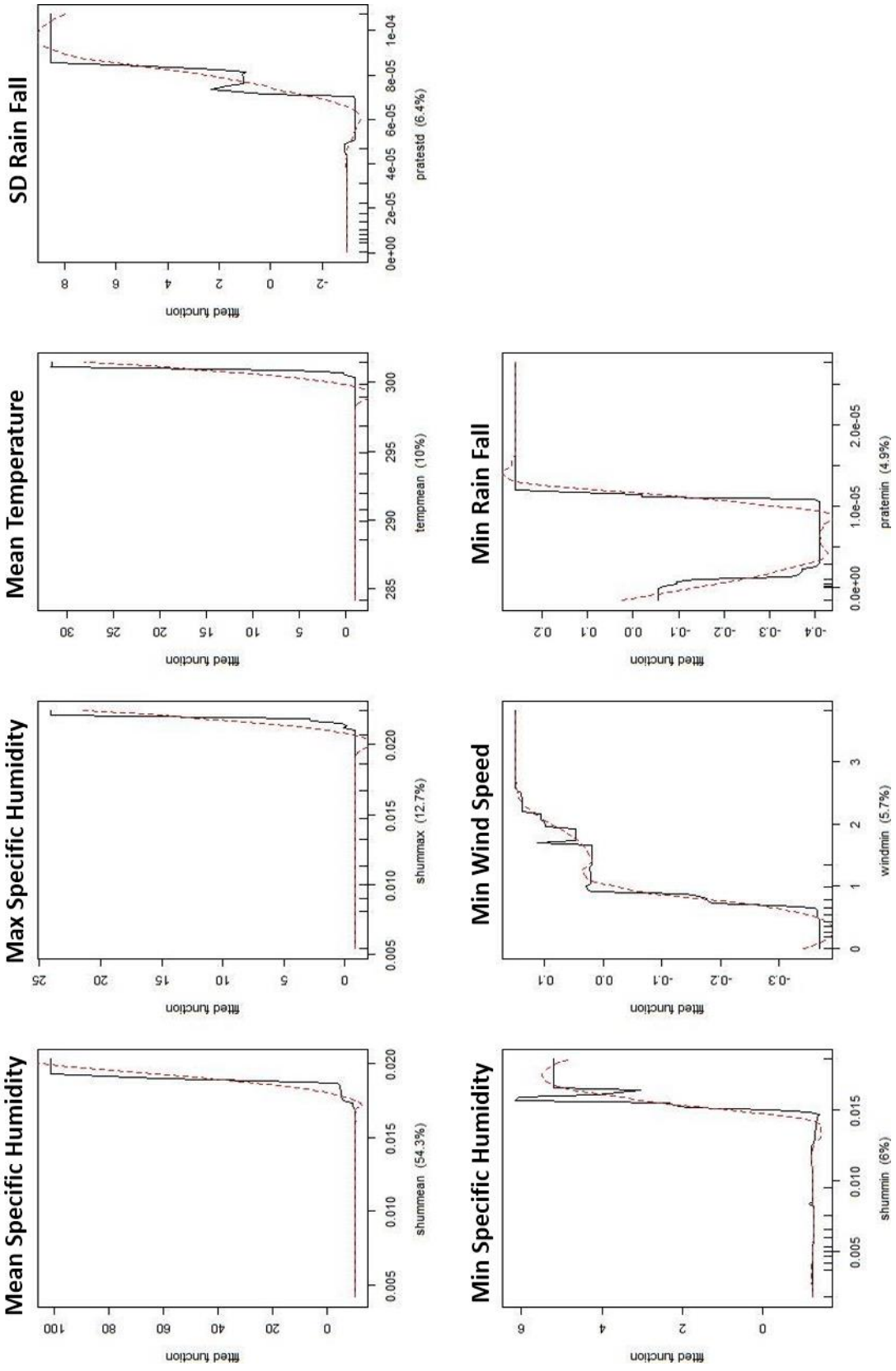


Figure 2.6: Melioidosis incidence forecast for 2017 based on composite grids. We projected our model on historical grids for each variable which most closely represented the 2017 forecasts.

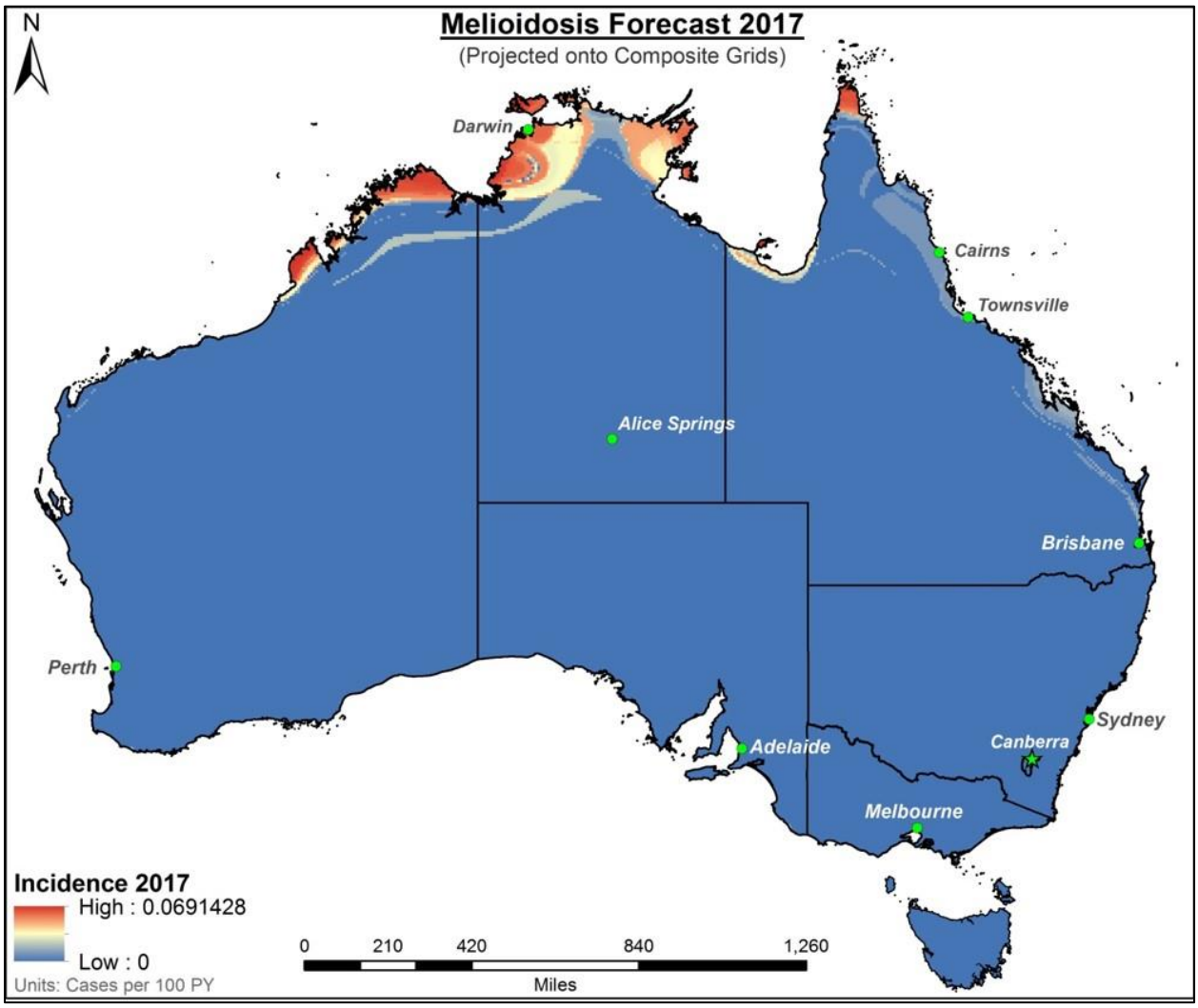


Figure 2.7: Melioidosis incidence forecast for 2017 at the original resolution of the NCEP forecast. The resolution is too low for the model output to be very useful.

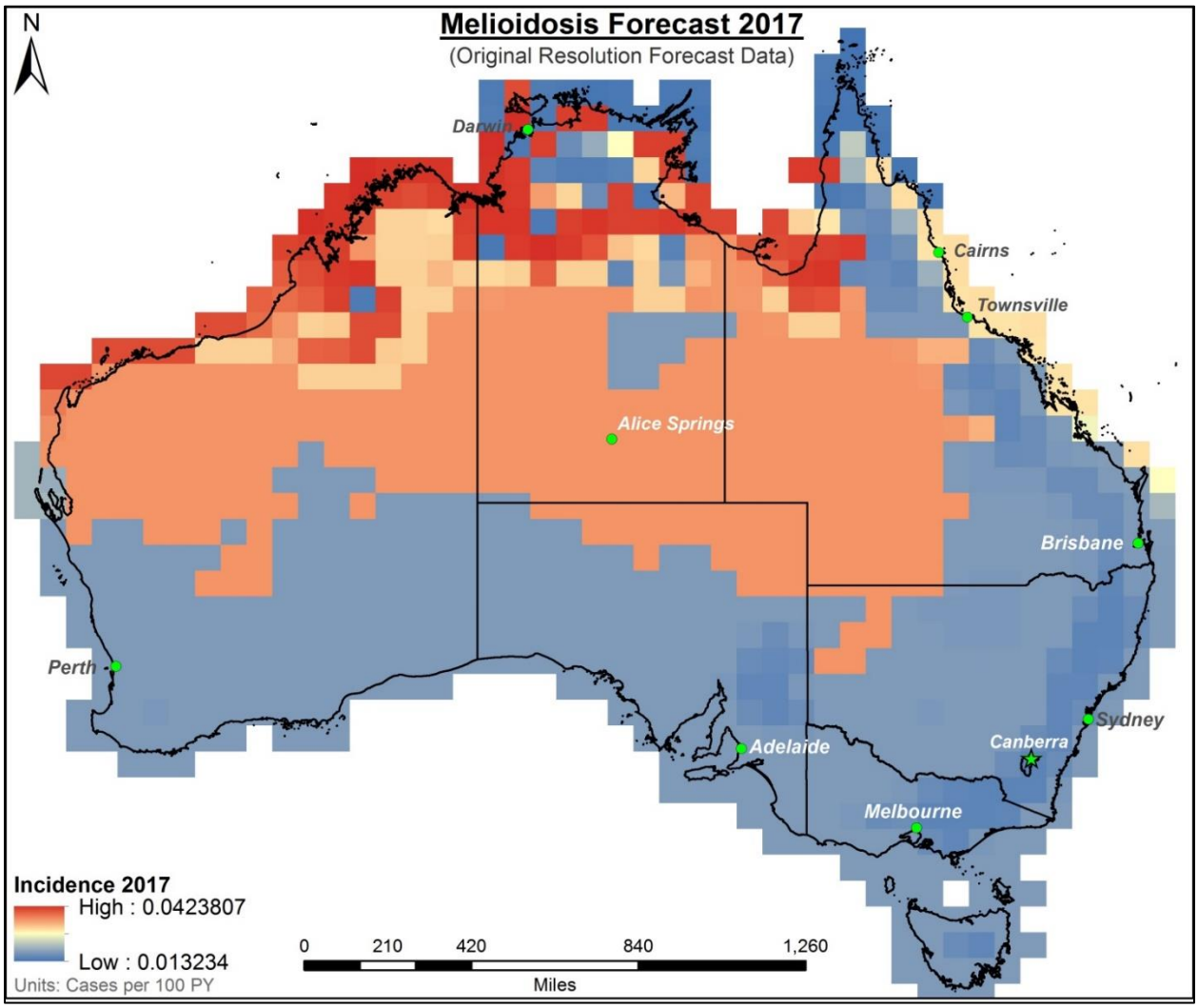


Figure 2.8: Melioidosis incidence forecast for 2017 based on NCEP weather predictions. Note that the cubic convolution resampling techniques used to resample the NCEP weather prediction data to match the resolution of the training data have added Moiré banding, a common artifact in image processing.

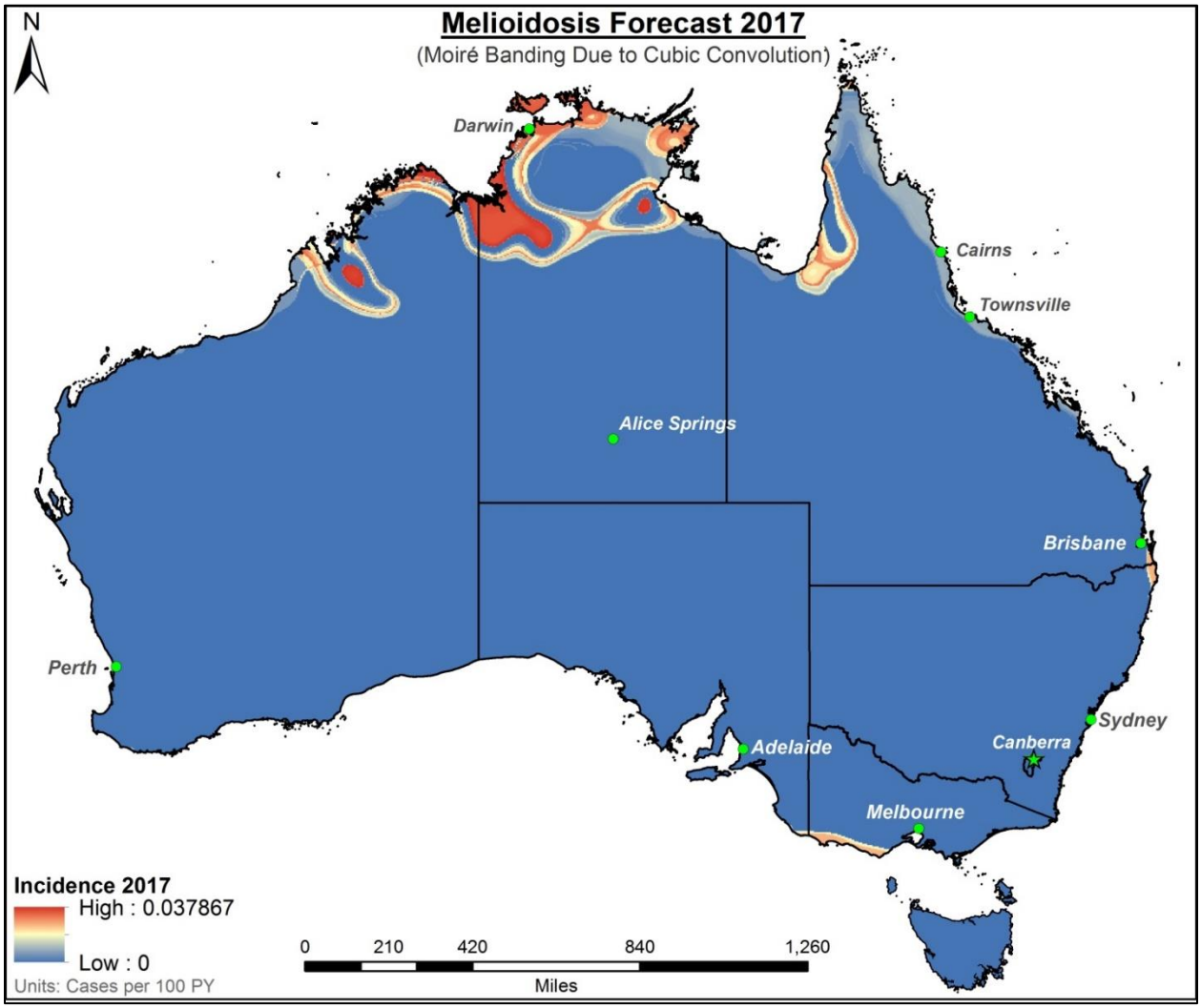


Figure 2.9: The density of model predicted melioidosis cases in 2017. Areas with no predicted cases are shown in green, providing contrast for case density seen in a spectrum of yellow to red, with red showing areas of the highest case density.

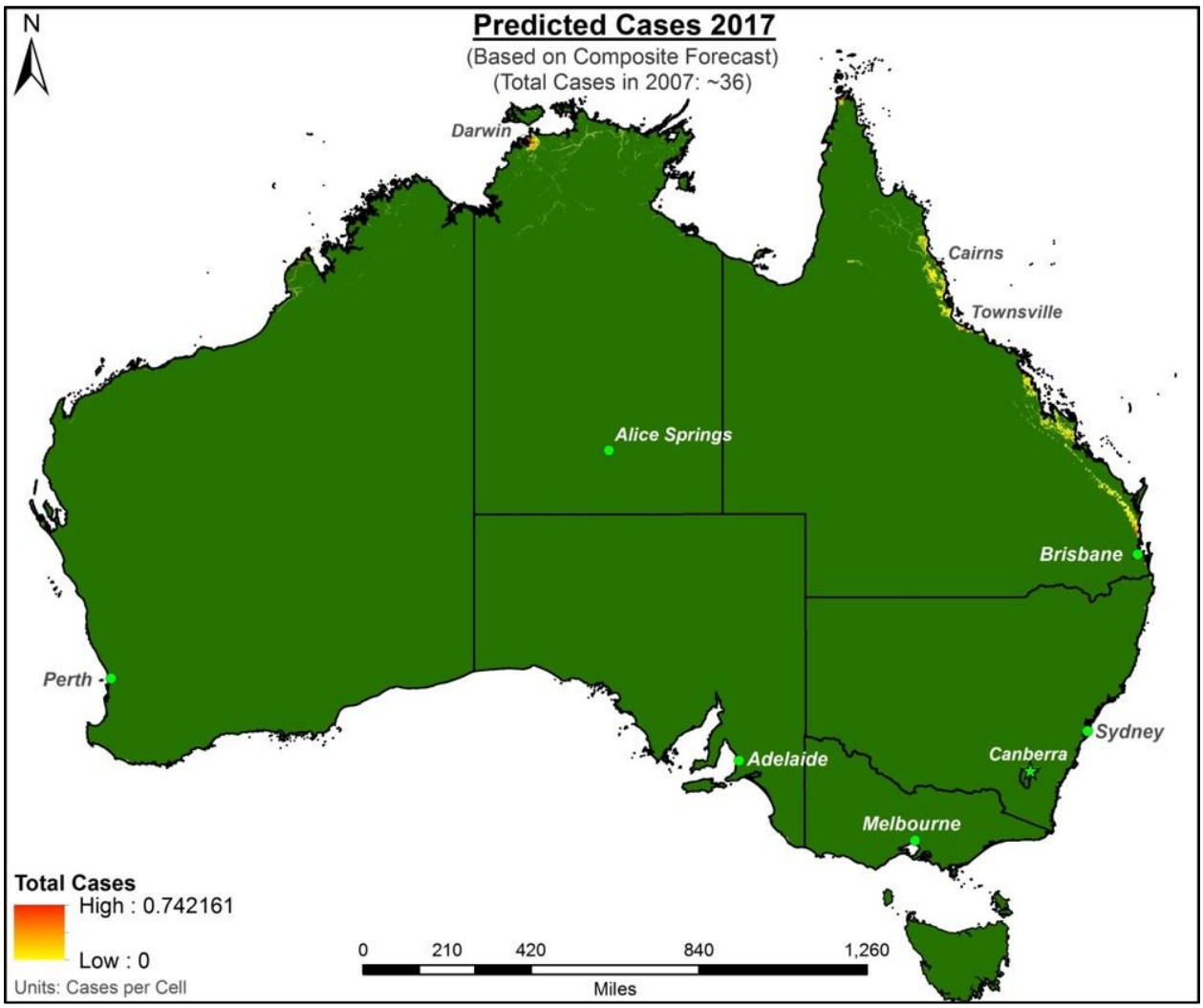


Figure 2.10: The estimated density of melioidosis cases in the population centers of Northern Territory (upper) and Queensland (lower). Areas with no predicted cases are shown in green.

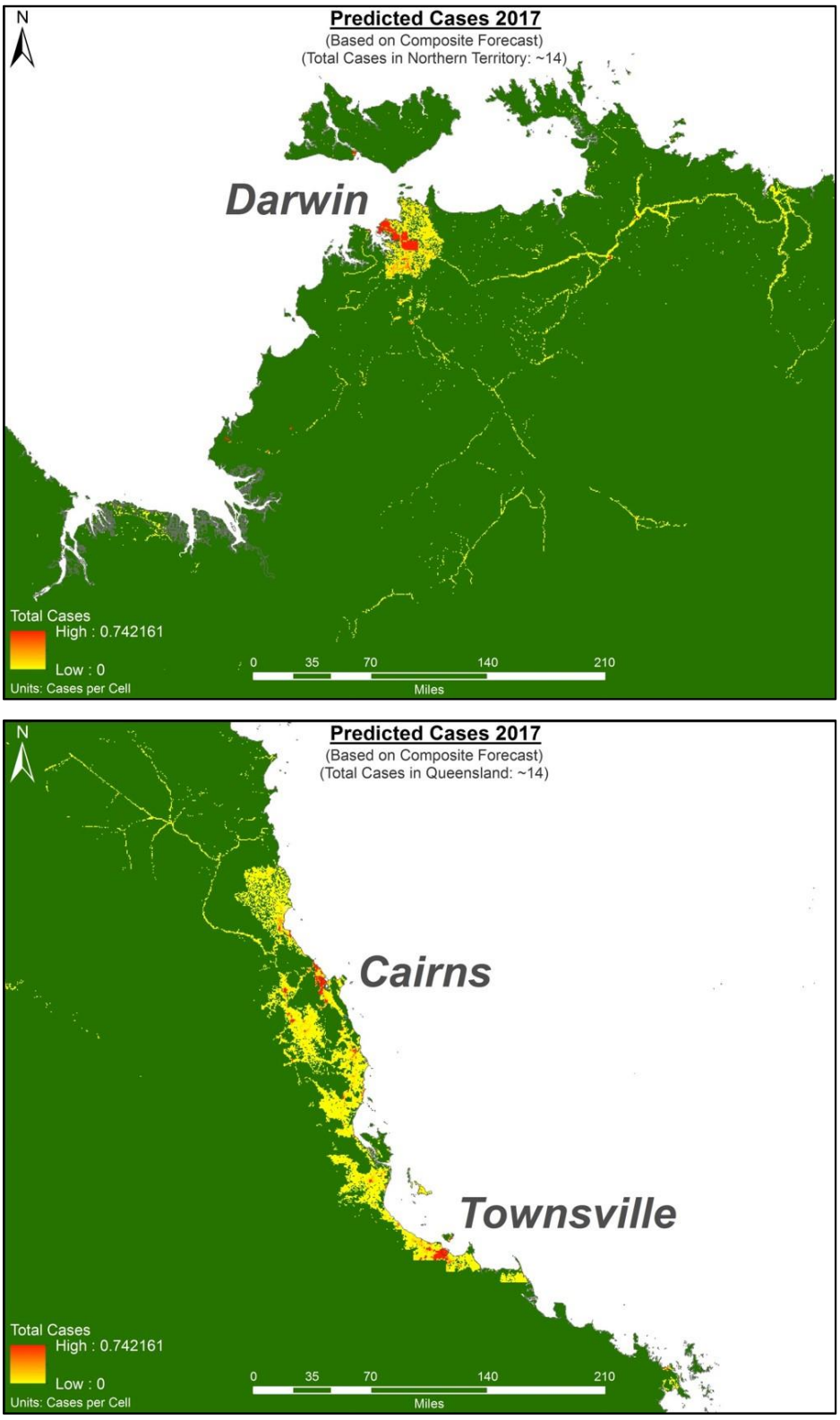


Figure 2.11: The melioidosis incidence for 2014 (upper), which we predicted would see moderate incidence across a wide area, and 2009 (lower) which we predicted experienced high incidence across the nation.

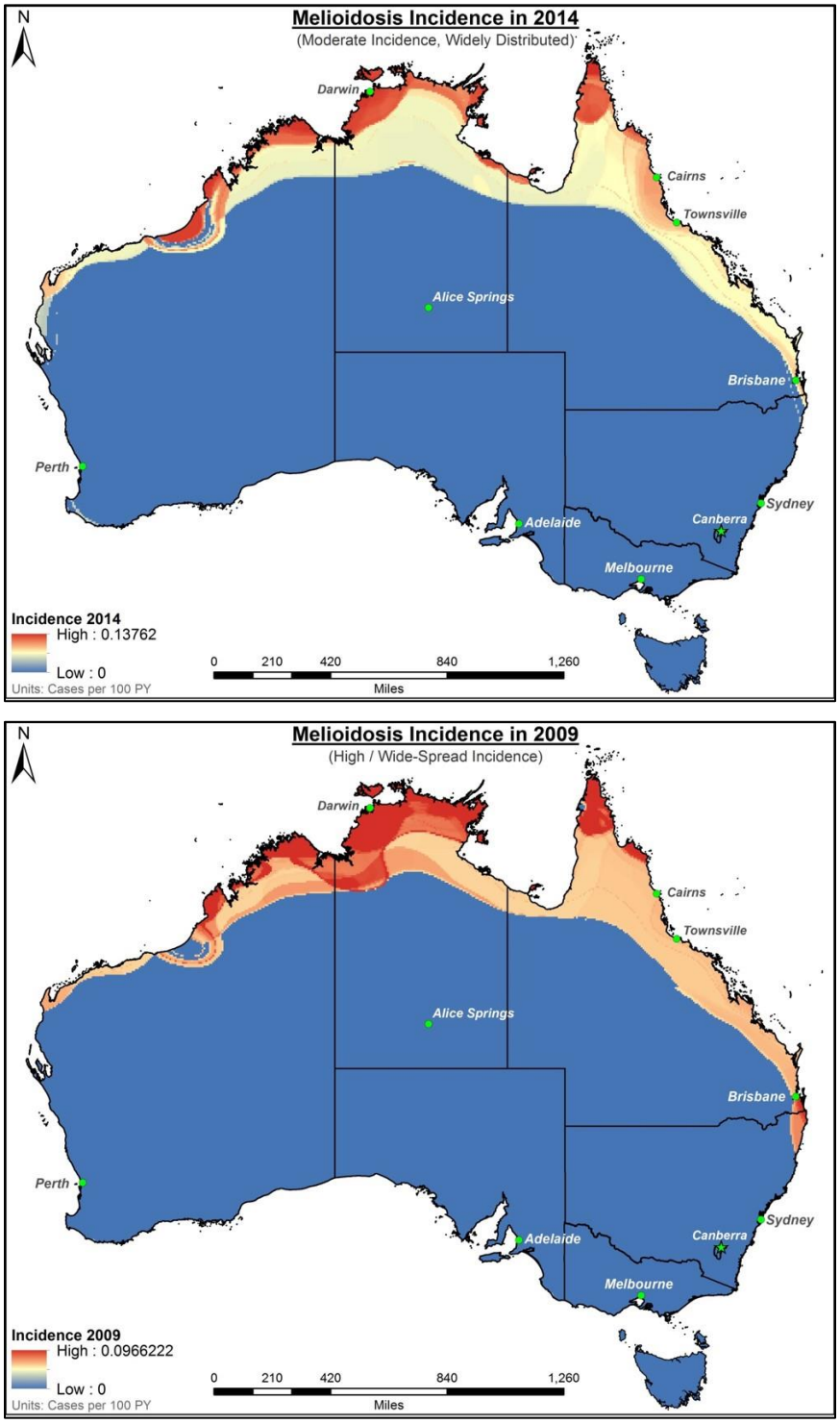


Figure 2.12: Jackknife derived variable importance for the Townsville model. In this case we see that soil type and elevation are both relevant. In fact, a model built solely on elevation would still have an AUC of 0.83.

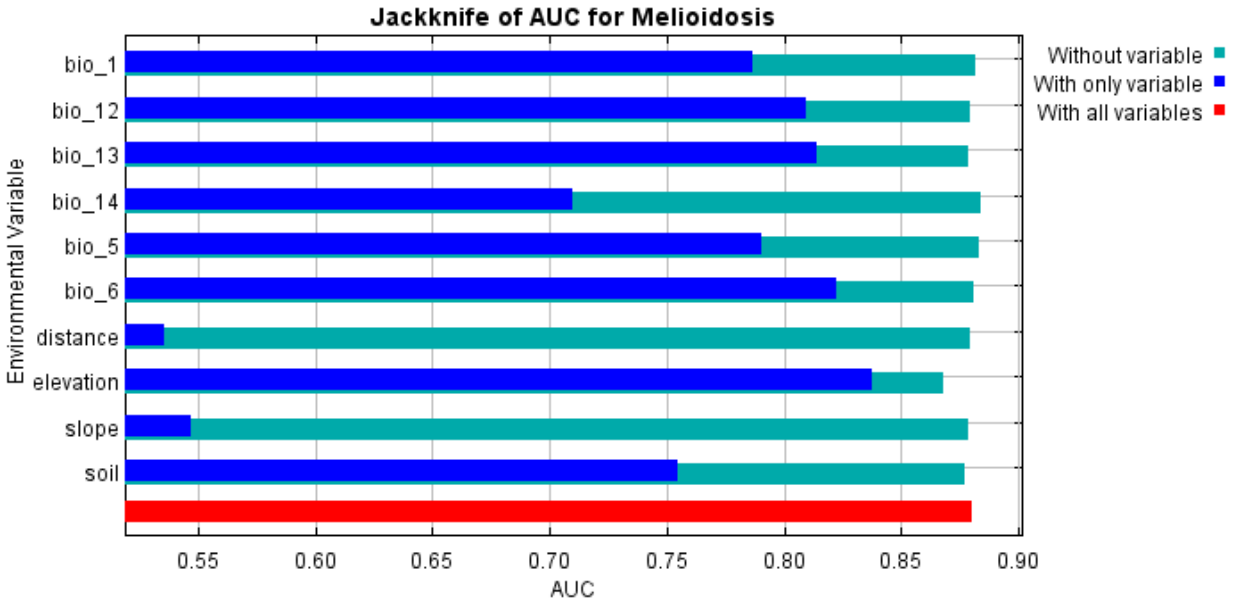


Figure 2.13: Selected response curves for the Townsville model. Here we see how elevation (upper), soil type (middle) and precipitation of wettest month (lower) affect melioidosis in our study area.

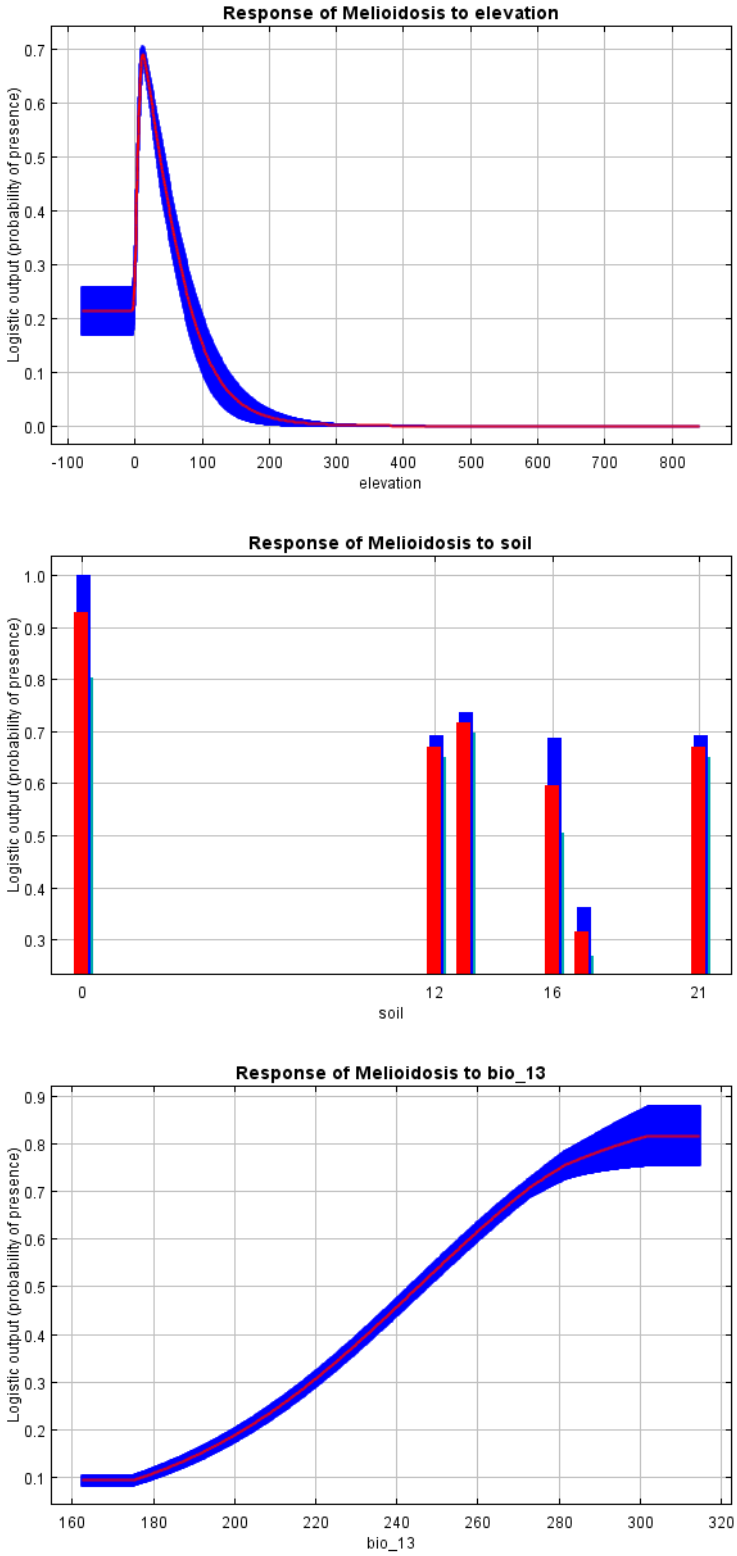


Figure 2.14: The differences between forecast and actual weather for the 2017 season.

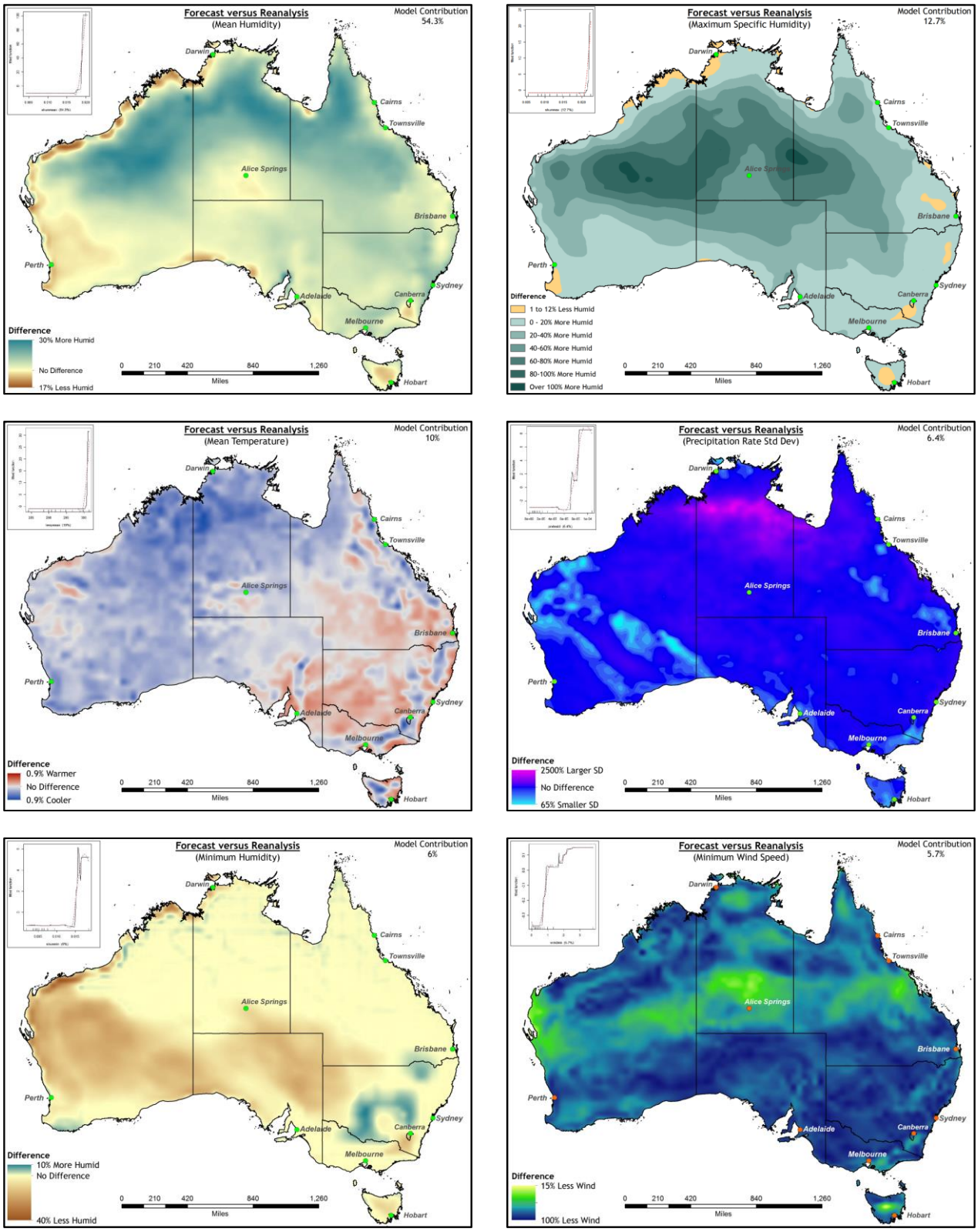


Figure 2.15: Our updated prediction of melioidosis incidence in 2017, made with the re-analysis data, after the close of the season. Though we are seeing some artifacts in Western Australia, these are generally underpopulated areas, and our expected cases were quite close to predicted values.

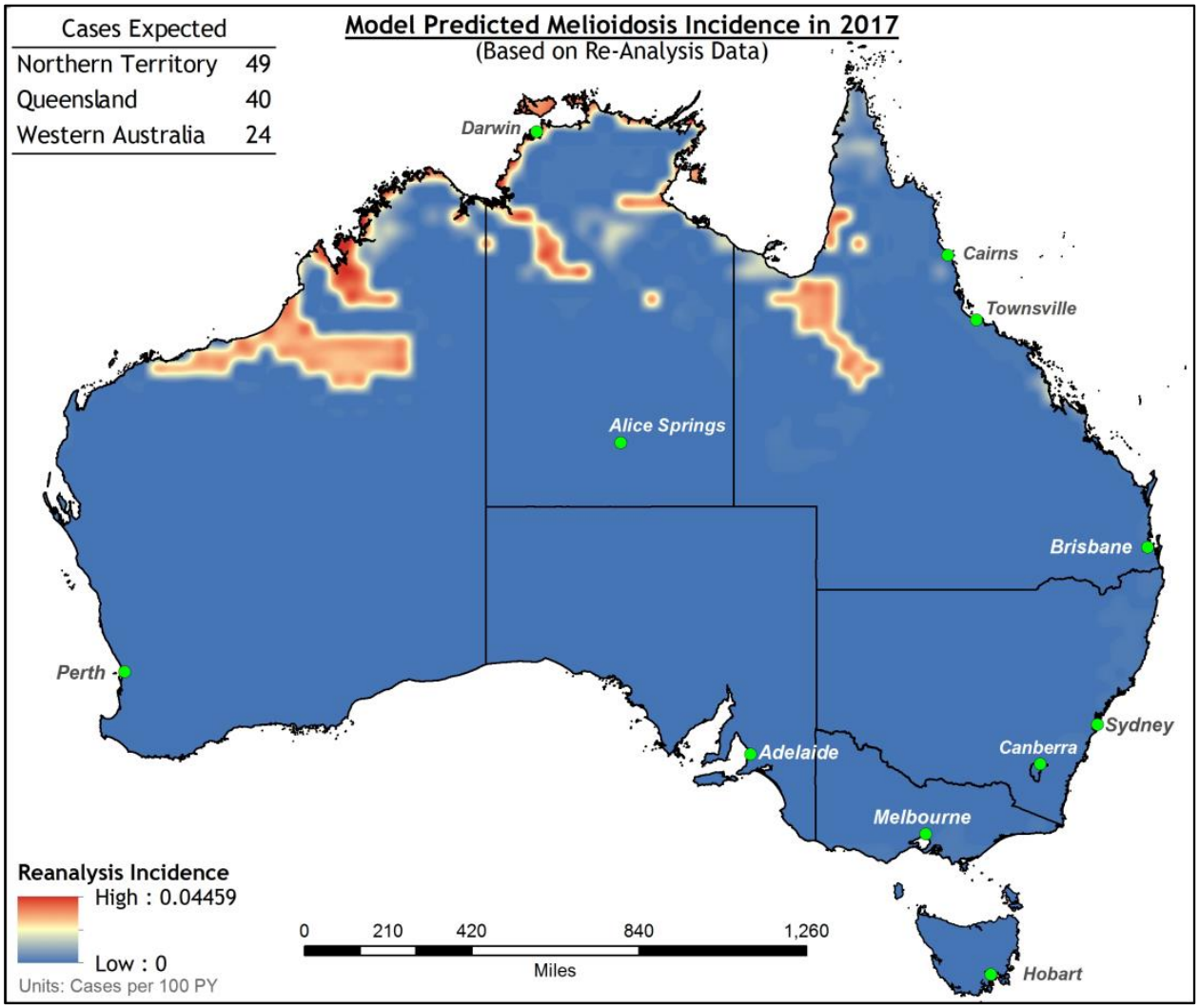


Table 2.1: Variable contributions for the initial melioidosis forecasting model. Note that the top ten variables account for 90% of the cumulative contribution.

Variable	Percent	Cumulative
Mean Specific Humidity	52.50%	52.50%
Max Specific Humidity	6.60%	59.20%
Mean Temperature	6.50%	65.70%
Elevation	6.30%	72.00%
Min Wind Speed	4.50%	76.50%
Landform Curvature	3.40%	79.90%
Min Rain Fall	3.10%	83.00%
Min Specific Humidity	2.90%	85.90%
SD of Rain Fall	1.70%	87.70%
Landform Slope	1.60%	89.30%
Max Rain Fall	1.60%	90.90%
Mean Lat. Heat Flux	1.30%	92.20%
SD of Wind Speed	1.20%	93.40%
Mean Wind Speed	1.10%	94.60%
Min Temperature	1.00%	95.60%
SD Temperature	0.90%	96.40%
Min Lat. Heat Flux	0.80%	97.30%
SD Specific Humidity	0.80%	98.10%
Max Lat. Heat Flux	0.70%	98.70%
Max Temperature	0.30%	99.00%
Max Wind Speed	0.30%	99.30%
Mean Precipitation Rate	0.20%	99.50%
SD Lat. Heat Flux	0.20%	99.70%
SD Soil Moisture	0.10%	99.80%
Max Soil Moisture	0.10%	100.00%
Min Soil Moisture	0.00%	100.00%
Mean Soil Moisture	0.00%	100.00%

Table 2.2: Variable contributions for our parsimonious forecasting model. Note that terrain-based variables were removed due to issues with confounding.

Variable	Percent	Cumulative
Mean Specific Humidity	54.3%	54.3%
Max Specific Humidity	12.7%	67.0%
Mean Temperature	10.0%	77.0%
SD Rain Fall	6.4%	83.4%
Min Specific Humidity	6.0%	89.4%
Min Wind Speed	5.7%	95.1%
Min Rain Fall	4.9%	100.0%

Table 2.3: To determine which year of historical climate data was most similar to the 2017 NCEP forecast for any given variable, we compared the Mean Square Error at 1000 randomly dropped points, as well as the correlation coefficient between the two rasters. Here we see this technique applied to mean humidity, clearly showing that 2003 is the most representative historical grid. Note that the Spearman's Rank-Order correlation coefficient between the two methods was 0.8679, suggesting a high degree of agreement between the two regarding which year is most appropriate.

Year	MSE	Rank	PCC	Rank
2000	0.004025	14	0.93448	14
2001	0.003875	13	0.94048	13
2002	0.002322	3	0.96013	4
2003	0.001869	1	0.96422	1
2004	0.002655	6	0.95424	5
2005	0.001903	2	0.96129	2
2006	0.003388	10	0.94513	11
2007	0.002705	7	0.94895	10
2008	0.002983	8	0.94197	12
2009	0.003111	9	0.95193	8
2010	0.003523	11	0.94926	9
2011	0.005919	15	0.92449	15
2012	0.002551	5	0.95306	7
2013	0.002404	4	0.96066	3
2014	0.003542	12	0.95394	6

Table 2.4: Shows variable contributions for the environmental niche models created to approximate the Limmathurotsakul (31) study which provided the bulk of our case data. Note that both modeling algorithms determined that rainfall of the wettest month was most significant.

Variable	BRT Importance	MaxEnt Importance
Rain of Wettest Month	63.3	29.4
Max Temp. Warmest Month	11.7	12.3
Annual Mean Temp.	10.4	2.8
Annual Precipitation	6.2	34.3
Min Temp. Coldest Month	5.0	15.7
Rain of Driest Month	3.0	0.8
Soil Type	0.8	3.9

Table 2.5: The number of melioidosis cases by state, as estimated by the forecast and reanalysis models, compared with actual case counts from public records. Note that total cases for Queensland were estimated using the seasonal epidemic curves from Stewart (87). News sources for the 2017 cases are included below.

State	Forecast	Reanalysis	2017 Cases	Source
Northern Territory	14	49	53	[1]
Queensland	14	40	37	[2]
Western Australia	4	24	8	[3]

Sources:

1: https://www.sbs.com.au/news/nt-rains-signal-deadly-melioidosis-mozzies

2: http://www.cairnspost.com.au/lifestyle/spike-in-amount-of-melioidosis-cases-in-far-north/news-story/6937dc9cd3c7368bda7655b452187370

3: http://ww2.health.wa.gov.au/Articles/N_R/Notifiable-infectious-disease-report?report=melioidosis

Table 2.6: The Townsville specific niche model incorporating fine-scale terrain metrics, which are shown in red. Though this model cannot be projected across the entire continent, it lends credence to the idea that elevation and soil are significantly associated with melioidosis incidence.

Variable	Percent Contribution	Permutation Importance
Elevation	33.9	38.6
Soil Type	30.9	6.2
Precipitation of Wettest Month	14.7	4.4
Slope	5.8	5.7
Min Temp Coldest Month	5.7	28
Annual Precipitation	3.3	3.1
Max Temp Warmest Month	2.2	8.9
Annual Mean Temperature	1.7	0.6
Precipitation of Driest Month	1.6	4.3
Distance to Water	0.2	0.2

Chapter

3. Methods for Rapid Mobility Estimation to Support Outbreak Response

Attribution

The following chapter is based on a manuscript under review by *Health Security*. Co-authors include Dr. Srinivasan Venkatramanan and Patrick Corbett, who were responsible for the meta-population patch model, and Dr. Bryan Lewis, who assisted with the modeling and supervised the efforts.

3.1 Abstract

Outbreak investigators often assume homogeneous mixing models for human and disease mobility when modeling infectious diseases in data poor regions. But recent outbreaks such as the 2014 Ebola outbreak of West Africa have shown the limitations of this approach in an era of increasing urbanization and connectivity. Both outbreak detection and predictive modeling depend on realistic estimates of human mobility, but these are difficult to obtain in a timely manner, especially when dealing with an emerging outbreak in an under resourced nation. Weighted travel networks with realistic estimates for population flows are often proprietary, expensive, or non-existent. In this paper, we propose a method for rapidly generating a mobility model from open-source data. As an example, we use road and river network data, along with population estimates, to construct a realistic model of human movement between health zones in Democratic Republic Congo (DRC). Using these mobility data, we then fit an epidemic model to real-world surveillance data from the recent Ebola outbreak in the Nord Kivu region of the DRC to illustrate a potential use of the generated mobility estimation. In addition to providing a way for rapid risk estimation, this approach brings together novel techniques to merge diverse GIS datasets that can then be used to address issues that pertain to public health and global health security.

3.2 Introduction

For those investigating neglected tropical disease outbreaks, the issue of disease mobility is a novel concern. Diseases like Ebola would emerge in an isolated community, raze the area, then fade away (105,106). The likelihood of an Ebola outbreak reaching a major urban area was small, and the threat posed to the security of developed nations was nearly inconsequential. Prior to the 2014, only a single Ebola epidemic garnered significant international concern (107). But recent outbreaks have shown that neglecting mobility during an ongoing outbreak response is both obsolete and dangerous. The 2014 West Africa Ebola outbreak provides a convincing example of this new paradigm. The initial spillover occurred near the village of Meliandou in southern Guinea (108). By our estimates, using LandScan 2013 (109), Meliandou has a population of roughly 100-200 people. But it lies just 12 km northeast of the city of Guéckédou, a town of an estimated 145,000. The surrounding area is one of the most highly urbanized parts of West Africa (110). Due to this high degree of urban connectivity, the virus quickly spread west to Conakry (108). International borders could not contain the virus, which temporal phylogeny shows soon reached the Kailahun District of Sierra Leone and Lofa County in Liberia (53). Eventually afflicting the metro areas of Freetown and Monrovia, the outbreak would go on to sicken 28,652, of whom 11,325 died (111). Though the threat of sustained autochthonous transmission in the West was perhaps overstated, nosocomial cases in the United States and Europe were documented, and the threat of further introductions by way of air travel was not insignificant, despite travel restrictions (112,113).

The Ebola outbreak was not alone in demonstrating the need for spatial modeling. The 2015-16 Zika virus epidemic showed a similar pattern. Wallowing in obscurity for nearly 70 years, the virus reached Brazil via French Polynesia in 2013 (54). There it found that which had eluded it thus far:

highly connected and urbanized population centers coinciding with the appropriate mosquito vector (29). First noticed in early 2015, by November of the same year, the virus had spread across the entire continent reaching as far as Mexico and parts of the United States (114).

At the time of writing, this theme threatens to repeat itself. Ebola is back in the international spotlight, having recently threatened Mbandaka, a major city in the Democratic Republic of Congo (DRC) and by way of the Congo River, Kinshasa, a city of nearly ten million (115–117). A second Ebola outbreak in the DRC, this time near Beni, population of 1.4 million, is currently ongoing (116). These two outbreaks are not outliers. On the contrary, the rapid growth of population centers and increasing interconnectivity in the developing world threaten to make these events commonplace in the future.

Models of the spread of the 2014 Ebola outbreak across West Africa showed that the strongest predictors for dispersion were population density and travel time (53). Both drive mobile epidemics, and both are growing at an impressive rate (116). The entire matter is made worse by a lagging healthcare infrastructure, which is a near ubiquitous problem in tropical developing nations (117).

3.2.1 Spatial Modeling

The conclusion is inexorable: epidemiologists must include mobility in their analyses. In all phases of outbreak investigation – from early detection to predictive modeling of growth and spread – some understanding of the spatial connectivity of the region is necessary.

Outbreak detection methods that make assumptions about mobility based on polygonal contiguity and geographic distances are improper. In truth, hierarchical diffusion dictates that the traffic between two distant urban centers connected by a major highway likely far exceeds the traffic to a smaller town that is geographically proximal. Likewise, the use of differential equation based compartmental models, which assume random mixing across an entire nation (45), are equally inappropriate. In both cases, accounting for mobility can greatly improve accuracy and realism.

For outbreak detection, we can account for mobility with a modified spatial weights matrix. Outbreak detection software allows users to manually define the relationships between sub-regions (e.g. health districts or census tracts). By using a mobility model to generate spatial weights instead of based them on geographic relationships alone, we can ensure that we have an accurate representation of interconnectivity between areal units.

For predictive modeling, mobility can be included with a metapopulation patch model (45). The traditional compartmental model for Ebola bins the study area population into familiar categories such as Susceptible, Exposed, Infectious, Recovered, Hospitalized, and Unburied (118). It is quick to deploy but assumes random mixing across the entire study area. Conversely, an agent-based model can realistically represent mobility (119), but requires calibrated synthetic population data and is too complex to deploy rapidly. The patch model offers an elegant compromise.

The patch model extends the compartmental model concept by breaking the study area into multiple patches, each representing a smaller region in which random mixing is a closer approximation. These connected patches can then be fit at both the overall population level as well as each patch individually (120).

3.2.2 Mobility Approximation

As both outbreak detection and predictive modeling are predicated upon a thorough understanding of the interactions between each region, mobility metrics are crucial. Though curated data exists for the United States, it is often unavailable for developing nations. In its absence, the modeler can estimate mobility using a gravity model (110,121) or radiation model (43,122). Both require a measure of population for each source and destination patch, and distance between them.

In small study areas one could make use of Euclidean “crow-flight” distance between the centroids of patches, but this is too simplistic for a national model. Across an area the size of the DRC, bisected by rivers and mountains, the use of network-based travel time is necessary to more accurately estimate mobility. Once must also factor multiple modes of transportation and differing speeds. In many developing nations, major waterways are critical to the movement of goods and people and cannot be ignored, along with rail lines, roads, walking paths, and flight connections (116).

In the absence of curated mobility data, mobility models should be generated using validated transportation network data, such as those provided by NavTeq/Here or Google. But these datasets are often proprietary and costly. In many cases the modeler has access to neither mobility data nor travel network data. In such cases, and when time and resources are limited, one can construct a reasonable approximation using open source data instead.

In this paper, we examine sources of such data, as well as the process by which one could rapidly create a travel network, and an accompanying gravity model of mobility for use in patch modeling or as a spatial weight matrix for outbreak detection. As a case study, we also explore the creation of a patch model for use in the investigation of an ongoing outbreak of Ebola in DRC.

3.2.3 Objective and Significance

The objective of this study is to demonstrate the feasibility of rapidly creating transportation networks and gravity models of mobility, in data sparse areas such as Central Africa. Such metrics are crucial to many public health problems including the detection of outbreaks with hot spot analyses, and modeling the diffusion of a disease through a community, the latter of which is vital for targeting and evaluation of potential public health responses. When time is of no consequence, or the researcher has access to professionally curated mobility data, the methods of this paper are superfluous. The value of these methods is more apparent in time critical situations when other sources of mobility data are unavailable. Though imperfect, they are a vast improvement over network-agnostic methods such as polygonal contiguity or Euclidean distance metrics. Accordingly, we feel that these methods may improve the public health response to epidemics in rural regions by giving researchers and public health professionals a more realistic view of the spatial complexities of the epidemic.

3.3 Methods and Materials

3.3.1 Choosing Open Source Data

OpenStreetMap (OSM) is perhaps the most well-known source of open geospatial network data. OSM products are incredibly detailed, but the process by which their network data is created often leaves small disconnects in the road network. To network analysis software like Esri's Network Analyst or Python NetworkX, these breaks appear impassable. Whether miles or millimeters, any disconnect makes a potential route invalid. With thousands of these gaps across a nation the size of the DRC, network analysis is significantly hampered. One could repair these manually, but not if the work is time-sensitive. One could automate the repair process by using topology rules to eliminate dangles, but this runs a risk of creating false connections. For example, a road that leads to a cliff or riverbank could be inappropriately connected to a road on the other side, provided both were within the preset distance threshold.

Though often less detailed than OSM, other open source alternatives exist. For our work, we chose to make use of the Digital Chart of the World (DCW) (85), available from diva-gis.org. The roads and rivers of DCW are contiguous, without disconnects, and like OSM the dataset covers the entire globe. While the road network detail of OSM is missing, DCW has the detail needed to model interactions between large areal units such as health districts.

The primary disadvantage is that DCW was last updated in 1992. Not only have major highways been constructed since, but the courses of major rivers have slightly changed. Nevertheless, upon comparing the DCW to OSM, as seen in **Figure 3.1**, we feel that DCW will suffice when modeling travel between large areal units such as health districts. DCW is in fact, more detailed than OSM when it comes to the river network. For those who prefer OSM, the methodology we outline in this paper could be applied to their products just as easily.

3.3.2 Inland Water Polygons

In DRC, the rail network suffers significant neglect, and large-scale commercial air travel is limited, but the Congo river remains a primary mode of transportation for goods and people (123). As such, we must take the river and its navigable tributaries into account. But those attempting to convert inland waterway data to a traversable network will run into an unexpected impediment: the use of polygonal data. Tributaries and small streams are represented by a shapefile containing polylines, but wide stretches of river, lakes, and areas with significant river islands are typically represented by a second file containing polygons. This is evident in **Figure 3.2** which shows the area around Lake Mai-Ndombe. Merging these two files together poses difficulties. If one were to convert these polygons to an edge-list network, travel would only be possible along the river's banks. An *in-silico* agent would not be able to cross the lake or river, only travel around it.

One could reduce these polygonal outlines to one dimensional network edges by dropping a centerline between the polygon's outer boundaries. The "Production Centerline" tool in ArcMap, or the "v.centerline" tool in GRASS, can accomplish this. However, there are limitations to this approach as well. If the tributaries are connected to the river and lake banks, they will not reach the newly created centerlines, and the network will be effectively broken. Consider again Lake

Mai-Ndombe. Reducing the polygonal outline of the lake to its centerline disconnects it from most of its tributaries, as seen in **Figure 3.3**.

One could manually fix these issues, but not for an area as large as the DRC. Furthermore, these tools also give unusual outputs when dealing with rounded features such as lakes, ponds, and reservoirs. Finally, when dealing with a waterbody as significant as the Congo River, which measures nearly 20 kilometers across at its widest point, reducing the width to zero results in an unrealistic representation of real-world travel.

Ideally, we would like to construct a network that 1. includes the centerline, representing travel up or down the river, 2. maintains the original banks and true width of the river, and 3. includes edges providing realistic river or lake crossings. We accomplish this with the following procedure in Esri ArcMap 10.5.1:

3.3.3 Water Polygons to Edges

We begin by converting the inland water polygons to points representing the outline of the water body using “Feature Vertices to Points” with all vertices selected. This is followed by a Voronoi tessellation using these points as input to “Thiessen Polygons”. This creates a centerline as well as numerous lines projected perpendicularly from the water’s edge to said centerline. To remove features outside of the silhouette of the original waterbody, we simply “Clip” the tessellated output with the original waterway polygon. We then use the “Multipart to Singlepart” tool to eliminate any resulting multipart polygons. An example of the output of this operation is shown in **Figure 3.4.**

As vertices are concentrated in areas with significant curvature, this procedure creates an overabundance of extremely thin polygons in sharp river bends called slivers. These polygons, which often run parallel to larger ones, become superfluous edges in the finished product.

Slivers can be identified as those with a very small Thinness Ratio, which is calculated using the “Polygon Sliver Check” tool. Barring access to this tool, one could manually calculate Thinness Ratio using the following formula, where T is the thinness ratio, A is the area of the polygon, and P the perimeter (124):

$$T = 4\pi \frac{A}{P^2}$$

We removed polygons with a thinness ratio of less than 0.3, and polygons with an area of less than 100 hectares, using “Eliminate (Data Management)”.

The polygons were then converted to lines using “Polygon to Line” and the resulting file was then merged with the file representing the tributaries from DCW. To ensure there was no overlap and that features were appropriately connected, we used “Integrate (Data Management)” with an XY tolerance of five meters. The finished river network file, an example of which can be seen in **Figure 3.5**, could be converted to an edge-list for use in NetworkX, or ArcMap Network Analyst. Instead we chose to combine it with road network data to create a network representing the two primary methods of travel in the DRC (123).

3.3.4 Combining with Road Data

For our modeling efforts we merged the river data constructed in the prior section with the road data provided by DCW. We assume that river crossings are common enough in the DRC that any

intersection between a major road and a large body of water will include a ford, ferry or dock. We do not incorporate a delay when switching between modes of transportation, though this could be done by using weighted barriers in ArcMap. Note that bridges should already be included in the road data and not affected by the merge.

After merging the files, we again ran an “Integrate” to ensure that the road and river lines intersected when nearby. This was followed by a “Planarize Lines” function to split features at their intersections. Finally, we estimated speeds for each mode of travel; three meters per second for river travel, nine m/s for road travel, and 20 m/s for highway travel. Using these speeds, we calculated the time to traverse each edge of the newly created bimodal travel network. The resulting file can be exported as an edge-list network or converted to a network database in Network Analyst.

3.3.5 Sources of Population Data

To estimate population for each health district we made use of LandScan 2017 (109), a gridded population data set at a resolution of 30x30 arc-seconds, and a map of Health Districts from the Humanitarian Data Exchange (125). An alternative source of population data is WorldPop 2015 (126), which includes most of Africa at a 3x3 arc-second resolution. The advantage of LandScan in this case was memory usage, as WorldPop is both 100 times the resolution and floating point rather than integer. This is not a significant hurdle when calculating the zonal statistics, but it significantly hampers population weighted travel time calculations.

3.3.6 Mean Centers

As with all patch models, the health districts in question are polygonal in nature. When calculating distances between them for use in our model, we need to reduce this complex structure to a point, or series of points. The most obvious means by which to do this is to simply take geometric centroids. The limitation of this approach is that the population centers of the districts are often not near the centroid, and the resulting district-to-district distance is unrealistic.

An alternative is to use a population grid and calculate the distance between each pair of cells then aggregate by district. This is extremely resource intensive, and as such we propose using population weighted centroids instead. This is a compromise in that it requires less intensive computation than district-to-district distance calculations but is significantly more realistic than geometric centroids. Though confounded by districts with two distinct population centers on opposing sides, we feel it is sufficient for situations requiring rapid deployment of models.

We begin by discarding all population grid cells with zero values for the sake of parsimony. After doing so, we converted the raster cells into points using the eponymously named tool. With a shapefile of DRC health districts, we assigned each point a Health District ID using a “Spatial Join”. Finally, we applied “Mean Center (Spatial Statistics)” with the raster values as the weight field, and health district ID as the case field. The output of this is a point file where each point represents the population weighted centroid of its respective health district.

3.3.7 Network Analysis

With the transportation network, the population data, and the population weighted centers in hand, our last task was to create an Origin-Destination travel cost matrix. To do so we created a network

dataset from the travel network, ensuring that travel time was a network attribute and used as an evaluator with the proper temporal unit.

After the network dataset is created, we used the “New OD Cost Matrix tool” with impedance set to travel time to create our travel matrix. These data can be extracted as a CSV and used to create a gravity or radiation model of mobility using standard methods (20, 21).

3.3.8 Risk Estimation for Network Epidemics

Network-based models in epidemiology (127) have gained prominence in recent times, thanks to their ability to model mobility-induced interaction mechanisms either at individual or population level, and the increasing availability of data pertaining to such interactions. Metapopulation models (45,128) provide an elegant framework to capture interactions between subpopulations, which are themselves homogeneously mixing.

Given the populations of each patch P_i (in this case, health districts) and the OD travel matrix among the patches $T_{i,j}$, we employ a gravity model to generate the flow matrix satisfying:

$$\text{(a) } F_{i,j} \propto \frac{P_i * P_j}{T_{i,j}^2} \text{ and (b) } \sum_j F_{i,j} = 1$$

We then simulate a metapopulation SEIR (Susceptible-Exposed-Infectious-Recovered) model among the patches, with $F_{i,j}$ representing the fraction of individuals of from patch i spending their day in patch j on any given day. This approach is mainly used to mimic commuter-like mobility between the patches. More details on the method can be found in literature (129).

Such a metapopulation model can be used for several purposes such as disease forecasting and planning. In this case study, we will estimate early risk. To do so, we initialize the infections in the metapopulation model appropriately (n_0 initial cases in a seed patch i_0) and run the model forward with different levels of transmissibility β and time horizon T. At the end of T, for each patch the cumulative risk is estimated as:

$$Risk(i) = \frac{R_i(T)}{\sum_j R_j(T)}$$

Where $R_i(T)$ is the number of recovered individuals in patch i at time T. When coupled with a forecast, the projected risk is obtained by taking the proportion of incremental cases, instead of cumulative recovered cases.

3.3.9 Patch Model Calibration

We created five separate configuration stages for the most pressing North Kivu health districts: Mabalako, Beni, Butembo, Katwa, Kalnguta, and Oicha. Each configuration utilized uniform alpha and gamma parameters, in this case 0.133 and 0.1 respectively, corresponding to an incubation period of 7.5 days and infectious period of 10 days (118). Meanwhile, variations were made to the beta parameters, the rate at which susceptible-infected contact results in new infections. These adjustments were made to approximate major inflection points of the different health districts, time points that likely coincided with newly introduced interventions or super-spreading events. We assume the true number of cases generally exceed those observed, thus the simulated case counts are generally higher than the observed ground truth being fitted.

In adjusting the beta parameter at each of the five stages, we found several interesting trends. First, we noticed a sharp increase in the rate of new infections in Beni from stage one (days 20-70) to stage two (days 70-110). To account for this, we increased the beta value by 92%. The ground data then experienced three approximate inflection points in stages three (days 110-140), four, and five. We accordingly decreased the beta value by around 50%, 33%, and 75%. This decrease illustrates a significant effect on the spread of Ebola in Beni.

3.4 Results

The resulting files from this methodological example will be included in the supplemental materials of this paper. These data will include the population weighted health district centroids, the travel network, an ArcMap compatible network dataset, and the OD cost matrix we used to generate our mobility model. **Figure 3.6** shows the influence of Beni on neighboring districts of Nord Kivu as a function of disease mobility. In this case, keeping with hierarchical diffusion, we expect Katwa, a major city in the south, to be the most likely destination for new Ebola cases derived from the Beni outbreak. For comparison, **Figure 3.7** shows the distance from Beni as a function of polygonal contiguity and Euclidean distance. Though the latter two are common metrics for a spatial weights matrices, they bear little resemblance to the interactions predicted by our mobility model. Using case counts updated as of January 23, 2019 the current outbreak is calibrated both at the national and health district level, as seen in **Figure 3.8.**

The mobility linked health zones are independently calibrated at different inflection points to account for variations in the force of infection within the health zones themselves. While the majority of cases have occurred in Beni, the activity there has slowed significantly in the preceding months, as has activity in Mabalako, the initial origin of this outbreak. Recent growth is mainly seen in Katwa and Butembo, which leads to the significant slope of the projected case counts in the simulations for day 180 to 210.

Ebola's stochastic nature makes it notoriously difficult to forecast, especially at the resolution of health districts and in the context of a volatile security situation which likely affects mobility itself. As such the uncertainty of any single health zone's trajectory may be large and should not be considered as independent forecasts. To provide an idea of what areas are at highest risk for future cases, we use the simulated case counts projected forward 30 days from January 23rd and look at the relative proportions of cases generated in the health zones. The top 10 relative proportions of future cases are presented in **Table 3.1**, which also shows distance from Beni as a function of polygonal and Euclidean distances. Again, the latter metrics are far from adequate representations of mobility.

3.5 Discussion

These methods provide a novel means to quickly estimate mobility patterns in information poor areas to support outbreak investigations. This allows us to put constraints on how people move around and transport diseases in our models and can improve both outbreak detection and predictive modeling.

In our experience, one could follow the above procedures and generate a transportation network in less than a day, with a mobility model deployed in less than a week. The results of such efforts would improve upon any model which assumes random mixing across a nation the size of the DRC. The method also improves upon outbreak detection methods which use Euclidean distance or polygonal contiguity. These procedures are most suitable to real-time epidemic modeling efforts when rapid turnaround is required to help inform the ongoing public health response. This is particularly true in developing nations where curated travel network data are scarce.

That said, there are a myriad of limitations with this methodology. It is not an adequate replacement for a professionally curated travel network, and even less a replacement for calibrated mobility data. If such data are available for the study area and can be sourced in a timely manner, it is highly preferable to make use of them. If, however, no such data exist, the method outlined above certainly improves upon spatially-unaware efforts.

3.5.1 Limitations

Estimating speed for rural roads in a developing world is difficult. In truth, the speed is probably highly variable depending on the weather and method of transportation. Even on a major highway, a moped might travel half the speed of an automobile. On a rural route, we may need to include bicycles and animal carts. Similarly, we have no way of estimating the speed variability in river traffic, or the effect of the river's current. Furthermore, unpaved rural routes are often washed out and impassable during periods of high rain, and we have no means by which to account for this.

Beyond that, the DCW data we made use of is out-of-date, meaning new highways constructed between growing urban areas and newly constructed bridges are unaccounted for. The method also assumes uniform rates of water-body crossings, when in reality the schedule of ferries may confound travel time metrics. We are also unable to account for the effects of sociopolitical boundaries, including tribal affiliation, ethnicity, religion, and political affiliation. Such boundaries may be as just impassable as any river or mountain, and territorial disputes between government and rebel forces in DRC are currently confounding the Ebola response (130–132). Finally, nothing in the methodology currently accounts for the fluctuations that an emerging infectious disease can introduce to human mobility. The outbreak may increase mobility as locals flee the area or may depress it as locals self-quarantine. Though many other methods of mobility estimation, including those based on historical cell tower records, also suffer the same challenges making real-time inferences.

3.5.2 Uncertainty and Sensitivity Analysis

Uncertainty is difficult to quantify in such a study. The purpose of this method is to approximate mobility in an area with significant data sparsity. As such there are no competing models for comparison. Certainly, there is substantial uncertainty in both the shape and speed of the

transportation network, but no obvious way to account for these in the absence of higher fidelity data. One could use competing transportation networks to increase confidence in our resulting OD cost matrix, but again, such data are often unavailable. Sensitivity could be modeled by introducing random variations in traversal speeds for each edge in the network and then repeating the analysis to determine how disease spread is affected. But rapid deployment is an integral feature of this methodology, and as such we did not delve further into sensitivity analysis.

3.5.3 Conclusion

This methodology allows one to rapidly create a transportation network and mobility model for use is epidemic response. It can be used to augment outbreak detection or forecast the spread of a disease, despite an absence of preexisting mobility data. We have demonstrated its use in the forecasting of Ebola spread from Beni in the DRC and identified Katwa as a probable destination for new cases originating from Beni. The method is not intended to replace curated mobility or transportation network data. However, in the event of an ongoing epidemic, when time is limited and data are scarce, it can significantly improve upon more traditional distance or polygonal-contiguity based analyses.

3.6 Funding and Acknowledgements

This study was supported by the Defense Threat Reduction Agency (DTRA) Comprehensive National Incident Management System (CNIMS) Contract HDTRA1-17-0118; and the National Institutes of Health (NIH) and National Institute of General Medical Sciences (NIGMS) Models of Infectious Disease Agent Study (MIDAS) Cooperative Agreement U01GM070694.

We thank Dr. Caitlin Rivers, Johns Hopkins University, and Sophie Meakin, University of Warwick, for compiling the Ebola situation reports with case counts into machine-readable datasets. We also acknowledge the WHO and CDC personnel who maintain the DRC health district shapefiles on the Humanitarian Data Exchange. The authors declare no conflicts of interest.

Figure 3.1: A comparison in the density and detail of the road and river networks provided by OSM and DCW. Both are detailed enough to model the interactions of health districts. Note that provinces of DRC are added in pastel for scale.

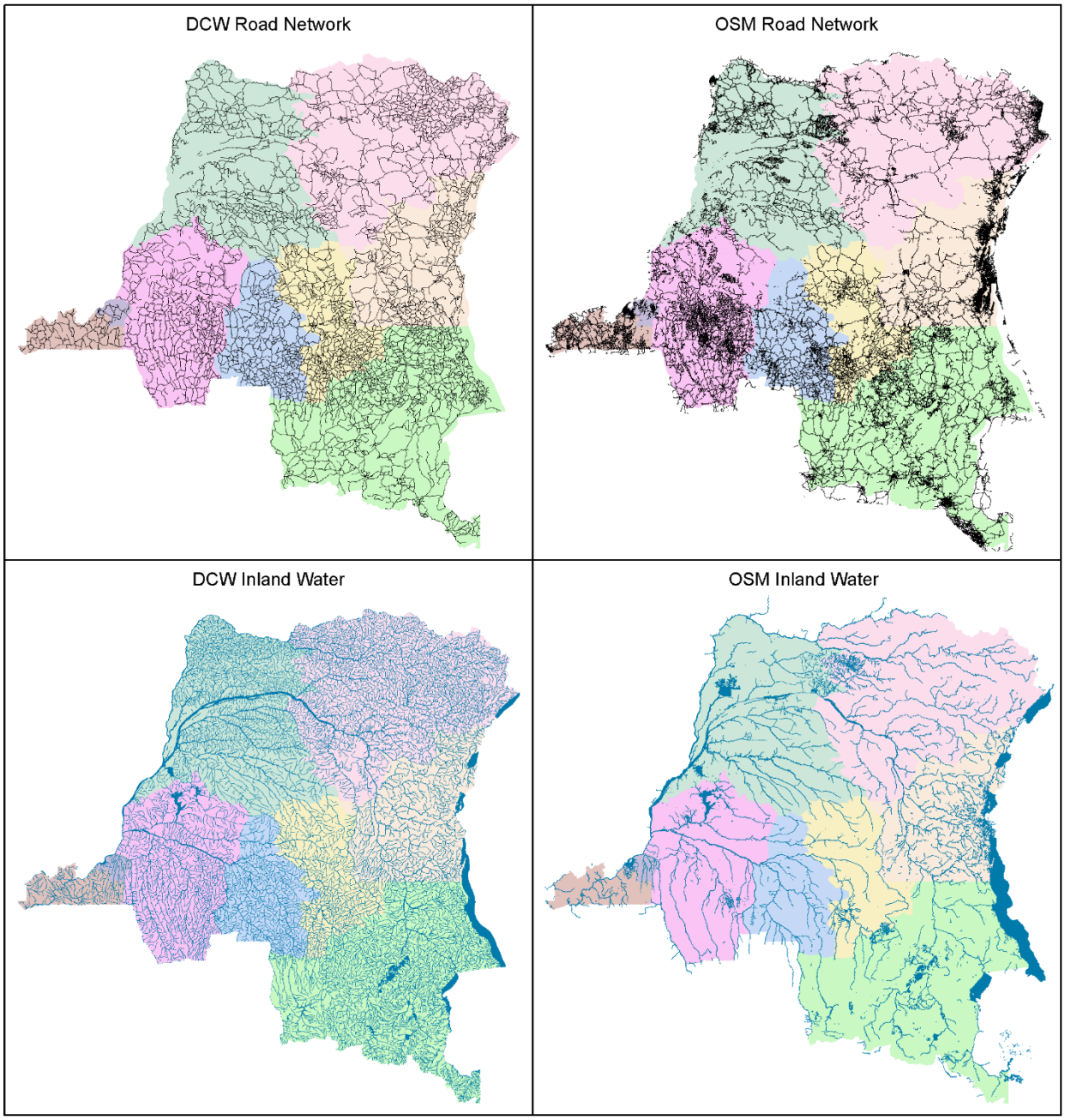


Figure 3.2: Most geographic representations of inland waterways represent streams as lines (blue) and lakes and rivers as polygons (yellow). This is evident when we look at the data corresponding to Lake Mai-Ndombe in western DRC. Combining these two formats poses a challenge to mobility modelers.

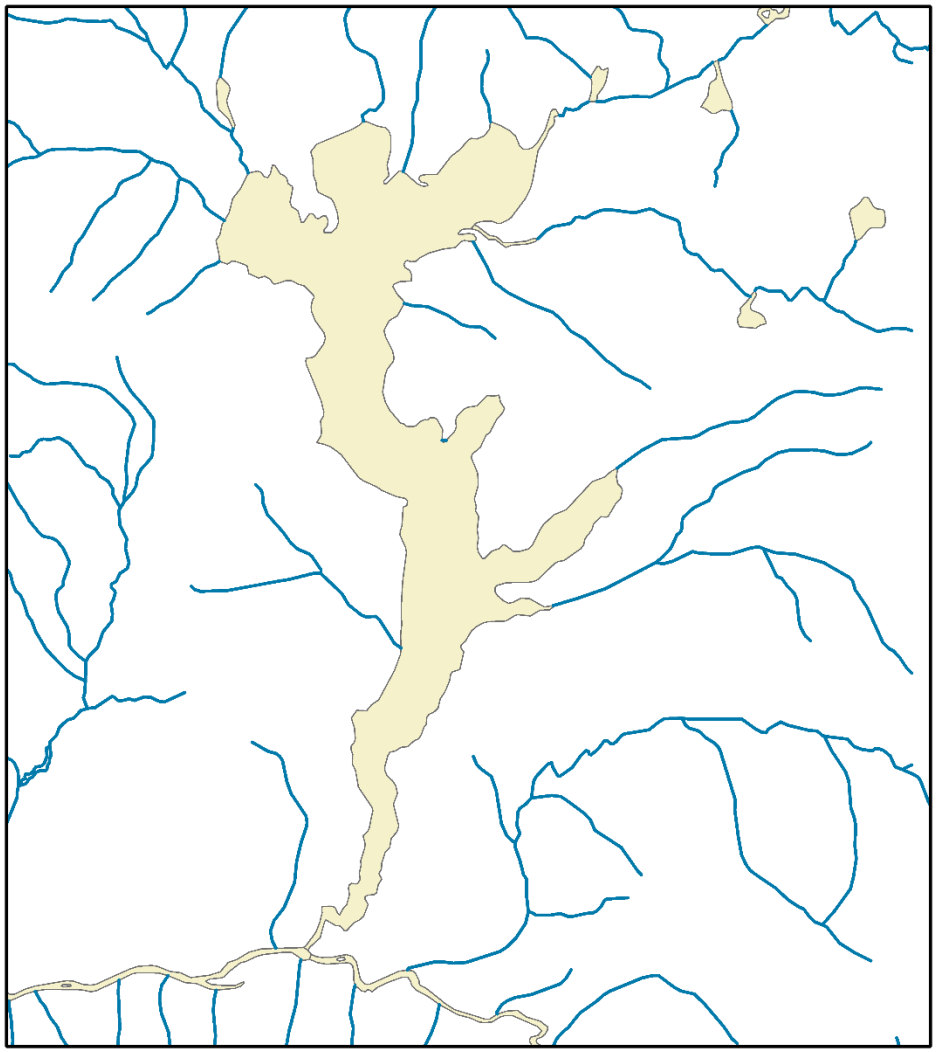


Figure 3.3: Reducing a river polygon to a network edge via a centerline operation creates artificial disconnects in the river network. Here we see the outline of Lake Mai-Ndombe in light purple, its centerline in dark purple, related streams in blue. Newly created disconnects are shown in red. If the researcher attempted to convert this into an edge-list network, none of these tributaries would be considered connected to the lake.

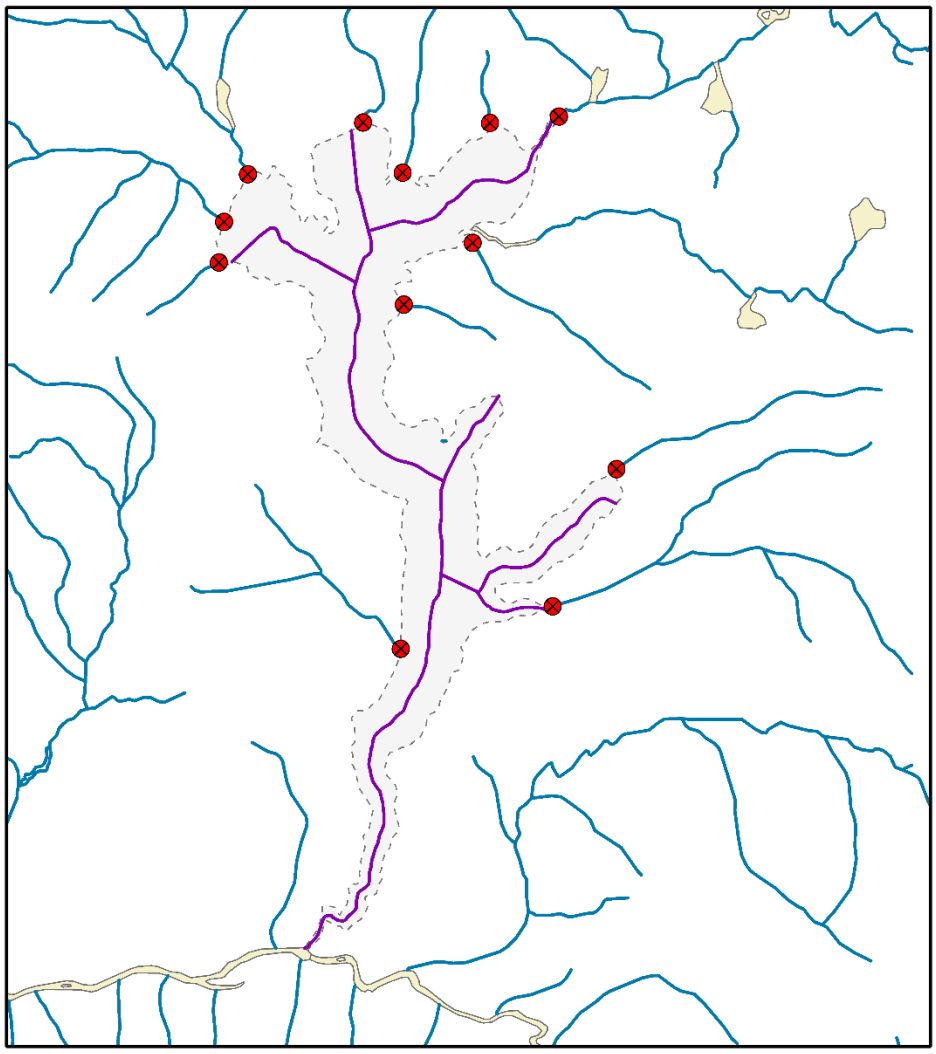


Figure 3.4: Here we see the outline of Lake Mai-Ndombe midway through our polygon to travel network procedure, having been converted to tessellated polygons. Note the numerous slivers at sharp bends which must be eliminated before proceeding.

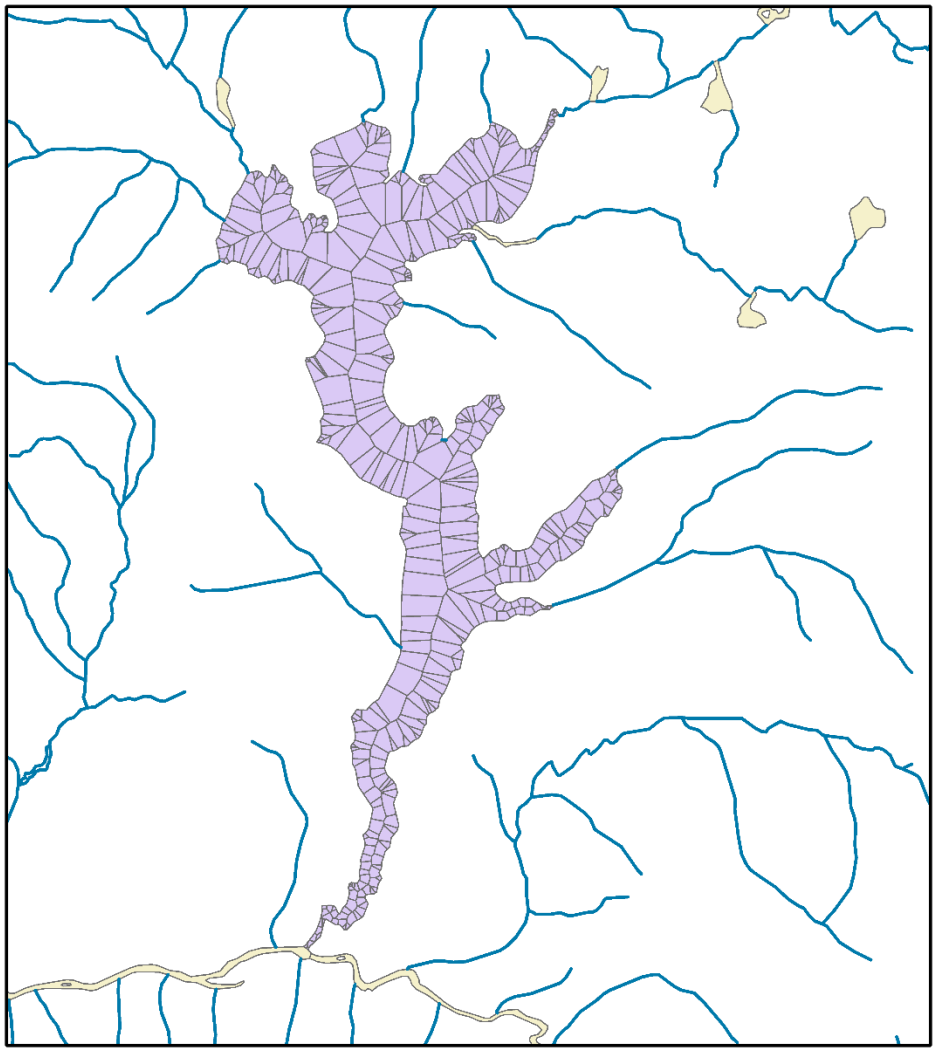


Figure 3.5: The same polygon from Figure 4.4, now having been fully converted to network edges along which an *in-silico* agent may travel.

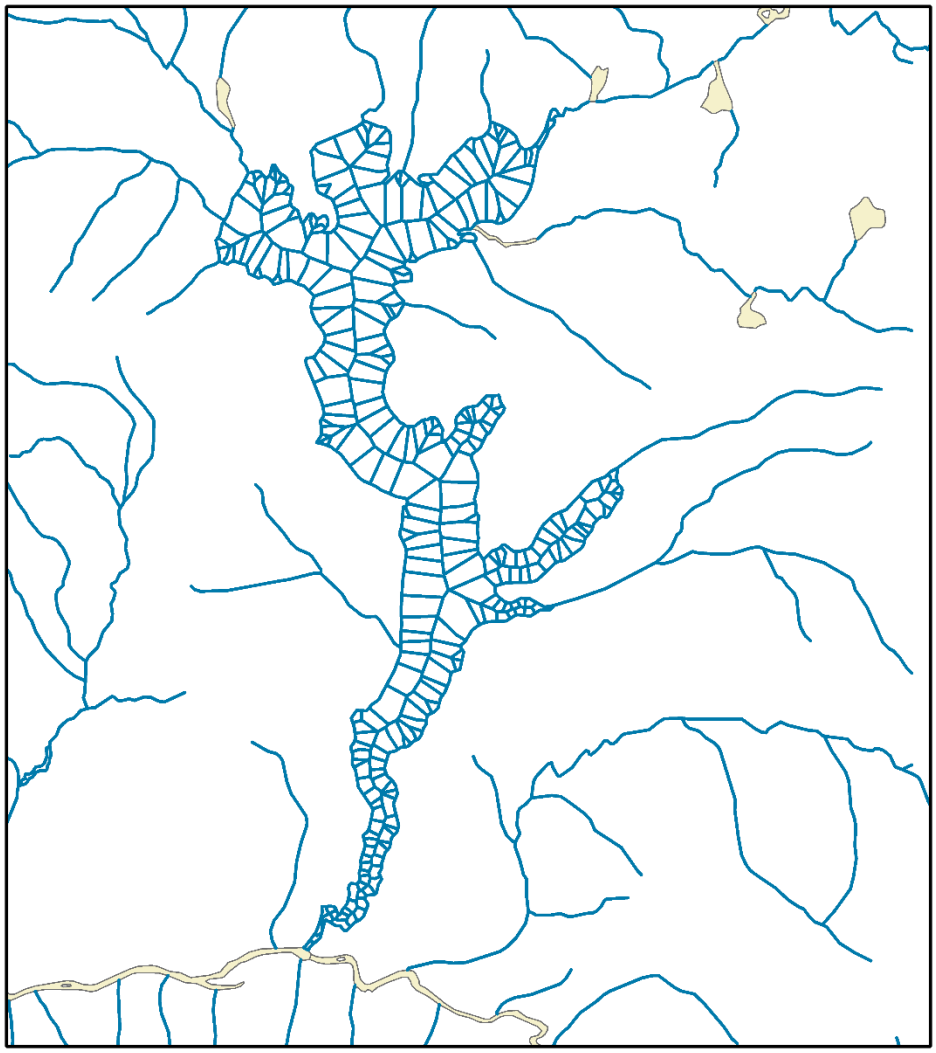


Figure 3.6: The influence that movement from Beni has on neighboring districts according to our completed mobility model. If Ebola does spread from Beni, we expect Katwa is the district most likely to suffer new cases.

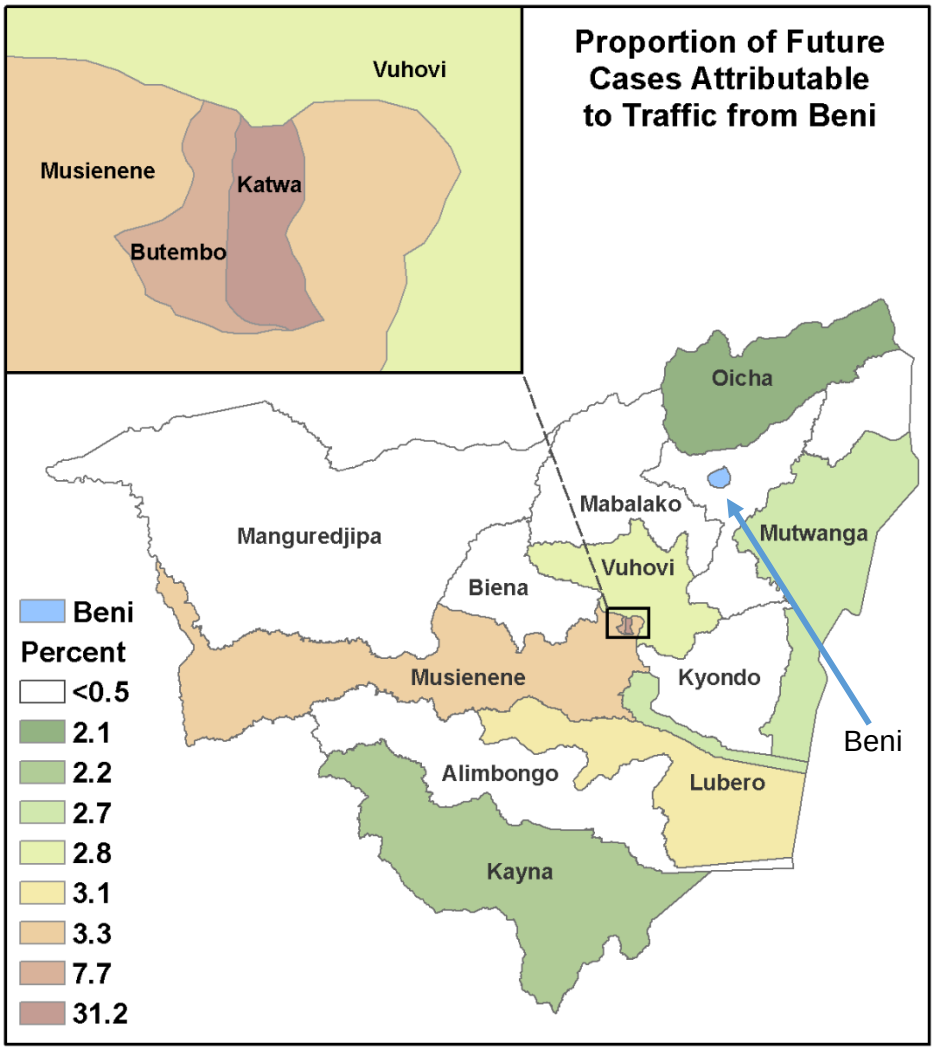


Figure 3.7: The distance from Beni as a function of polygonal contiguity (left), and Euclidean distance (right). Neither is an accurate representation of mobility when modeling the spread of disease or detecting outbreaks. Note that Katwa and Butembo are not considered at great risk in these models.

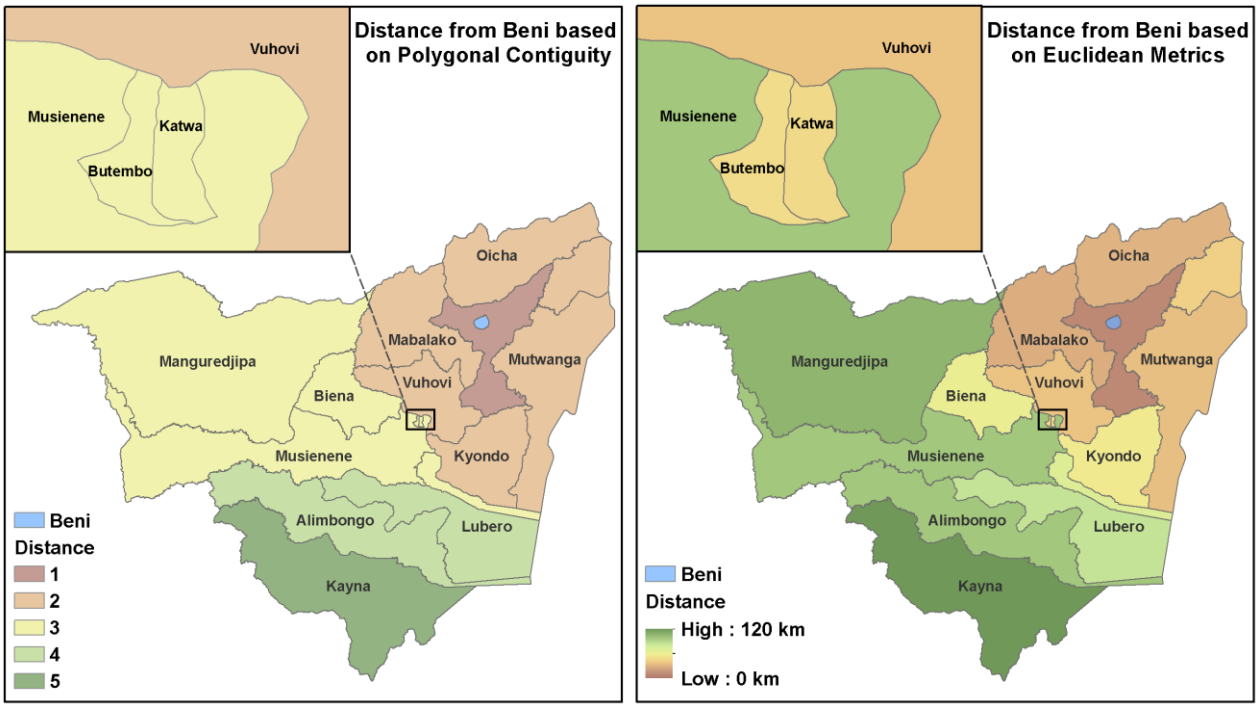


Figure 3.8: Cumulative case counts as reported through WHO situation reports are shown with a solid line, calibrated simulation curves shown with dashed lines approximate the course of the epidemic within each health district. Simulations are run 30 days past the last observation on Jan 24th (marked with grey vertical line at day 179) to produce 30-day risk estimates.

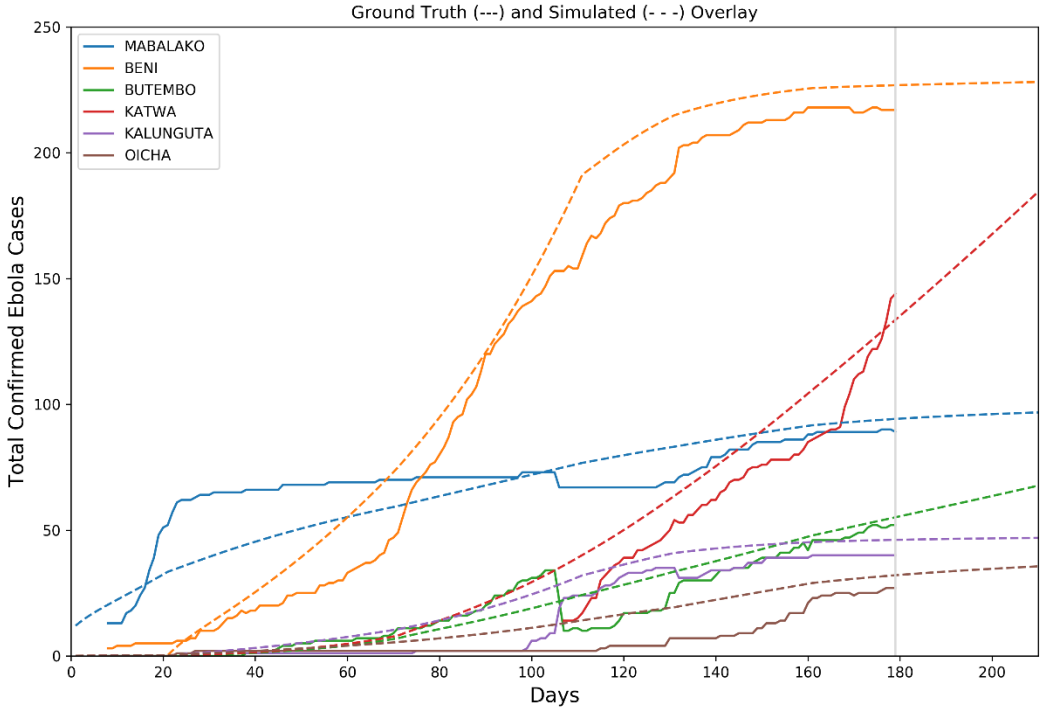


Table 3.1: The proportion of future cases for the most likely health districts, as well as polygonal and Euclidean distances from Beni. Neither distance metric is sufficient to realistically estimate the risk posed by the outbreak in Beni.

Health District	Future Case Proportion	Polygonal Distance	Euclidean Distance (km)
Katwa	31.20%	3	46.8
Butembo	7.70%	3	47.6
Musienene	3.30%	3	97.0
Kyondo	3.20%	2	54.9
Lubero	3.10%	4	84.4
Vuhovi	2.80%	2	36.4
Mutwanga	2.70%	2	34.9
Masereka	2.70%	3	68.6
Kayna	2.20%	2	120.1
Oicha	2.10%	2	29.6

Chapter

4. Augmenting Common Epidemiological Analyses with Network Derived Mobility Metrics

4.1 Abstract

Mobility metrics are frequently left out of epidemiological efforts or approximated using Euclidean distances. This is a reasonable simplification for the purpose of modeling some zoonotic or free-living diseases, but inappropriate for disease which is spread by human contact. In such cases, we need some understanding of how people move about the study area, for the disease often goes with them. In this work we demonstrate the creation of a mobility model and its applications in common epidemiological studies. Specifically, we will consider a hotspot analysis and geostatistical incidence model for the hepatitis C virus (HCV) in southwest Virginia.

We obtained case data from the Virginia Department of Health and geocoded each case to its corresponding census tract. Our mobility model was then created using a road network from NavTeq/Here. After placing population-weighted centroids in each tract, we created an origin-destination cost matrix, which, in conjunction with population data from LandScan, allowed us to fit a gravity model of mobility. We then verified the reasonableness of our model with commuter data from the US Census Bureau.

With these data in hand, we ran hotspot analyses using traditional spatial conceptualizations as well as our mobility model. We then created a geostatistical model, trained on data from 2012-2015, to estimate incidence in 2016. Included in this model were both spatial and temporal autocovariate terms, the former of which was generated by our mobility models, as well as numerous metrics for educational attainment, socioeconomic status, and financial burden.

We found that the addition of the mobility model greatly improved the detection rate for the hotspot analysis. We also found that while the temporal covariate was the most significant predictor of HCV incidence, the spatial autocovariate was both significant and improved the model's predictive power. Moreover, a global Moran's I test showed that the inclusion of the spatial autocovariate reduced autocorrelation in the model residuals.

We feel these results are quite plausible given the impact of person-to-person transmission on HCV rates. The inclusion of mobility data greatly improved the efficacy of both example analyses. Such efforts can improve health departments' ability to detect outbreaks as well as improve predictive power of modeling efforts.

4.2 Introduction

All too often, mobility is overlooked in epidemic response. Even in studies rife with spatiotemporal analyses, mobility is routinely seen as a secondary consideration, or one easily accounted for by simple approximations. Researchers focused on sylvatic zoonoses like Lyme disease (133), are far more concerned with the range and density of the vectors, reservoirs, the reproductive hosts, the dilution hosts, and the underlying abiotic factors affecting each. Researchers concerned with diseases caused by free-living pathogens like melioidosis, focus primarily on the underlying environment, variables like soil type, rainfall, elevation, temperature, and wind speed. And while there is significant spatiotemporal variation in each, and spatial autocorrelation as well, it is

generally assumed that a model built on high quality data will capture the majority of this. When spatial conceptualizations are needed, Euclidean distance metrics often suffice. After all, free-living bacteria are at the mercy of their surroundings, moving along with soil, wind, and water without any say in the matter. Vector-borne diseases are constrained by natural barriers, but otherwise move in any given direction so long as that movement benefits the vector or reservoir.

This cannot be said for diseases in which humans are part of the transmission chain, such as Zika virus or dengue virus, or diseases that are exclusively transmitted person-to-person, such as the hepatitis C virus (HCV). In these cases, human mobility is paramount. Natural barriers may be irrelevant, easily bypassed by a major highway, and natural distances have less consequence than travel times do. It is in these cases where traditional conceptualizations of spatial interdependence begin to fail.

Geospatial analysts make use of terms like Rook and Queen contiguity, terms borrowed from chess, to indicate the permissible directions of travel along a grid (134). This makes sense in a large city, but at a broader scale, Euclidean distances are meaningless, as human travel is bound to the transportation network of the area. An understanding of the network involved is critical, but even this is not sufficient on its own. In order to fully understand mobility in the area, we must also account for differing speeds along this network, unusual connectivity, and the attractive force of different population centers.

With this in mind, it becomes apparent that the use of polygonal contiguity and Euclidean distance metrics are insufficient analogs of spatial interactions, particularly when investigating disease outbreaks. The solution to this issue is straightforward. Rather than conceptualize spatial relationships with distance or contiguity, we do so with a mobility model. One could use these data to model the interactions between patches of a meta-population patch model, or to inform the interactions among synthetic populations in an agent-based model. But they could also augment a variety of common epidemic analyses, including outbreak detection and incidence modeling. For this paper, we would like to demonstrate the latter two, using a case study of hepatitis C in rural southwest Virginia.

4.2.1 Augmenting Outbreak Detection

A common fixture of the spatial epidemiologist's toolbox, hotspot analyses aid in the detection of outbreaks by identifying clustered high and low values. Though a variety of statistical measures exist, the most common is the Getis-Ord Gi* (12,40). Mathematically the operation is very similar to traditional measures of spatial autocorrelation such as Moran's I. Though, the latter is concerned with the global distribution of values – clustered, distributed, or random – while the hotspot analysis identifies areas where high or low values are too tightly clustered to be the result of chance alone. To do so, the analysis tool requires some understanding of the spatial relationships of these areas. Here we find the traditional spatial conceptualizations: polygonal contiguity by edge or overlap, inverse-distance and inverse-distance squared, network distance, and others. But none of these are appropriate for a pathogen like HCV, especially in a rural area like southwest Virginia.

In our case, it is not uncommon to find two rural towns separated by a mountain or forest. On a map of census tracts, these two towns may appear adjacent, separated by only a small distance,

and closely connected in terms of polygonal contiguity. But in truth, significant interactions between these two towns are unlikely due to the presence of barriers. On the other hand, a distant city, connected by a major highway, and with a substantial population, would appear quite distant to any traditional spatial conceptualization scheme, but we'd certainly expect that city to interact significantly with its surroundings as it may be the only major urban center in the vicinity. By using mobility to identify neighbors by degree of interaction, rather than by spatial relationships, we can more properly apply this common epidemiological tool to a pathogen like HCV.

4.2.2 Augmenting Geostatistics

While nearly all spatial data exhibits some form of spatial autocorrelation, a model that exhibits autocorrelated residuals is suspect (41). This implies that the model is either missing a critical explanatory variable, which itself is autocorrelated, or has failed to fully capture the spatial structure of underlying explanatory variables. The presence of autocorrelated residuals also violates the assumption of normally distributed independent errors, a key assumption to any regression analysis (41). A model with autocorrelated residuals is also more likely to generate false positive errors and mislead the researcher into rejecting the null hypothesis without good cause (135,136). There are two primary sources of residual autocorrelation: forcing by underlying variables and community-driven processes (137). In cases dominated by the former, it is occasionally possible to eliminate autocorrelation entirely by adding the appropriate variables (136). In cases where the latter is significant, we need to take measures to address it. A common tool to do so is the use of regression kriging, which fits a regression model then spatially kriges the residuals to eliminate autocorrelation (138,139). Another alternative is to include a spatial autocovariate term. Doing so should not only reduce residual autocorrelation but should improve predictive power as well (41).

Given the community impact on HCV, we feel the inclusion of an appropriate autocovariate is a more prudent approach. The autocovariate term for each area is a calculated as function of the values of its neighbors and the degree to which they interact. As with the hotspot analysis, the question of what constitutes a neighbor must be considered. Again, we feel that distance and polygonal contiguity are insufficient to model the interaction between these cells, and this is where a mobility model can augment this analysis.

4.2.3 HCV and the Social Transmission Niche

The case studies we present both focus on HCV , a pathogen that kills 20,000 per year, which is more than all of the other 60 nationally reportable infectious diseases combined (140). Estimates are that over 18 million infections occurred worldwide per year since 2006 (141). It is also one of the only reportable infectious diseases in the United States whose incidence rate continues to climb (140). As the name suggests, HCV attacks the liver, causing cirrhosis and occasionally hepatocellular carcinoma (142,143), for which it is the primary causative agent (144). HCV is also the primary driver of liver transplant surgeries in the world (145), and accounts for roughly four billion dollars in economic damages annually (146). Though modern treatment modalities have improved considerably with the introduction of next generation medicines like Ledipasvir / Sofosbuvir, the cost of such medications is exorbitant. HCV poses an even more significant threat

when the patient is coinfected with HIV. In such cases, the rapid onset of liver failure is predictably fatal (147).

Driven primarily by percutaneous events such as injection drug use (148–150) and tattooing (151), and potentially some sexual acts (150,152), HCV spreads exclusively person-to-person. There are no reservoirs of HCV outside of the human population, no vectors, and no amplifier hosts. The pathogen's survival is unaffected by climate, and movement is unrestricted by natural barriers. So, at first glance it may seem unusual for us to use a niche modeling tool like BRT to model HCV. Afterall, we certainly wouldn't expect to find a significant response to abiotic factors like mean summer temperature or rainfall. At first glance we are tempted to suggest that HCV's niche is bound only by the presence of humans.

But that is not the case. Decades ago, when the primary risk factors for acquiring HCV were receiving donated blood products or organ transplants (153), there were no obvious social determinants. The disease was randomly distributed among all demographics. But today the disease is primarily spread by injection drug use, amateur tattooing, or sexual contact involving blood to blood contact. It depends upon these risky behaviors just as a disease like Lyme or malaria depends on a compatible vector. And just like a vector, these risky behaviors are not randomly distributed, but clearly delineated in the environment. Moreover, these behaviors are as highly associated with the underlying environment as a vector would be. Rather than temperature or elevation, we are now concerned with social factors such as poverty, disenfranchisement, and crime; but these can be mapped as easily. As such, we posit that we can apply the same niche modeling techniques used to model vector-borne diseases to model HCV instead.

In infectious disease ecology, a zoonotic transmission niche is an area where zoonotic spillovers are plausible. This is defined by a confluence underlying abiotic and biotic factors that give rise to the interactions necessary for such spillovers to occur (32). Extending this framework, we propose that HCV has a specific "social transmission niche", where the social demographics give rise to the risk factors necessary for sustained HCV transmission. In this chapter, we intend to model it. As in **Chapter 2**, we will not be concerned with presence or absence alone, but rather incidence. But first we must consider the underlying factors that drive HCV and are associated with its risk factors at a population level.

Among these, gender ratio must be included due to inherent differences in HCV acquisition rates between the sexes. While women are 400% more likely than men to acquire HCV during sexual contact (92), men are 60% more likely to present with HCV, even if one accounts for differing rates of injection drug use (93). Furthermore, it is thought that 60% of American HCV cases are male (94). Age is also a consequential factor (154), with HCV being four times more prevalent in the over-40 demographic than among those under-30 (93). Additionally, HCV exhibits a peak in prevalence between 40 and 50 years of age (94). We must likewise consider educational attainment, which is itself associated with HCV, as well as most of the common risk factors. Those who fail to complete high school have nearly triple the risk of acquiring HCV as someone who did (95). Those who finish college experience about one quarter the risk of those with high school degrees, which is less than a tenth the risk of those who did not complete high school (95). This

relationship is plausible, as lower educational attainment is also strongly associated with injection drug use (96).

Similarly, we must consider metrics of unemployment and economic hardship, as HCV is strongly associated with low socioeconomic status (150). Unemployment is also correlated with violent crimes (97) incarceration, and drug-use (98) which are themselves strongly correlated with HCV acquisition. Finally, we will consider measures of childhood poverty, as it is strongly associated with adolescent drug use (99).

4.2.4 Objective and Significance

The objective of this study is to demonstrate the feasibility of generating mobility metrics and incorporating them into common epidemic analyses such as hotspot detection and geostatistical modeling. To do so greatly augments the efficacy of both. A mobility conscious hotspot analysis can detect outbreaks clustered not by space but by human connectivity. This allows the public health infrastructure to more readily detect epidemics and target interventions.. One could use such models to estimate incidence, identify areas of underreporting or gaps in surveillance, and identify problem areas caused by local forcing. The inclusion of mobility metrics in such models also improves their predictive power and reduces autocorrelation in the model residuals, and the accompanying risk of type I errors.

4.3 Methods and Materials

4.3.1 Study Area

The study area in question is the Virginia Department of Health (VDH) Southwest Region (Region III), which is seen in **Figure 4.1**. The region includes 29 counties and 11 independent cities spread across southwestern Virginia. Encompassing nine subordinate health districts, the region falls under the jurisdiction of a single regional epidemiologist, which limits the confounding that may arise from differing surveillance priorities of multiple epidemiologists. In personal communication the epidemiologist also noted that hepatitis C is a primary concern of hers, and accordingly we expect the surveillance in question is of the highest caliber.

The region includes the majority of Virginian counties considered part of Appalachia and suffers from some of the nation's highest HCV incidence rates. Mountainous and predominantly rural, the region is dominated by pastures and deciduous forests, as seen in **Figure 4.1**. It is home to roughly 1.35 million inhabitants, but the population is spread thinly. Only two major cities with urban populations exceeding 100,000 can be found in the region, Roanoke and Lynchburg. The more rural parts of the region are economically troubled. At the start of our study period, 2012, the region saw 10 of 29 counties listed as "at-risk" by the Appalachian Regional Commission Distressed Counties Program. As of today, four of those counties have become "distressed", while a sole county in the study area is listed as "competitive". The region is also in the throes of the opioid epidemic, which, along with economic distress, is closely associated with the hepatitis C epidemic (155). This is likely the result of the association between prescription opioids and drugs such as heroin, which sometimes take the place of said opioids when supplies are limited or

prescriptions withheld(156,157). Given the uniformity of surveillance in the area, and the impact that this pathogen is having on the region, we feel it is ideal for our analyses.

4.3.2 Acquisition of Case Data

Case data were sourced from the VDH and sanitized by removing all personally identifiable information under the oversight of the Institutional Review Boards of both the VDH and Virginia Tech. Specifically, we queried all health department investigations within the study area mentioning the word “Hepatitis” in any database field over the latest available five-year period (2012 to 2016). At the time of our data request, 2017 data had been compiled, but neither verified nor approved for release by the VDH. As shown in **Figure 4.2**, this resulted in 14,327 investigations for consideration in our spatial analyses. Of these we removed 3,044 for having non-viable addresses; often these were incomplete addresses, post office boxes, or points outside of our study area. We removed a further 1,662 records for listing addresses other than the home of the individual, such as a prison or diagnosing hospital. As our search was intentionally broad, we captured 1,533 records caused by the unrelated hepatitis A and hepatitis B viruses, or not properly classified. All of these were also eliminated from our dataset. Of the resulting 8,088 hepatitis C investigations, we found that 1,280 were ruled “not a case” during the investigation, and 34 were left undiagnosed. After removing these, we retained 6,774 confirmed or probable HCV cases within the study area and study timeframe, a capture rate of 47.3%.

Among these we identified 105 confirmed acute cases (1.5%), 5,983 confirmed chronic cases (88.3%), 7 probable acute cases (0.1%), and 679 probable chronic cases (10.3%). As per the CDC’s 2012 case-definition (158) used by the VDH during our study period, cases are confirmed by positive enzyme immunoassay for anti-HCV antibodies above a set signal to cut-off ratio, positive HCV recombinant immunoblot assays, or genotype testing which detects HCV RNA. Probable cases are those with positive enzyme immunoassays, but which have neither met the signal to cut-off ratio requirement, nor been verified by more specific testing. Acute cases are those in which the patient recalls a specific onset of symptoms and has either elevated alanine aminotransferase levels or presents with jaundice. All laboratory confirmed or probable cases that do not fit the explicit definition for acute cases, are considered chronic, though the actual date of exposure is nigh impossible to estimate.

For our purposes we will combine all four of these categories and use all 6,774 points in our study. Though we run the risk that some of the chronic cases may be many years old, the acute cases represent less than 2% of our total study population and are too sparse for spatial analysis.

4.3.3 Geocoding in a Rural Environment

Accurately converting rural addresses to geocoordinates is a common impediment in such studies. Geocoding services are commonly far less accurate in rural regions than they are in highly developed urban areas (159,160). The most accurate means of geocoding is to use a verified database such as the county’s E911 emergency services database, then manually verify each point using aerial orthoimagery (161). Such an approach is infeasible with nearly seven-thousand cases, and E911 databases sometimes differ by municipality. As an alternative, we turned to more commonly known commercial geocoders, making use of several, as no single geocoder we tested was able to accurately capture all points in question.

After formatting the address data into the common "*street number, street name, town name, state abbreviation, and zip code*" format, we used the Geocoder extension for Python (available online at: geocoder.readthedocs.io), to geocode case addresses using the following services: ArcGIS, Bing Maps, Google Maps, HERE/NavTeq, MapBox, MapQuest, OpenCage, and TomTom. These represent all the compatible geocoding services that cover the appropriate area, except for GeocodeFarm's service, which had too austere a rate limit for our purposes.

In addition to geocoordinates, each of these services returns confidence and accuracy values. The former represents the quality of match between the geocoder's database and the queried address. Any ambiguity reduces this score, which is typically in a range from 1-10. For example, the existence of both "Cherokee Drive" and "Cherokee Trail" in Blacksburg, Virginia would thoroughly confound a geocoder that encountered the address "`500 Cherokee, Blacksburg, VA 24060`". In that case, the geocoder would return a low confidence value. The same is true if the name is misspelled or the house number doesn't match the known range for a specific street. As the name suggests, the accuracy value represents how clearly the geocoder can pinpoint an address, such as to the "rooftop", "interpolated" or "approximate" location in Google Maps parlance. In cases where the geocoder cannot locate an address, it often returns the centroid of the corresponding zip code, which is useless for our purposes.

After geocoding all our points using all eight services, we discarded any return which did not have both a very high confidence score (9 or higher on a 10-point scale), and a street address level of accuracy. Though on average we lost approximately 10-15% of the results from each geocoding service, there was substantial overlap between them. Of the 6,774 cases, 6,732 (99.8%) had high accuracy and high confidence results from at least three different geocoding services. Of the remaining 42 cases, 28 had suitable results from two services, and an additional 10 from a single service. Though we would prefer redundancy, these addresses were included in our analyses. Four addresses could not be geocoded with sufficient confidence and accuracy by any of the eight services. We examined each of these manually but had no more success than the geocoding services. We suspect they may represent fake addresses, intentionally given to health department workers by embarrassed patients, or addresses that were so malformed or misspelled as to be nonsensical. Regardless, they were removed from the final dataset, leaving us a geocoded patient population of 6,770.

Our geocoded cases were then converted to points and assigned to a census tract by spatially joining with the US Census Bureau's TIGER/Line Shapefiles in Esri ArcMap 10.51. Since the geocoders all returned slightly different results, each of which were assigned to a census tract independently, the possibility existed of cases being assigned to multiple tracts. In such cases, we used a majority filter to assign the case to whichever tract was most favored by the different geocoding services. The data were then aggregated, giving us the total number of cases by census tract and year. Using population estimates from the US Census Bureau, we also calculated the incidence in cases per 100,000 person-years for each year.

4.3.4 Census Tract Centroids

Following the procedure set forth in **Chapter 3**, we created a gravity model to estimate the interaction between census tracts. Rather than use geometric centroids, we again found

population-weighted centroids using data from the LandScan gridded population dataset (109). As our study period covers a five-year period, we averaged the corresponding 2012 - 2016 LandScan grids, then converted the cells to points. Using a spatial join, we assigned each point to a census tract, then used a mean centers calculation to find the population-weighted centroids.

This methodology works quite well for larger areal units such as census tracts, but it may not be appropriate for smaller units. For example, in disease ecology, road-bounded polygons are commonly used as areal units, as the roads form natural barriers to wildlife (162). In such situations, LandScan cells, may be too large to be of use. An alternative may be to utilize land cover data. To demonstrate this, we imported land cover data from the Virginia Geographic Information Network (VGIN)at a 1x1 meter resolution and resampled it to 3x3 meters using a majority filter. We then discarded all cells that were not categorized as high density impervious and converted the remaining cells to points. As the VGIN data does not distinguish between roads and built structures, we imported a road network from the Here/NavTeq 2017 dataset and eliminated all points within 10 meters of a road polyline. The remaining points, which should represent structures alone, were then subject to a “mean centers” calculation.

4.3.5 Travel Network

Our travel network was built using the Here/NavTeq 2017 road dataset. Here/NavTeq lists speed limits for only a fraction of the roads included in the dataset, but all entries are given a speed category. By averaging the recorded speed limits for each category, we were able to fill in the remaining speed limits. Though not exact, these categories very closely approximated 65 to 25 mph in 10 mph increments, and we are confident that our estimates are reasonable. We then removed all roads that were listed as having no public access – in this case mostly found on government property such as the Radford Army Ammunition Plant – and projected the data into UTM Zone 17N. This allowed us to accurately calculate the length of each edge, and after dividing by speed, to calculate the time to traverse each edge in seconds. We then converted these data to an ArcGIS network dataset with travel time as the primary accumulator. Dropping a random point in the center of the network, we created a service area with unlimited distance bands and discarded all roads that were not included. Doing so allows us to find superfluous roads that are not properly connected to the main network and could confound our analysis. As it turns out the majority of these were extremely rural routes and one highway rest stop.

4.3.6 Gravity Model of Mobility

With the parsimonious road network in hand, we created an origin-destination (OD) cost matrix, calculating travel time between each census tract centroid and its peers. Pulling in mean population over the five year study period, we then fit the gravity model using the formula (122):

$$F_{ij} = G\frac{p_i^{\alpha} p_j^{\beta}}{d_{ij}^{\gamma}}$$

Where the interaction between points i and j (F_{ij}) is represented by a constant G times the population of point i (p_i) to an exponent (α) times the population of point j (p_j) to an exponent (β), all divided by the distance between points i and j (d_{ij}) to an exponent (γ). If this were a common physics problem, the G would be the gravitational constant of the universe, the alpha and beta

parameters set to one, and the gamma two. But for gravity models of mobility, these terms (seen in red), can be better fitted with calibration data.

To calibrate our gravity model, we made use of the US Census Bureau's American Community Survey (ACS) 2009-2013 5-year Commuting Flows, the latest available. These data show an estimate of the number of commuting workers between surveyed counties and serve as reasonable approximation of interaction. Though our data are at a census tract level, we expect that a gravity model trained on county-level data in our study area, will be more realistic and representative of real-world interactions than an uncalibrated model. In either case, the model should be general enough to function at the smaller census tracts.

As the ACS data covered the years 2009-2013, we calculated population for each county by averaging LandScan grids for the corresponding years. We then repeated our population-weighting procedure to find centroids for each county. And using the same road network as before, we then created an OD cost matrix representing travel time from each county to its peers.

Using MiniTab 18 we fitted a non-linear regression model relating distance and population to commuter flow, finding the attraction constant and exponents for our trained gravity model. While fitting the regression model we noticed that our response variable, the flow of commuters between counties, violated the assumption of normality. As such, it was necessary to transform it, which we did using the standard Box-Cox transform resulting in a $\lambda = -0.08$. This drastically improved the distribution, reducing the Anderson-Darling statistic from 106 to about 2.5, though even the transformed response did not fully conform to a normal distribution. We also included a natural log transformation that produced nearly identical results. As this was far easier to incorporate into our later models, we used the natural log transformation in favor of the Box-Cox transform for later work. Once the parameters were fitted, we applied the completed gravity model equation to our OD cost matrix and population data to estimate interaction between census tracts.

For the sake of comparison, we also created a traditional gravity-model using the standard "$\alpha = 1$, $\beta = 1$, $\gamma = 2$" exponent values, as well as a BRT model relating commuter flow to population and distance by ensembled decision trees. For our purposes the actual flow of commuters is immaterial. Both our hotspot analyses and geostatistical models make use of proportional flow rates. As such, to determine if the models produced realistic results when applied to census tracts, we calculated census to census tract flows using each model, then aggregated these up to the county level. We then we calculated the proportion of traffic that each county received from its peers and compared these values to the real-world flow proportions calculated from the ACS data. After comparing the three methods, we selected the most accurate to conceptualize spatial relationships for future analyses.

4.3.7 Mobility Enhancement of Hotspot Analysis

The first case study we present will be the augmentation of the common hotspot analysis using mobility. To do so we followed the typical procedure, beginning with a Global Moran's I test to ensure that our data are indeed autocorrelated. Following this, we ran an Incremental Spatial Autocorrelation, which repeats a local Moran's I test at pre-set distance intervals. This allows us to determine the optimal distance band for use in a traditional Getis-Ord Gi* analysis. We then ran four Getis-Ord Gi* analyses, with four distinct spatial relationships: inverse distance, inverse

distance squared, polygonal contiguity only, and one with our custom spatial weights matrix, based on the gravity model calculated earlier. This will allow us to detect hotspots covering multiple census tracts, which are not necessarily in close geographically proximity, but are likely to have high levels of interaction given connectivity.

4.3.8 Demographic Data for Estimating Incidence

The second case study we present will be the use of mobility to augment a predictive geostatistical model, in this case an effort to predict HCV incidence at the census tract level as a function of demographic characteristics. To do so we made use of Boosted Regression Trees (BRT), which is typically used in ecological niche modeling. Given the numerous methodologies to approximate incidence, such as geographically weighted regression (163), hierarchical Poisson modeling via Bayesian inference Using Gibbs Sampling (BUGS) (164) and integrated nested Laplace approximations (INLA) (165), our use of BRT may seem like an unusual choice. But the problem of estimating HCV incidence is nearly identical to that of estimating the density of a free-living pathogen like melioidosis as seen in **Chapter 2**. In both cases the species in question is inexorably tied to the underlying strata, whether physical or demographic, and the geospatial variations therein. Given that incidence is a continuous variable, we feel a Gaussian BRT model is appropriate.

Demographic variables were sourced from the US Census American Community Survey, at the census tract level and for each appropriate year in the study period, 2012 to 2016. Given the literature on HCV risk, we included the following variables: male to female ratio, median age, median family income, percent of families in poverty, percent of seniors in poverty, percent of children in poverty, and percent of populace on public support including supplemental nutrition. Also included were unemployment, part-time employment, and full-time employment rates. In our study, area unemployment and full employment exhibited a -0.5 correlation, suggesting we would lose significant information if we discarded either. Age-dependency ratios for children, seniors, and total population were also included, as well as percentage of each census tracts' residents in each ACS educational category, from less than ninth grade to graduate degree holder.

We must also account for the existing spatiotemporal distribution, as well as the fact that HCV is transmitted by person-to-person contact. The HCV rates of each census tract depend heavily on the density of cases in prior years, as well as the incidence in highly connected neighboring tracts. To account for these effects, we included spatial and temporal autocovariate terms.

4.3.9 Spatiotemporal Autologistic Modeling

In models which include autologistic terms, the value of the autocovariate is typically the average or maximum value of each spatial unit's neighbors (41). What constitutes a neighbor differs by methodology. In many instances, researchers make use of simple polygonal contiguity, while others weight the influence of each neighbor as a function of inverse distance or distance squared. In our methodology, we weight the influence of each neighbor as a function of gravity model predicted population flow from that neighbor. This takes the form of:

$$A_i = \sum_{j \in S} I_j \frac{F_{ji}}{\sum_{j \in S} F_{ji}}$$

Where A_i is the spatial autocovariate value for census tract i, I_j is the incidence of tract j, and F_{ji} is the flow of commuters from j to i, while $\sum_{j \in S} F_{ji}$ represents the total flow of commuters from all tracts into tract i. Once calculated for each tract, these values are incorporated into the incidence model, as an explanatory variable, specifically, the spatial autocovariate. Conversely, the temporal covariate term is simply the incidence on a one-year lag. One would expect significant consistency year to year, given that all new HCV cases are consequent to contact with existing cases.

4.3.10 Model for Estimating Incidence

With our measures of historical incidence, demographic data, and autocovariates, we fit a model relating the former to the latter two. Using the Dismo package for R (23), and following the procedure outlined in **Chapter 1**, we trained a ten-fold Gaussian BRT model on incidence from 2013 to 2015, then projected it onto 2016 to validate the results. Note that while most spatial models of disease burdens utilize a Poisson distribution, incidence is continuous rather than discrete, and we are required to assume a Gaussian distribution. As we are including a temporal autocovariate, we cannot include 2012 data in any form other than as the temporal autocovariate term for 2013.

4.3.11 Evaluating the Effect of Mobility

The autocorrelation of spatial data is expected, but the presence of significantly autocorrelated model residuals is a red flag. As our autologistic efforts were meant to combat this, a good means by which to evaluate the efficacy of these methods is to analyze the global autocorrelation of our residuals. To do so we projected our trained model onto the 2016 demographic data with, and without, the spatial autocovariate. We then compared the predictions of the two models to the actual incidence values using mean squared error and Pearson correlation coefficients. This will give us an indication of whether the autocovariate term improved the predictive power of the models. Finally, we ran a Global Moran's I test on the residuals of each model to determine if the autocovariate reduced the autocorrelation of the model residuals.

4.4 Results

4.4.1 Geocoding and Census Tracts

A summary of our geocoding results can be seen in **Table 4.1**. As stated earlier, the quality of a geocoded result depends on two factors: confidence and accuracy. The former refers to the likelihood that the service has matched the input address to the appropriate address within its database. The latter refers to the accuracy of the coordinates themselves. We defined high quality results as those with both high confidence and accuracy figures. By our definition, all but 42 of our case addresses received high quality results from at least three services. Of those, four were not properly resolved by any service. Most of our cases, 4,465, had high results from all eight geocoders. In our testing, Bing Maps was the most successful at resolving rural addresses, with a 98.4% success rate. This was followed by MapQuest, ArcGIS, and NavTeq/Here all in the mid-97% range. MapBox and OpenCage were the least effective geocoders in this task, finding about 84% of our addresses with both confidence and accuracy.

To reduce these coordinates to census tract case counts, we used a spatial join in Esri ArcMap 10.5.1. When the geocoders were in agreement, the points were generally spread-out over an area of about 200-meters in diameter, as seen in **Figure 4.3**. This seems substantial but is markedly smaller than a census tract. Much of the disagreement in tract assignment that we encountered was the result of wildly differing geocoordinates.

As there was no way to easily verify nearly seven thousand points, we then assigned the case to whichever tract held the majority of geocoder generated points for the corresponding address. The results of these efforts are also shown in **Table 4.1**. As expected, there was significant agreement between geocoding services for most cases, with all eight geocoders agreeing on the census tract for 3,917 of our addresses. Recall that only 4,465 addresses had high quality results from all eight geocoders, as such this represents a success rate of (87.7%). Twenty-four of our cases had no consensus, with multiple conflicting results. These were discarded from our dataset. A further nine were found by only one geocoder and were retained.

As for tract assignment consistency, we found a similar ranking to that of the performance of geocoders. ArcGIS, Bing Maps, and NavTeq/Here again led the pack in terms of frequency of participation in the majority solution, each exceeding 98%. MapBox and OpenCage were the least successful in this effort, at 84 and 79% respectively. Details of these rankings are also seen in **Table 4.1**.

Given the difficulties encountered, we suspect that no single geocoder could successfully resolve all the addresses in a rural study area like ours. Given the success of the top three geocoders however, we feel that an ensemble of the three would be adequate.

4.4.2 Census Tract Centroids

An example of the output of the weighted centroid methodology can be seen in **Figure 4.4**. When compared to street maps or population density maps, it is evident that the weighted method produces far more realistic results than geometric centroids. The latter is prone to putting points in the middle of wild terrain without a single road nearby, while the weighted points all fell within densely populated and well-connected areas.

The land cover weighting method also produced convincing results. **Figure 4.5** contrasts the results of the population density weighting and land cover weighting methodologies. We suspect that the population density methodology is more credible, as one cannot easily distinguish between businesses, apartments, and single-family homes from land cover data alone. Density is also not readily distinguished from land cover. Nevertheless, it is a substantial improvement over geometric centroids, and a valid alternative in areas where LandScan cells are too large to be used. Given that VGIN data is offered at a one-meter resolution, this method could conceivably be used to find the population centroid of something as small as a standard quarter-acre residential lot. In either case, the use of weighting is highly recommended when attempting to estimate flow between the centers of population bearing areal units like census tracts or counties.

4.4.3 Mobility Modeling

After manually inspecting our travel network we compared travel times from a single origin to 20 random destinations. The results, shown in **Table 4.2**, were promising, with the average disparity between Google Maps and our estimated travel time within about 10% of each other. The correlation coefficient between the two was 0.981. As such we are confident that the network model was representing travel times with sufficient accuracy for our purposes.

Our non-linear regression fitted the gravity model with the following parameters:

$$F_{ij} = 0.00361996 \frac{p_i^{0.537453} p_j^{0.942304}}{d_{ij}^{1.00921}}$$

Where F_{ij} represents the flow of commuters from i to j, p_i is the population of origin tract i, p_j is the population of the destination tract j, and d_{ij} represents the travel time between them in minutes. We also fitted a Boosted Regression Tree model, relating flow to population and travel time via decision tree ensemble.

To evaluate these models, we applied them at a census tract level, estimating flow between each, then aggregated these results back to the county level for which we have measured flow rates from the ACS. For our application, we are not so concerned about specific flow rates, but rather the proportion of travelers from each tract to each other. This is a measure of their interaction, and influence on each other. We compared the aggregated county proportions generated by our models to proportions calculated from the ACS using correlation and mean squared error (MSE) statistics. An example of this is seen in **Table 4.3**. Surprisingly, the calibrated model was the least accurate, with an MSE of 4.823 versus 1.698 for the BRT model, and 1.690 for the untrained gravity model. The calibrated model was also the least well correlated with real world proportions, with a correlation coefficient of 0.829, versus 0.936 for the BRT model, and 0.939 for the untrained gravity model. As the crux of this paper involves the use of a mobility model to augment traditional epidemiological analyses, rather than the mechanics of calibrating such a model, we did not continue with the trained model. As the unweighted gravity model produced the most realistic results, we used it in the following analyses.

4.4.4 Hotspot Analyses

The five-year incidence values for our study area were calculated from population data and mapped in **Figure 4.6**. Using a standard inverse-distance conceptualization we found that our five-year averaged incidence was spatially autocorrelated at a global level, with a Moran's I of 0.274 with a z-score of above 14.05. By inverse-distance squared method we found a Moran's I of 0.524 with a z-score of 11.79, and by polygonal contiguity we found a Moran's I of 0.382 with a z-score of 11.61. All accounts suggest that incidence is indeed autocorrelated, and that further examination is warranted. Following the standard procedure of incremental local autocorrelation to identify the ideal distance threshold, we found a peak z-score of 25.90 at 136 km. Using this figure in our inverse-distance and inverse-distance squared Getis-Ord Gi* tests we found only four census tracts meeting the definition for hotspots. The same four, specifically in Pulaski County and Roanoke City, appeared in both test results. The vast majority of the census tracts in the study area do not show significant autocorrelation.

A far more striking result was seen after incorporating a spatial weights matrix constructed with our gravity model of mobility. In this case, also seen in **Figure 4.6**, we find numerous hot and cold spots that went undetected by traditional methods. Without a realistic representation of mobility in the area and the interaction between census tracts resulting from this mobility, epidemiologists investigating the HCV epidemic may have missed these hotspots.

4.4.5 Incidence Modeling

Our BRT model for HCV incidence, trained on data from 2013 to 2015, achieved a training data correlation of 0.870, with a cross-validation test correlation of 0.634. It performed similarly when projected onto 2016 data, with a correlation of 0.698 between observed incidence and predicted incidence. Variable contributions are shown in **Table 4.4**. As expected, the temporal autocovariate was the most significant predictor, representing 37.1% of the total model. Percent of the population on public assistance was the second strongest contributor followed by the spatial autocovariate term in third place. These contributed 6.6% and 5.7% respectively. Variable response curves for the top six variables are shown in **Figure 4.7**. HCV incidence had a strong positive correlation with five of the six of the top contributors, excluding percent with associate degree which was generally protective. The response curve to the temporal autocovariate plateaued at about 300 / 100,000 PY, while the curve for the spatial autocovariate did so at about 200 / 100,000 PY. HCV incidence was strongly associated with percent on public assistance above 25%, and percent high school educated above 45%. If one were to combine the various factors into various bins, the temporal autocovariate would remain the most consequential variable at 37.1%, followed by measures of socioeconomic status and employment at 28.9%, educational metrics at 21.9%, and the spatial autocovariate terms at 7.3%. Age and sex ratio were nearly irrelevant to our models.

After removing the spatial autocovariate from the model, we see a minor reduction in predictive power. The correlation between predictions for 2016 and actual incidence values drops to 0.666. The total mean-squared error also increases by 11.29%. Though it is less significant than the temporal autocovariate, the spatial autocovariate is certainly relevant to these modeling efforts.

4.4.6 Effects of Autocovariate

The addition of our spatial autocovariate term did indeed reduce the autocorrelation of the HCV model residuals when projected onto the year 2016, as seen in **Figure 4.8**. By the traditional inverse-distance squared method, the residuals of the HCV model with mobility included showed a Moran's I of 0.169 and a Z-score of 4.10, versus a Moran's I of 0.246 and a Z-score of 5.91 for the mobility agnostic model. When one includes our spatial weights matrix developed to conceptualize the spatial relationships between census tracts, we see even smaller figures. The model which included mobility saw a Moran's I of 0.068, with a Z-score of 4.778, while the model without mobility recorded a Moran's I of 0.106 and a Z-score of 7.293. Though the model residuals were still autocorrelated to some degree, even after including the spatial autologistic term, the effect was significantly diminished and the residuals far closer to randomly distributed than before.

4.5 Discussion

4.5.1 Geocoding in Rural Environments

As expected, geocoding in extremely rural regions such as the western part of our study area, posed a significant challenge. But we feel that our efforts were successful, and deployable far more rapidly than more traditional E911-orthoimagery methodologies. Nevertheless, significant sources of uncertainty remain. When given a street address, most geocoding services will place a point near the eponymous street. This is evident in **Figure 4.2**, where our three most successful geocoders assumed the home was effectively on the shoulder of the road. In truth, many rural homes and most farmhouses are quite distant from the nearest passing street. In our case, we do not feel this is an issue, since census tract borders typically follow roads or natural barriers, also see in **Figure 4.2**. It is unlikely that they separate very many homes from their mailboxes. Though the mailboxes of some homes are on the opposite side of the street, only a very small fraction of streets act as census tract borders. Nevertheless, when doing more detailed modeling with discrete points rather than census tracts, this is an issue which remains unresolved and deserves attention.

Another concern involved our use of a majority filter to assign points to census tracts. We assumed that agreement among several disparate geocoding services was a strong indication of accuracy. In our work, it appears that ArcGIS, Bing, Here/NavTeq, and MapQuest dominate the rest of the pack when it comes to rural geocoding, generating the most high-quality results and agreeing on census tract assignment most of the time. Perhaps they are all collectively wrong. We cannot eliminate the possibility these services created their address database from the same faulty source, ensuring we get similar results from each. Perhaps one of the geocoders we characterized as less reliable was in fact, the most accurate. Manual verification, as described in Mazumdar (161), may be the only way to elucidate this.

4.5.2 Centroid Finding Methodologies

Though population weighting has served us well, we have merely scratched the surface of land cover weighting. At present, our greatest limitation is the inability to distinguish between commercial land, single-family homes, and high-density apartment blocks. This could be solved, and the method thoroughly augmented, by the inclusion of residential zoning data. Unfortunately, no single repository of such data exists, and collecting it for each municipality in a study area the size of ours would be tedious. In the absence of such data, one could instead distinguish between single-family homes and apartments by the size of the footprint. The footprint of a common apartment in our study area often exceeds 600 square meters, and larger blocks approach 1,000 m^2. By comparison, most homes are substantially smaller, averaging 250 m^2 less. Yet another method of distinguishing may be the use of road network density, and network centrality, both of which are strongly associated with population density. Low density residential areas typically are on the periphery of the road network, while high density areas are well connected and central. In fact research has shown that density is strongly correlated with straightness, betweenness, and closeness centrality measures (166). Network data may provide a reasonable approximation for population density when neither land cover nor population density is available.

4.5.3 Thoughts on Mobility Modeling

Though it was not the crux of our paper, the results of our trained gravity model were less than optimal. Upon further inspection, a peculiarity emerged. The untrained model had significantly higher mean-squared error when compared to the actual commuter flows, and a lower Pearson correlation coefficient as well. This is to be expected, given the fact that the non-linear regression algorithm is seeking to minimize this error. However, the untrained model was superior at estimating proportionality of inbound traffic by origin. When each of the 40 counties and cities were considered individually, the trained gravity model was more accurate in only four cases, including the largest three counties, where it had a substantial advantage. We suspect that a least-squares estimator is not ideal for this application, as it may overfit to the larger values associated with high population counties, at the expense of the far more numerous smaller counties.

On the other hand, our endeavor to fit these data with BRT was surprisingly successful. It was nearly as efficient as the gravity model. Nevertheless, we are concerned about clamping: projecting the model onto test data with values well outside of those found in the training data. Decision tree models are not as generalizable as traditional regression methods, and do not easily support changes of scale. Given that most census tracts are smaller in population and far closer together than counties are, we felt it was inappropriate to use the BRT model of mobility, despite the seemingly impressive results.

Though we were aiming for the most parsimonious mobility model, it is entirely possible that significant improvements could be achieved by incorporating other terms. Population is not the only attractive force pulling outlying individuals to the city center. The economic status of the area, as well as the availability of jobs must be substantial modifiers of commuting flow. We would expect a distressed origin to send significant workers to a more economically prosperous destination, but flow in the opposite direction is likely limited. The same can be said for economically motivated travelers looking to buy goods or enjoy recreational activities. It is unlikely that such travelers are traveling to distressed areas. We could potentially account for this by incorporating economic metrics into our mobility model, factoring in the disparities in income and employment metrics. Future efforts should also include population density, as compact areas with substantial economic activity are strong attractors. Finally, we can incorporate some metrics of transportation capabilities and preferences when estimating mobility. Surely the proportion of families with access to an automobile will affect the degree of interaction between rural tracts.

4.5.4 Limitations

The most striking limitation is of course, the inability to distinguish between chronic and acute cases. In personal correspondence with the regional epidemiologist, we were told that the health department suspects that most of the new cases identified by their personnel were recently acquired. Often cases are detected during the initial period of malaise, or by routine screenings. In these cases, it is entirely plausible that the case was acquired in the year of diagnosis and near the address given. But there is no way to verify this. Given the propensity of HCV to lay dormant for decades after the initial sickness, it is entirely probable that many of the cases in our data were acquired years earlier and in a different part of the state, if not the country. There remains no

definitive method to resolve this issue. However, the error this introduces should be randomly distributed in both space and time. Surely, we do not expect the ratio of relevant contemporary cases to deleterious historical cases to change from year to year. Given that cases meeting the acute definition are such a small minority of the case dataset, we are left with no choice but to make use of all the cases detected.

Another source of confounding we must consider is the discrepancy between site of infection and home address. Considering that injection drug use is often the source of infection, we can fairly assume that many cases are indeed domestically or peridomestically acquired, but certainly not all of them. Many must be infected at their place of employment, at medical facilities, at the homes of friends or suppliers, perhaps even in public spaces. This should not affect our analyses as we are not attempting to determine the site of infection, but rather where the HCV positive individuals themselves live. Still the possibility remains that our first-order mobility model is not properly capturing the impact of centralized infection points. Consider the possibility that an area known for injection drug-use exists in a small town, and attracts individuals living on either side of the town. The interaction of these individuals at this central point is not properly captured by a model of flow between their two home census tracts. In this case, a second-order mobility model would be far more useful and should be considered for future analyses.

Another issue to consider regarding mobility metrics is that the movement of commuting workers may not accurately capture the movement of HCV through the environment. As we saw in our incidence model, employment is negatively correlated with HCV, as is educational attainment. It is possible that the people most likely to acquire HCV have drastically differing mobility profiles than those commuting for work or education.

We must also consider the impact of surveillance disparities. Though the entire region is administered by a single regional epidemiologist, who specifically prioritizes HCV surveillance, the priorities of the district epidemiologists and their staffs must have some impact. Furthermore, the entire surveillance system depends on timely reporting from physicians and labs. Though it is a reportable disease, it is not absurd to imagine differing rates of testing by physicians in dissimilar areas. In fact, health department reporting may be partially self-fulfilling, as local physicians who are alerted to the local impact of the epidemic may increase their testing efforts and detect even more cases. Conversely, in areas with few detected cases, these physicians may consider testing superfluous, and their partialities may be confirmed by reports of low incidence rates.

Finally, when modeling the incidence of HCV, we are limited by the absence of a substantial predictive variable: a history of incarceration. The act of incarceration generally concentrates those already at risk alongside those with the disease and forces them into situations that are ideal for HCV transmission, whether by violence, tattooing, or sexual contact. In our study area, 1,406 (9.8%) of the detected cases reported their addresses as prison or jail facilities, which suggests that many more acquired the disease while incarcerated and brought it home with them. Unfortunately, a measure of formerly incarcerated individuals per census tract is not easily found. Data on current inmates are available, but these individuals will not affect their home tracts for months or years. This also confounds our mobility analyses. If a sizable portion of new cases are acquired in prison, then dispersed into the surrounding areas on release, this cannot be captured

by our census tract mobility metrics. In theory a history of incarceration is itself associated with many demographic variables, and the HCV incidence model should capture some of this. But the only way to properly account for the dissemination of HCV from point-sources like prison is with an agent-based model.

4.5.5 The Impact of Mobility

The inclusion of the spatial weights matrix substantially improved the detection rate of our hotspot analysis. We feel this is a plausible and expected result given that HCV spreads person-to-person and that commuter flow between tracts cannot be inferred from distance alone. As such, any algorithm designed to detect clustered high values cannot make do without some measure of this interaction. Without an understanding of mobility, an algorithm like Getis-Ord Gi* may find erroneous clusters, coincidentally neighboring tracts with high values but which do not actually interact. It is also liable to miss legitimate clusters, such as distant tracts found along the same major highway and experiencing significant interaction. Consequently, we feel this method significantly augments the traditional hotspot-based outbreak detection methods.

The impact of the temporal autocovariate on the HCV incidence model was expected. For a disease that depends exclusively on contact with existing cases, the case density from prior years will surely be the strongest contributor to future cases, regardless of the environment. That said, our spatial autocovariate was also significant. The third largest contributor to the model, the spatial autocovariate improved model fitness and reduced the spatial autocorrelation of the model's residuals. Given the importance of avoiding residual autocorrelation, we consider this a success and recommend it for similar analyses. Indeed, we suspect that the spatial autocovariate is even more significant in the modeling of more virulent and rapidly spreading pathogens such as Influenza or Ebola.

We should also note that during the course of our work, we encountered a prime example of how one could use these models to identify abnormal local forcing. The census tract with the highest overall incidence rates and one of the largest model residuals was in Pulaski County. The model continuously underpredicted incidence for this tract. The VDH then confirmed that an HCV-positive individual, known for giving tattoos at parties with homemade equipment, was responsible for nearly two dozen cases in the tract during our study period. Though we cannot claim credit for this individual's detection, it confirms our hypothesis that large model residuals may be the result of unusual local occurrences and likely merit further investigation. In this case, the residual was not the result of a flawed model, but a local black swan event that the model could not have predicted from demographic data alone.

4.5.6 Uncertainty and Sensitivity Analysis

There is substantial uncertainty in all health reporting data, and little way to account for the variation in underreporting rates through space and time. We do however have a means by which to account for the uncertainty in the regressors of both models, as the US Census Bureau routinely publishes confidence intervals for their data. We could incorporate uncertainty of commuter flows while fitting our gravity model and include the uncertainty for each demographic variable in our incidence models as well. Our response curves and variable contribution figures allow us a good understanding of the sensitivity that each model has to fluctuations in the regressors. As such we

can estimate the effects that uncertainty may have and generate confidence intervals for both our commuter flows and our incidence estimates. We could also generate maps showing low and high estimates of predicted incidence, as well as examine the effect that commuter flow uncertainty has on our hotspot analyses.

4.5.7 Conclusion

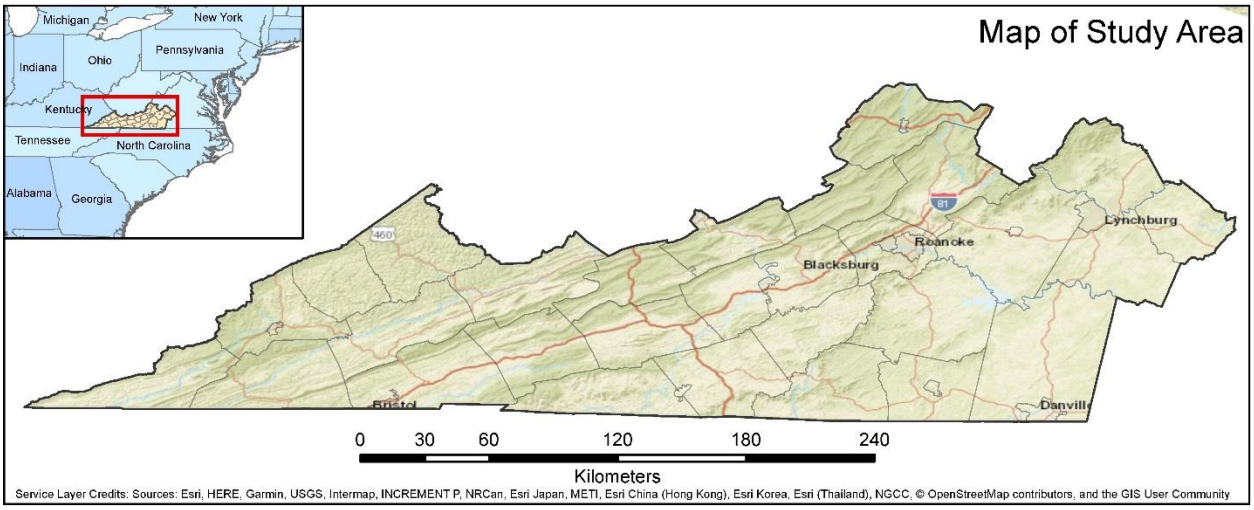
In conclusion, we have shown that a mobility derived spatial autocovariate term can improve the predictive power of geostatistical models of incidence and reduce the spatial autocorrelation of the model residuals. Such autocorrelation violates the assumption of normally distributed errors critical to regression analyses and increases the likelihood of type I error. After taking mobility into account, these models could be used to find areas with poor surveillance or unusual local forcing, prompting health department investigations or enhancing intervention targeting. We have also shown that incorporating mobility can augment hotspot analyses by in integrating realistic connectivity and commuter flows, rather than assuming connectivity as a function of geographic proximity. This could allow health departments to identify unusually shaped hotspots which may be geographically disconnected, such as those which follow a major highway.

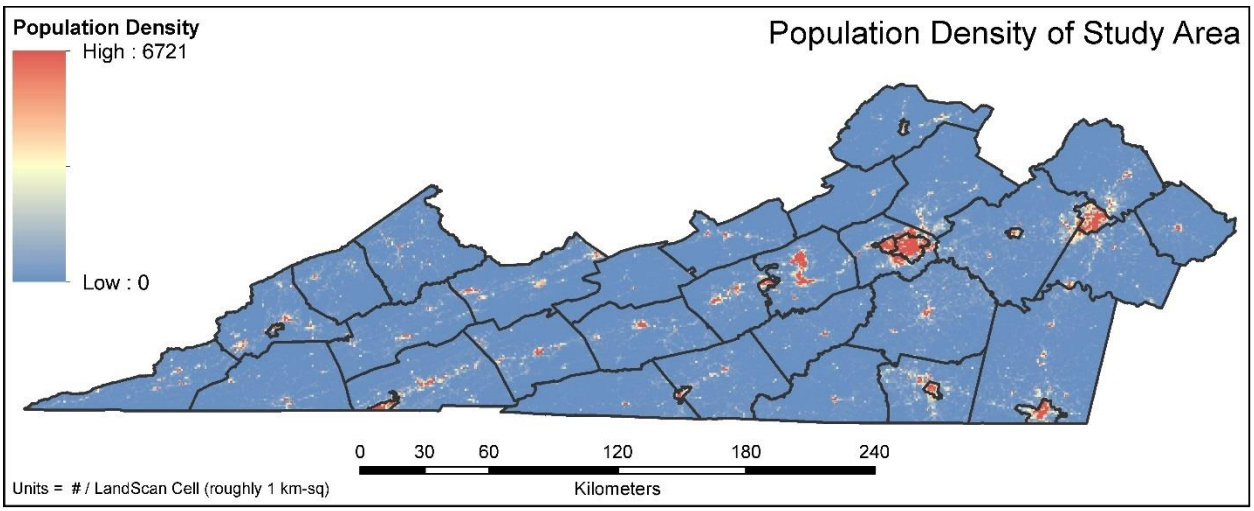
4.6 Funding and Acknowledgements

This study was supported by the Defense Threat Reduction Agency (DTRA) Comprehensive National Incident Management System (CNIMS) Contract HDTRA1-17-0118; and the National Institutes of Health (NIH) and National Institute of General Medical Sciences (NIGMS) Models of Infectious Disease Agent Study (MIDAS) Cooperative Agreement U01GM070694.

We thank Regional Epidemiologist Paige Bordwine, of the Virginia Department of Health, for her assistance with everything hepatitis C related and her extensive experience in the study area.

Figure 4.1: Shows our study area: VDH region III (top). The area is sparsely populated, with only a few major cities (middle). Mostly deciduous forest and farmland (bottom), and with few major highways, the use of Euclidean distances to model interaction between areas is not ideal.





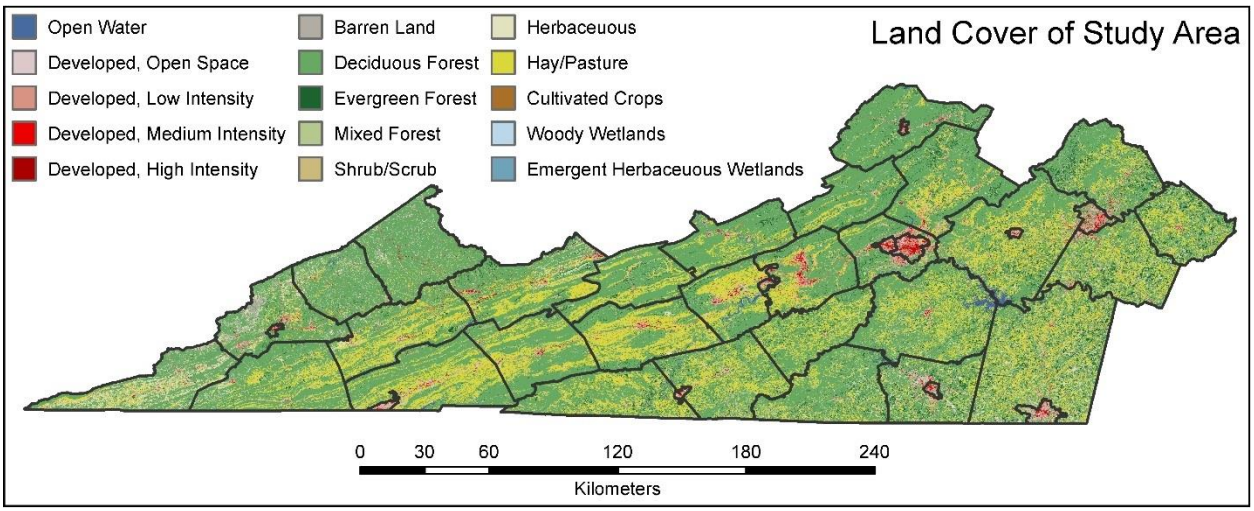


Figure 4.2: The study flow diagram for case data acquisition. After removing unsuitable data, we were left with 6774 geocoded hepatitis C cases in our study area. Note this includes 105 confirmed acute cases, 5983 confirmed chronic cases, 7 probable acute cases, and 679 probable chronic cases.

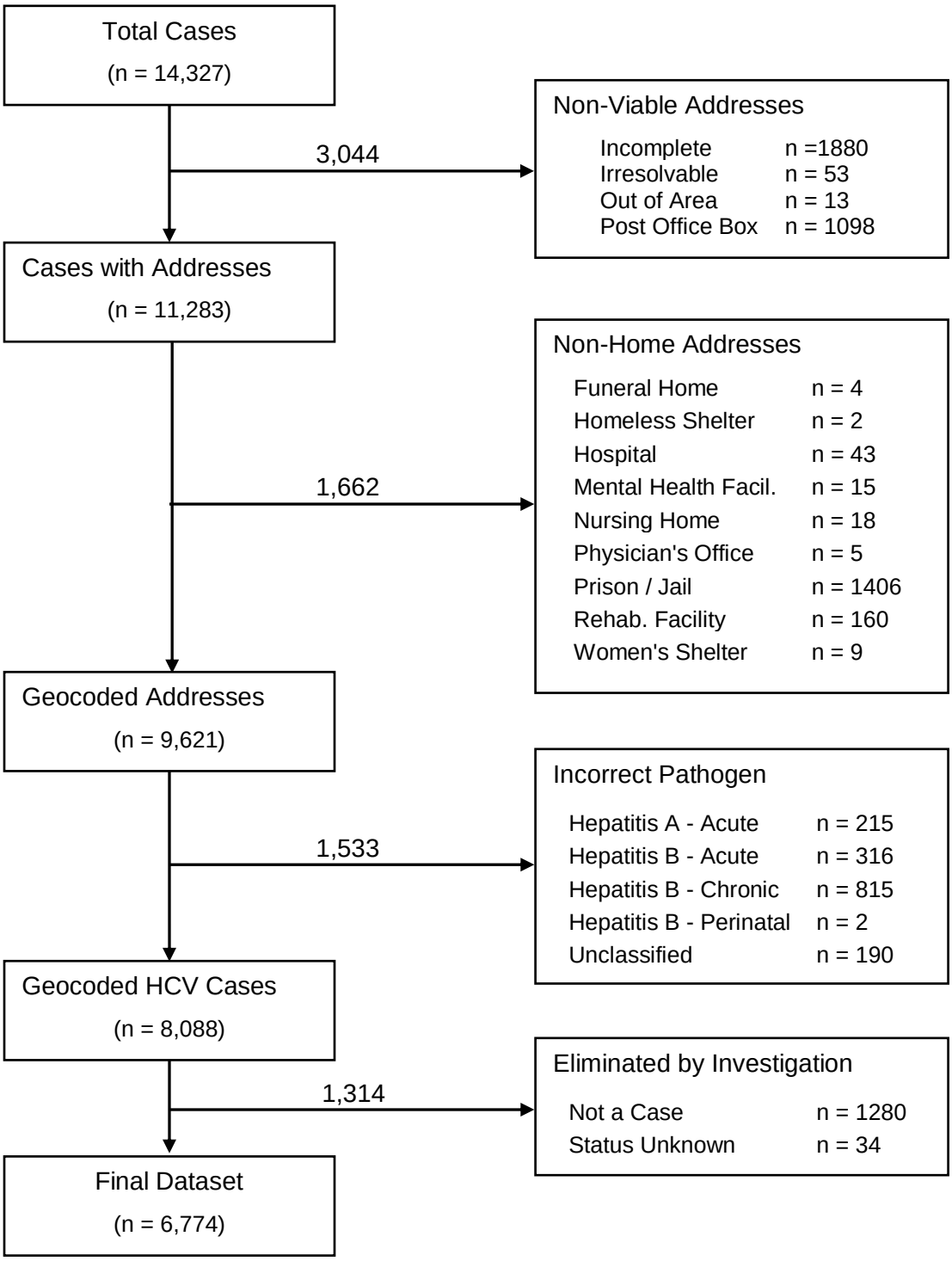


Figure 4.3: Though each geocoder produced a slightly different result, the disparity between these results is generally inconsequential when compared to the size of a rural census tract. In this case all eight geocoders placed the test address well within the same census tract (above), despite the points being spread out across a 200-meter section of road (below).

Note that this address was selected at random from Google Maps. It is representative of the rural addresses in our study, but it is not part of our HCV case data which are HIPPA protected.

Service Layer Credits: Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Figure 4.4: Illustrates the difference between the geometric centroids (blue points) and population weighted centroids (green triangles) of the census tracts around Blacksburg, VA. Downtown areas such as tract 1 are reasonably well represented by geometric centroids. But population weighting is far more realistic in rural areas such as tract 2. This becomes apparent when considering the road network (upper) and population density (lower). The geometric centroid in tract 2 lies in a sparsely populated area and would be snapped to a poorly connected rural route, confounding mobility analyses.

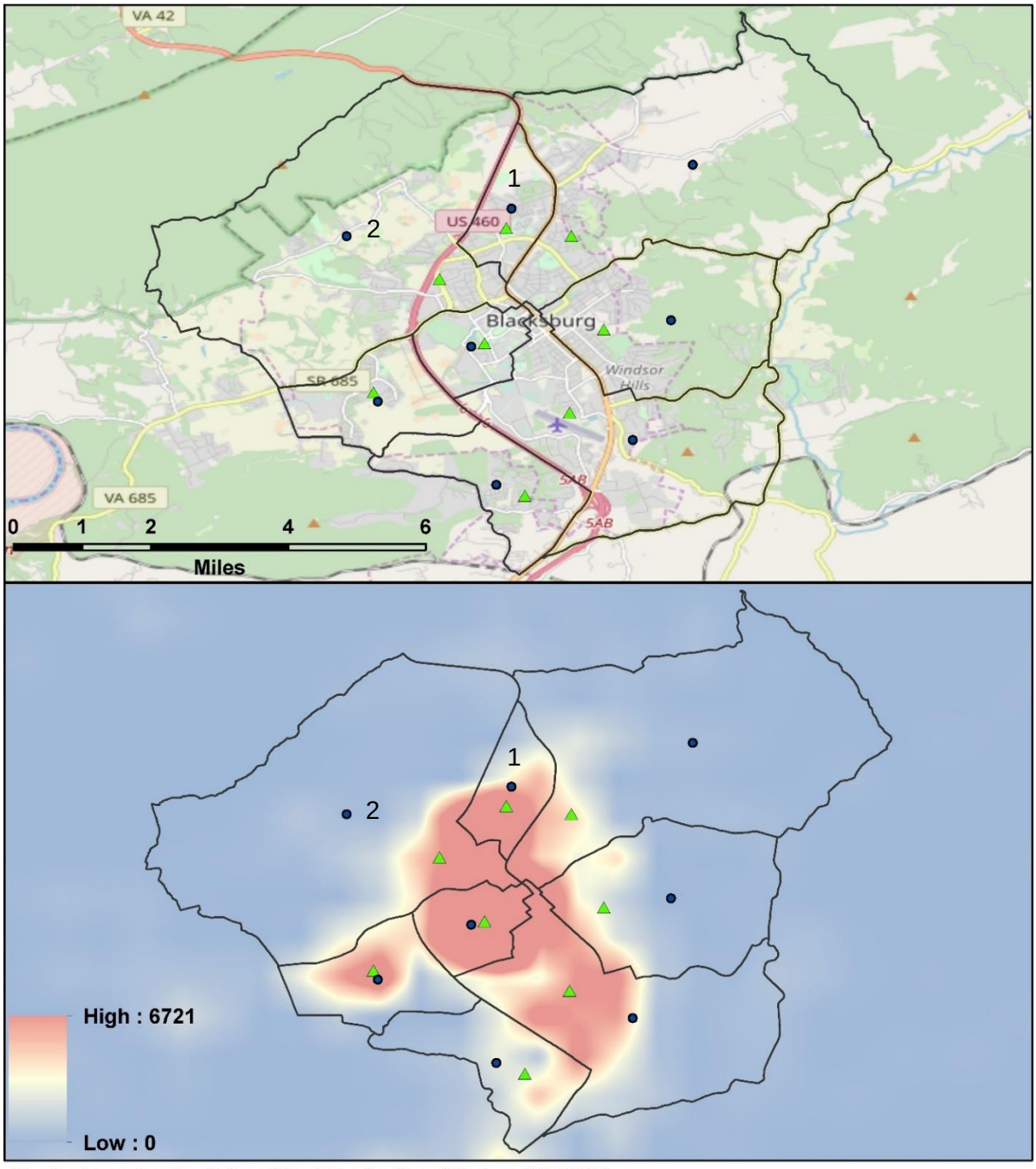


Figure 4.5: Compares the geometric centroids (blue circles) and population-weighted centroids (green triangles) from Figure 4.4, with land cover-weighted centroids (orange crosshatch). When considering both population density (above) and land cover (below). Though not as accurate as population weighting, the land cover method still greatly improves upon geometric centroids.

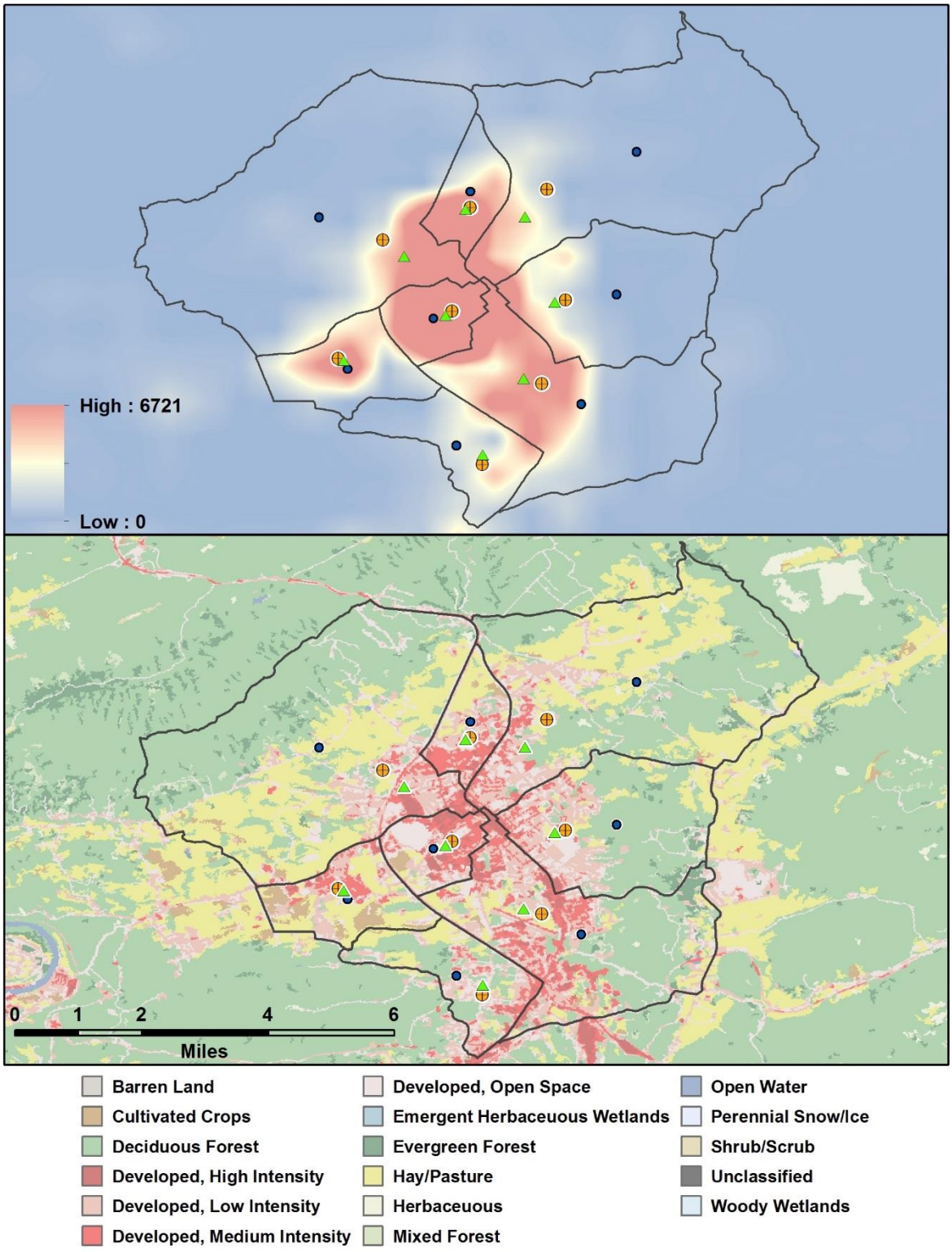


Figure 4.6: Shows the incidence of HCV in the study area averaged over the study period (top). We also see the results of our hotspot analyses, that used inverse-distance squared (center) and gravity model of mobility (bottom) to conceptualize spatial relationships between tracts.

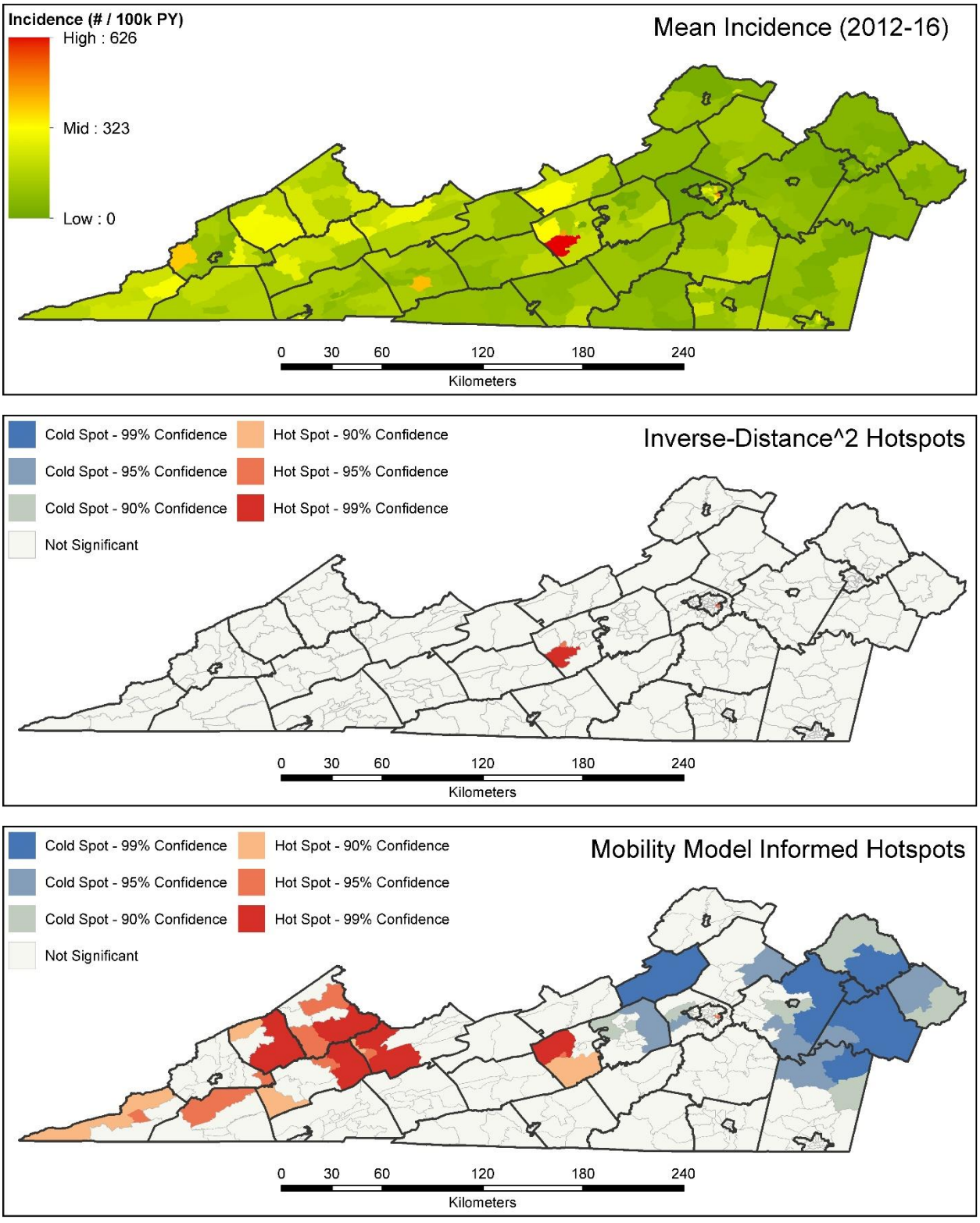


Figure 4.7: Shows response curves for the six most significant contributors to the BRT based incidence model. As expected, the temporal autocovariate is the most significant.

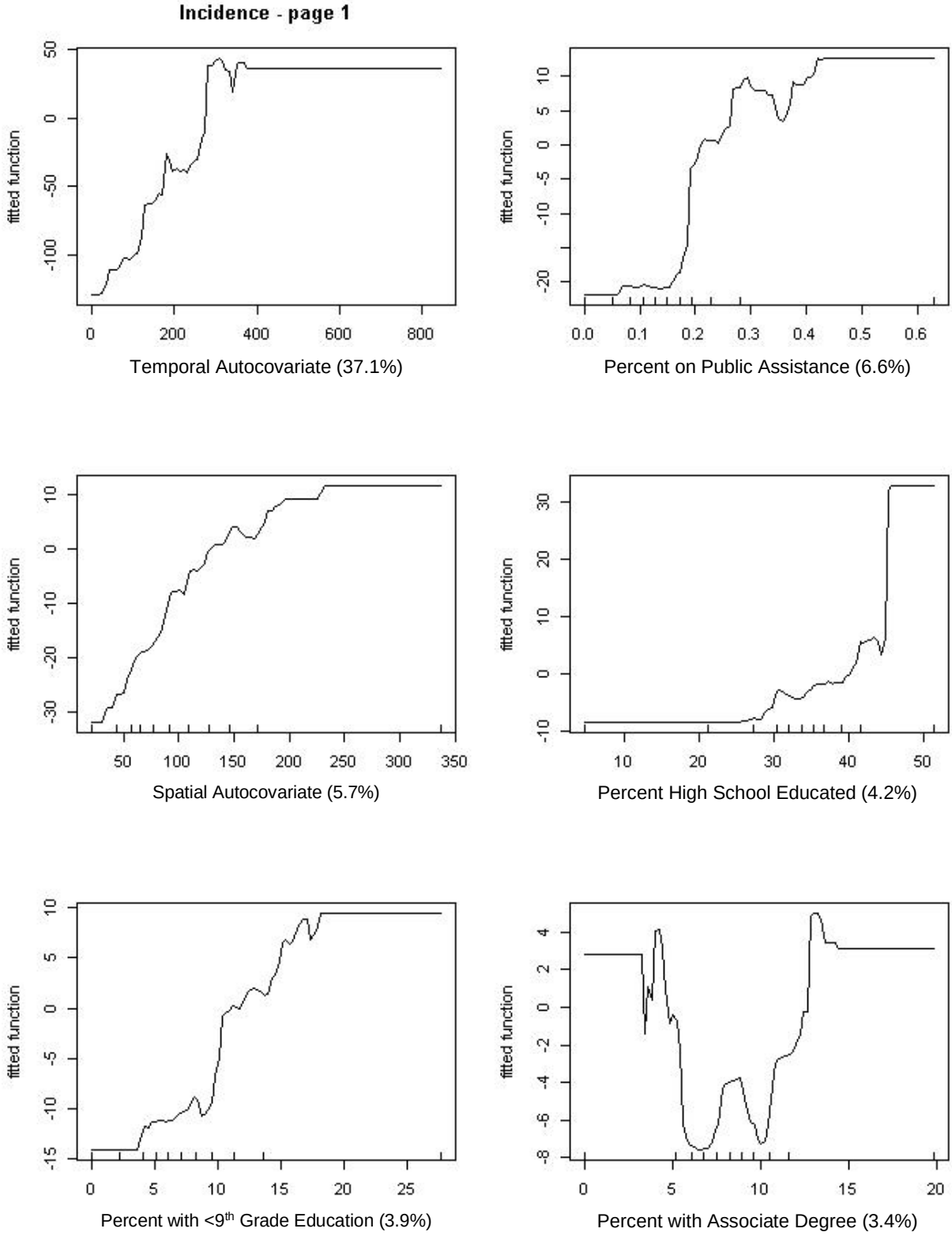


Figure 4.8: Shows the spatial autocorrelation of HCV model residuals with and without mobility. Areas in red saw more incidence in the real world than the model predicted, while blue areas experienced a lesser rate than expected. Inverse-distance squared (ID^2) and spatial weights matrix (SWM) methods both showed that the mobility model reduced residual autocorrelation.

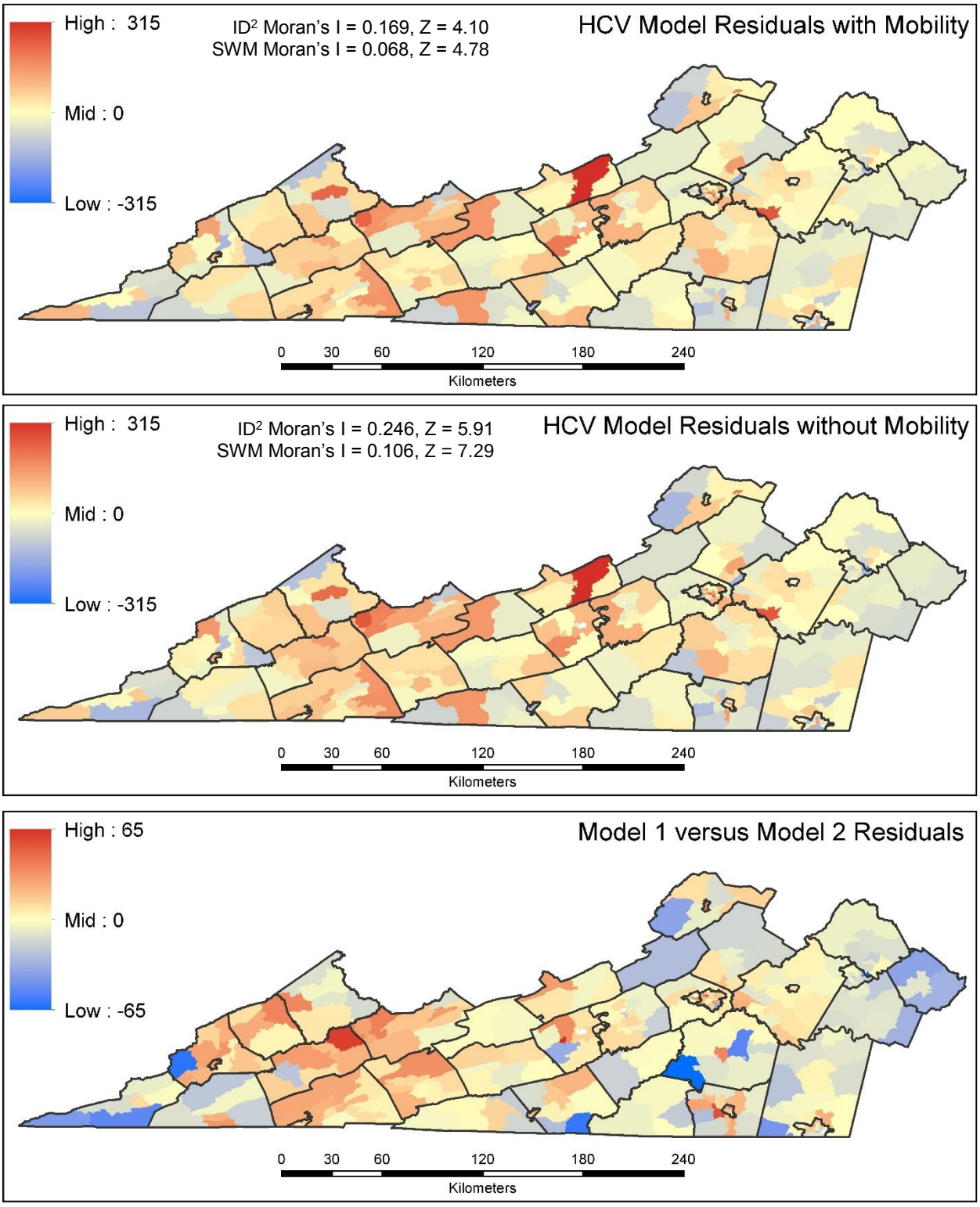


Table 4.1: Shows the results of our geocoding efforts. In the first section (top) we see the number of points by the number of high-quality results (HQR) from the eight different geocoders. For example, 4465 of the addresses we geocoded had HQR from all eight services, while ten points had just a single reliable result. In the second section, we see the number of HQR returned by each geocoding service. Bing Maps leads the pack with 98.4% of all inputs returning an HQR. The third section shows us agreement between geocoders, in terms of number of addresses placed in the same census tract by the eight different geocoding services (GS). For example, we see that all eight geocoders agreed on census tract assignment for 3917 of the addresses we input. Finally, the fourth section (bottom), shows us how often each service was part of the majority. For example, when it comes to assigning an address to a census tract, the ArcGIS geocoding service was part of the majority 98.52% of the time.

Number of Addresses x Number of High-Quality Geocoding Results

8 HQR	7 HQR	6 HQR	5 HQR	4 HQR	3 HQR	2 HQR	1 HQR	0 HQR
4,465	1,460	545	150	75	37	28	10	4
65.9%	21.6%	8.0%	2.2%	1.1%	0.5%	0.4%	0.1%	0.1%

Number of High-Quality Results per Geocoding Service

ArcGIS	Bing	Google	Here	MapBox	MapQuest	OpenCage	TomTom
6,605	6,668	6,149	6,603	5,657	6,612	5,668	6,474
97.5%	98.4%	90.8%	97.5%	83.5%	97.6%	83.7%	95.6%

Number of Address x Geocoders in Agreement on Census Tract Assignment

8 GS	7 GS	6 GS	5 GS	4 GS	3 GS	2 GS	1 GS	0 GS
3,917	1,596	701	232	127	67	32	9	24
58.6%	23.9%	10.5%	3.5%	1.9%	1.0%	0.5%	0.1%	0.4%

Frequency of Membership in the Majority Solution in Census Tract Assignment

ArcGIS	Bing	Google	Here	MapBox	MapQuest	OpenCage	TomTom
6,582	6,578	6,057	6,569	5,601	6,310	5,283	6,332
98.52%	98.46%	90.66%	98.32%	83.83%	94.45%	79.07%	94.78%

Table 4.2: Shows a comparison of census tract to census tract travel times as estimated by our network model (OD Cost) and Google Maps (Google). In this case we randomly selected an origin and twenty random destinations. With a roughly 10% difference between the two, we consider our network model to be sufficiently accurate for our purposes.

Origin	Destination	OD Cost	Google	Difference
51121020100	51121021300	11.0	18	39%
51121020100	51121021500	20.7	20	4%
51121020100	51161030500	35.9	40	10%
51121020100	51063920200	55.0	50	10%
51121020100	51023040502	62.8	55	14%
51121020100	51035080300	100.3	70	43%
51121020100	51005080202	87.7	100	12%
51121020100	51005080100	101.8	90	13%
51121020100	51185020100	95.9	100	4%
51121020100	51089010500	114.3	120	5%
51121020100	51143010200	113.0	110	3%
51121020100	51173030702	106.5	95	12%
51121020100	51031020102	120.4	120	0%
51121020100	51009010300	123.9	110	13%
51121020100	51143010801	130.2	130	0%
51121020100	51143010802	150.8	150	1%
51121020100	51027010600	147.7	140	6%
51121020100	51167030401	136.1	130	5%
51121020100	51195930800	230.0	200	15%
51121020100	51105950300	203.5	200	2%
			Mean	10.5%
			Pearson's ρ	0.981

Table 4.3: An example comparing the accuracy of our mobility model predicted commuter flows. Here we see the (real) fraction of ACS derived inbound commuters to Montgomery County, VA by origin. We also see the estimated fractions based on the Boosted Regression Trees model (BRT), the unweighted gravity model (Std-Grav) and the calibrated gravity model (Cal-Grav). By comparing the correlation coefficients (Pearson's ρ) and mean squared error (MSE) we see that the standard gravity model was the most accurate. Though this is only a single county used as an example, the results are mirrored in the full dataset, with the unweighted gravity model exhibiting a 0.939 correlation coefficient and by far the smallest mean-squared error when compared to the training data.

Origin	Real	BRT	Std-Grav	Cal-Grav
Carroll County	0.2%	1.3%	1.6%	3.3%
Bland County	0.2%	0.4%	0.4%	0.9%
Grayson County	0.3%	0.8%	0.4%	1.7%
Tazewell County	0.3%	1.7%	1.3%	4.0%
Botetourt County	0.9%	2.1%	3.5%	5.1%
Lynchburg City	1.3%	2.2%	2.8%	7.3%
Wythe County	1.8%	1.3%	2.5%	3.7%
Floyd County	2.0%	1.3%	2.3%	2.6%
Giles County	3.2%	10.5%	5.0%	4.6%
Salem City	15.8%	6.2%	9.8%	7.0%
Roanoke County	16.0%	13.2%	19.4%	18.0%
Radford City	17.1%	18.1%	16.4%	8.8%
Pulaski County	18.9%	31.9%	16.9%	13.1%
Roanoke City	21.9%	9.1%	17.7%	19.9%
	Pearson's ρ	0.753	0.959	0.859
	MSE	4.95	0.86	2.73

Table 4.4: Variable contributions for our BRT model of incidence. Note that while the temporal autocovariate is the most substantial predictor, the spatial autocovariate is also significant.

Variable	Influence	Cumulative
Temporal Autocovariate	37.1%	37.1%
Percent on Public Assistance	6.6%	43.7%
Spatial Autocovariate	5.7%	49.4%
Percent Finished High School	4.2%	53.6%
Percent <9th Grade Education	3.9%	57.5%
Percent with Associates Degree	3.4%	60.9%
Percent of Seniors in Poverty	3.3%	64.2%
Percent 9-12th Grade Education	3.3%	67.5%
Median Income	3.1%	70.6%
Sex Ratio	2.7%	73.3%
Percent with Bachelor's Degrees	2.6%	75.9%
Age-Dependency Ratio	2.5%	78.4%
Percent of Labor Force	2.4%	80.8%
Percent Some College	2.4%	83.2%
Age-Dependency Ratio (Seniors)	2.3%	85.5%
Median Age	2.2%	87.7%
Percent with Graduate Degrees	2.1%	89.8%
Age-Dependency Ratio (Children)	1.9%	91.8%
Unemployment Rate	1.9%	93.7%
Percent Living in Poverty	1.9%	95.6%
Spatial Autocovariate (Lagged)	1.6%	97.2%
Percent Fully Employed	1.4%	98.6%
Percent of Children in Poverty	1.4%	100.0%

Chapter

5. Conclusion

5.1 Theoretical Contributions

Over the course of this work, we have made several contributions to our theoretical understanding of both disease modeling and the driving factors of certain diseases. We have contributed further evidence that melioidosis is strongly driven by the environment. While most studies show associations with rainfall and temperature, we have demonstrated that humidity is also a possible contributor. Though further study is needed before one can claim causality, it is a plausible conclusion, given *B. pseudomallei's* abhorrence of desiccation and requirement for wet loose soil. We have also shown that wind speed is a factor, lending credence to the theory that the pathogen can be aerosolized and transported by dust and particulate debris. This is also plausible given that many other soil-borne pathogens, such as coccidioidomycosis, are affected by wind speed (167).

At a finer scale, we have also demonstrated a strong link between melioidosis and abiotic environmental factors, such as elevation and soil type. Again, this is well within the realm of biological plausibility, as elevation is strongly correlated with temperature and humidity, and one would expect a free-living saprophyte to have preferences for certain soils. Finally, we have demonstrated that it is possible to combine models trained on historical climatic data with weather forecasts, to estimate future incidence rates. Regrettably, we have also demonstrated the difficulties in this, given the challenge of forecasting weather so far into the future. Nevertheless, the concept can be extended to diseases that are not so reliant on future conditions, and we feel this remains a constructive contribution to the theory of disease modeling.

Our efforts in the Democratic Republic of Congo have shown that it isn't necessary to rely on curated mobility data when modeling an epidemic; in a pinch, one can simply generate their own mobility metrics from open source data. We have also shown that such mobility data can be used to inform metapopulation models and generate far more realistic results than traditional Euclidean distance metrics. Such mobility data also augments the detection of hotspots by properly conceptualizing the relationship between areas that interact despite geographic proximity or lack thereof. Finally, we have demonstrated that including a mobility derived spatial autocovariate term in a geostatistical model of a disease spread person-to-person, not only improves the predictive power, but reduces autocorrelation in the model residuals. A useful addition given that such autocorrelation violates the assumption of independent errors and increases the likelihood of type I errors.

5.2 Methodological Contributions

In **Chapter 2** we demonstrated the use of Boosted Regression Trees to create geostatistical models of incidence. While the algorithm is traditionally used with data following a Bernoulli distribution, and an expected logistic output, we have shown that it is effective at predicting continuous outcomes on a Gaussian distribution. We have also demonstrated a methodology for combining weather forecasts from the National Centers for Environmental Prediction with historical data to estimate future incidence cases, as well as a method for using georeferenced case data to examine the effects of fine-scale environmental variations.

In **Chapter 3**, we presented a method for creating a transportation network from open source river and road data from Digital Chart of the World. We also demonstrated how to incorporate varying edge speeds in the network and create a network dataset for use in mobility modeling. We then provided a method for finding the population-weighted centroids of health districts. Finally, we showed how one might create a gravity model of interaction and use this to inform a metapopulation patch model, and how doing so improves realism in model predictions.

In **Chapter 4**, we provided a method by which one could geocode a large number of rural addresses with high reliability. We demonstrated the construction of a mobility metric given an existing transportation network provided by Here/NavTeq, and how one might evaluate the accuracy of such a model given real-world commuter flow data from ACS. We then provided an alternative methodology for finding population centroids using land cover data. We demonstrated the use of hotspot analysis using a mobility model generated spatial weights matrix. And finally, we showed how to create a mobility weighted spatial autocovariate term for use in a geostatistical model to improve predictive power and reduce residual autocorrelation.

5.3 Applied Contributions

The methods and theoretical contributions we have presented here have broader implications in public health and disease modeling. Our travel network and centroid finding methodologies have applications in a variety of work. Mobility metrics depend on accurate travel networks, and in situations where neither mobility nor curated road and river data exist, our method allows a researcher to create their own. Such a travel network and the associated mobility models can be used to create metapopulation patch models, informing the public health response to emerging and ongoing epidemics. They can be used to estimate long distance interactions between synthetic agents in an agent-based model, improving the accuracy of such efforts. They can be used to augment hotspot detection by finding geographically disparate but connected areas, allowing health departments to detect outbreaks and focus their attention and resources more properly. And finally, they can be used to augment geostatistical models.

Our work with geostatistical modeling shows that such techniques can be used to identify regions of underreporting, weaknesses in surveillance coverage, and areas with unusual localized forcing. Doing so can allow a health department to target interventions such as educational outreach and focus surveillance efforts. It may also convince governments to take the threat of a specific pathogen more seriously and plan legislative countermeasures. It can be used to inform local physicians of the threats in their community. By combining these models with weather forecasts, we can also estimate future incidence rates, allowing local health departments time to prepare for epidemics and to ensure they have allocated resources appropriately. It is conceivable that this technique could one day be available to the public, and local disease risk may be as easily accessible as weather on a smart phone.

5.4 Future Direction

The future of our geostatistical modeling lies with Lyme disease. We have sought more accurately resolved melioidosis cases for several years and are not convinced that such data are available. Moreover, we have demonstrated that weather forecasts nine-months into the future are highly

suspect and unreliable. Lyme disease offers the opportunity to apply these methodologies to a disease that is primarily driven by historical conditions. Lyme is also a stage two pathogen in the path towards human endemicity (133), and there is no human-to-human transmission to confound our analyses or require the additional consideration of human mobility. We feel this is an excellent demonstration of this methodology and will likely produce more accurate predictions than our melioidosis work did. We intend to forecast Lyme incidence in the VDH Southwest Region for the 2019-2020 Lyme season, using road-bound polygons as our areal units.

Future work on mobility will focus on more accurately training gravity models, and an investigation into the use of radiation models which many studies have shown more accurately predict mobility by accounting for the proclivity of larger areas to divert or intercept travelers intended for more distant destinations (43). Consider two towns separated by an hour of travel time, the scale of interaction between them would certainly be affected by the presence or absence of a larger city situated between them. A gravity model has no way to account for this, while a radiation model does so with ease. We would also like to incorporate other explanatory variables into our mobility models, such as population density, and economic factors such as unemployment rate and median income. Network metrics such as betweenness and redundancy may also affect mobility. The addition of more accurate calibration data may also be useful. Many smartphone apps collect data on the mobility of users, as does Google itself. Such data would be invaluable in improving the training of such models.

We would also like to continue to investigate the means by which to accurately reduce polygonal areas, such as health districts or census tracts, to points for the creation of the origin-destination cost matrices used in our mobility models. We have presented the use of population and land cover-weighting, but other alternatives exist. Instead of trying to find a single point, perhaps it is more realistic to divide each areal unit into multiple points, calculate interactions between them, and then aggregate these up. One could do so by using a cluster analysis to find population centers, then a Voronoi tessellation to divide the areal unit into subunits around said population centers. This method could also provide the self-potential and internal mixing rates of each areal unit.

Another avenue of future exploration would be the use of agent-based modeling to examine the effects of prisons on the hepatitis C epidemic. As we discussed in Chapter 4, there are no effectual means by which to account for the diffusion of HCV cases out of prisons and back into the surrounding environment. But this could be approximated with an agent-based model, if we have some understanding of the service area of each prison and the distribution of the origin of inmates in that service area.

Finally, on a more personal note, my career seems to be tending towards the research track. But I would also like to continue working to advance the collaboration between the fields of public health and geography, to explore pedagogical methods for spanning the divide between dedicated geographers and formally trained epidemiologists, and to encourage cross-discipline inquiry. Geography offers remarkable insight into disease dynamics, and public health benefits immensely from geospatial analysis. I hope that this work has demonstrated this, if ever so slightly, and perhaps has inspired others to join in bridging the ever-shrinking gap between these fields.

Bibliography

1. Meade MS, Emch M. Medical Geography. Guilford Press; 2010.
2. Hill AB. The environment and disease: association or causation? SAGE Publications; 1965.
3. Retief F, Cilliers L. Malaria in Graeco-Roman times. Acta Cl. JSTOR; 2004;127–37.
4. Curran J. The Yellow Emperor's classic of internal medicine. British Medical Journal Publishing Group; 2008.
5. Snow J. On the mode of communication of cholera. John Churchill; 1855.
6. Gorgas WC. Sanitation in Panama. Appleton; 1915.
7. Gould P. Geography 1957--1977: the Augean period. Ann Assoc Am Geogr. Wiley Online Library; 1979;69(1):139–51.
8. Cliff AD, Ord K. Spatial autocorrelation: a review of existing and new measures with applications. Econ Geogr. Taylor & Francis; 1970;46(sup1):269–92.
9. Cooper L. Location-allocation problems. Oper Res. INFORMS; 1963;11(3):331–43.
10. Krämer A, Kretzschmar M, Krickeberg K. Modern infectious disease epidemiology: Concepts, methods, mathematical models, and public health. Springer; 2010.
11. Lai S, Zhou H, Xiong W, Gilbert M, Huang Z, Yu J, et al. Changing epidemiology of human brucellosis, China, 1955--2014. Emerg Infect Dis. Centers for Disease Control and Prevention; 2017;23(2):184.
12. Ord JK, Getis A. Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal. Wiley Online Library; 1995;27(4):286–306.
13. Aarabi S, Sidhwa F, Riehle KJ, Chen Q, Mooney DP. Pediatric appendicitis in New England: epidemiology and outcomes. J Pediatr Surg. Elsevier; 2011;46(6):1106–14.
14. Coleman M, Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN. Using the SaTScan method to detect local malaria clusters for guiding malaria control programmes. Malar J. BioMed Central; 2009;8(1):68.
15. Liu W, Yang K, Qi X, Xu K, Ji H, Ai J, et al. Spatial and temporal analysis of human infection with avian influenza A(H7N9) virus in China, 2013. Euro Surveill. 2013 Jan;18(47):1–8.
16. Loonsk JW. BioSense—a national initiative for early detection and quantification of public health emergencies. MMWR Morb Mortal Wkly Rep. 2004;53(Suppl):53–5.
17. Bradley CA, Rolka H, Walker D, Loonsk J. BioSense: implementation of a national early event detection and situational awareness system. MMWR Morb Mortal Wkly Rep. 2005;54(Suppl):11–9.
18. Schlitt JT, Lewis B, Eubank S. ChatterGrabber: A lightweight easy to use social media surveillance toolkit. Online J Public Health Inform. University of Illinois at Chicago Library; 2015;7(1).
19. Peterson AT. Mapping disease transmission risk: enriching models using biogeography and ecology. JHU Press; 2014.
20. Phillips SJ, Dudík M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography (Cop). Wiley Online Library; 2008;31(2):161–75.
21. Warren DL, Seifert S. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecol Soc Am. 2011;21(2):335–42.
22. Elith J, Phillips SJ, Hastie T, Dudík M, Chee YE, Yates CJ. A statistical explanation of MaxEnt for ecologists. Divers Distrib. 2011;17(1):43–57.
23. Elith J, Leathwick J. Boosted Regression Trees for ecological modeling. R Doc Available https//cran r-project org/web/packages/dismo/vignettes/brt pdf. 2017;
24. Elith J, Graham CH. Do they? How do they? WHY do they differ? on finding reasons for differing

performances of species distribution models. Ecography (Cop). 2009;32(1):66–77.

25. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008;77(4):802–13.
26. Stockwell D. The GARP modelling system: problems and solutions to automated spatial prediction. Int J Geogr Inf Sci. Taylor & Francis; 1999;13(2):143–58.
27. Escobar LE, Qiao H, Lee C, Phelps NBD. Novel Methods in Disease Biogeography: A Case Study with Heterosporosis. Front Vet Sci. 2017;4(July).
28. Guo Q, Kelly M, Graham CH. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecol Modell. Elsevier; 2005;182(1):75–90.
29. Kraemer MUG, Sinka ME, Duda KA, Mylne AQN, Shearer FM, Barker CM, et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. Albopictus. Elife. 2015;4(JUNE2015):1–18.
30. Mak S, Morshed M, Henry B. Ecological niche modeling of Lyme disease in British Columbia, Canada. J Med Entomol. The Oxford University Press; 2010;47(1):99–105.
31. Limmathurotsakul D, Golding N, Dance DAB, Messina JP, Pigott DM, Moyes CL, et al. Predicted global distribution of Burkholderia pseudomallei and burden of melioidosis. Nat Microbiol. 2016;1(January):1–5.
32. Pigott DM, Golding N, Mylne A, Huang Z, Henry AJ, Weiss DJ, et al. Mapping the zoonotic niche of Ebola virus disease in Africa. Elife. eLife Sciences Publications Limited; 2014;3:e04395.
33. Shcheglovitova M, Anderson RP. Estimating optimal complexity for ecological niche models: A jackknife approach for species with small sample sizes. Ecol Modell. Elsevier B.V.; 2013;269:9–17.
34. De Araújo CB, Marcondes-Machado LO, Costa GC. The importance of biotic interactions in species distribution models: A test of the Eltonian noise hypothesis using parrots. J Biogeogr. 2014;41(3):513–23.
35. Mullins JC, Garofolo G, Van Ert M, Fasanella A, Lukhnova L, Hugh-Jones ME, et al. Ecological niche modeling of Bacillus anthracis on three continents: evidence for genetic-ecological divergence? PLoS One. 2013 Jan [cited 2014 Jun 9];8(8):e72451.
36. Peterson AT. Predicting the Geography of Species' Invasions via Ecological Niche Modeling. Q Rev Biol. 2003;78(4):419–33.
37. Brownstein JS, Holford TR, Fish D. Effect of climate change on lyme disease risk in North America. Ecohealth. 2005;2(1):38–46.
38. Carvalho BM, Rangel EF, Vale MM. Evaluation of the impacts of climate change on disease vectors through ecological niche modelling. Bull Entomol Res. Cambridge University Press; 2017;107(4):419–30.
39. Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, et al. Global mapping of infectious disease. Philos Trans R Soc. 2013;368(1614):20120250.
40. Getis A, Ord JK. The analysis of spatial association by use of distance statistics. In: Perspectives on Spatial Data Analysis. Springer; 2010. p. 127–45.
41. Crase B, Liedloff AC, Wintle BA. A new method for dealing with residual spatial autocorrelation in species distribution models. Ecography (Cop). 2012;35(10):879–88.
42. Buscema M, Grossi E, Breda M, Jefferson T. Outbreaks source: A new mathematical approach to identify their possible location. Phys A Stat Mech its Appl. Elsevier B.V.; 2009;388(22):4736–62.
43. Ren Y, Ercsey-Ravasz M, Wang P, González MC, Toroczkai Z. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. Nat Commun. 2014;5.
44. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, Vespignani A. Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci. National Acad Sciences; 2009;106(51):21484–9.

45. Alexander KA, Lewis BL, Marathe M, Eubank S, Blackburn JK. Modeling of Wildlife-Associated Zoonoses: Applications and Caveats. Vector-Borne Zoonotic Dis. 2012;12(12):1005–18.

46. Bisset KR, Chen J, Feng X, Kumar VS, Marathe M V. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of the 23rd international conference on Supercomputing. 2009. p. 430–9.

47. Barrett CL, Bisset KR, Eubank SG, Feng X, Marathe M V. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In: SC'08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing. 2008. p. 1–12.

48. Mohan J. Location-allocation models, social science and health service planning: An example from North East England. Soc Sci Med. Elsevier; 1983;17(8):493–9.

49. Yeh AG-O, Chow MH. An integrated GIS and location-allocation approach to public facilities planning—An example of open space planning. Comput Environ Urban Syst. Elsevier; 1997;20(4):339–50.

50. ReVelle CS, Swain RW. Central facilities location. Geogr Anal. 1970;2(1):30–42.

51. Church R, Velle CR. The maximal covering location problem. Pap Reg Sci. Wiley Online Library; 1974;32(1):108–18.

52. Kumar N. Changing geographic access to and locational efficiency of health services in two Indian districts between 1981 and 1996. Soc Sci Med. Elsevier; 2004;58(10):2045–67.

53. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature. 2017;544(7650):309–15.

54. Faria NR, Do Socorro Da Silva Azevedo R, Kraemer MUG, Souza R, Cunha MS, Hill SC, et al. Zika virus in the Americas: Early epidemiological and genetic findings. Science (80-). 2016;352(6283):345–9.

55. Clements ACA, Moyeed R, Brooker S. Bayesian geostatistical prediction of the intensity of infection with Schistosoma mansoni in East Africa. Parasitology. 2006;133(6):711–9.

56. Vounatsou P, Raso G, Tanner M, N'Goran EK, Utzinger J. Bayesian geostatistical modelling for mapping schistosomiasis transmission. Parasitology. 2009;136(13):1695–705.

57. Diggle PJ, Thomson MC, Christensen OF, Rowlingson B, Obsomer V, Gardon J, et al. Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty. Ann Trop Med Parasitol. Taylor & Francis; 2007;101(6):499–509.

58. Raso G, Schur N, Utzinger J, Koudou BG, Tchicaya ES, Rohner F, et al. Mapping malaria risk among children in Côte dIvoire using Bayesian geo-statistical models. Malar J. 2012;11:1–11.

59. Smith DL, Guerra CA, Snow RW, Hay SI. Standardizing estimates of the Plasmodium falciparum parasite rate. Malar J. 2007;6:1–10.

60. Watson SC, Liu Y, Lund RB, Gettings JR, Nordone SK, McMahan CS, et al. A Bayesian spatio-temporal model for forecasting the prevalence of antibodies to Borrelia burgdorferi, causative agent of Lyme disease, in domestic dogs within the contiguous United States. PLoS One. San Francisco: Public Library of Science; 2017 May;12(5).

61. Anggraeni W, Nurmasari R, Riksakomara E, Samopa F, Wibowo RP, Condro LT, et al. Modified Regression Approach for Predicting Number of Dengue Fever Incidents in Malang Indonesia. Procedia Comput Sci. Elsevier B.V.; 2017;124:142–50. Available from: https://doi.org/10.1016/j.procs.2017.12.140

62. Kaestli M, Grist EPM, Ward L, Hill A, Mayo M, Currie BJ. The association of melioidosis with climatic factors in Darwin, Australia: A 23-year time-series analysis. J Infect. Elsevier Ltd; 2016;72(6):687–97.

63. Cheng AC, Currie BJ. Melioidosis: epidemiology, pathophysiology, and management. Clin Microbiol Rev. 2005;18(2):383–416.

64. Kaestli M, Mayo M, Harrington G, Watt F, Hill J, Gal D, et al. Sensitive and specific molecular detection of Burkholderia pseudomallei, the causative agent of melioidosis, in the soil of tropical northern Australia. Appl Environ Microbiol. 2007;73(21):6891–7.

65. Currie BJ, Ward L, Cheng AC. The epidemiology and clinical spectrum of melioidosis: 540 cases from the 20 year darwin prospective study. PLoS Negl Trop Dis. 2010;4(11).

66. Currie BJ, Fisher D a, Howard DM, Burrow JN, Lo D, Selva-Nayagam S, et al. Endemic melioidosis in tropical northern Australia: a 10-year prospective study and review of the literature. Clin Infect Dis. 2000;31(4):981–6.

67. Currie BJ, Jacups SP. Intensity of Rainfall and Severity of Melioidosis, Australia. Emerg Infect Dis. 2003;9(12):1538–42.

68. Baker AL, Warner JM. Burkholderia pseudomallei is frequently detected in groundwater that discharges to major watercourses in northern Australia. Folia Microbiol (Praha). 2015;1–5.

69. Baker AL, Ezzahir J, Gardiner C, Shipton W, Warner JM. Environmental attributes influencing the distribution of Burkholderia pseudomallei in Northern Australia. PLoS One. 2015;10(9):1–11.

70. State of Queensland. Melioidosis in Queensland – 2012-2016 [Internet]. 2017. Available from: https://www.health.qld.gov.au/__data/assets/pdf_file/0026/671183/melioidosis-qld-2012-2016.pdf

71. Currie BJ, Jacups SP, Cheng AC, Fisher DA, Anstey NM, Huffam SE, et al. Melioidosis epidemiology and risk factors from a prospective whole-population study in northern Australia. Trop Med Int Heal. 2004;9(11):1167–74.

72. Dance DAB. Melioidosis: the Tip of the Iceberg? Clin Microbiol Rev. 1991;19(57):42.

73. Parameswaran U, Baird RW, Ward LM, Currie BJ. Melioidosis at royal darwin hospital in the big 2009-2010 wet season: Comparison with the preceding 20 years. Med J Aust. 2012;196(5):345–8.

74. Cheng AC, Jacups SP, Gal D, Mayo M, Currie BJ. Extreme weather events and environmental contamination are associated with case-clusters of melioidosis in the Northern Territory of Australia. Int J Epidemiol. 2006;35(2):323–9.

75. Dobson JE, Bright EA, Coleman PR, Durfee RC, Worley BA. LandScan: a global population database for estimating populations at risk. Photogramm Eng Remote Sensing. 2000;66(7):849–57.

76. Barbet-Massin M, Jiguet F, Albert CH, Thuiller W. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol Evol. Wiley Online Library; 2012;3(2):327–38.

77. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. Wiley Online Library; 2017;37(12):4302–15.

78. Nachtergaele F, van Velthuizen H, Verelst L, Batjes NH, Dijkshoorn K, van Engelen VWP, et al. The harmonized world soil database. In: Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010. 2010. p. 34–7.

79. Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. Taylor & Francis Group; 1996;5(3):299–314.

80. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Modell. Elsevier; 2006;190(3):231–59.

81. Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, et al. The NCEP climate forecast system version 2. J Clim. 2014;27(6):2185–208.

82. Elith J, Leathwick J. Boosted regression trees for ecological modeling. 2016;

83. De'ath G. Boosted trees for ecological modeling and prediction. Ecology. 2007;88(1):243–51.

84. Corkeron ML, Norton R, Nelson PN. Spatial analysis of melioidosis distribution in a suburban area. Epidemiol Infect. 2010;138(9):1346–52.

85. Danko DM. The digital chart of the world project. Photogramm Eng Remote Sens. 1992;58:1125–8.

86. Mortlock TR, Metters D, Soderholm J, Maher J, Lee SB, Boughton G, et al. Extreme water levels, waves and coastal impacts during a severe tropical cyclone in northeastern Australia: A case study for cross-sector data sharing. Nat Hazards Earth Syst Sci. 2018;18(9):2603–23.

87. Stewart JD, Smith S, Binotto E, McBride WJ, Currie BJ, Hanson J. The epidemiology and clinical features of melioidosis in Far North Queensland: Implications for patient management. PLoS Negl Trop Dis. 2017;11(3):1–15.

88. Cheng AC, Jacups SP, Ward L, Currie BJ. Melioidosis and Aboriginal seasons in northern Australia. Trans R Soc Trop Med Hyg. 2008;102(SUPPL. 1).

89. Birch CPD, Oom SP, Beecham JA. Rectangular and hexagonal grids used for observation, experiment and simulation in ecology. Ecol Modell. 2007;206(3–4):347–59.

90. Levi T, Kilpatrick AM, Mangel M, Wilmers CC. Deer, predators, and the emergence of Lyme disease. Proc Natl Acad Sci. 2012;109(27):10942–7.

91. Berger KA, Ginsberg HS, Dugas KD, Hamel LH, Mather TN. Adverse moisture events predict seasonal abundance of Lyme disease vector ticks (Ixodes scapularis). Parasites and Vectors. 2014;7(1):1–8.

92. Subak S. Effects of Climate on Variability in Lyme Disease Incidence in the Northeastern United States. Am J Epidemiol. 2003;157(6):531–8.

93. Ogden NH, Bigras-Poulin M, O'Callaghan CJ, Barker IK, Lindsay LR, Maarouf A, et al. A dynamic population model to investigate effects of climate on geographic range and seasonality of the tick Ixodes scapularis. Int J Parasitol. 2005;35(4):375–89.

94. Allan BF, Keesing F, Ostfeld RS. Effect of forest fragmentation on lyme disease risk. Conserv Biol. 2003;17(1):267–72.

95. Brownstein JS, Skelly DK, Holford TR, Fish D. Forest fragmentation predicts local scale heterogeneity of Lyme disease risk. Oecologia. 2005;146(3):469–75.

96. Nupp TE, Swihart RK. Effect of forest patch area on population attributes of white-footed mice (Peromyscus leucopus) in fragmented landscapes. Can J Zool. NRC Research Press; 1996;74(3):467–72.

97. Das A, Lele SR, Glass GE, Shields T, Patz J. Modelling a discrete spatial response using generalized linear mixed models: Application to Lyme disease vectors. Int J Geogr Inf Sci. 2002;16(2):151–66.

98. Li J, Kolivras KN, Hong Y, Duan Y, Seukep SE, Prisley SP, et al. Spatial and Temporal Emergence Pattern of Lyme Disease in Virginia. Am J Trop Med Hyg. 2014;91(6):1166–72.

99. Seukep SE, Kolivras KN, Hong Y, Li J, Prisley SP, Campbell JB, et al. An Examination of the Demographic and Environmental Variables Correlated with Lyme Disease Emergence in Virginia. Ecohealth. Springer US; 2015;2011(Figure 1).

100. Glass GE, Schwartz BS, Morgan JM, Johnson DT, Noy PM, Israel E. Environmental risk factors for Lyme disease identified with geographic information systems. Am J Public Health. 1995;85(7):944–8.

101. Ostfeld RS, Canham CD, Oggenfuss K, Winchcombe RJ, Keesing F. Climate, deer, rodents, and acorns as determinants of variation in Lyme-disease risk. PLoS Biol. 2006;4(6):1058–68.

102. Koenig WD, Knops JMH, Carmen WJ, Stanback MT, Mumme RL. Acorn production by oaks in central coastal California: influence of weather at three levels. Can J For Res. 1996;26(9):1677–83.

103. Monaghan AJ, Moore SM, Sampson KM, Beard CB, Eisen RJ. Climate change influences on the annual onset of Lyme disease in the United States. Ticks Tick Borne Dis. Elsevier GmbH.; 2015;6(5):615–22.

104. Hu W, Tong S, Mengersen K, Oldenburg B. Rainfall, mosquito density and the transmission of Ross River virus: A time-series forecasting model. Ecol Modell. 2006;196(3–4):505–14.

105. Pourrut X, Kumulungui B, Wittmann T, Moussavou G, Délicat A, Yaba P, et al. The natural history of Ebola virus in Africa. Microbes Infect. 2005;7(7–8):1005–14.

106. Breman JG, Piot P, Johnson KM, White MK, Mbuyi M, Sureau P, et al. The epidemiology of Ebola hemorrhagic fever in Zaire, 1976. Ebola virus Haemorrh fever. Elsevier/North Holland; 1978;103–24.

107. Thacker PD. An Ebola epidemic simmers in Africa. JAMA. American Medical Association; 2003;290(3):317–9.

108. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of Zaire Ebola Virus Disease in Guinea - Preliminary Report. N Engl J Med. 2014;371(15):1418–25.

109. Bright EA, Rose AN, Urban ML. LandScan 2013 Global Population Database. East View Inf Serv ORN Lab. 2013;

110. Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ. Commentary: containing the Ebola outbreak-the potential and challenge of mobile network data. PLoS Curr. Public Library of Science; 2014;6.

111. Bell BP. Overview, control strategies, and lessons learned in the CDC response to the 2014--2016 Ebola epidemic. MMWR Suppl. 2016;65.

112. Poletto C, Gomes MFC, y Piontti AP, Rossi L, Bioglio L, Chao DL, et al. Assessing the impact of travel restrictions on international spread of the 2014 West African Ebola epidemic. Euro Surveill Bull Eur sur les Mal Transm Eur Commun Dis Bull. NIH Public Access; 2014;19(42).

113. Gomes MFC, y Piontti AP, Rossi L, Chao D, Longini I, Halloran ME, et al. Assessing the international spreading risk associated with the 2014 West African Ebola outbreak. PLoS Curr. Public Library of Science; 2014;6.

114. Gatherer D, Kohl A. Zika virus: A previously slow pandemic spreads rapidly through the Americas. J Gen Virol. 2016;97(2):269–73.

115. Green A. Ebola outbreak in the DR Congo: lessons learned. Lancet. Elsevier Ltd; 2018;391(10135):2096.

116. Munster VJ, Bausch DG, de Wit E, Fischer R, Kobinger G, Muñoz-Fontela C, et al. Outbreaks in a Rapidly Changing Central Africa—Lessons from Ebola. N Engl J Med. Mass Medical Soc; 2018;

117. Gostin LO. New Ebola outbreak in Africa is a major test for the WHO. JAMA - J Am Med Assoc. 2018;320(2):125–6.

118. Legrand J, Grais RF, Boelle PY, Valleron AJ, Flahault A. Understanding the dynamics of Ebola epidemics. Epidemiol Infect. 2007;135(4):610–21.

119. Bissett K, Cadena J, Khan M, Kuhlman CJ, Lewis B, Telionis PA. An integrated agent-based approach for modeling disease spread in large populations to support health informatics. In: Biomedical and Health Informatics (BHI), 2016 IEEE-EMBS International Conference on. 2016. p. 629–32.

120. Riley S. Models of Infectious Disease. Science (80). 2007;316(5829):1298–301.

121. Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLoS Comput Biol. 2012;8(10).

122. Simini F, González MC, Maritan A, Barabási AL. A universal model for mobility and migration patterns. Nature. 2012;484(7392):96–100.

123. Foster V, Benitez DA. The Democratic Republic of Congo's infrastructure: a continental perspective. The World Bank; 2011.

124. Wu Q, Merchant F, Castleman K. Microscope image processing. Elsevier; 2010.

125. HDX. Democratic Republic of Congo and Neighboring Countries Health Boundaries. 2018.

126. Linard C, Gilbert M, Snow RW, Noor AM, Tatem AJ. Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS One. Public Library of Science; 2012;7(2):e31743.

127. Marathe M V, Vullikanti AKS. Computational epidemiology. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014. p. 1969.

128. Daley DJ, Gani J. Epidemic modelling: an introduction. Vol. 15. Cambridge University Press; 2001.

129. Venkatramanan S, Chen J, Gupta S, Lewis B, Marathe M, Mortveit H, et al. Spatio-temporal optimization of seasonal vaccination using a metapopulation model of influenza. In: Healthcare Informatics (ICHI), 2017 IEEE International Conference on. 2017. p. 134–43.

130. Moran B. Fighting Ebola in conflict in the DR Congo. Lancet. Elsevier; 2018;392(10155):1295–6.

131. Arie S. Rebel attacks and political scaremongering raise risk of Ebola spreading in the Congo. British Medical Journal Publishing Group; 2018.

132. Cousins S. Violence and community mistrust hamper Ebola response. Lancet Infect Dis. Elsevier; 2018;18(12):1314–5.

133. Wolfe ND, Dunavan CP, Diamond J. Origins of major human infectious diseases. Nature. 2007;447(7142):279–83.

134. Lloyd C. Spatial data analysis: an introduction for GIS users. Oxford university press; 2010.

135. Lennon JJ. Red-shifts and red herrings in geographical ecology. Ecography (Cop). Wiley Online Library; 2000;23(1):101–13.

136. Diniz-Filho JAF, Blackburn TM, De Marco P, Mauricio Bini L, Hawkins BA. Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. Ecography (Cop). 2007;30(3):375–84.

137. Legendre P. Spatial Autocorrelation : Trouble or New Paradigm ? Author (s): Pierre Legendre Published by : Ecological Society of America SPATIAL AUTOCORRELATION : TROUBLE OR NEW PARADIGM ? Ecology. 1993;74(6):1659–73.

138. Hengl T, Heuvelink GBM, Rossiter DG. About regression-kriging: From equations to case studies. Comput Geosci. Elsevier; 2007;33(10):1301–15.

139. Kumar S, Lal R, Liu D. A geographically weighted regression kriging approach for mapping soil organic carbon stock. Geoderma. Elsevier; 2012;189:627–34.

140. Ly KN, Hughes EM, Jiles RB, Holmberg SD. Rising mortality associated with Hepatitis C virus in the United States, 2003-2013. Clin Infect Dis. 2016;62(10):1287–8.

141. Petruzziello A, Marigliano S, Loquercio G, Cozzolino A, Cacciapuoti C. Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes. World J Gastroenterol. 2016;22(34):7824–40.

142. Saito I, Miyamura T, Ohbayashi A, Harada H, Katayama T, Kikuchi S, et al. Hepatitis C virus infection is associated with the development of hepatocellular carcinoma. Proc Natl Acad Sci. National Acad Sciences; 1990;87(17):6547–9.

143. de Oliveria Andrade LJ, D'Oliveira A, Junior RCM, De Souza EC, Silva CAC, Parana R. Association between hepatitis C and hepatocellular carcinoma. J Glob Infect Dis. Wolters Kluwer--Medknow Publications; 2009;1(1):33.

144. Goossens N, Hoshida Y. Hepatitis C virus-induced hepatocellular carcinoma. Clin Mol Hepatol. Korean Association for the Study of the Liver; 2015;21(2):105.

145. Davis GL, Albright JE, Cook SF, Rosenberg DM. Projecting future complications of chronic hepatitis C in the United States. Liver Transplant. Wiley Online Library; 2003;9(4):331–8.

146. Kim WR. The burden of hepatitis C in the United States. Hepatology. Wiley Online Library; 2002;36(S1):S30--S34.

147. Salmon-Ceron D, Lewden C, Morlat P, Bévilacqua S, Jougla E, Bonnet F, et al. Liver disease as a major cause of death among HIV infected patients: Role of hepatitis C and B viruses and alcohol. J Hepatol. 2005;42(6):799–805.

148. Thomas DL, Factor SH, Kelen GD, Washington AS, Taylor E, Quinn TC. Viral hepatitis in health care personnel at the Johns Hopkins Hospital: the seroprevalence of and risk factors for hepatitis B virus and hepatitis C virus infection. Arch Intern Med. American Medical Association; 1993;153(14):1705–12.
149. Shepard CW, Finelli L, Alter MJ. Global epidemiology of hepatitis C virus infection. Lancet Infect Dis. Elsevier; 2005;5(9):558–67.
150. Alter MJ. Epidemiology of Hepatitis C. Hepatology. 1997;26(S3):62S–65S.
151. Tohme RA, Holmberg SD. Transmission of hepatitis C virus infection through tattooing and piercing: a critical review. Clin Infect Dis. Oxford University Press; 2012;cir991.
152. Tohme RA, Holmberg SD. Is sexual contact a major mode of hepatitis C virus transmission? Hepatology. Wiley Online Library; 2010;52(4):1497–505.
153. Thompson ND, Perz JF, Moorman AC, Holmberg SD. Nonhospital health care--associated hepatitis B and C virus transmission: United States, 1998--2008. Ann Intern Med. Am Coll Physicians; 2009;150(1):33–9.
154. Akbar N, Basuki B, Garabrant DH, Sulaiman A, Noer HMS, MULYANTO. Ethnicity, socioeconomic status, transfusions and risk of hepatitis B and hepatitis C infection. J Gastroenterol Hepatol. Wiley Online Library; 1997;12(11):752–7.
155. Asher AK, Patel RC, Zibbell JE, Ward JW, Kupronis B, Holtzman D, et al. Increases in Acute Hepatitis C Virus Infection Related to a Growing Opioid Epidemic and Associated Injection Drug Use, United States, 2004 to 2014. Am J Public Health. 2017;108(2):175–81.
156. Kolodny A, Courtwright DT, Hwang CS, Kreiner P, Eadie JL, Clark TW, et al. The prescription opioid and heroin crisis: a public health approach to an epidemic of addiction. Annu Rev Public Health. Annual Reviews; 2015;36:559–74.
157. Jones CM. Heroin use and heroin use risk behaviors among nonmedical users of prescription opioid pain relievers--United States, 2002--2004 and 2008--2010. Drug Alcohol Depend. Elsevier; 2013;132(1):95–100.
158. Zibbell JE, Iqbal K, Patel RC, Suryaprasad A, Sanders KJ, Moore-Moravian L, et al. Increases in hepatitis C virus infection related to injection drug use among persons aged≤ 30 years-Kentucky, Tennessee, Virginia, and West Virginia, 2006-2012. MMWR Morb Mortal Wkly Rep. Centers for Disease Control and Prevention; 2015;64(17):453–8.
159. Bonner MR, Han D, Nie J, Rogerson P, Vena JE, Freudenheim JL. Positional accuracy of geocoded addresses in epidemiologic research. Epidemiology. LWW; 2003;14(4):408–12.
160. Ward MH, Nuckols JR, Giglierano J, Bonner MR, Wolter C, Airola M, et al. Positional accuracy of two methods of geocoding. Epidemiology. JSTOR; 2005;542–7.
161. Mazumdar S, Rushton G, Smith BJ, Zimmerman DL, Donham KJ. Geocoding accuracy and the recovery of relationships between environmental exposures and health. Int J Health Geogr. BioMed Central; 2008;7(1):13.
162. Jackson L, Levine J, Hilborn E. A comparison of analysis units for associating Lyme disease with forest-edge habitat. Community Ecol. 2006;7(2):189–97.
163. Lopez D, Gunasekaran M, Murugan BS, Kaur H, Abbas KM. Spatial big data analytics of influenza epidemic in Vellore, India. Proc - 2014 IEEE Int Conf Big Data, IEEE Big Data 2014. 2014;19–24.
164. Christiansen CL, Morris CN. Hierarchical Poisson regression modeling. J Am Stat Assoc. Taylor & Francis Group; 1997;92(438):618–32.
165. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser b (statistical Methodol. Wiley Online Library; 2009;71(2):319–92.
166. Wang F, Antipova A, Porta S. Street centrality and land use intensity in Baton Rouge, Louisiana. J Transp Geogr. Elsevier Ltd; 2011;19(2):285–93.
167. Weaver EA, Kolivras KN. Investigating the Relationship Between Climate and Valley Fever (Coccidioidomycosis). Ecohealth. Springer; 2018;15(4):840–52.