5-2013

Landscape Epidemiology and Machine Learning: A Geospatial Approach to Modeling West Nile Virus Risk in the United States

Sean Gregory Young
University of Arkansas, Fayetteville

LANDSCAPE EPIDEMIOLOGY AND MACHINE LEARNING: A GEOSPATIAL APPROACH TO MODELING WEST NILE VIRUS RISK IN THE UNITED STATES

LANDSCAPE EPIDEMIOLOGY AND MACHINE LEARNING: A GEOSPATIAL APPROACH TO MODELING WEST NILE VIRUS RISK IN THE UNITED STATES

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Arts in Geography

By

Sean Gregory Young
Brigham Young University
Bachelor of Science in Geography, 2011

May 2013
University of Arkansas

ABSTRACT

The complex interactions between human health and the physical landscape and environment have been recognized, if not fully understood, since the ancient Greeks. Landscape epidemiology, sometimes called spatial epidemiology, is a sub-discipline of medical geography that uses environmental conditions as explanatory variables in the study of disease or other health phenomena. This theory suggests that pathogenic organisms (whether germs or larger vector and host species) are subject to environmental conditions that can be observed on the landscape, and by identifying where such organisms are likely to exist, areas at greatest risk of the disease can be derived. Machine learning is a sub-discipline of artificial intelligence that can be used to create predictive models from large and complex datasets. West Nile virus (WNV) is a relatively new infectious disease in the United States, and has a fairly well-understood transmission cycle that is believed to be highly dependent on environmental conditions. This study takes a geospatial approach to the study of WNV risk, using both landscape epidemiology and machine learning techniques. A combination of remotely sensed and *in situ* variables are used to predict WNV incidence with a correlation coefficient as high as 0.86. A novel method of mitigating the small numbers problem is also tested and ultimately discarded. Finally a consistent spatial pattern of model errors is identified, indicating the chosen variables are capable of predicting WNV disease risk across most of the United States, but are inadequate in the northern Great Plains region of the US.

This thesis is approved for recommendation
to the Graduate Council.

Thesis Director:

_____

Dr. Jason A Tullis

Thesis Committee:

_____

Dr. Jackson Cothren

_____

Dr. Bart Hammig

THESIS DUPLICATION RELEASE

I hereby authorize the University of Arkansas Libraries to duplicate this thesis when needed for research and/or scholarship.

Agreed _____
Sean Gregory Young

Refused _____
Sean Gregory Young

ACKNOWLEDGEMENTS

This study would not have been possible without the data generously provided by the Centers for Disease Control and Prevention (CDC). I also wish to thank each member of my thesis committee for their encouragement, support, and patience.

DEDICATION

This thesis is dedicated to my loving wife Brittney and to our wonderful children.

CONTENTS

1. Introduction and Background

Landscape epidemiology, sometimes called spatial epidemiology, is a sub-discipline of medical geography that uses environmental conditions as explanatory variables in the study of disease or other health phenomena. Machine learning is a sub-discipline of artificial intelligence that can be used to create predictive models from large and complex datasets. West Nile virus (WNV) is a relatively new infectious disease in the United States, and has a fairly well-understood transmission cycle that is believed to be highly dependent on environmental conditions. This study takes a geospatial approach to the study of WNV, using both landscape epidemiology and machine learning techniques.

"Given that the transmission of pathogens leading to disease requires the close juxtaposition of a susceptible individual with an infected conspecific, vector, or environmental source of pathogens, transmission dynamics are inherently spatial processes" (Ostfeld, Glass, & Keesing, 2005, p. 328). While closely related to various medical, public heath, and geographic approaches, the landscape epidemiology approach to disease research is unique in many ways. To understand the tradition of landscape epidemiology, it is helpful to briefly review its history and that of its progenitor, medical geography.

1.1. Medical Geography

Medical Geography, or Health Geography as it is sometimes called, by its very nature has always been a cross-disciplinary field of study. It has at various times been associated most closely with applied medicine, landscape ecology, regional geography, cartography, and spatial statistics among other fields. In its modern form it is most commonly associated with the medical discipline of epidemiology, a field of study "concerned with the distribution and determinants of health and diseases, morbidity, injuries, disability, and mortality in

populations"(Friis & Sellers, 2009, p. 6). Sometimes called nosogeography (from the Greek *nósos* for disease or sickness), medical geography has been studied in one form or another since long before the famous quantitative revolution beginning in the 1950s in geography and even before the birth of scientific medicine in the late nineteenth century (Numbers, 2000).

Dr. Jacques M. May, often considered the father of modern medical geography, recognized the long tradition of medical and geographic knowledge intertwining when he pointed out that "the idea that such an approach should be made….was understood by Hippocrates" (May, 1950). The "Father of Modern Medicine," Hippocrates of Cos was an ancient Greek physician born around 460 BC and well regarded for his many writings and for the founding of the Hippocratic School of Medicine (Grammaticos & Diamantis, 2008). Although he is best known for the Hippocratic Oath that physicians today still take – "the aim of the physician should be to do good to his patient, or, at least, to do no harm" (Hippocrates, 1849, p. 341) – he is also regarded as one of the first to recognize the connection between health and place. His treatise "On Airs, Waters, and Places" discusses the impacts "different seasons, the winds, the various kinds of water, the situation of cities, the nature of soils, and the modes of life, exercise upon the health" (Hippocrates, 1849, p. 181). It seems clear that he was at least aware of the complexities of human health and its relationship to physical environmental factors (Meade & Emch, 2010). He passed this understanding, along with much of his medical expertise, down to his students. Although similar ideas are also to be found in the teachings of other ancient scholars such as Plato, very little was added to this basic premise for nearly two millennia.

It has been said, "the earliest physicians knew little of the cause of diseases beyond the fact that certain ones seemed to be found in certain localities only" (James & Jones, 1954, p.

453). Tropical diseases were so classed because they occurred primarily in the tropics. Some diseases, such as respiratory illness, appeared to improve or even be cured by patients simply moving to higher elevations. There are many more examples, but the pattern was consistent. Although the root causes of disease were illusive, the correlation between the environment and illness was easily observed, if not always readily understood. Indeed, much of classical medicine was devoted to primarily geographic inquiry, identifying where diseases commonly occurred and among whom.

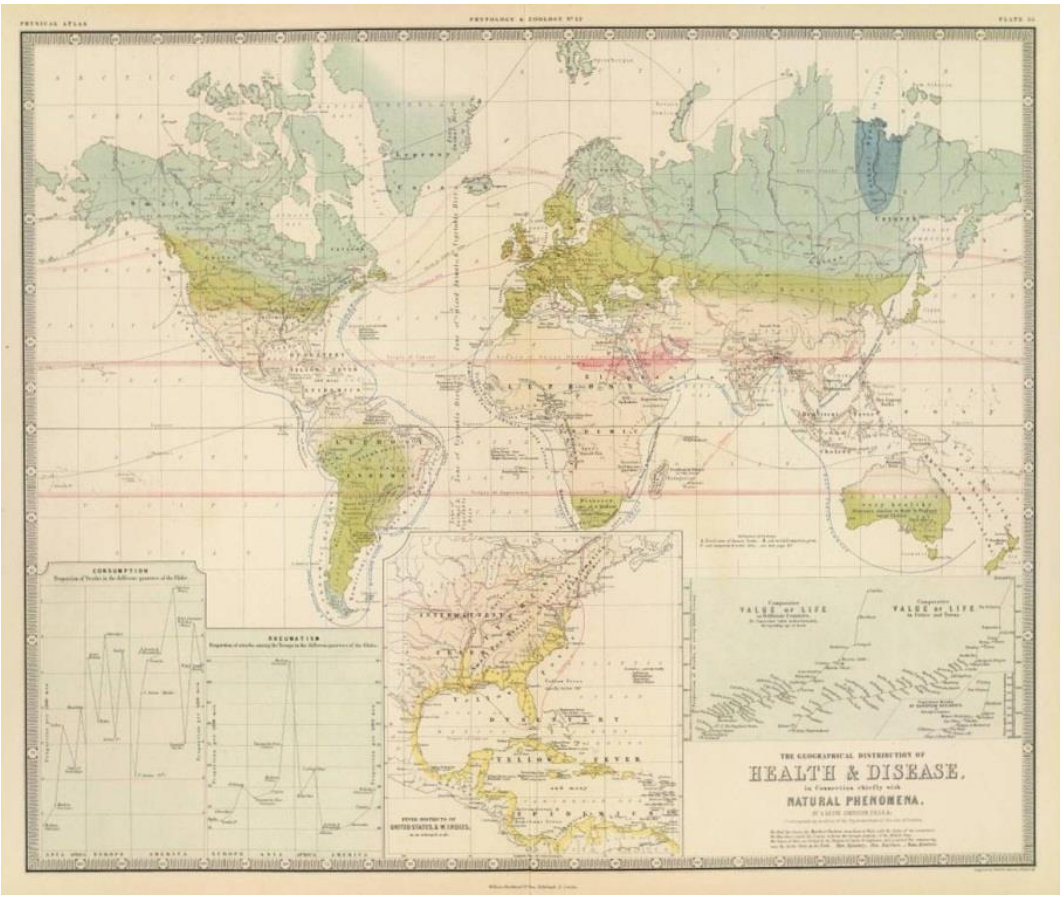

Figure 1 – Alexander Johnston's "Geographical Distribution of Health and Disease in Connection Chiefly With Natural Phenomena" (Johnston, 1856).

Before the advent of the 'germ theory of disease,' there was little in the way of scientific explanation for the cause of disease outside of observable environmental conditions. "Without a

'germ' or 'virus' to assign any direct cause to, the epidemiologists were left with the need to produce disease maps, in order to relate these deaths and cases to every natural event and condition out there – be these wind, sun, weather, topography, or population and transportation features" (Altonen, 2002). The earliest explicit disease maps focused on endemic disease boundaries which tended to remain fairly stationary over long periods of history. The famous "Health and Disease" world map of Alexander Johnston is probably the best surviving example (see Figure 1). It is essentially an isotherm map with particular diseases assigned to climatic bands, with local variations from "his review of endemic and epidemic prone regions around the world based on exploration, travel and migration history" (Altonen, 2002).

Despite the persistent lack of verifiable explanation of why such region-disease relationships existed, geographical analysis was considered an established part of medical research "until the Pastorian discoveries turned attention to the study of pathogenic organisms" (James & Jones, 1954, p. 453) and away from spatial relationships. "With the rise of bacteriology and the germ theory of disease in the late nineteenth century, medical geography went into decline….the new laboratory medicine of Claude Bernard and Louis Pasteur did indeed strip medical geography of the cachet it once enjoyed" (Numbers, 2000, p. 219). By the early part of the 1900s, medical researchers had largely left geography behind, and geographers had failed to keep pace with advances in medicine. In the 1922 text "Principles of Human Geography" the discussion of health seems more closely tied to Hippocrates than Pasteur: "The geographical distribution of health and energy depends upon climate and weather more than on any other single factor" (Huntington & Cushing, 1922, p. 248).

1.1.1. LANDSCAPE EPIDEMIOLOGY AND THE QUANTITATIVE REVOLUTION

Evgeny Nikanorovich Pavlovsky was a Russian parasitologist who recognized the same connection between disease and place observed since Hippocrates, but he may have been the first to accurately describe *why* such a relationship should exist. He correctly described that the pathogenic organisms responsible for disease were themselves profoundly sensitive to environmental conditions. He also recognized the importance of disease vectors such as ticks and mosquitoes, and that they also were subject to environmental variables that could be observed on the landscape. He formalized his observations into a new scientific field which he called "landscape epidemiology" (Pavlovsky, 1965). In essence "the theory behind landscape epidemiology is that by knowing the vegetation and geological conditions necessary for the maintenance of specific pathogens in nature, one can use the landscape to identify the spatial and temporal distribution of disease risk" (NASA, 2001).

Around this same time the field of geography was undergoing "a radical transformation of spirit and purpose" otherwise known as the "Quantitative Revolution" (Burton, 1963). The American Geographical Society started developing an "atlas of disease" in 1944 (American Geographical Society, 1944), the Communicable Disease Center (later to be renamed the Centers for Disease Control and Prevention, but consistently referred to as the CDC) was organized in 1946 with the primary purpose of locating and killing malarial mosquitoes (CDC, 2010), and two years after that the World Health Organization was founded (World Health Organization, 2012).

In the midst of this revival and transformation, Dr. Jacques M. May emerged as an innovator and a leader in the newly reborn field of quantitative medical geography. He helped redefine the field, literally, when as a member of the International Geographical Union's Commission on Medical Geography he defined it as "the study of the distribution of manifested and potential diseases over the earth's surface and of factors which contribute to disease

(pathogens) followed by the study of the correlations which may exist between these and the environmental factors (geogens)" (T. Brown & Moon, 2004, p. 751). May is sometimes considered the father of modern medical geography, not because he created the field, but because he integrated the field. He understood the power of applying quantitative analysis to spatial phenomena, especially when backed by a thorough understanding of the underlying pathogen ecology of the diseases being examined. Under his leadership "this vision of a 'new' medical geography was realized, because May was able to carefully interweave the two 'sciences' of medicine and geography within his disease ecology perspective" (T. Brown & Moon, 2004, p. 759).

Today medical geography is a cross-disciplinary approach to studying health and well-being, disease, illness and other spatially distributed health phenomena (Association of American Geographers, 2011). The Internet has allowed for the collection and mass-dissemination of medical and health related data (often connected to a specific geographic location) on a scale never before possible, as exemplified by the WHO's Global Health Atlas, the CDC's ArboNET and others (CDC, 2012b; World Health Organization, 2007). Remote sensing has emerged as a "fundamental geospatial analysis tool" (Quattrochi, Walsh, Jensen, & Ridd, 2004, p. 377) that provides vast amounts of data of both the physical and human landscape, much of which can be used in landscape epidemiological studies such as this one (Hay, 2000). The development of geographic information systems (GIS) has allowed for greater integration than ever before of both data and analysis techniques, including advanced spatial statistics applicable to health-related research (Abler, 1987; Goodchild, 1992), and is especially well suited to "establishing relationships between disease rates and exposures to environmental factors" (Rushton, 2003, p.

51). The legacies of May, Pavlovsky, and even Hippocrates are apparent in the continuing traditions and emerging practices of this rapidly advancing field.

1.2. Machine Learning

Machine learning is a branch of artificial intelligence concerned with computer systems capable of learning, or using logical inductive inference to improve performance (Quinlan, 1986). In the simplest task-oriented or "engineering approach" to machine learning, the system is trained on a set of data and creates algorithms to classify or categorize the data, then uses those algorithms to categorize new data based on what it "learned" from the training data. "The inductive inference machine [machine learning program] takes categories that have been useful in the past and, by means of a small set of transformations, derives new categories that have reasonable likelihood of being useful in the future" (Solomonoff, 1956, p. 1). This process is often cyclical, resulting in the system "learning" and improving its accuracy over time. In geospatial studies, machine learning is sometimes used in place of simple statistical techniques like linear regression in an attempt to better model complex relationships with multiple interacting variables, such as the relationship between disease the environment.

One very common machine learning technique involves the use of hierarchical decision trees to discriminate among classes of objects (Carbonell, Michalski, & Mitchell, 1983). A binary partitioning algorithm selects the variables by which to split the data into categories at each level of the hierarchy, and the resulting tree is used to classify each object in the dataset. Such trees can be thought of as having object attributes at the nodes, alternative values of these attributes along the edges, and leaves corresponding to sets or classes of objects with matching attributes. One could use a decision tree to manually decide at each node which category or group of categories a particular data object is most like, and then traverse, or move through the

tree making decisions based on the object's attribute values until a final classification is reached (Quinlan, 1993). Figure 2, adapted from Quinlan (1986), is an example of a simple decision tree with only two possible classes, P and N, corresponding to Positive Instances and Negative Instances. In this example, certain weather conditions are considered appropriate for some unspecified activity, class P, and by traversing the tree starting at the root (the top of the diagram), one could determine if current conditions qualified or not (Quinlan, 1986). It is fairly simple to extend this concept to include any number of classes and variables. This form of analysis is well established in remote sensing classification studies, and is sometimes referred to as CART (classification and regression tree) analysis (Congalton, 2010).

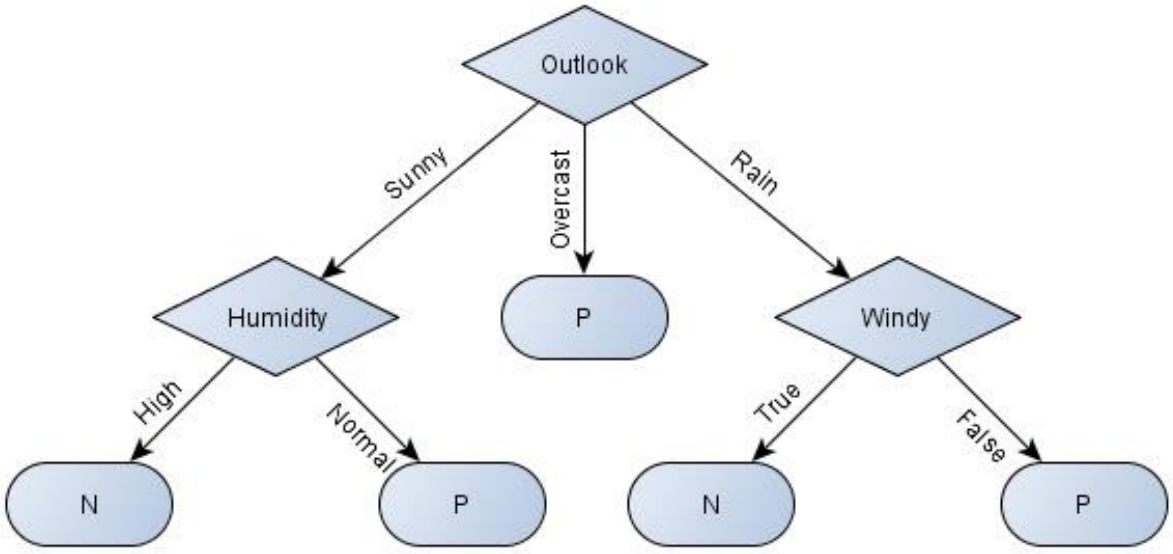

Figure 2 – A simple decision tree, adapted from (Quinlan, 1986, p. 87).

Another common machine learning approach involves the creation of production rules that take the form of if-then statements (Carbonell et al., 1983). If certain conditions are met in the data, then specific action can be taken on that data, perhaps placing it in a class or applying a particular algorithm to create quantitative output (Jensen, 2005). Multiple rules can be applied

to the same dataset to try and fit the model to the data more closely than possible with more basic statistical techniques. Cubist is an inductive machine learning program created by RuleQuest that develops decision trees for data mining purposes (RuleQuest Research, 2012). It can also convert those trees into production rules consisting of if-then statements, which are much easier to understand and interpret (J. R. Jensen, Hodgson, Garcia-Quijano, Im, & Tullis, 2009). In this study Cubist was used to model the impact of complex environmental variables on West Nile virus disease incidence.

1.3. West Nile virus

West Nile virus (WNV) was first identified in a patient from the West Nile district of northern Uganda in 1937 (Campbell, Marfin, Lanciotti, & Gubler, 2002; C. G. Hayes, 2001). The disease has been endemic in various parts of Africa, Asia, Europe and Australia since that time, but only recently made the oceanic leap to the New World. The first known appearance of the disease in the Western hemisphere was in New York City, NY, USA in 1999 (Nash et al., 2001; Petersen & Roehrig, 2001). Since then it has spread across the country and has resulted in "the largest epidemics of neuroinvasive WNV disease ever reported" (E. B. Hayes et al., 2005, p. 1167). It is now widely considered "the dominant vector-borne disease in this continent" (Kilpatrick, Kramer, Jones, Marra, & Daszak, 2006, p. 0606).

The chief premise of the field of epidemiology is that disease is not a random occurrence, but occurs "in patterns that reflect the operation of underlying factors" (Friis & Sellers, 2009, p. 142). Epidemiological studies can be divided into two broad categories, descriptive and analytical, the former generally preceding the latter. Analytical epidemiology studies are more concerned with the etiology, or causes, of disease, and how to better predict and/or manage disease occurrence. Understanding the descriptive epidemiology of a disease is an important

prerequisite to ensure sound inferences and analytical techniques are employed. When studying a particular disease, the first question is often "what is it?" but is quickly followed by the "who, where, when, and why/how" of the disease. In this section I will briefly outline some of the descriptive epidemiological characteristics of WNV, starting with the Natural History of the disease (the what), followed by the categories of Person, Place, and Time (who, where, and when) and conclude with the Etiology and Transmission (the why and how). Note: these will be reported as they apply to WNV in the US only, and may differ in some ways from characterizations of WNV in the Old World.

1.3.1. NATURAL HISTORY

West Nile virus is a flavivirus, related to Saint Louis, Japanese, Kunjin, and Murray Valley encephalitis viruses (CDC, 2003). Like all viruses, WNV is an obligate parasite that depends on the cells it infects for replication (Campbell et al., 2002; Oldstone, 1998). The strain introduced into the Western hemisphere via New York in 1999, identified as NY99, is closely related to the lineage I strain found in Israel in 1998, both notable for their increased pathogenicity among birds (C. G. Hayes, 2001; Lanciotti, 1999). The virus has "subsequently undergone subtle genetic alteration" but remains a highly virulent threat to both avian and human hosts (W. Reisen & Brault, 2007, p. 642).

For humans there are two broad categories of disease that can result from WNV infection (see Figure 3). Neuroinvasive WNV, sometimes called "severe" WNV disease or West Nile virus neuroinvasive disease (WNND), is a potentially life-threatening class of diseases including West Nile meningitis, West Nile encephalitis, and acute flaccid paralysis (CDC, 2012b). Common symptoms include fever, movement disorders, tremors, myoclonus, and Parkinsonism, with fatigue, headache, and myalgias often persisting for several months. Some neuroinvasive

patients “have good long-term outcome, although an irreversible poliomyelitis-like syndrome may result” (Sejvar et al., 2003, p. 511).

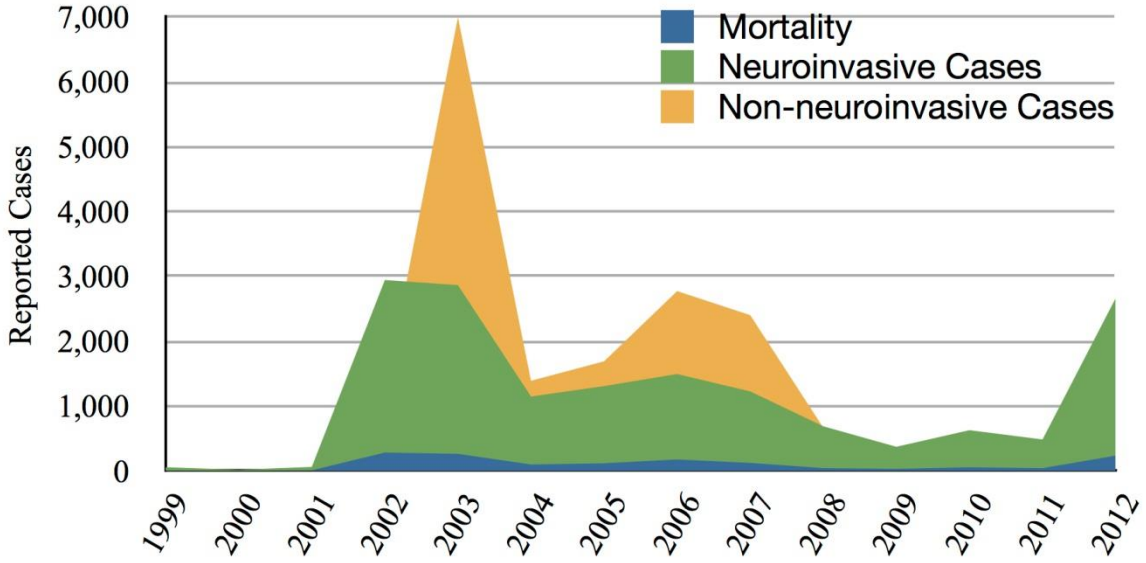

Figure 3 – Reported human cases of WNV (Neuroinvasive and Non-neuroinvasive) and Deaths in the US from 1999-2012, as reported to the CDC (CDC, 2012b).

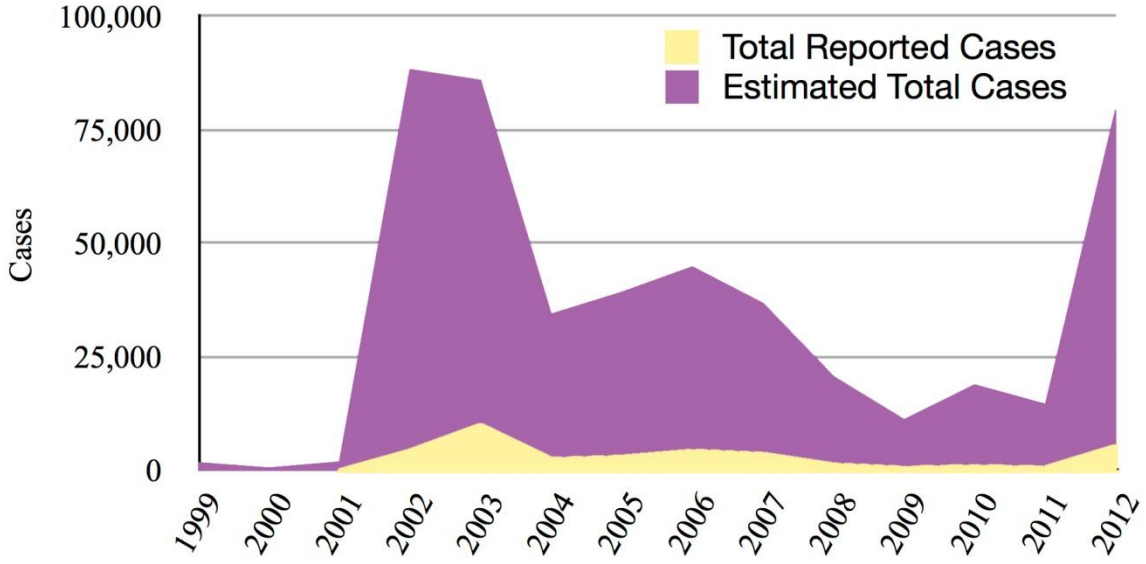

Figure 4 – Total reported human cases of WNV in the US from 1999-2012, compared to estimated total cases derived from serosurvey results by Mostashari et al. (2001) and others.

It is estimated that only about 1 in 150 infected persons will develop a severe form of the disease (CDC, 2012b). Another 20% or so develop relatively mild symptoms generally termed West Nile Fever, or non-neuroinvasive WNV, and the remaining roughly 80% are completely asymptomatic (Mostashari et al., 2001). Non-neuroinvasive symptoms are often so mild that many cases are likely misdiagnosed and unreported every year (Sejvar et al., 2003), however there is evidence that West Nile fever may in fact be more severe than generally acknowledged with around 30% requiring hospitalization and nearly 80% missing work or school due to the illness with a median absence of 10 days (Watson et al., 2004). It is estimated that over 1 million Americans have likely been infected (W. Reisen & Brault, 2007). Serosurvey results from Mostashari et al. (2001) and others were used to calculate estimates of total WNV infections (see Figure 4), demonstrating the severity of presumed underreporting of WNV. The case-fatality rate (see Figure 5) for non-neuroinvasive cases is below 1%, and for neuroinvasive cases it ranges from 3% to 15% (CDC, 2012b).

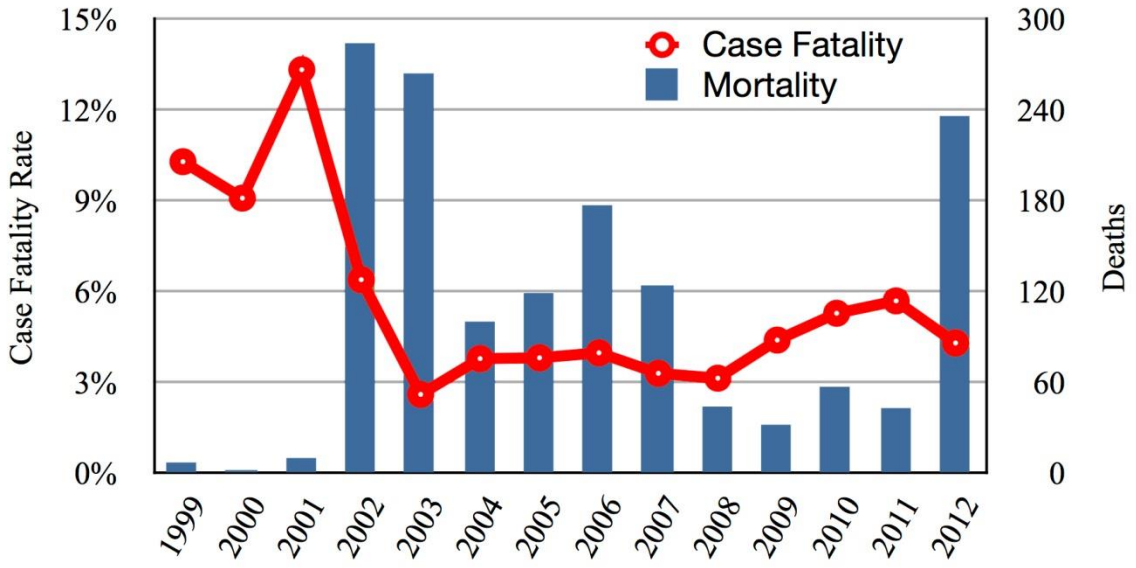

Figure 5 – Case Fatality Rate (red line) plotted over total deaths from WNV (blue bars) in the US from 1999-2012, as reported to the CDC (CDC, 2012b).

1.3.1.1. Case Definition

While clinical symptoms and patient history may indicate WNV infection, diagnosis requires laboratory testing (CDC, 2012b). Initial testing includes an enzyme-linked immunosorbent assay (ELISA) for IgM antibodies (CDC, 2012a). There are now at least four FDA-cleared WNV serological kits available commercially for presumptive diagnosis used widely by state public health and private commercial laboratories, however these tests can produce false positives due to cross-reactive antibodies from similar viruses, so they require confirmation using plaque reduction neutralization tests (PRNT) which were established by the CDC as the gold standard (Janusz, Lehman, Panella, Fischer, & Staples, 2011).

For a case to be classified as confirmed neuroinvasive, it must meet the clinical criteria of fever above 100.4° F, acute signs of central or peripheral neurological dysfunction (meningitis, encephalitis, acute flaccid paralysis, or other signs) documented by a physician, and the absence of a more likely clinical explanation. It must also meet laboratory criteria of virus isolation from tissue, blood, or body fluid; or the ELISA and PRNT tests described above. If the virus-specific IgM antibodies are confirmed via ELISA but no other testing is performed the case is classified as probable. Similarly for a case to be classified as confirmed non-neuroinvasive, it must meet clinical criteria of fever above 100.4° F, absence of neuroinvasive disease and absence of more likely clinical explanation, and must undergo the same laboratory testing as the neuroinvasive cases. Probable cases likewise receive ELISA but no confirmatory testing (CDC, 2012a). The ArboNET system (see Section 3.3.2) only records cases that have received laboratory testing, but does not distinguish between confirmed and probable cases (CDC, 2003).

1.3.2. PERSON CHARACTERISTICS

1.3.2.1. Age

Generally considered "the most important factor" (Friis & Sellers, 2009, p. 146) among personal attributes, age is perhaps less important in WNV transmission than might be expected, at least when total cases are observed together. In fact, all ages are considered equally susceptible to infection, since exposure to infected mosquitoes can occur at any age. The development of disease in response to infection, however, is related to age. Neuroinvasive cases in particular appear to be strongly associated with advancing age, especially between 60-89 years (E. B. Hayes et al., 2005; O'Leary et al., 2004). Using data from the 2002 epidemic, median age for fever cases was in the 40s, while median age of severe cases was in the 60s. Deaths resulting from all forms of WNV infection during that same year had a median age in the 70s (O'Leary et al., 2004). From 1999-2007, the case fatality ratio of neuroinvasive WNV disease was around 1% for most age cohorts, but jumped to 14% for adults over age 50 (Staples, 2009).

Unsurprisingly then, although all ages are susceptible to infection, older persons appear more likely to develop a severe form of disease and are also more likely to die as a result of either form of disease (Sejvar, Lindsey, & Campbell, 2011). At least one study observed an inverse correlation between age and non-neuroinvasive WNV disease, but the findings may have been subject to volunteer bias among the participants (J. A. Brown et al., 2007).

1.3.2.2. Sex

Similar to age, sex does not appear to be linked to WNV morbidity when looking at total cases (Campbell et al., 2002). For example, again making use of the 2002 epidemic figures, male infection rates hovered around 50% for most forms of the disease (O'Leary et al., 2004). Incidence of neuroinvasive cases, however, were markedly higher among males than females, with .51 and .36 per 100,000 respectively from 1999-2007 (Staples, 2009). Mortality also

appears to be more common among men, with males making up nearly 70% of the total deaths in 2002 (O'Leary et al., 2004).

1.3.2.3. Race/Ethnicity

Data on race and ethnicity with regards to WNV is scarce and non-exhaustive, but some does exist. According to the CDC, incidence rates from 1999-2007 for neuroinvasive WNV only (total incidence was not reported by race) was approximately .37 for Caucasians, .3 for African Americans, and .09 for "other" per 100,000 (Staples, 2009). Unfortunately, race and ethnic data of disease cases was not available for this study.

1.3.2.4. Socio-Economic Status

Socio-Economic Status is an interesting and somewhat confounding factor with regards to WNV infection. For reasons not entirely understood, some studies indicate strong relationships between low socio-economic areas and high WNV incidence (Harrigan et al., 2010), while others report the strongest relationships between middle class suburban neighborhoods and high WNV rates (Rochlin, Turbow, Gomez, Ninivaggi, & Campbell, 2011). Interestingly, these differences appear to be somewhat regional in nature, although the differences themselves have not yet been studied. I will not address this in my study, but it is a potentially interesting topic for future research.

1.3.2.5. Other Person Factors

Other person variables are hard to come by for WNV. There does not appear to be any data (at least not publicly available) on the effects of marital status, religion, family size, blood type, personality traits, or occupation on either WNV morbidity or mortality. That's not to say there are not important considerations involving these factors, merely that the current data is incomplete on the subject. For example, "human behaviors, such as smoking and dog walking,

that bring humans outside the protection of screened or air-conditioned homes at night elevate the risk of infection" by increasing their risk of exposure to infected mosquitos, but such data is not collected for WNV patients, so quantitative assessments are problematic (W.K. Reisen, 2010, p. 474). Laboratory experiments have indicated that the protein CCR5 may be a protective factor against WNV, therefore those with genetic mutations that lack functional CCR5, including an estimated 1% of North American Caucasians, "may be at greater risk of fatal encephalitis from WNV infection" (W. G. Glass et al., 2005, p. 1095). There are indications that immunocompromised hosts may also have increased susceptibility (W. G. Glass et al., 2005), but according to the CDC it is currently still unknown if these individuals are indeed at increased risk (CDC, 2012b).

1.3.3. PLACE CHARACTERISTICS

Many studies have examined the place characteristics of WNV disease within the United States. Since this is also the primary focus of this research, these will be examined in more detail in the Literature Review (see Section 2). This section will look only at general place factors.

Within the United States, WNV was first identified in New York City in 1999, and quickly spread out in a manner similar to contagious diffusion, first spreading to nearby areas and then reaching farther and farther out from the initial place of introduction. In 2000, Rappole et al. warned that viremic migratory birds could spread the virus very rapidly over long distances, as had been observed in the Old World WNV movements (J. H. Rappole, Derrickson, & Hubálek, 2000). Luckily this did not appear to take place, at least not to a significant degree (J. H. Rappole et al., 2006). After just a few short years, however, the disease managed to spread across the continent, although it still has not penetrated much into the colder northern regions of Canada and Alaska, likely due to the inhospitable conditions for the necessary mosquito vectors.

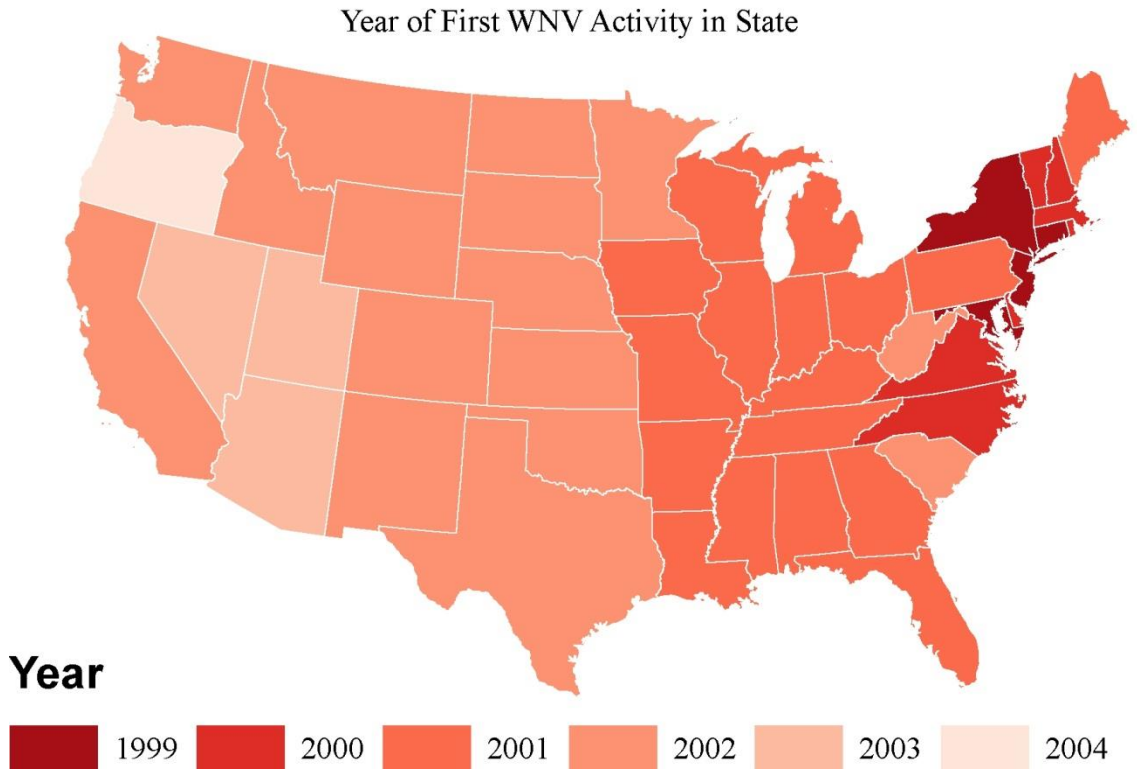

Figure 6 – Year during which WNV activity was first detected in each state, as reported to the CDC.

While the disease has spread from coast to coast and is now considered endemic (A. Townsend Peterson, Robbins, Restifo, Howell, & Nasci, 2008) there does appear to be significant spatial clustering as evidenced by extremely high spatial autocorrelation values measured using both global and local Moran's I tests (Anselin, 1995; Moran, 1950) across the continental United States (Young & Jensen, 2012). Furthermore, for reasons not entirely understood, the states and counties with the highest cumulative incidence and most pronounced incidence rates normalized by population are primarily clustered together in the northern Great Plains (Lindsey, Kuhn, Campbell, & Hayes, 2008). When analyzing clustering of incidence rates, which normalize the disease data by population, the northern Great Plains, as well as southwest Idaho, stand out as persistent hotspots (Sugumaran, Larson, & DeGroote, 2009; Young & Jensen, 2012).

1.3.4. TIME CHARACTERISTICS

West Nile virus depends on mosquito vectors, and as such WNV infection is seasonal and cyclic in nature. There was some speculation early on that without migratory bird diffusion to bring the virus back and forth from tropical regions where it can thrive year-round, the virus might die out over the winter (J. Rappole & Hubalek, 2003), but it has been shown to overwinter in *Culex* mosquitos (Nasci et al., 2001), allowing it to reappear seasonally without reintroduction via migratory birds. Although the virus remains in mosquitoes over the winter months, human cases do not normally occur during this time. Human cases tend to occur (in the US at least) from mid-summer to mid-autumn, or in other words, mosquito season. When weather conditions are right, this time can extend from as early as April to as late as December, but the majority of cases occur (meaning symptoms first manifest) between July and September (Staples, 2009).

Cx. pipiens, or the common house mosquito, a prominent WNV vector across the continent "demonstrate a late-summer shift" in feeding behavior from primarily birds to primarily humans and other mammals (Kilpatrick et al., 2006, p. 0608). This behavior, while limiting the time during which human disease transmission generally occurs, actually amplifies WNV epidemics. By feeding primarily on birds capable of carrying WNV in the early summer as opposed to "wasted" feedings on humans and other dead-end hosts, the intensity of the epidemic in mosquitoes is amplified. This in turn likely leads to a greater number of human infections after the feeding shift than would have occurred if the mosquitoes fed on humans year-round (Kilpatrick et al., 2006).

Reisen and Brault (2007) identified a "three year epidemic pattern" in North America, whereby WNV is quietly introduced into a new area with "low avian depopulation rates and few human cases," successfully overwinters, and then undergoes "explosive epidemic amplification"

in its second season in the area, followed by subsidence possibly associated with herd immunity. Despite the tendency to subside following the epidemic year, they also noted that avian herd immunity "appears to be transient owing to antibody decay, rapid population turnover and perhaps enhanced brooding success due to reduced population density and increased resources," resulting in the possibility of renewed viral amplification and further epidemics (W. Reisen & Brault, 2007, p. 643).

1.3.5. ETIOLOGY AND TRANSMISSION

West Nile virus is a zoonotic disease of birds, which are the primary and amplifying hosts, and is considered an arbovirus (arthropod-borne virus, transmitted by blood-sucking insects (Mosby, 2009)) transmitted primarily by mosquito vectors, although bird-to-bird transmission has been demonstrating in laboratory settings (McLean et al., 2001). While there are over 160 bird species and at least 36 mosquito species involved in the viral transmission cycle, the most common culprits include corvids (crows and jays) and Culex mosquitoes (CDC, 2003). The CDC has recommended examination of dead American Crows in particular as an effective surveillance strategy (CDC, 2003; LaDeau, Calder, Doran, & Marra, 2010).

Interestingly, "American Crows were found to be significantly underrepresented in the blood meals of the ornithophilic [bird-feeding] mosquitoes" compared to abundance measurements, possibly suggesting bird-to-bird transmission occurs in the wild (Apperson et al., 2004, p. 80). While the virus has also been found in horses, reindeer, sheep, deer, bears, and feral swine (Gibbs et al., 2006), the author is not aware of any confirmed cases of transmission to humans from any host or vector other than mosquitoes. Infected mammals, including humans, do not develop sufficient viral levels in their blood to infect biting mosquitoes, making them so-called "dead-end" or incidental infections and not a part of the normal transmission cycle (E. B.

Hayes et al., 2005; W.K. Reisen, 2010). With the rare exceptions of mother-to-child transmission and contaminated blood transfusions or organ transplantations (all donated blood in the US is screened for WNV RNA (CDC, 2012b)), human-to-human transmission does not appear to take place, making quarantines unnecessary for infected patients (Ciota et al., 2008; Kumar et al., 2004).

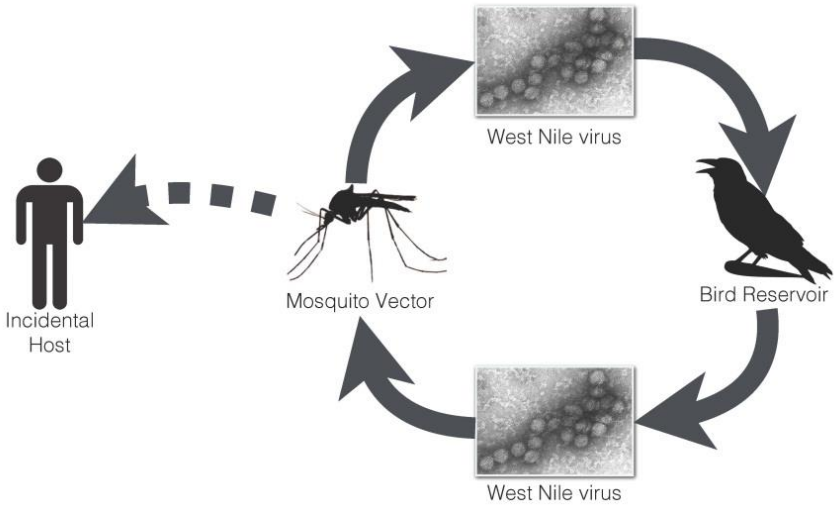

Figure 7 – WNV Transmission Cycle, adapted from (CDC, 2012b).

1.4. Statement of the Problem

Many existing studies seeking to define areas of WNV risk required data that is both expensive and challenging to obtain such as dead bird surveillance or mosquito collection and testing, usually requiring lots of man-hours in the field (Mostashari, Kulldorff, Hartman, Miller, & Kulasekera, 2003). While such data is very useful, it is not collected in a standardized or uniform manner and sampling is extremely sparse at best, with large portions of the country not collecting these data at all. Further, as Allen and Wong (2006) noted, "some counties…are contemplating the idea of not gathering and testing dead birds in the future, partly for financial reasons, and partly because previous WNV-cases already confirmed the presence of the virus in

the region" (Allen & Wong, 2006, p. 263). In fact, even the CDC reports on its WNV Q&A page, in response to the question about why some areas stop collecting dead birds – "Some states and jurisdictions are no longer collecting dead birds because they have sufficiently established that the virus is in an area, and additional testing will not reveal any more information. Shifting resources away from testing of dead birds allows those resources to be devoted elsewhere in surveillance and control" (CDC, 2012b). While this is true, it ignores the usefulness of this data to studies such as this one. Since such data has very poor geographic coverage and is in increasingly short supply, other methods must be found that will work in the absence of extensive field data.

In addition, some researchers feel that the underlying geography of the United States is far too diverse to permit the creation of national-scale models of WNV risk, relying instead on customized regional or state-level models (DeGroote & Sugumaran, 2012; Winters et al., 2008). Others feel their results indicate some level of generalizability and that a country-wide model might be feasible (A. Townsend Peterson et al., 2008; Shaman, 2009). This study seeks to contribute to this discussion.

Finally, there are two common problems with spatial studies of disease, the small numbers problem and the modifiable areal unit problem (MAUP). The small numbers problem "is probably the most pervasive problem in disease mapping" occurring when the number of cases of disease in an area is small, or when the population of the area is small (Pringle, 1995, p. 343). This can be due small areas or rare disease or both, but when such small numbers are used to calculate rates the results can be very misleading. Relatively minor changes in the data can appear very significant due to the small numbers involved in the calculations, often exaggerating or otherwise confounding results. The small numbers problem is apparent in the WNV incidence

data, both due to the variable sizes of counties and due to the fact that it is still a relatively new disease and is not fully endemic across the contiguous United States. Since only about 20% of infections cause sickness, the data we have are only a small sample of total infections. Furthermore, the fact that many WNV cases (those that do get ill) are likely misdiagnosed, the actual case count is probably severely underreported (see Figure 4), which only amplifies the small number problem. The MAUP is a systemic geographical research problem affecting all studies that use arbitrary and modifiable zones. When data are examined at different geographical scales or levels of aggregation, the results can change dramatically, casting doubt on the validity of the spatial statistics or other models used (Gehlke & Biehl, 1934).

1.4.1. Research Questions and Hypotheses

The primary research question under investigation in this study is if WNV risk can be quantified and predicted with acceptable accuracy across the continental United States using remotely sensed environmental variables. I hypothesize that the environmental variables of NDVI, elevation, land cover, precipitation, and temperature are spatially related to WNV incidence strongly enough to allow for predictive risk modeling at the national level. A second question looks at a finding of Landesman et al. (2007) that prior-year precipitation measurements were stronger predictors of disease incidence than concurrent year precipitation (Landesman et al., 2007). A third research question is if there are clear regional variations that impact model performance spatially, and if so, what might be causing them. A fourth and final question is if my method of mitigating the small numbers problem, described in section 3.6.2, is effective or not based on its impact on model performance.

2. Literature Review

With regards to landscape epidemiology, GIS, remote sensing, and spatial risk models, Eisen and Eisen (2011) stated clearly in their article titled "Using Geographic Information Systems and Decision Support Systems for the Prediction, Prevention, and Control of Vector-Borne Diseases" that "it should be noted that the literature is too extensive for exhaustive reviews of all related published papers; therefore, [I] present only selected, representative publications as examples" (L. Eisen & Eisen, 2011, p. 42).

Researchers looking at Lyme disease, the most frequent vector-borne disease in the US prior to WNV's introduction, determined that "there is a need to extend risk analysis to larger, less well defined areas while reducing the expenditure of time and resources" and concluded that geographic information systems could provide the needed framework to "rapidly identify risk factors of zoonotic disease over large areas" (G. E. Glass et al., 1995, p. 944). One persistent challenge with landscape epidemiological studies, and indeed epidemiological studies in general, is the question of what potential risk factor variables to investigate and which to ignore. When dealing with environmental variables, there is almost no end to the list of factors that could prove to be significantly correlated with disease.

2.1. Habitat Models

Many existing studies have endeavored to simplify this list of potential variables by focusing on the WNV transmission cycle, namely the distribution of either the bird reservoir hosts or the mosquito vectors. Cooke et al. (2006) modeled mosquito habitats using environmental variables to predict WNV risk. Their models indicated that 67% of human cases occurred in areas predicted as high-risk. They also noted that "dead bird occurrences are correlated with human WNV risk and can facilitate the assessment of environmental variables

that contribute to that risk" (Cooke, Grala, & Wallis, 2006, p. 36). Trawinski and Mackay (2008) found that WNV vector mosquito abundance is spatially autocorrelated, indicating it can be predicted for unsampled locations (Trawinski & Mackay, 2008). Beck et al. (1994) used remote sensing in a landscape epidemiological study of malarial mosquitoes in Mexico, and found landscape elements could predict vector abundance with an overall accuracy of 90% (Beck et al., 1994).

2.1.1. Ecologic Niche Models

Peterson et al. (2004) recognized existing vector habitat maps are "at best incomplete, if not actually misleading" and instead used ecological niche modeling using rule-set prediction algorithms to more accurately predict Leishmaniasis vector presence (A. Townsend Peterson, Pereira, & De Camargo Neves, 2004, p. 10). Ecologic niche modeling (ENM) relates "known occurrences of species across landscapes" to environmental variables across the same landscape to identify the ecologic distribution of the species, which can then be used to predict potentially suitable habitats at locations where species occurrence is not known (A. T. Peterson, 2006, p. 1822).

Ecologic niche modeling uses rule-based machine learning algorithms to characterize "general environmental regimes under which species or phenomena may occur" but has seldom been applied to disease transmission studies (A. T. Peterson, 2006, p. 1823). While my data is not of fine-enough spatial resolution to be considered ecological niche modeling, the same basic techniques were applied.

2.2. Human-Environmental Models

Others researchers ignore the transmission cycle entirely and focus instead on descriptive epidemiological factors such as socioeconomic status or land use without attempting to model

the habitats of birds or mosquitoes. For example, outbreaks in the Chicago and Detroit areas showed positive correlation between infection and socioeconomic factors including income and age of housing (M. O. Ruiz, Tedesco, McTighe, Austin, & Kitron, 2004; M. Ruiz, Walker, Foster, Haramis, & Kitron, 2007). Studies in north-central US have found middle class suburban neighborhoods tend to be at highest risk, while similar studies for the southern and western US indicate highest risk is in low income areas (Rochlin et al., 2011). Brown et al. (2008) found urban land covers in the Northeastern US to have the highest odds of above-median disease incidence (H. E. Brown, Childs, Diuk-Wasser, & Fish, 2008). Harrigan et al. (2010) examined a disease hotspot in Orange County, CA and found both mosquito and human WNV prevalence were best explained (their models explained as much as 95% of the variation) using economic variables and "anthropogenic characteristics of the environment" including neglected swimming pool density (Harrigan et al., 2010, p. 1).

Researchers working in Florida found positive correlations between hydrology models and WNV infection, which they believed could be generalized to the national level (Shaman et al., 2009). Human land-use and WNV infection rates among American crows were strongly correlated in the northeastern United States (LaDeau et al, 2010). Gates and Boston (2009) identified a very strong relationship between irrigation and both human and equine WNV occurrence at the county level over a three-year period, presumably due to irrigation increasing available mosquito habitat and therefore increasing risk of disease transmission. They found as irrigation rose as a percentage of total land area by only 0.1% that the WNV incidence rate would increase by 50% for humans and 63% for horses (Gates & Boston, 2009). Liu et al. (2008) similarly found WNV outbreaks in Indianapolis were influenced by percentages of agriculture and water (Liu, Weng, & Gaines, 2008). Bowden et al. (2011) analyzed human

WNV incidence and land cover across the US and identified regional variations. In Northeastern regions urban land covers were positively associated with WNV disease while in the Western US agricultural land covers had the strongest positive association. They theorized the regional differences they observed can be explained by behavioral differences between the prominent mosquito vectors for the respective regions (Bowden, Magori, & Drake, 2011).

2.3. Climate Models

Climatological data is commonly used due to its close relationship with reservoir and vector habitats. One national scale study found a potential correlation between decreased WNV infection rates and below average summer temperatures (Reisen, 2009). Renneboog et al. (2009) also found that as low temperatures increased, mosquito abundance increased (Renneboog et al., 2009). Another study hypothesized that increasing water flow in catch basins would positively impact breeding conditions for *Culex* mosquitos, and alternately drought years would negatively impact mosquito populations. These impacts would in turn presumably decrease infection rates (Ghosh, 2011). While these conclusions make sense with what is known of optimal mosquito habitats, one study in Florida found that droughts can actually amplify the disease in a manner similar to that observed for the related St. Louis encephalitis virus (Shaman, Day, & Stieglitz, 2002). The theory is that "drought brings avian hosts and vector mosquitoes into close contact" as they are forced to cluster around the less-abundant water sources which "facilitates the epizootic cycling and amplification of the arboviruses within these populations" (Shaman, Day, & Stieglitz, 2005, p. 134). This theory could potentially explain why the hot and dry year 2012 was the worst epidemic year since 2003, but the data for 2012 was not yet available from the CDC as of this writing, so the theory could not be tested in this study. Interestingly, Landesman et al. (2007) found that human WNV incidence was associated more with precipitation from the

preceding year than the concurrent year, with above-average rainfall in the eastern US and below-average rainfall in the western US both preceding outbreaks. Monthly precipitation and 3-month "seasons" of precipitation were found to be highly variable and generally not as well correlated with WNV incidence as simple annual precipitation data. They also noted that in some species of mosquitoes, "droughts can facilitate population outbreaks…in the following year" (Landesman, Allan, Langerhans, Knight, & Chase, 2007, p. 337).

Soverow et al. (2009) looked at a number of weather variables in connection with WNV incidence from 2001-2005 across 17 states and discovered warmer temperatures, elevated humidity, and heavy precipitation all increased human infection rates independently of one another (Soverow, Wellenius, Fisman, & Mittleman, 2009). Researchers looking at WNV infection in Culex mosquitoes in northeast Illinois found increased air temperature was the strongest predictor of increased infection, and that precipitation and temperature alone could explain up to 79% of the spatial variability (M. O. Ruiz et al., 2010). Reisen et al. (2006) found that the virus itself has trouble replicating within mosquitoes when temperatures are below around 14° C, and further discovered that above-average summer temperatures appear to be linked closely to the epidemic summers of 2002-2004 (William K. Reisen, Fang, & Martinez, 2006).

2.4. Remote Sensing and GIS

Remotely sensed data has long been recognized by epidemiologists, biogeographers, conservationists and others as a useful tool for estimating habitat extents of both flora and fauna, among other uses (Cline, 1970; Washino & Wood, 1994). Swatantran et al. (2012) successfully used remote sensing and machine learning methods to map migratory bird habitats (Swatantran et al., 2012). Hayes et al. (1985) demonstrated that imagery from Landsat 1 and 2 could be used

to identify mosquito larval habitats (R. O. Hayes, Maxwell, Mitchell, & Woodzick, 1985). Dambach et al. (2012) performed a similar study in rural West Africa, finding remotely sensed precipitation, temperature, and vegetation indices could be used to predict vector densities (Dambach et al., 2012). NDVI (Normalized Difference Vegetation Index) in particular, when used in combination with topographic data, has "proven to have excellent predictive ability" of WNV risk (A. Townsend Peterson et al., 2008, p. 343). NDVI is an index, originally created by Rouse et al. (1974), derived from spectral reflectance values in the near infrared and red portions of the electromagnetic spectrum, and designed to be proportional to photosynthetic activity (John R Jensen, 2007; Rouse, Haas, Schell, & Deering, 1974). It is a key component of many WNV risk models due to its close association with vegetation that can serve as habitat for both birds and mosquitos, and its association with water, necessary specifically for mosquito breeding and egg-laying (H. Brown, Duik-Wasser, Andreadis, & Fish, 2008).

Many researchers have attempted to use remotely sensed environmental datasets in predictive spatial epidemiological models (R. J. Eisen & Eisen, 2008). Perhaps one of the best examples of a successful predictive disease risk model was the Rift Valley fever risk map created by Anyamba et al. (2009) which provided 2-6 weeks of warning of an outbreak in the Horn of Africa. Using a combination of remotely sensed environmental data, they accurately prospectively predicted the spatial and temporal disease activity with enough notice to facilitate response and mitigation efforts (Anyamba et al., 2009).

GIS research also has a "long history" with studies of human health and well-being (Foody, 2006), with many proponents pointing back to John Snow's famous London Cholera epidemic map of 1850 (Snow, 1855). While Snow obviously didn't use a GIS in his analysis, the basic methods he employed of spatial thinking and pattern recognition continues today (Brody,

Rip, Vinten-Johansen, Paneth, & Rachman, 2000). With regards to vector borne diseases such as WNV, the link between the climate, the environment, and disease outbreaks is of increasing interest among researchers. So much of the relevant data is inherently spatial that a means of integrating and analyzing such data in an explicitly-spatial context is becoming imperative. Shuchman et al. (2002) suggest a GIS-based system with a remote sensing component "could significantly improve the management of vector borne disease events" by providing, among other things, "an improved prediction capability based on climate and environmental models" (Shuchman, Malinas, & Edson, 2002, p. 305).

2.5. The Small Numbers Problem and the Modifiable Areal Unit Problem (MAUP)

There are basically three methods of dealing with the small number problem. The first is to use spatial smoothing techniques that compute a location's value based on the values of that location's neighbors, reducing spatial variability (Wang, 2006). The second is to aggregate values to larger geographic areas until sufficiently high values are reached, but this approach again introduces the challenges associated with the MAUP (Wang, Guo, & McLafferty, 2012). The MAUP can be tested by comparing different levels of aggregation, but this dramatically increasing the complexity of the model and in turn diminishes the interpretability of results. Unfortunately there is no simple fix for this problem, although data normalization and the consistent use of the same areal units (e.g. counties) can help mitigate its effects (Openshaw, 1984). The final method commonly used to address the small numbers problem is to aggregate values over time (Wang, 2006). While this method is fairly straightforward and does not exacerbate the MAUP, it does require all explanatory variables to likewise be aggregated over the same temporal range, and it limits the amount of time-series comparisons that can be made from the data.

3. Methods and Materials

3.1. Study Area

As of 2012, human cases of West Nile virus have been reported from all of the lower 48 United States, while no cases have yet been reported in either Alaska or Hawaii (CDC, 2012b). This study restricts its focus to these contiguous states. For reasons of confidentiality, disease incidence data is only made available aggregated to the county level, so the county is the basic unit in this study.

While a study of human-environment interactions such as this would be better served by smaller, naturally defined regions and boundaries as opposed to political ones, the research is limited by the resolution of the available data. Furthermore, although the data and methods here employed could be used to create regional models instead of a single national model, one of the research questions being investigated was if a single national model with suitable accuracy could be created despite the obvious regional differences across the study area. Such region-specific models can and probably should be created as well for improved predictive accuracy, but for the scope of this study, a single study area was selected.

3.2. Remote Sensor Data

3.2.1. Normalized Difference Vegetation Index (NDVI)

NDVI (Normalized Difference Vegetation Index) is a measure of "greenness" and contains information on both vegetation and surface water. It was selected, as previously mentioned, for its relationship to both avian and mosquito habitat. NDVI data was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument aboard the Terra satellite (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2012). The MOD13A3 data was used, which gives monthly NDVI values at a spatial resolution of 1 × 1km

(Solano, Didan, Jacobson, & Huete, 2010). Figure 8 shows the NDVI in January (on the left) and July (on the right) of 2003, demonstrating how "greenness" changes throughout the year.

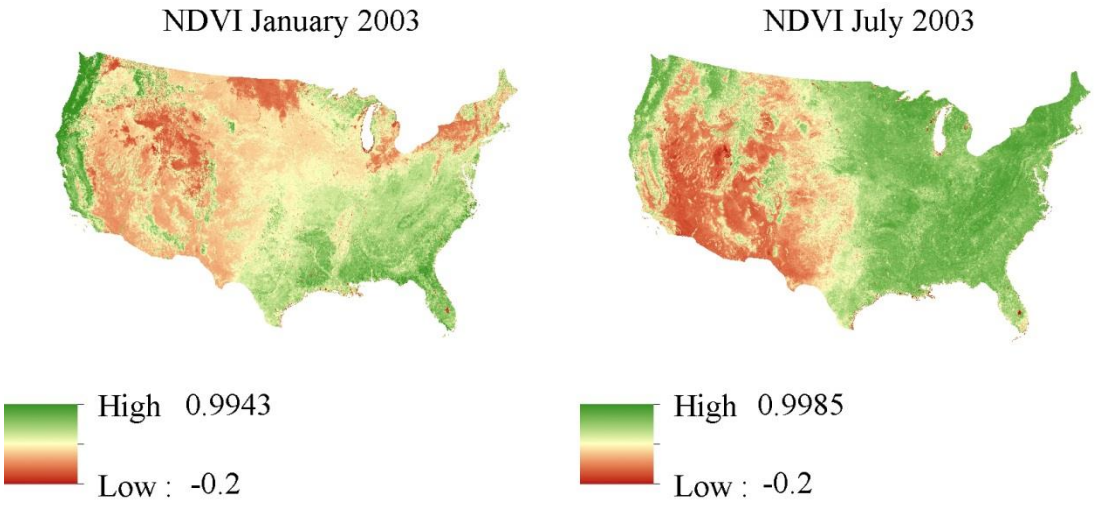

Figure 8 – Normalized Difference Vegetation Index (NDVI) values for the continental US in 2003, derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument (NASA Land Processes Distributed Active Archive Center (LP DAAC), 2012).

3.2.2. ELEVATION

Topographic data (see Figure 9) was obtained from Global Land Cover Facility SRTM (Shuttle Radar Topography Mission) global mosaic at 30 × 30 arc second resampled to 1 × 1km resolution, originally collected in February 2000 (USGS, 2006). Elevation generally changes very slowly, and data that covers the entire study area is infrequently collected, so the same SRTM data was used for the entirety of the study period.

Elevation derivatives of slope and aspect (see Figure 10) were created using the Spatial Analyst toolbox in ArcMap 10.1 (*ArcGIS Desktop*, 2012). Slope was measured in degrees. Aspect was measured in degrees, and was then reclassified into 8 categories, representing the four cardinal and four inter-cardinal compass directions, sometimes called "D8."

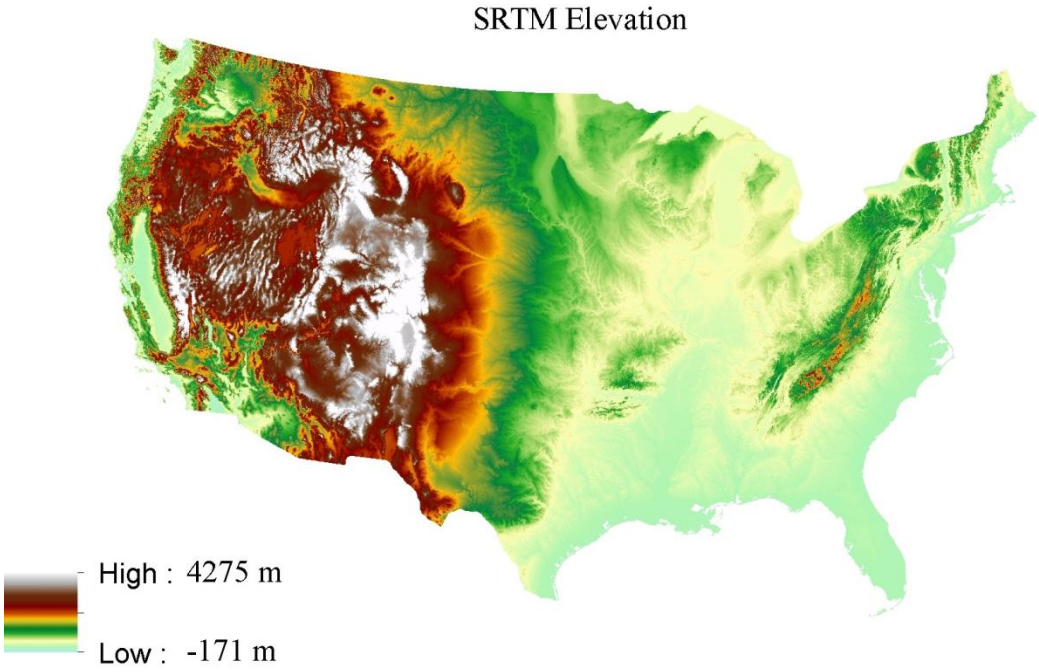

Figure 9 – Elevation data for the continental US, derived from the Shuttle Radar Topography Mission (SRTM) of 2000 (USGS, 2006).

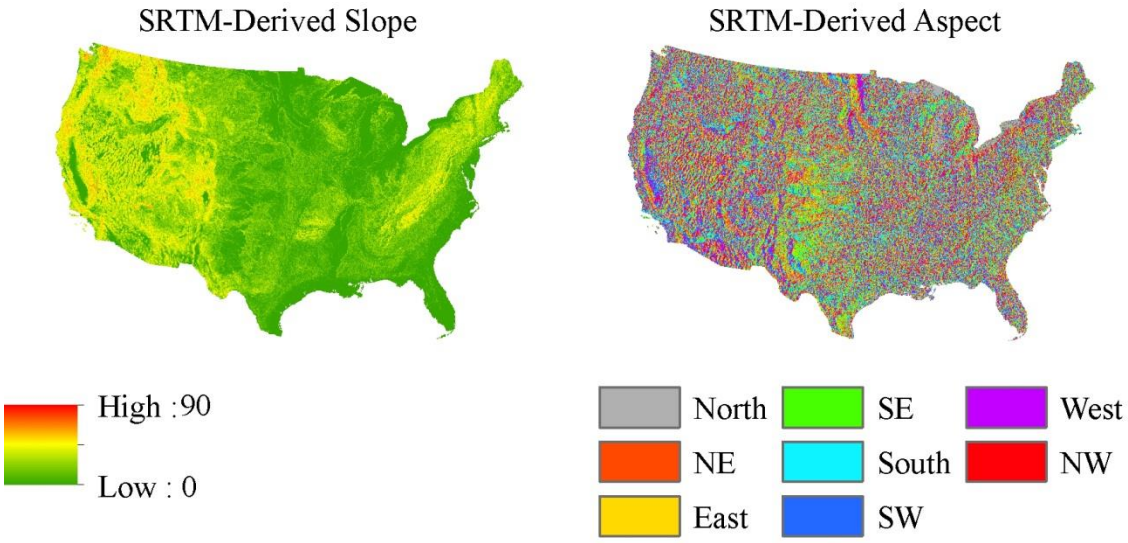

Figure 10 – Derivatives of SRTM Elevation data for the continental US (USGS, 2006). Slope measure in degrees, and Aspect reclassified into 8 principal compass directions (D8).

3.2.3. *Land Cover*

Land Cover data (see Figure 11) was obtained from the National Land Cover Database 2006 (NLCD2006), maintained by the Multi-Resolution Land Characteristics Consortium (MRLC) and the US Geological Survey (USGS). This dataset covers the conterminous United States with 16 classes of land cover (not counting some Alaska-only classes that were not present in this dataset), and was created using primarily unsupervised classification from Landsat ETM+ satellite imagery at a nominal spatial resolution of 30 × 30 meters (Fry et al., 2011).

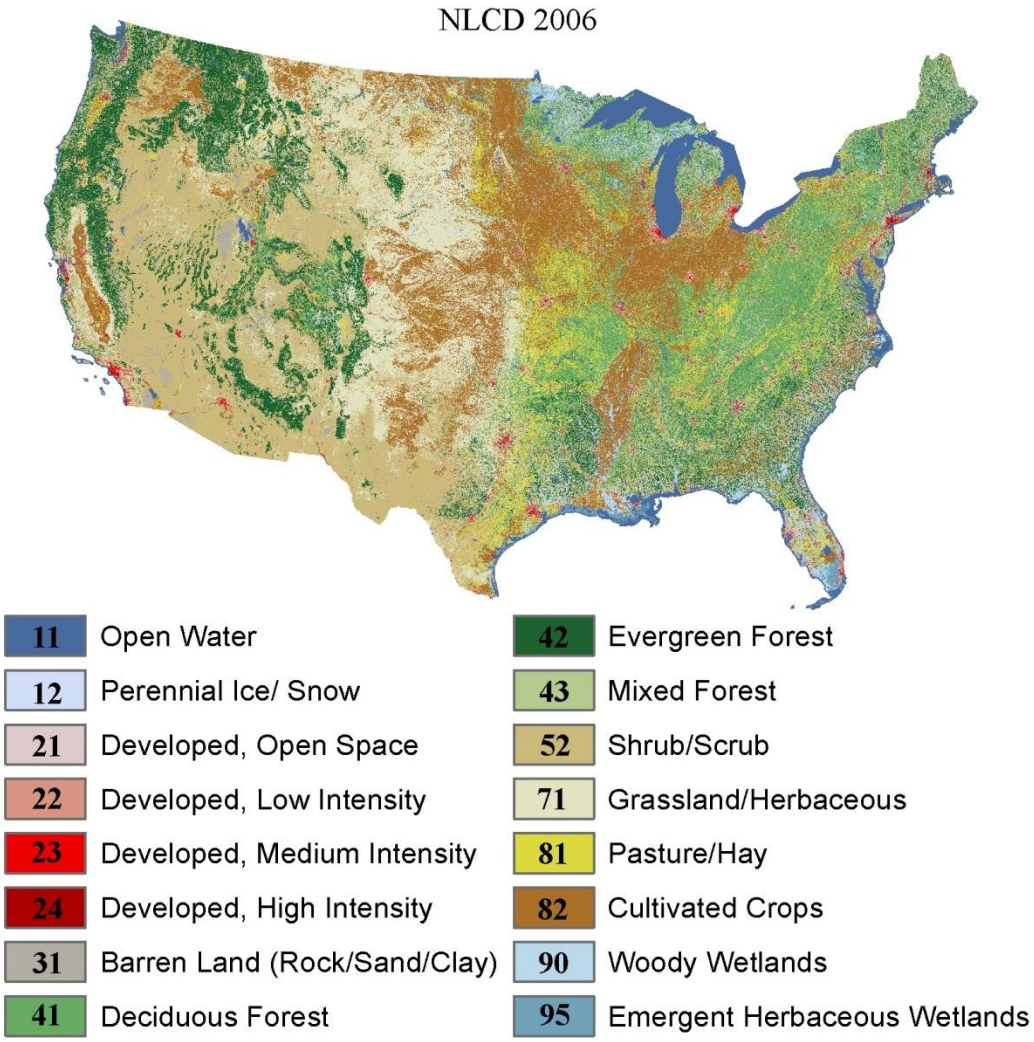

Figure 11 – National Land Cover Database of 2006 with Legend, from the MRLC and USGS (Fry et al., 2011).

3.3. *In Situ* and Ancillary Data

3.3.1. Climate Data

The in situ data used in this study included temperature and precipitation data from Oregon State University's PRISM (Parameter-elevation Regressions on Independent Slopes Model) Climate Group database which uses point data and underlying grids such as digital elevation models and 30 year climatological averages to improve interpolation accuracy, especially in mountainous terrain (Oregon State University, 2012). Precipitation data (see Figure 12) was provided in an ARC/INFO gridded ASCII format in the form of 30 year normals from 1981-2010 and annual precipitation measurements at a nominal spatial resolution of 30 × 30-arcseconds measured in millimeters.

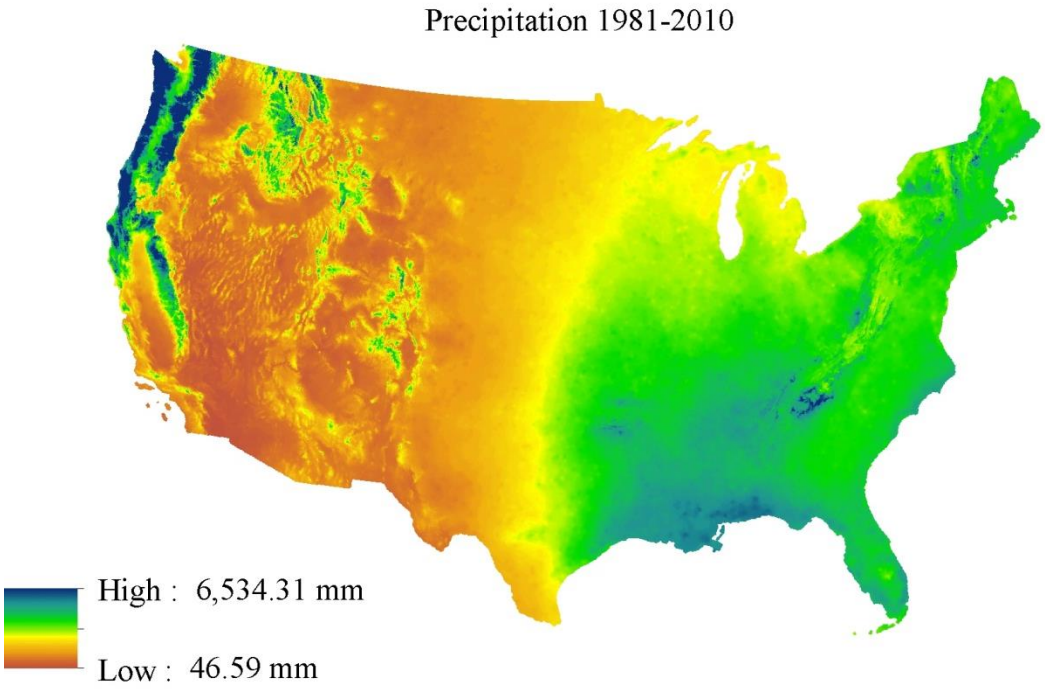

Figure 12 – 30-year precipitation normal for the continental US, from the PRISM Climate Group (Oregon State University, 2012).

Temperature data (see Figure 13) was in the form of 30 year normal, average monthly and annual maximum and minimum temperature values, also at a resolution of 30 × 30-arcseconds from 1981-2010, measured in degrees Celsius. Temperature maximums and minimums were in separate grids, so they were processed separately and combined later.

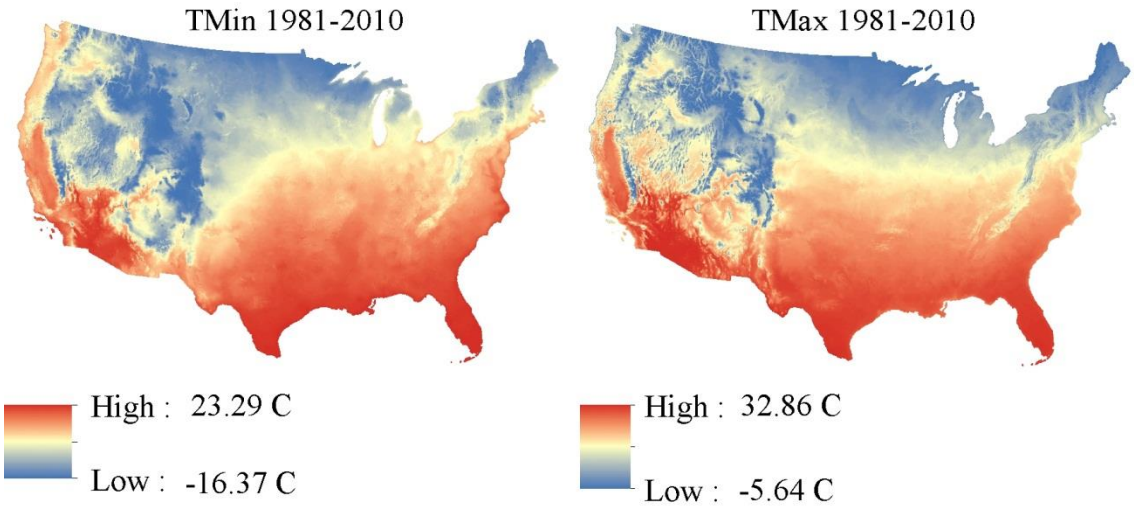

Figure 13 – 30-year temperature normal for the continental US, from the PRISM Climate Group (Oregon State University, 2012). TMin represents average minimum temperatures and TMax represents average maximums.

3.3.2. Disease Incidence Data

Data on reported WNV infections were obtained from the CDC's ArboNET system. In response to WNV's introduction and rapid spread through the US, ArboNET was developed by the Centers for Disease Control and Prevention (CDC) in 2000 to monitor and track WNV cases and other human arboviral diseases within the US (CDC, 2003). Neuroinvasive WNV is included in the list of nationally notifiable diseases maintained by the Council of State and Territorial Epidemiologists (CSTE) in consultation with the CDC (CSTE, 2012). This list indicates which diseases must be reported by law to the CDC and within what time frame the report must be made. While non-neuroinvasive cases are not included in the list of nationally

notifiable diseases, the CDC strongly encourages states to report them to ArboNET anyway (CDC, 2003). With the ArboNET system, local clinics and health workers report suspected or confirmed cases of WNV infection to their State Health Department. The State Health Department should then upload the data directly into the ArboNET system within the next normal reporting cycle (usually 7 days), which is then managed and processed by the CDC and made available to the public as aggregated data at the county level (Young & Jensen, 2012).

The first few years of WNV in the US were atypical due to the spatially restricted nature of the virus as a new emerging pathogen in a new environment. It wasn't until 2002 that WNV was detected west of the states bordering the Mississippi River, but during that year it spread all the way to California and Washington. By 2003 it had occurred in 47 of the lower 48 states, so for this reason 2003 was selected as the first year from which incidence data was used in model generation in this study. The study period chosen was the six-year period from 2003-2008.

3.3.3. CENSUS DATA

Population data were obtained from the US Census Bureau ("Census Bureau Homepage," 2013). Intercensal population estimates are created by the Federal State Cooperative Program for Population Estimates (FSCPE) and are also distributed by the US Census Bureau. The vector GIS county data for the US was created by Esri, derived from Tele Atlas data and was provided with the ArcGIS 10.1 software (*ArcGIS Desktop*, 2012).

3.4. SOFTWARE PROGRAMS AND TOOLS

3.4.1. ARCGIS DESKTOP 10.1

ArcGIS for Desktop Advanced 10.1 (formerly known as ArcInfo) is a suite of software programs created by Esri (Environmental Systems Research Institute), and was the GIS of choice for this study. The ArcMap program was used extensively during almost all stages of the

analysis, from pre-visualization and data exploration to analysis and final map creation. All maps created for this study were made with ArcMap (*ArcGIS Desktop*, 2012).

3.4.2. CUBIST 2.07 GPL EDITION

Cubist is a data mining program built by RuleQuest Research that uses machine learning to build rule-based predictive models. These models, while created using complex decision trees, are expressed as collections of rules, each with an associated multivariate linear model, to maximize interpretability. When data matches a specific rule's condition, the model associated with that rule is used to calculate a predicted value (RuleQuest Research, 2012). Cubist also supports model testing on independent subsets of the data that can then be imported back into GIS software for further visualization and testing. Cubist can also create the model based on a randomly sampled subset of the input data, and use the remainder for testing. A single-threaded Linux version of Cubist 2.07 is available under the GNU GPL (general public license) free of charge, and this was the version used in this study (*Cubist*, 2012).

RuleQuest also provides free C source code for a companion program called simply "Sample.c" meant primarily as an illustration for how Cubist models can be used in other programs (RuleQuest Research, 2012). Sample.c takes as input the model created by Cubist, and a ".cases" file containing data matching the format of the data used to create the model, and as output produces predicted values for those cases. Since Cubist only outputs predictions for test cases, Sample.c was used to generate the model's predicted values for all cases (aka counties). This small program was designed for use with models generated by Cubist 2.08 or later, so minor modifications were made to the code to allow it to recognize models created with Cubist 2.07 which was used in this study.

3.4.3. OTHER PROGRAMS AND TOOLS

Other programs and tools used in this study included Microsoft Excel 2010, part of the Microsoft Office Suite, LibreOffice v3.6.5 Calc for Linux (*LibreOffice*, 2013), Python v2.7 (*Python*, 2013), R: A language and environment for statistical computing v2.14.2 (R Development Core Team, 2012), and Notepad++ v6.2.3 (Ho, 2012). Excel and Calc were used to calculate simple ratios such as disease incidence rates and to format data tables for use in Cubist. Python 2.7, and the associated IDLE (Interactive DeveLopment Environment) were used extensively during data preprocessing to create and run geoprocessing scripts using the ArcPy site package of tools from the ArcGIS software (see Appendix A). R was used to calculate an odds ratio (OR) using the "epitools: Epidemiology Tools" package (Aragon, 2012). Notepad++ was used primarily to format input data for Cubist and to view output from Cubist. All of the above tools, with the exception of Excel, are available online free of charge.

3.5. Study Design

In traditional epidemiological studies descriptive epidemiology generally precedes analytic, first asking who, what, when and where, and then moving on to the why or the etiology of the disease. In the case of WNV the etiology of the disease itself is fairly well understood already, but due to the complex nature of human-environment interactions, the spatial distribution or "where" of the disease is not. Rather than investigate the spatial aspects of the disease in an effort to understand the underlying etiology, this study takes a reverse approach using the already known etiology of the disease to further investigate the geographical components.

This study takes both an ecological and retrospective study design approach. It is ecological merely by nature of the data, which has been aggregated to the level of counties by the CDC for the sake of privacy considerations. Retrospective study designs in epidemiology

investigate the association between a disease and past exposures to risk factors for the disease among a cohort. Since it is difficult to measure actual exposure to mosquitoes among disease cases, we instead use the environmental variables previously discussed as proxies for bird and mosquito habitats likely to be involved in WNV transmission. This landscape epidemiological approach allows us to treat the environmental variables as the "exposures" in a traditional retrospective design, although we already know the environmental variables themselves are not causative. Machine learning algorithms were used to examine and compare these exposures, and predictive models were created that could potentially be used in a prospective study design. One pseudo-prospective design was approximated using odd years to create the model and then testing it on even years during the study period.

3.6. Analytical Techniques

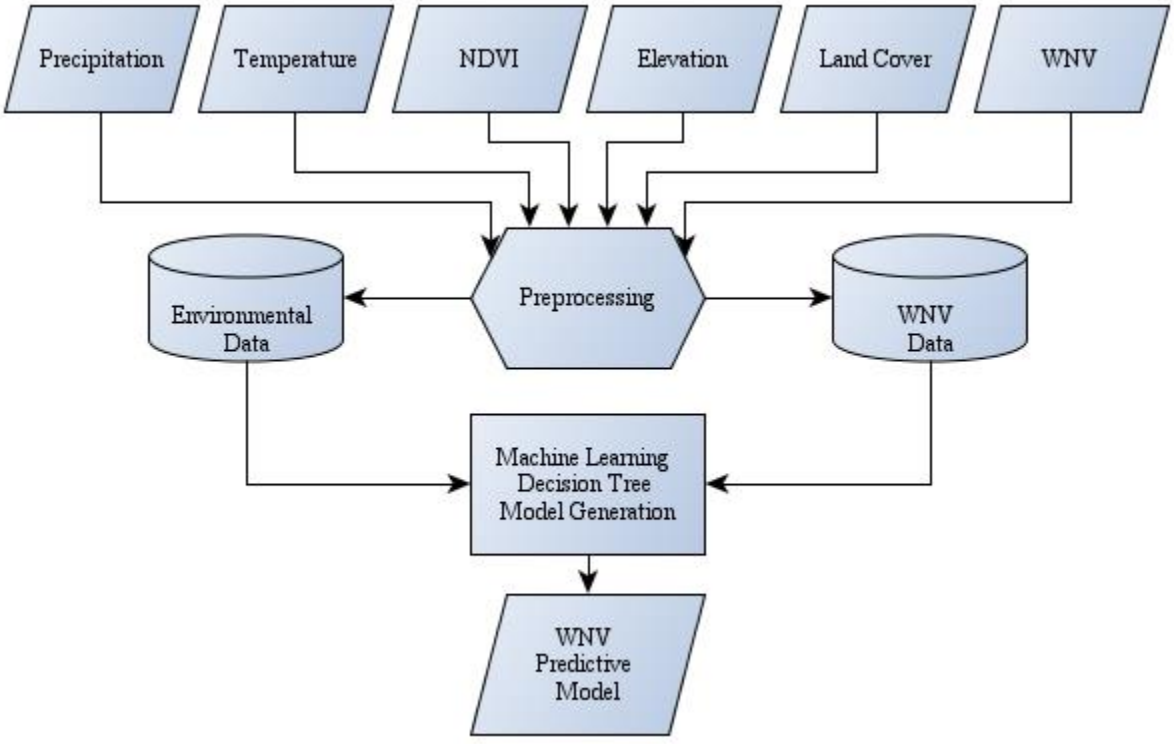

Figure 14 – A flowchart of methods and materials used in this study, simplified and condensed.

A simplified flowchart of this study can be found in Figure 14. The input datasets underwent preprocessing (described in section 3.6.1), then were fed into Cubist for model generation (see section 3.6.3), which output predictive models. The models were then evaluated according to the criteria found in section 3.6.4.

3.6.1. DATA PREPROCESSING

The West Nile virus data from the CDC was provided in Excel spreadsheets, and contained total cases by county, separated into neuroinvasive and non-neuroinvasive totals. Population data from the Census were used to change the raw case counts into incidence rates using Excel. NDVI data from the MODIS sensor came in georeferenced raster tiles that had to be mosaicked together. Each month of each year of the study period required 15 tiles to cover the entire continental United States, equaling a total of 1,080 tiles. An Esri ModelBuilder workflow was created to mosaic the first set of 15 tiles (corresponding to January 2003), which was then exported to a Python geoprocessing script using Esri's ArcPy site package. This script was then modified to allow rapid preprocessing of all the MODIS data. Similar scripts were created and used to preprocess the PRISM climate data. Each of the monthly environmental datasets – temperature, precipitation, and NDVI – were aggregated to the county level using zonal statistics in ArcMap. The zonal statistics tool was used to calculate a mean from all of the input pixels that fell within each zone, in this case counties, for each input dataset. This process was also scripted to save time (see Appendix A).

The remaining datasets of SRTM elevation and the NLCD2006 were snapshot datasets, as opposed to the multi-temporal nature of the previously discussed datasets. Preprocessing of these datasets was fairly straightforward and was performed using ModelBuilder models in ArcMap, shown in Figure 15. Since these datasets only needed to be processed once (as opposed

to the multi-temporal data that required lots of identical preprocessing), scripts provided no enhanced speed or other utility, and were therefore not created.

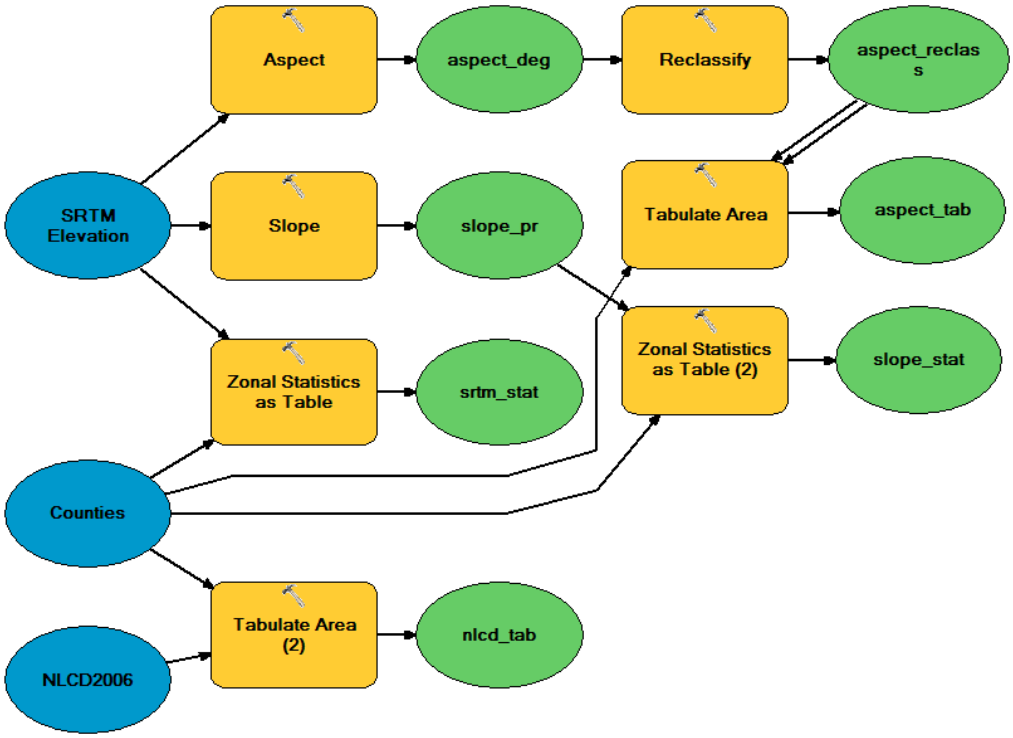

Figure 15 – Esri ModelBuilder diagram of Elevation and Land Cover preprocessing. Blue ovals correspond to raw input datasets, yellow rectangles to geoprocessing tools, and green ovals to derived datasets.

Elevation derivatives of slope and aspect were creating using tools from the Spatial Analyst toolbox. Elevation and slope were then aggregated to the county level using the same Zonal Statistics as Table tool from the Spatial Analyst toolbox used on the previously discussed datasets. Aspect and NLCD2006 were the only qualitative (categorical) variables used and required a slightly different aggregation technique. Aspect was first reclassified from degrees to 8 classes corresponding to the 4 cardinal and 4 inter-cardinal compass directions: North, North-East, East, South-East, South, South-West, West, and North-West. Aspect and NLCD were then both aggregated to the county level using the Tabulate Area tool from the Spatial Analyst

toolbox which allows for the areal amount of each variable occurring in the zone of interest to be preserved. The Zonal Statistics tool expects only quantitative data and calculates means and other statistics, while the Tabulate Area tool works with qualitative variables without losing data (as would occur for example when using the majority statistic in the Zonal Statistics tool which eliminates all variables in the zone of interest except the one covering a majority of the area).

The input data layers were saved to tables and joined together by FIPS (Federal Information Processing Standard) codes manually in ArcMap and Excel, as well as using python scripts (see Appendix A) into a single data table containing 3,105 rows and 283 columns for input into Cubist (see section 3.4.2). Two other tables were also created that temporally aggregated the data for all of the odd years and the even years of the study period.

3.6.2. Addressing the Small Numbers Problem

The small numbers problem was addressed in two separate ways. First, incidence data was converted to incidence rates, otherwise known as being normalized by population. All of the data was then temporally aggregated to attain more reliable rates. NDVI, and climate data that are measured monthly were averaged by corresponding months between years. Land cover and elevation data remained unmodified due to a lack of reliable change information for those datasets over the study period.

A second approach to mitigating the small numbers problem with regards to WNV, similar to a technique used by Biggerstaff and Petersen (2003), was also evaluated (Biggerstaff & Petersen, 2003). As discussed in Section 1.3.1, neuroinvasive disease cases represent only about 1 out of every 150 human infections. While non-neuroinvasive disease is "probably significantly underdiagnosed" due to its mild symptoms and clinical similarity to other diseases (CDC, 2003, p. 19), neuroinvasive cases are generally quite severe and it seems reasonable to

assume reporting of neuroinvasive cases is much closer to 100% than reporting for non-neuroinvasive WNV. Further, the amount of underreporting of non-neuroinvasive cases varies from year to year (see Figure 4), making a single underreporting adjustment impractical. Neuroinvasive cases were treated as a 150 infections each, resulting in a theoretical 30 non-neuroinvasive cases per neuroinvasive case (following the estimate that 20% of infections result in non-neuroinvasive disease) per year. These estimated values for WNV disease incidence (only counting the neuroinvasive and non-neuroinvasive cases, not total estimated infections) were then used to calculated estimated incidence rates. These two values for WNV incidence (raw/reported and estimated) were both run through Cubist separately so their relative strengths and weaknesses could be compared.

3.6.3. Predictive Model Generation

The above-mentioned datasets were aggregated to a common geographic scale (US Counties) and compiled into data tables for input into Cubist (see Appendix B). The WNV incidence data (either raw or estimated, depending on the specific test setup) served as the dependent variable in the equation, with the environmental data serving as the independent (or explanatory) variables in a manner conceptually similar to multiple regression.

Cubist contains a number of optional and advanced settings when creating rule-based predictive models. These optional settings include the use of unbiased rules, composite models, committee models, sampling, seeding, case weighting, cross-validation, extrapolation constraints and setting the maximum number of rules to be generated. Detailed explanations of these settings are unnecessary here, but the interested reader is referred to RuleQuest's website (www.rulequest.com) for more information (RuleQuest Research, 2012). Suffice it to say here

that after much research and testing, the default settings were deemed most appropriate for this study, with two exceptions outlined below.

For all but one of the experimental setups tested sampling was used, whereby Cubist uses a pseudo-random number generator to divide the input data cases into two groups, one for training and model generation, and the other for testing or model validation. Given the large number of cases (3,105 counties), it was heuristically determined that 80% of the data should be assigned to training and the remaining 20% reserved for testing.

The other optional setting used was seeding, whereby a specific seed value is provided by the user for Cubist to use in its pseudo-random number generator during sampling. This feature allows the selection of training and test cases during subsequent test runs with varying settings to be held constant. This removes the variable of chance associated with changing training sets between similar tests, allowing the variables that were changed to be evaluated with less uncertainty. This was only used during variations within single experimental setups as will be described in more detail in the Results section. Between experimental setups, the seed value was changed.

3.6.4. MODEL EVALUATION

"The most common mistake among machine learning beginners is to test on the training data and have the illusion of success" (Domingos, 2012, p. 2). Care was taken to maintain strict separation between training and test data, either with sampling as described above, or through the use of independent datasets. Cubist reports statistical accuracy measures for each model it creates, consisting of Average |Error|, Relative |Error|, and a Correlation Coefficient. The Average |Error|, or average error magnitude, is simply the mean absolute difference between the predicted values and the actual values. This is simple enough to interpret, as smaller values

would indicate less error and therefore a stronger model, although some datasets could contain large average error numbers and still be relatively strong models due to the nature and distribution of the input data. The Relative |Error|, or relative error magnitude, is the ratio of the average error magnitude divided by the error magnitude that would result from every predicted value being equal to the mean value. The relative error magnitude ought to be less than 1 for useful models. This provides a more comparable metric across models. Finally, the correlation coefficient is the Pearson's product-moment, or Pearson's r, measure of linear dependence, measured by dividing the covariance of the predicted and actual values by the product of their standard deviations. Values for the correlation coefficient will always fall between 1 and -1, with values near 1 indicated a near perfect correlation. This would indicate the model "fits" the real world data very well and is in effect a good predictive model. A value of 0.9, for example, could be interpreted as "the model explains 90% of the observed variation" or in other words, "the model is around 90% accurate at predicting X." Interpretations of correlation coefficient values as high, medium, or low, etcetera are somewhat arbitrary, but for the purposes of this study correlation coefficients higher than 0.7 were considered good fits, suitable for disease prediction.

Cubist also computes predicted disease incidence for each county in the test dataset from each model. Using Cubist's companion "Sample.c" program and an optional ".cases" file, predicted disease incidence rates were computed for both test and training cases. These predicted incidence values were reimported into ArcGIS and joined back to US counties vector data via FIPS code for mapping. This was done to allow a visual analysis of the model's predictive power and regional effectiveness. These prediction maps were created using a standard deviation classification technique applied to the difference between predicted values

and actual values (i.e. the model errors) to easily display where the model over- or under-predicted disease. Care was taken to ensure the neutral category in these maps (the areas within one-half positive or negative standard deviation from the mean) contained the true zero value of model errors, so that over- and under-prediction reading of the maps was accurate.

From the models created by Cubist, an odds ratio (OR) – a common epidemiological metric that looks at the association between exposure and outcome – was also computed (Friis & Sellers, 2009). The OR was designed for use among populations of individuals classified as exposed/not-exposed and diseased/not-diseased, and is not intended for use with aggregated data such as counties serving as cases. Some minor modifications had to be made, which likely impacted the odds ratio's effectiveness. Taking the first rule, which is the most explanatory, from the Cubist model with the highest correlation coefficient, environmental variables were identified that served as the "exposures" of interest. The first rule from Cubist was also used to determine threshold values for these variables so that counties could be classified as either exposed or not-exposed. Counties were classified as either disease present or not-present. With counties thus classified as exposed/not-exposed and disease present/not-present, an OR was calculated using the formula "(AD)/(BC), where A is the number of [counties with] the disease and have been exposed, B is the number who do not have the disease and have been exposed, C is the number who have the disease and have not been exposed, and D is the number who do not have the disease and have not been exposed" (Friis & Sellers, 2009, p. 661). This can be thought of as measuring whether or not the exposures of interest increase the chances of disease or not.

Normally this ratio would be equal to 1 if there was no difference between disease incidence among those exposed and those not-exposed. The OR should be higher than 1 for exposures that increase risk of disease, and values lower than 1 may indicate protective factors

that help prevent disease. The OR was computed using R: A language and environment for statistical computing, and the epitools: Epidemiology Tools package (Aragon, 2012; R Development Core Team, 2012). This provided a quantitative method for model evaluation outside of Cubist's reported statistical accuracy, although the necessary modifications to the measure made its results somewhat suspect.

4. RESULTS

There were eight experimental setups, each with between 2 and 4 variations, making a total of 30 runs using Cubist. The first setup (R1) included all six years of the study period. R1a and R1b used 30-year normals for temperature and precipitation data and the NDVI data was averaged by month over the study period. R1c and R1d did not use normals or averages, but looked at all of the explanatory variables for the entire study period – this was referred to as the "Everything but the Kitchen Sink" approach. As outlined by Jensen (2005), there is a large body of evidence demonstrating that machine learning can "deal effectively with tasks that involve highly dimensional data" and that "the new thinking is to let the geographic data itself 'have a stronger voice'" as opposed to using data reduction techniques before analysis (John R Jensen, 2005, p. 421).

The second setup (R2) used data averaged across the odd years of the study period to create the model, and was tested against the data averaged over the even years. This was the only experimental setup where two full datasets (all 3105 counties) were used, eliminating the need for sampling. Odds and evens were chosen simply to avoid any confusion with time-series analysis that may have arisen from a chronological ordering.

The third through eighth setups (R3-R8) each involved data from a single year of the study period, 2003-2008. The "a and b" runs were, as above, merely differentiating between whether raw ("a" runs) or estimated ("b" runs) WNV rates were being used as the predicted variable. The "c and d" runs for R3-R8 used precipitation data from the previous year instead of the year under study to test the findings of Landsman et al. (2007) that precipitation from the previous year was more strongly associated with disease outbreak than precipitation during the concurrent year (Landesman et al., 2007).

Table 1 below summarizes the variations on the eight experimental setups just described. Results from each of the eight experimental setups and their variations will be reported in turn. All of the predictive models created using Cubist were tested on data independent of the data used for training, or model generation. The average error, relative error, and correlation coefficient values reported below are the evaluations of the test data, while the maps show results from both training and test data for ease of readability.

Table 1 – Various Cubist Runs, or experimental setups. Abbreviations used: avg. = average; RR = Raw Rates of WNV Incidence; ER = Estimated Rates of WNV Incidence; EV = Explanatory/Environmental Variables; PPT = Precipitation; and " = Ditto (same as preceding).

Run#	Experimental Setup Notes	Predicted Variable
R1a	Avg. of all 6 years of EV	Avg. RR for all 6 years
R1b	"	Avg. ER for all 6 years
R1c	All 6 years (not avg.) of EV	Avg. RR for all 6 years
R1d	"	Avg. ER for all 6 years
R2a	Avg. of Odd Years used to create model, avg. of Even Years used to test.	Avg. RR for Even Years
R2b	"	Avg. ER for Even Years
R3-8a	Single Year (2003-08) EV only	RR for Single Year
R3-8b	"	ER for Single Year
R3-8c	Single Year (2003-08) EV with prior-year PPT	RR for Single Year
R3-8d	"	ER for Single Year

4.1. R1: Entire Study Period

The first experimental setup, R1, resulted in the highest correlation coefficients for any of the models created in this study, with 0.84 for R1a and 0.86 for R1c (see Table 2). Average and relative error magnitudes both increased when predicting estimated WNV rates (b and d), and the associated correlation coefficients correspondingly went down. The most used variables in rule conditions for R1 models included precipitation and the NLCD06 land cover class 41 – Deciduous Forest.

Table 2 – Results from setup R1(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R1a	20.7	0.39	0.84	Precipitation (’03-’08)
R1b	218	0.66	0.55	NLCD-41
R1c	18.9	0.35	0.86	Precipitation (’02)
R1d	185.2	0.56	0.56	Precipitation (’02)

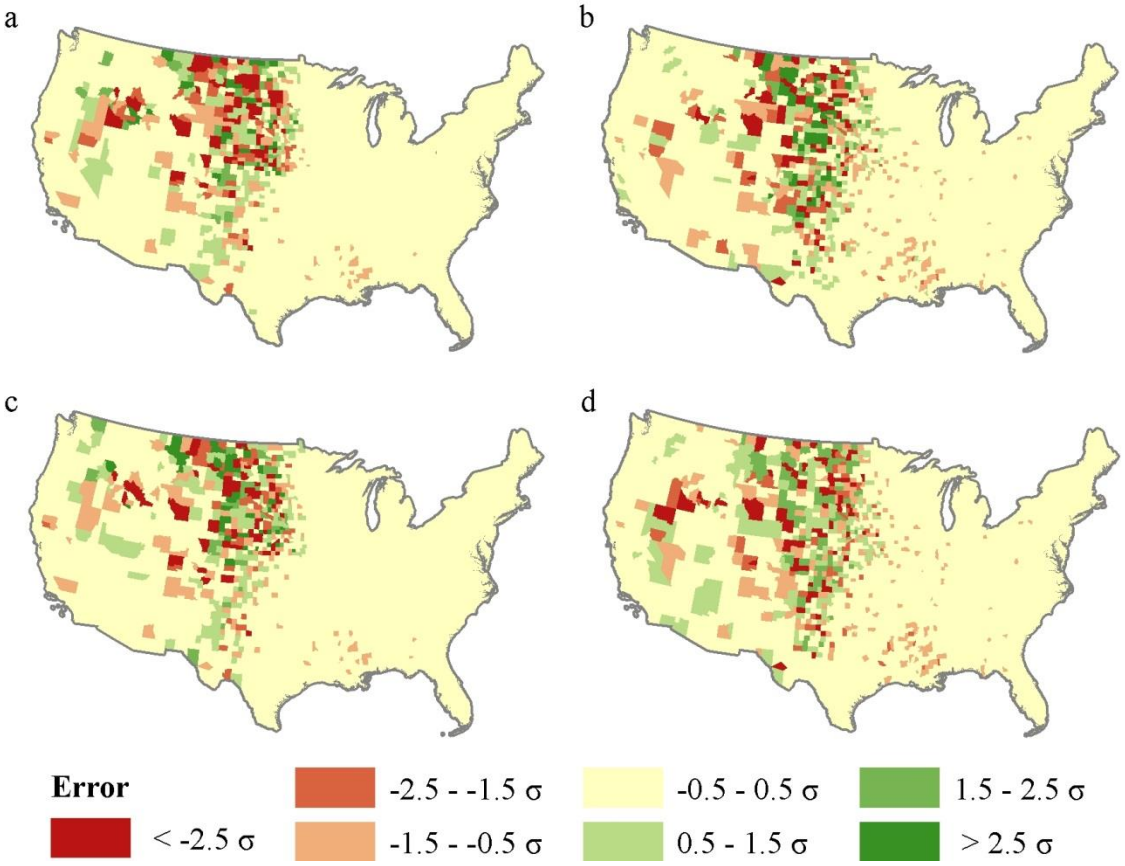

Figure 16 – Distribution of model errors from R1(a-d), categorized by Standard Deviations.

The spatial distribution of model errors (see Figure 16) exhibits a pattern. Roughly halfway across the country from East to West, all four R1 models start to exhibit evidence of significant over- and under-predicting of WNV incidence, both raw (a and c) and estimated (b and d). All four exhibit limited cases of under-predicted errors in the eastern half of the US, but little or no cases of over-predicting. This region of increased errors becomes less concentrated through most of the Mountain West and fades before reaching the West coast.

4.1.1. ODDS RATIO

As setup R1 resulted in the highest correlation coefficients, it was chosen for odds ratio (OR) calculation, as discussed in section 3.6.4. Based on the relatively simpler data and production rule conditions, model R1a was chosen over R1c, despite a slightly lower correlation coefficient. The first rule's conditions were used to identify key environmental variables and suitable thresholds for the data in order to classify counties as "exposed" or "not-exposed." The conditions used were: (Average Precipitation (30-year normal) > 754.649 mm) AND (Average NDVI for December > 0.3238802). Of the 3,105 counties in the study area, approximately two-thirds (2,156 counties) met the "exposed" criteria. The threshold for WNV incidence was set at zero, resulting in a little over half (1,698 counties) being classified as "disease-present" (see Table 3). The OR was calculated as (A*D)/(B*C), or (961*212)/(1195*737) = **0.23**. The R package "epitools: Epidemiology Tools" (Aragon, 2012) calculated a 95% confidence interval of 0.19 to 0.28.

Table 3 – County categorizations and totals used to calculate the Odds Ratio (OR) for model R1a.

	Disease Present	**Disease Not Present**
Exposed	A) 961	B) 1195
Not-Exposed	C) 737	D) 212

4.2. R2: ODD YEARS MODEL TESTED ON EVEN YEARS

Setup R2 resulted in correlation coefficient values of 0.34 for R2a and 0.13 for R2b (see Table 4). Average error magnitude increased when predicting estimated WNV rates (b), but relative error magnitude actually decreased, demonstrating the risk in interpreting the average error magnitude out of context. Despite a lower relative error, the correlation coefficient for R2b was much lower than R2a. The most used variables in rule conditions for R2 models included average minimum temperatures in December and average NDVI values in November.

Table 4 – Results from setup R2(a-b).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R2a	28.5	1.03	0.34	TMin Dec
R2b	176.2	0.93	0.13	NDVI Nov

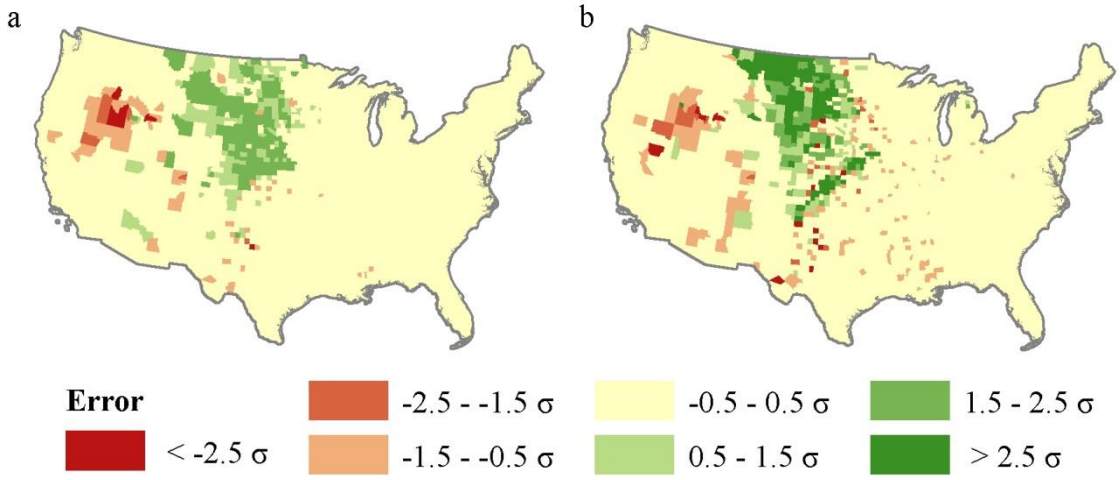

Figure 17 – Distribution of model errors for R2(a-b), categorized by Standard Deviations.

The spatial distribution of model errors (see Figure 17) exhibits a clear regional pattern in both models, although R2a appears to have less inconsistencies than R2b. The northern Great Plains region tends to exhibit over-prediction errors in both models, while a smaller region centered approximately in southern Idaho exhibits under-prediction errors in both models.

4.3. R3: 2003

Setup R3 resulted in relatively high correlation coefficient values of 0.84 and 0.8 for a and c respectively, and 0.55 and 0.58 for b and d (see Table 5). Average and relative error magnitudes increased when predicting estimated WNV rates (b and d) and correlation coefficients correspondingly went down compared to their raw WNV rates counterparts (a and c). The most used variable in rule conditions for R3 models was precipitation, both concurrent and prior-year. Results for R3a and c (as well as for b and d) were very similar, indicated little difference between model effectiveness between those variations.

Table 5 – Results from setup R3(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R3a	11.5	0.35	0.84	Precipitation ('03)
R3b	112.9	0.58	0.55	Precipitation ('03)
R3c	11.5	0.35	0.8	Precipitation ('02)
R3d	106.8	0.55	0.58	Precipitation ('02)

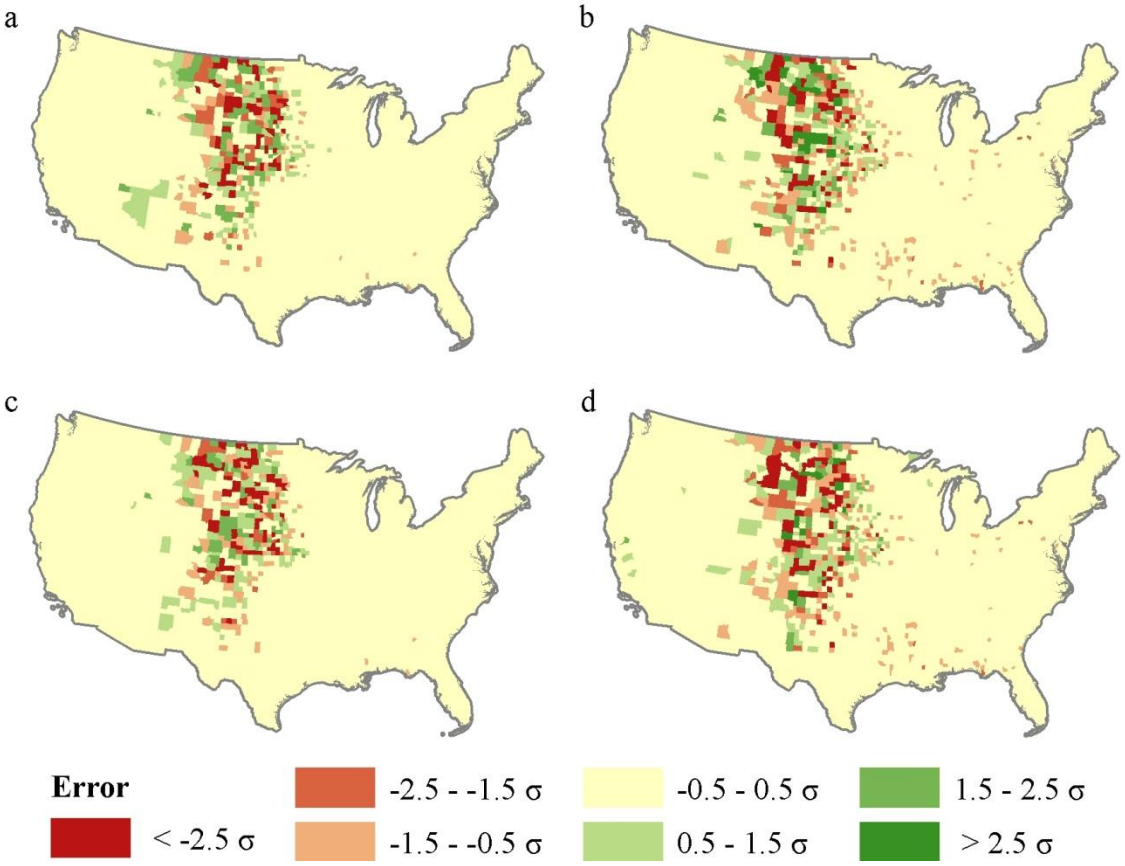

Figure 18 – Distribution of model errors for R3(a-d), categorized by Standard Deviations.

The spatial distribution of errors in the R3 models (see Figure 18) display a pattern very similar to that seen in the R1 models. The northern Great Plains region and the western Great Plains bordering on the Rocky Mountains together form an apparent cluster of errors, both over- and under-predicting WNV in all four models. The estimated WNV models (b and d) both exhibit scattered cases of mild under-prediction in the eastern US.

4.4. R4: 2004

Setup R4 resulted in low correlation coefficient values of 0.18 to 0.26 (see Table 6). Average error magnitudes increased when predicting estimated WNV rates (b and d), while relative error magnitudes decreased. Correlation coefficients were lower for the models predicting estimated WNV rates than those predicting raw rates. The most used variables in rule conditions for R4 models included NDVI in June and December, prior-year precipitation, and the land cover class 81 – Pasture/Hay.

Table 6 – Results from setup R4(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R4a	1.8	0.73	0.26	NDVI Dec
R4b	14.3	0.56	0.22	NLCD-81
R4c	1.6	0.64	0.26	Precipitation ('03)
R4d	14.3	0.56	0.18	NDVI June

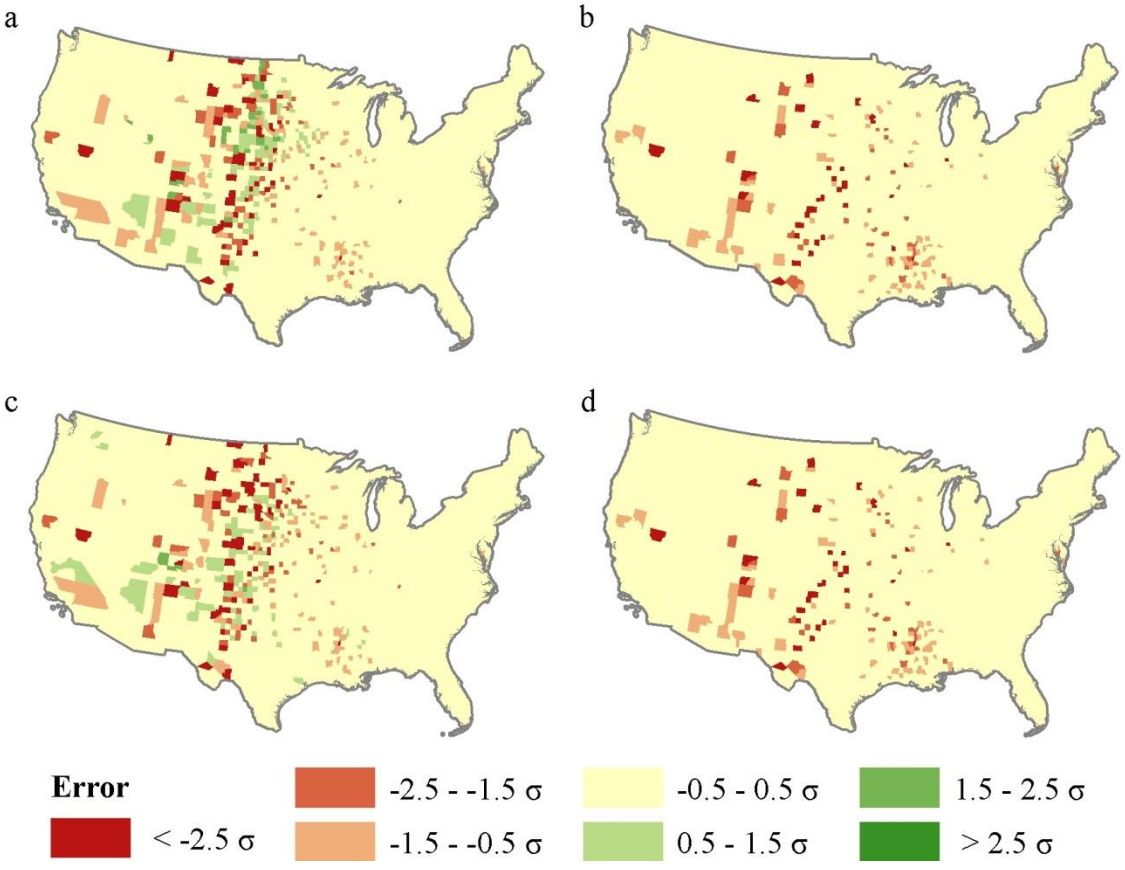

Figure 19 – Distribution of model errors for R4(a-d), categorized by Standard Deviations.

The spatial distribution of R4 model errors (see Figure 19) bears a passing resemblance to that seen previously, although the pattern is less pronounced in a and c than in previous models, and almost non-existent in models b and d. For reasons unknown, the estimated rates models (b and d) predicted very low values for all counties, resulting in the extremely low correlation coefficient values, as well as the lack of over-prediction errors.

4.5. R5: 2005

Setup R5 resulted in correlation coefficient values as high as 0.56 and as low as 0.12 (see Table 7). Average and relative error magnitudes both increased when predicting estimated WNV rates (b and d) and correlation coefficients correspondingly went down. R5a was the best model, explaining about 10% more of the data than the next best model, R5c. The most used variables in rule conditions for R5 models included average minimum temperatures for October and December, mean elevation, and land cover class 41 – Deciduous Forest.

Table 7 – Results from setup R5(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R5a	3.1	0.61	0.56	TMin Dec
R5b	42.5	0.82	0.12	NLCD-41
R5c	3.3	0.66	0.46	Elevation
R5d	40.3	0.77	0.12	TMin Oct

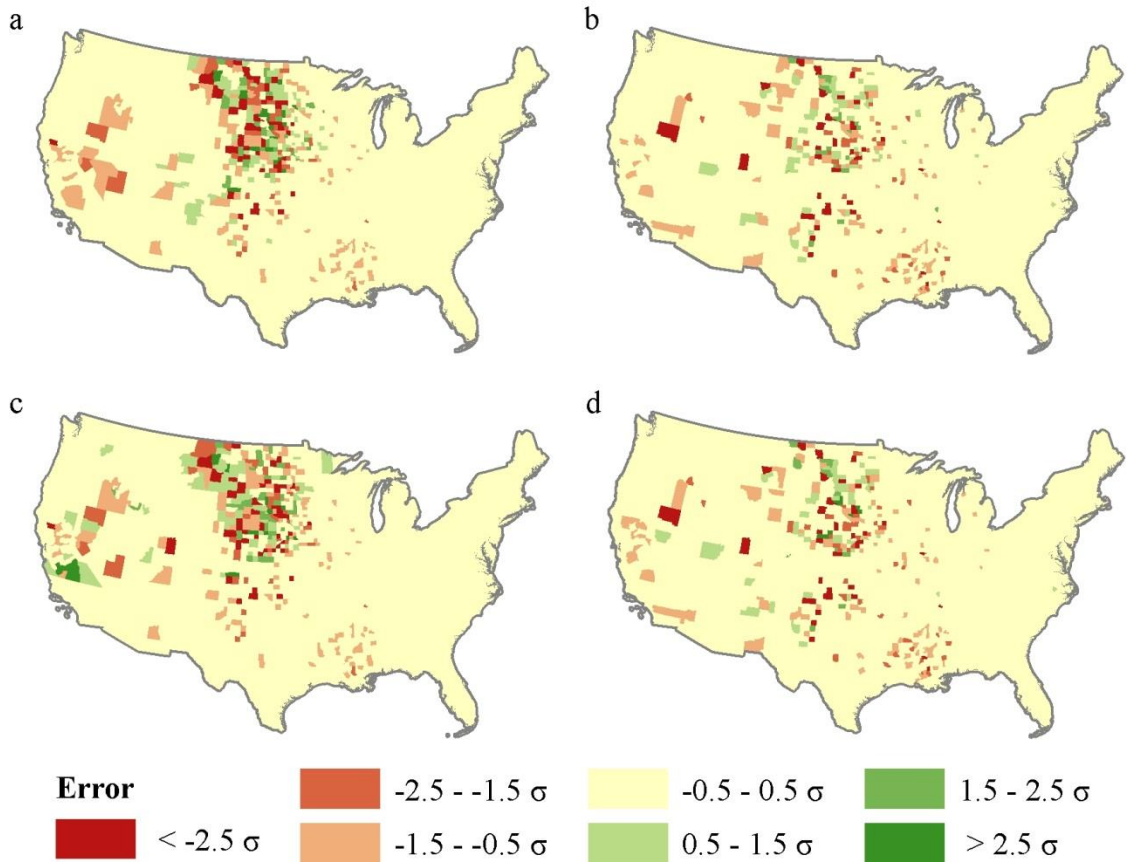

Figure 20 – Distribution of model errors for R5(a-d), categorized by Standard Deviations.

The spatial distribution of R5 model errors (see Figure 20) again resembles the trend previously observed of both over- and under-prediction errors clustering in the northern Great Plains and scattered throughout the West, with relatively few errors observed in the eastern US. The pattern is less apparent in the estimated WNV models (b and d), although it does still appear to be present to some extent.

4.6. R6: 2006

Setup R6 resulted in correlation coefficient values for a-d of 0.6, 0.41, 0.58, and 0.48 respectively (see Table 8). Average error magnitudes increased when predicting estimated WNV rates (b and d), while relative error magnitudes either stayed the same or went down compared to

raw rate models (a and c). The most used variables in rule conditions for R6 models included NDVI values for March and December.

Table 8 – Results from setup R6(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R6a	4.2	0.61	0.6	NDVI Dec
R6b	37.7	0.61	0.41	NDVI Mar
R6c	4.2	0.62	0.58	NDVI Dec
R6d	36.3	0.59	0.48	NDVI Mar

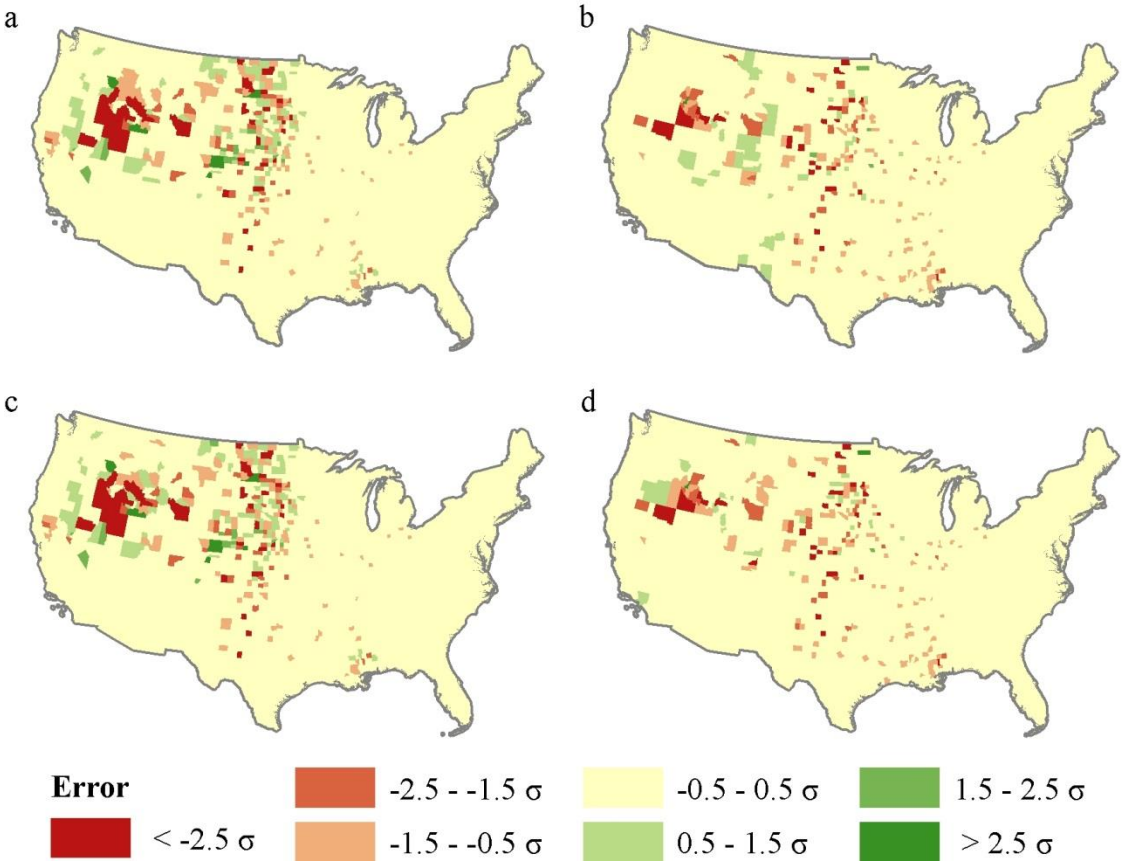

Figure 21 – Distribution of model errors for R6(a-d), categorized by Standard Deviations.

The spatial distribution of R6 model errors (see Figure 21) is similar to the general trend observed previously, with the notable exception that the northern Great Plains appears somewhat less error prone and the region centering on southern Idaho appears to be much more pronounced

in a and c with under-prediction errors in the core and over-prediction around the periphery. This pattern is present, but not as pronounced in the estimated WNV models (b and d).

4.7. R7: 2007

Setup R7 resulted in the widest range of correlation coefficients within a single experimental setup, ranging from 0.28 to 0.75 (see Table 9). Average and relative error magnitudes increased when predicting estimated WNV rates (b and d) and those models' correlation coefficients went down. The most used variables in rule conditions for R7 models included average minimum temperatures for October, and NDVI values for January.

Table 9 – Results from setup R7(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R7a	5.2	0.52	0.75	TMin Oct
R7b	37.2	0.67	0.28	NDVI Jan
R7c	5.3	0.53	0.75	TMin Oct
R7d	36.7	0.66	0.28	NDVI Jan

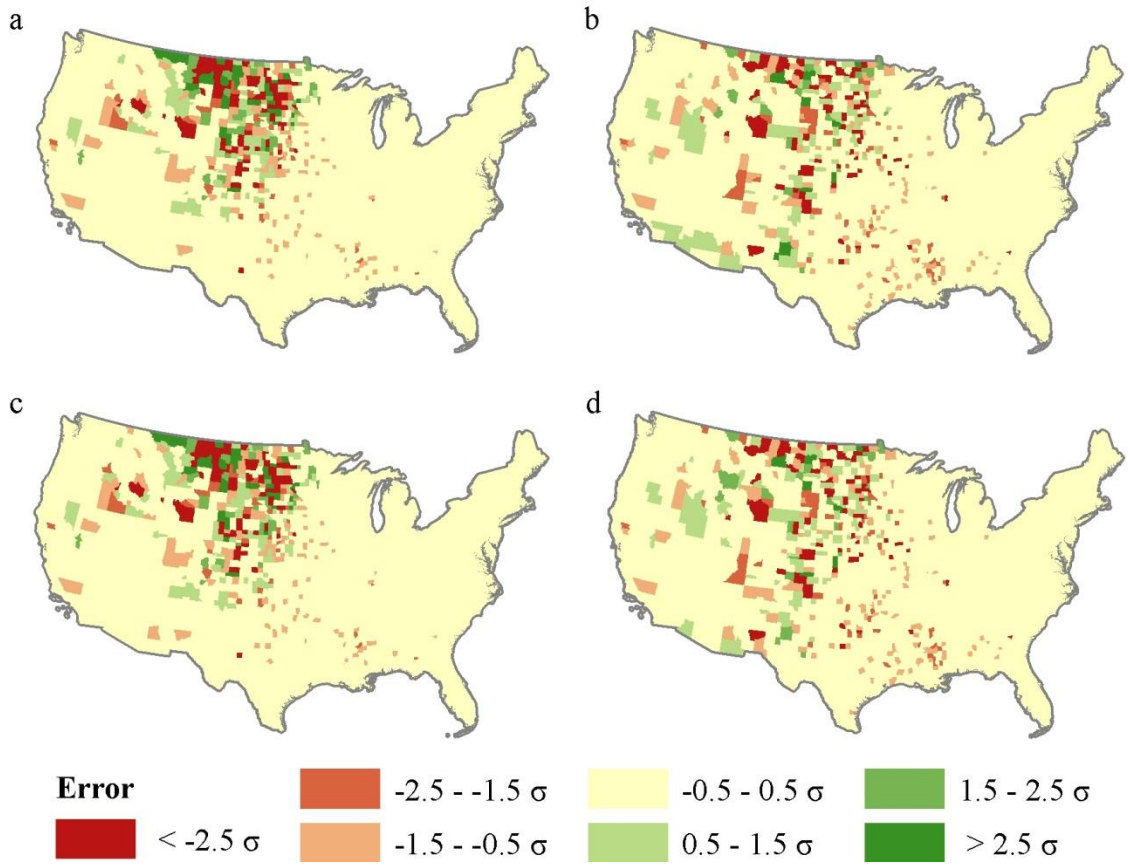

Figure 22 – Distribution of model errors for R7(a-d), categorized by Standard Deviations.

The spatial distribution of R7 model errors (see Figure 22) is yet again strikingly similar to the predominant trend of both under- and over-prediction errors clustering in the northern Great Plains region, with less significant errors scattered around the region, mostly fading before the West coast and the eastern Great Plains. This pattern is again most apparent in the raw WNV models (a and c), although still evident in the estimated WNV models (b and d). Again, as above, the estimated WNV models' errors tend to extend further east and exhibit less spatial clustering than their raw WNV model conterparts.

4.8. R8: 2008

Setup R8 resulted in correlation coefficient values of 0.29 and 0.3 for runs a and c, and values of 0 and 0.02 for b and d (see Table 10). Average error magnitudes increased when

predicting estimated WNV rates (b and d), but relative error magnitude decreased. The most used variables in rule conditions for R8 models included average minimum temperatures in April, and land cover class 71 – Grassland/Herbaceous.

Table 10 – Results from setup R8(a-d).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable
R8a	1.5	0.72	0.29	TMin Apr
R8b	13.3	0.62	0	NLCD-71
R8c	1.5	0.72	0.3	TMin Apr
R8d	13.2	0.62	0.02	NLCD-71

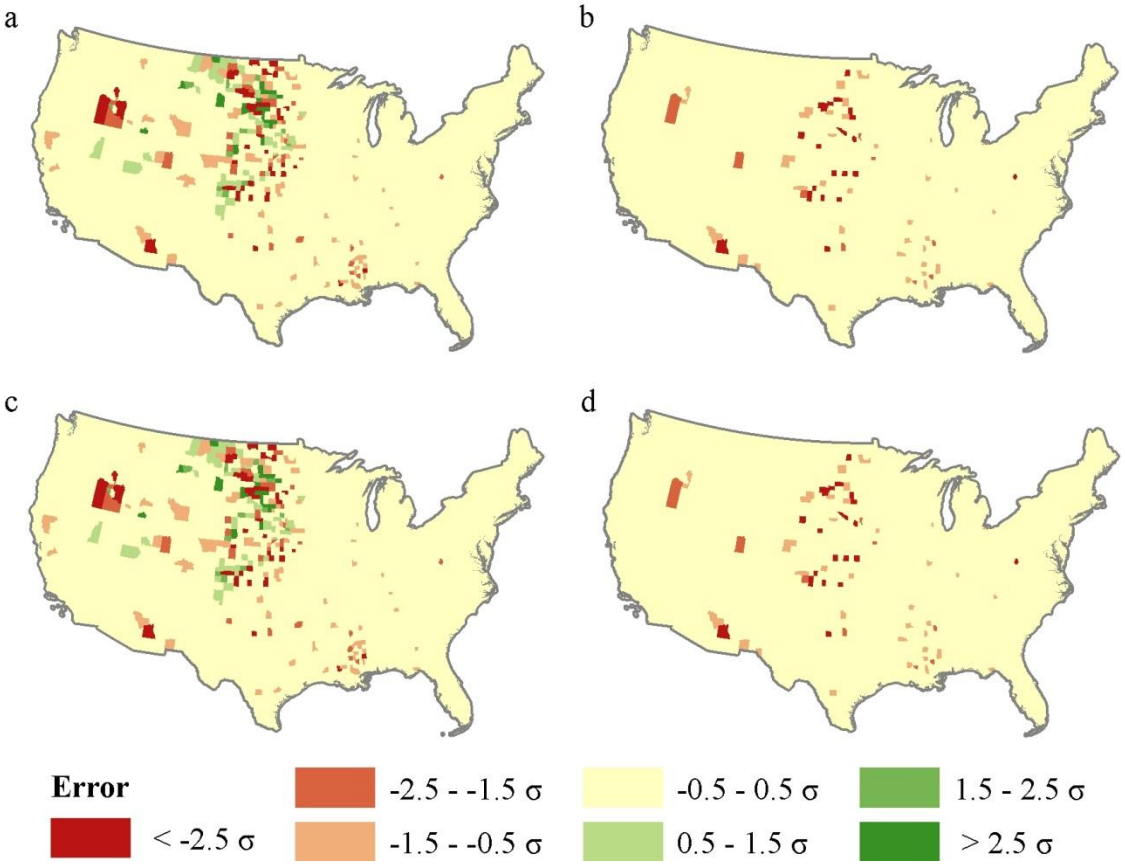

Figure 23 – Distribution of model errors for R8(a-d), categorized by Standard Deviations.

The spatial distribution of R8 model errors (see Figure 23) again resembles the general trend of clustered errors in the northern Great Plains and southern Idaho regions, although the pattern is somewhat less pronounced in 2008, and again almost non-existent in the estimated

WNV models (b and d). For reasons unknown, just as occurred with R4, the estimated WNV models (b and d) predicted very low values for all counties, resulting in the extremely low correlation coefficient values, as well as the lack of over-prediction errors.

4.9. Northern Great Plains Model

While the stated study area included the entire continental US, the consistent spatial pattern observed in the "northern Great Plains" (NGP) region was deemed of sufficient interest to warrant a follow up model. The region was delineated with the help of a local Moran's I map (see Figure 26) that identified a large cluster of high WNV incidence rate counties (see Figure 24). The data for these selected counties were then exported to a table and run through Cubist to generate a ninth, region-specific model, referred to here as the NGP model.

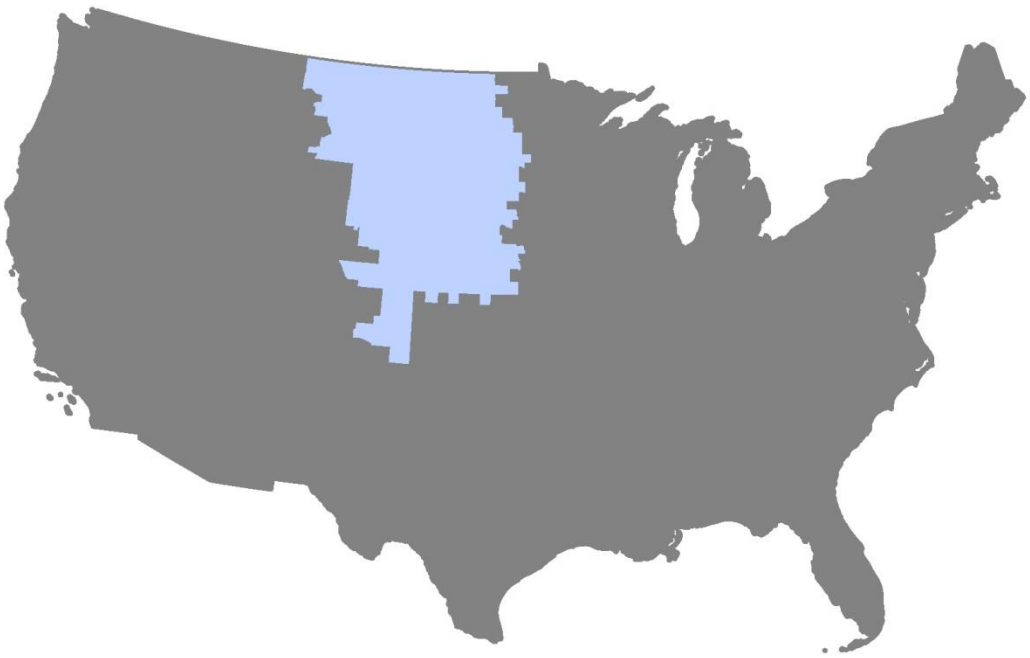

Figure 24 – The "Northern Great Plains" (NGP) region extracted for further modeling in Cubist.

The input data and Cubist parameters used during NGP model generation were matched as closely as possible to those used for R1c, the model with the highest correlation coefficient.

NGPa predicted raw WNV incidence rates while NGPb predicted estimated WNV incidence rates. The NGP models resulted in correlation coefficient values of 0.5 for a, and 0.27 for b (see Table 11). Average and relative error magnitudes increased when predicting estimated WNV rates (b) and the correlation coefficient correspondingly went down. The most used variables in rule conditions for NGP models included precipitation, average maximum temperatures in June, September, December and annually, and average minimum temperatures in May and July.

Table 11 – Results from setup NGP(a-b).

Run#	Average |Error|	Relative |Error|	Correlation Coefficient	Most Used Variable(s)
NGPa	144.6	0.84	0.5	Precipitation ('06) TMax June, Sep, Dec, Yr TMin May and July
NGPb	1097	0.99	0.27	TMin May

As with prior models, results were markedly poorer when predicting estimated WNV rates compared to raw incidence rates. In order to make an accurate comparison between national model R1c and regional model NGPa, the NGP counties were run through Sample.c using R1c's model, and a correlation coefficient of 0.7 was found, compared to the 0.5 achieved with the regional model, indicating the national model predicted the NGP region more accurately than the regional model did.

5. Discussion and Conclusion

The first research question asked in this study was if remotely sensed environmental variables could be used to predict WNV incidence rates with acceptable accuracy across the US. While the question was intentionally vague on what would be considered "acceptable accuracy," the results of the R1 models, specifically a and c, seem to justify the conclusion that they can. That being said, there were several obvious shortcomings with the many models, some of which were evidently spatial in nature. The calculated odds ratio (OR) of 0.23, for example, if interpreted in the traditional way would imply the selected environmental conditions identified by Cubist are in fact protective factors against WNV disease. This interpretation clearly does not harmonize with the other evaluations of the R1 models. It seems apparent that the adjustments to the OR measurement necessary for its use on aggregated data invalidated its effectiveness as an evaluation tool. The low OR score could alternatively evidence the complexity of the Cubist predictive models by demonstrating that the conditions used in a single rule cannot predict WNV alone, while the entire model together performs quite well.

The machine learning decision trees algorithms used by Cubist are spatially ignorant, or in other words, locations and spatial relationships were not variables used during model generation. The patterns and correlations identified within the data, used to create the predictive models, are all location unaware. With this in mind, the observed spatial pattern of model errors, remarkably consistent across the 30 different models, is intriguing. The figures above show consistently that the various models are least accurate in the northern Great Plains (NGP), Rocky Mountains, and southern Idaho areas. Perhaps even more interesting than this apparent regional clustering of errors is the fact that these regions appear to be subject to both under- and over-prediction at the same time, with the notable exception of the R2 models which spatially

segregated the under- and over-predicted areas. After initial model evaluation presented in the results above, most of the subsequent analysis was devoted to attempting to explain these patterns.

Figure 25 was created to evaluate what impact spatial sampling bias might have had on the observed results. US counties are notoriously disparate in both size and composition, which often causes problems in spatial studies such as this one. The eastern US is composed of generally very small counties, that are nevertheless highly populated. Much of the western US, by comparison, is made up of very large counties with generally lower populations, except for the coastal regions. For these reasons, almost any sampling scheme using US counties as a base unit is practically guaranteed to exhibit spatial bias. Sampling for the predictive models was performed aspatially within Cubist using a pseudo-random number generator with a seed value that was kept consistent within experimental setups.

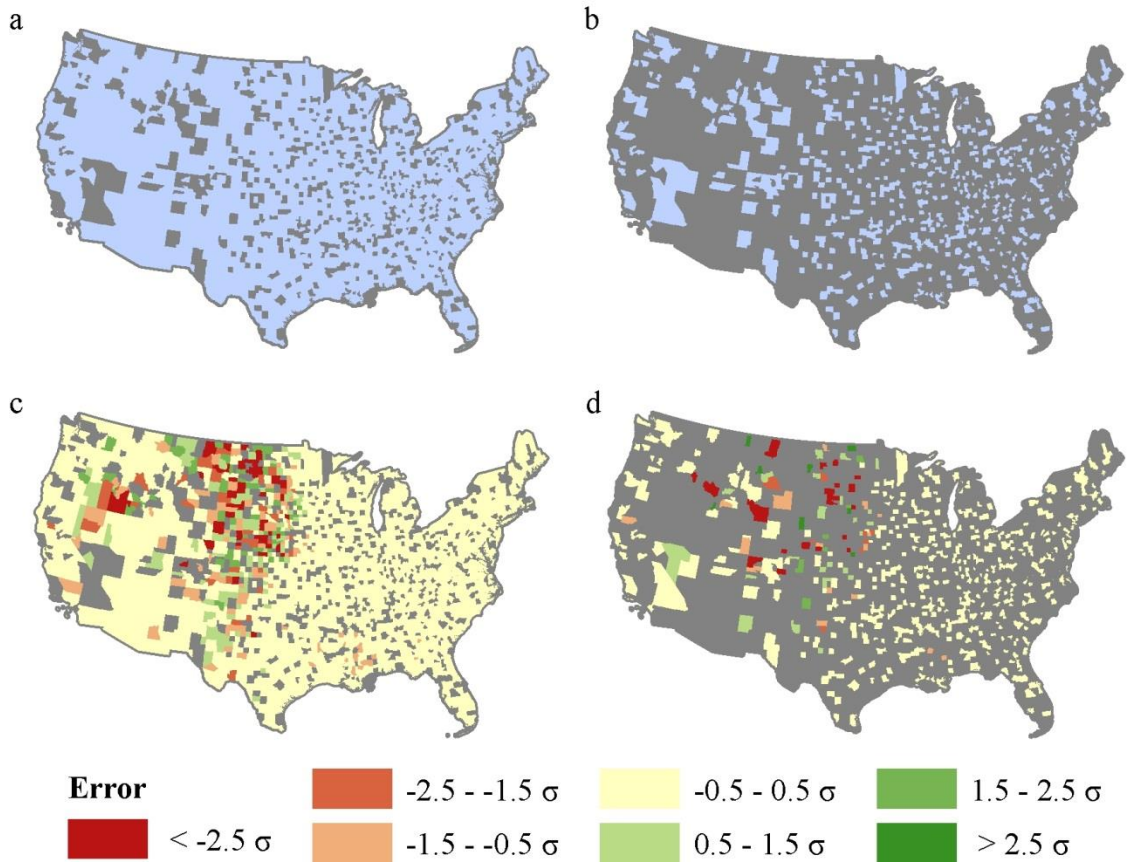

Figure 25 – Training and Test cases for Experimental test setup R1. a) shows Training cases in blue, b) shows remaining test cases in blue, c) and d) show predictive model R1a Errors overlayed on training and test cases.

Figure 25a shows the counties selected by Cubist as training cases for the R1 models, and Figure 25b shows the remaining counties used for testing. Visual inspection confirms that the sampling is about as unbiased as possible given the constraints already discussed. Figure 25c and d show the same counties as a and b, this time overlaid with the model errors from R1a to demonstrate that the observed spatial pattern of model errors is apparent in both training (c) and test (d) datasets, indicating spatial sampling bias was not a significant factor in producing the pattern.

The regions of poor model performance seem closely related to the regions previously identified by Young and Jensen (2012) and others as the areas with the most pronounced

clustering of disease incidence. Figure 26 shows the Anselin Local Moran's I, a spatial cluster and outlier analysis tool that measures spatial autocorrelation (Anselin, 1995), of the disease incidence data for the study period (a) compared to the errors or residuals from predictive model R1a (b).

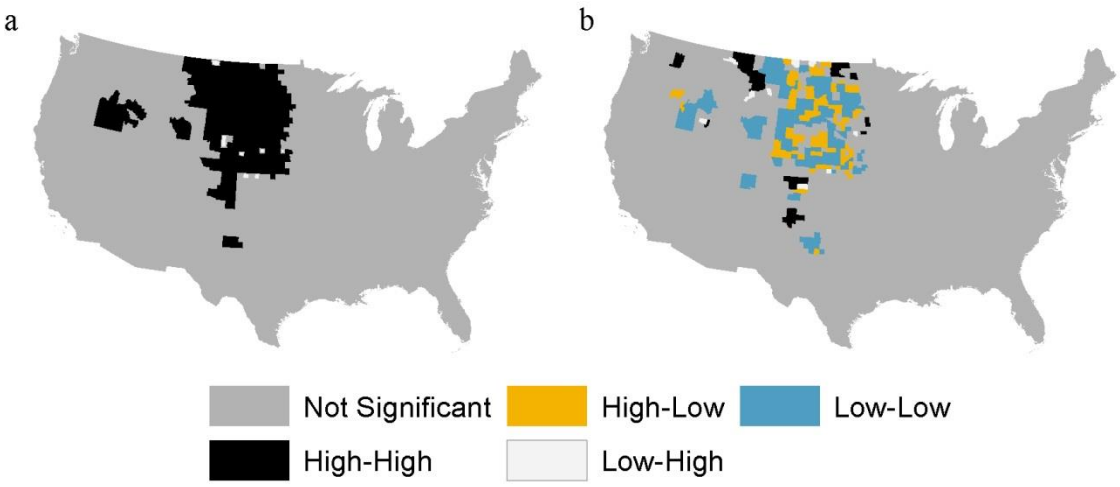

Figure 26 – Anselin Local Moran's I maps. a) shows clustering of WNV Incidence Rates for the entire 6-year study period, and b) shows clustering of residuals from predictive model R1a.

The similar spatial pattern indicates a connection, although it is primarily conjecture at this point as to what exactly that connection is. Perhaps the high incidence values in the region are a reflection of the relatively small populations, and thus a result of the small numbers problem discussed earlier. Figure 27 may offer some support of this theory, showing the lowest population counties in lighter shades of blue, which appear to match fairly well with the regions suffering from the most model errors, with the possible exception of the low-population counties in the eastern US.

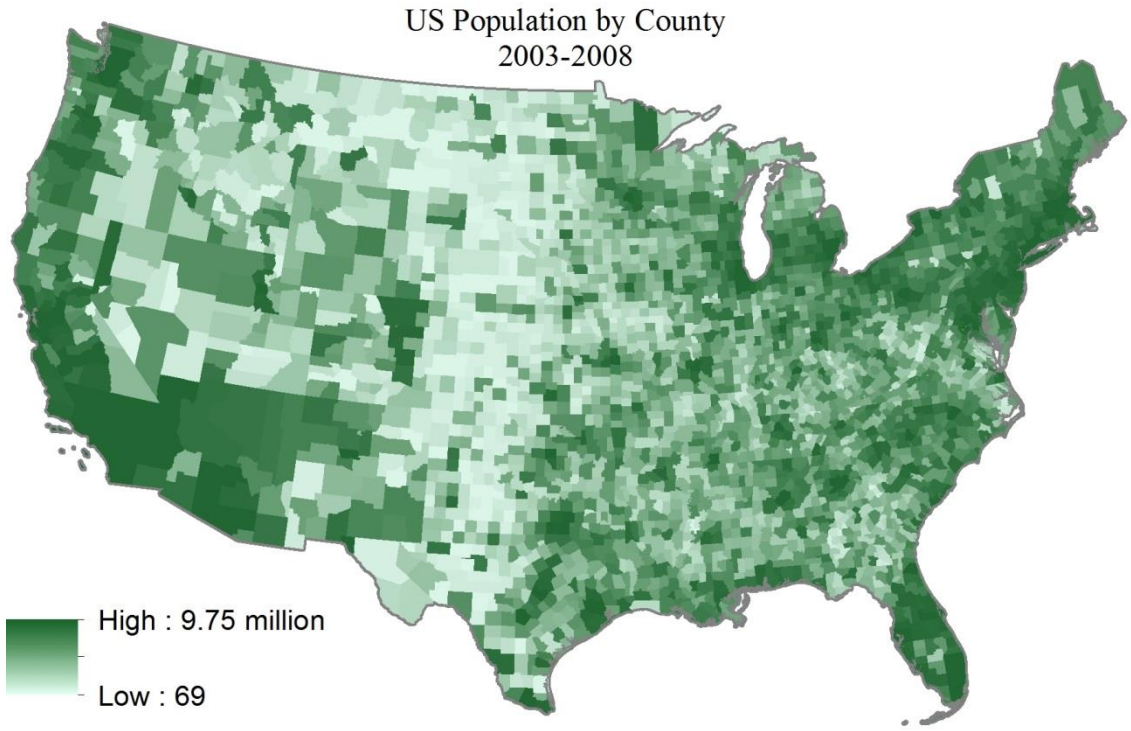

Figure 27 – Average Population during the study period (2003-2008). Notice the central, western, and "Northern Great Plains" (NGP) regions tend to have lower populations, making them more susceptible to the Small Numbers Problem.

The strongest evidence for this theory, that model errors are associated with low-population counties, is presented in Figure 28, which shows the standard deviations of the model errors for R1a and R1b plotted against county population. The resulting scatterplot shows that the models both perform admirably at a large range of population values, with the majority of the errors occurring in counties with relatively lower populations. This is most apparent in the top two scatterplots in Figure 28.

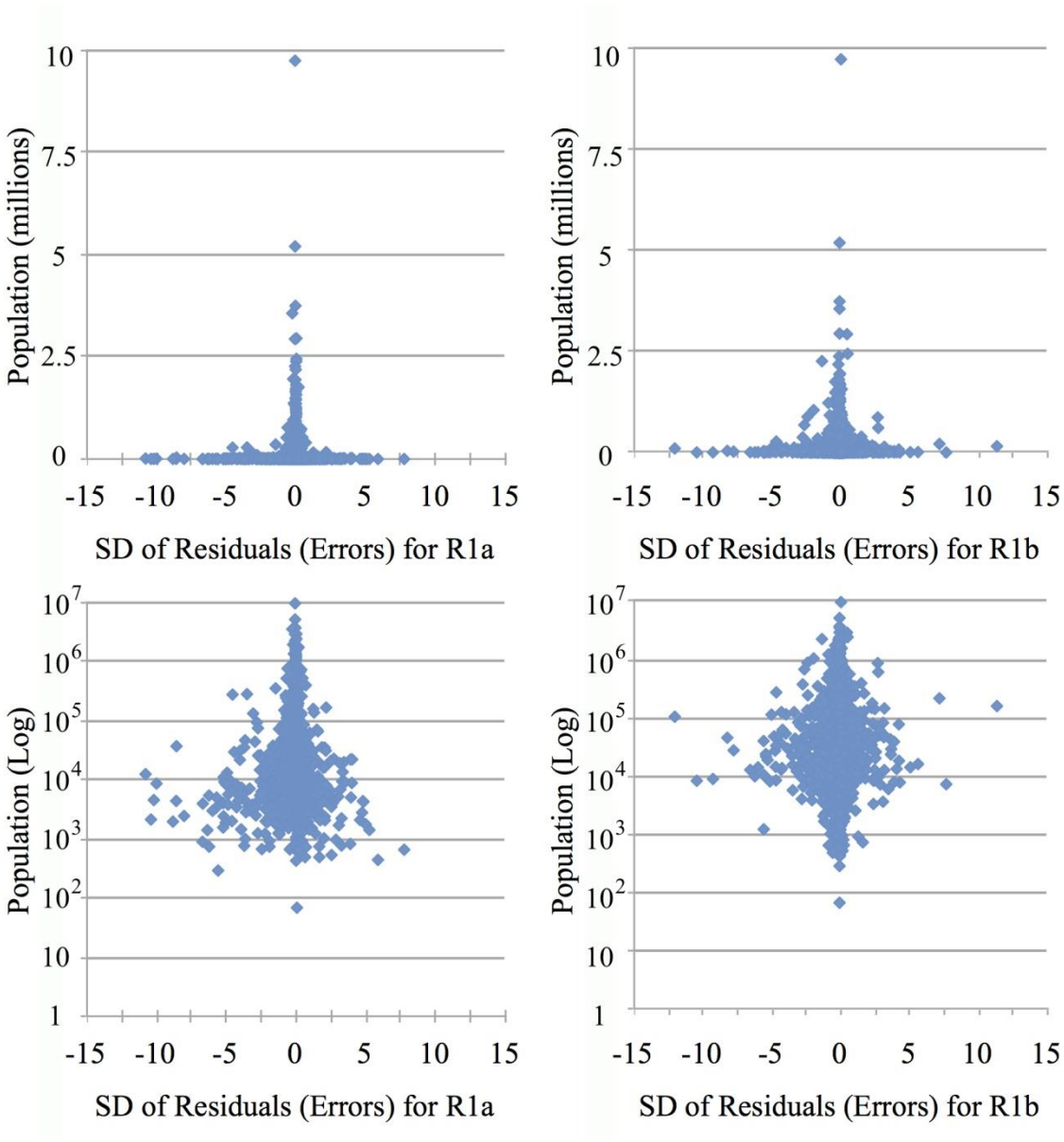

Figure 28 – Scatterplots of models R1a and R1b residuals (as Standard Deviations) against County Population. The top row shows population in millions, with the bottom plots show the same data on a logarithmic Y axis to better visualize the distribution.

The lower two plots in Figure 28 present the same data, but using a logarithmic scale for population, which emphasizes the differences between R1a and R1b. R1b errors, while they appear to follow the same pattern as R1a of being most common in low-population counties, can also be seen to "shift" the errors up, meaning counties with slightly higher populations are more

likely to be over- or under-predicted when using the estimated WNV model. This was an unexpected consequence as the estimated WNV models were created specifically to attempt to mitigate the small numbers problem, but they appear to have in effect expanded it to a "small and not-as-small numbers" problem. Between the consistently lower (often embarrassingly lower) correlation coefficients and the scatterplots of Figure 28, I have concluded that my method of estimating WNV incidence rates to mitigate the small numbers problem was not successful.

The R1 and R2 models were the most comprehensive in that they covered the entire study period, but the R3-R8 models served to illustrate temporal variations in the validity of the methods used in this study. Model strength varied widely by year, indicating a pronounced sensitivity to temporal changes, and a corresponding lack of consistent environmental conditions that can be used as reliable predictors. Said another way, the fact that the yearly models varied so much in effectiveness implies that there is not a single set of environmental conditions that will always indicate WNV disease presence. If there were, the Cubist models would have been expected to identify similar rule sets each year, resulting in models that performed equally well from year to year.

The R3-R8 models were also used to investigate the research question of whether prior-year precipitation was a better predictor than concurrent-year precipitation. The results, shown in the c and d models, were decidedly mixed. In all, 4 models showed decreased efficacy using prior-year precipitation, 4 showed increased efficacy, and 4 stayed the same, as measured by their correlation coefficients compared to their corresponding concurrent-year models. The average decrease was 0.05, while the average increase was 0.0325. It was concluded that the results in relation to the prior versus concurrent precipitation research question were inconclusive.

With some exceptions, the region centered on southern Idaho appeared in many models to be a secondary cluster of model errors. Unlike the NGP region however, the southern Idaho cluster almost always exhibited under-prediction errors at its core. There is likely some set of environmental conditions, or perhaps behavioral conditions among the population, responsible for this pattern that were not included in this study.

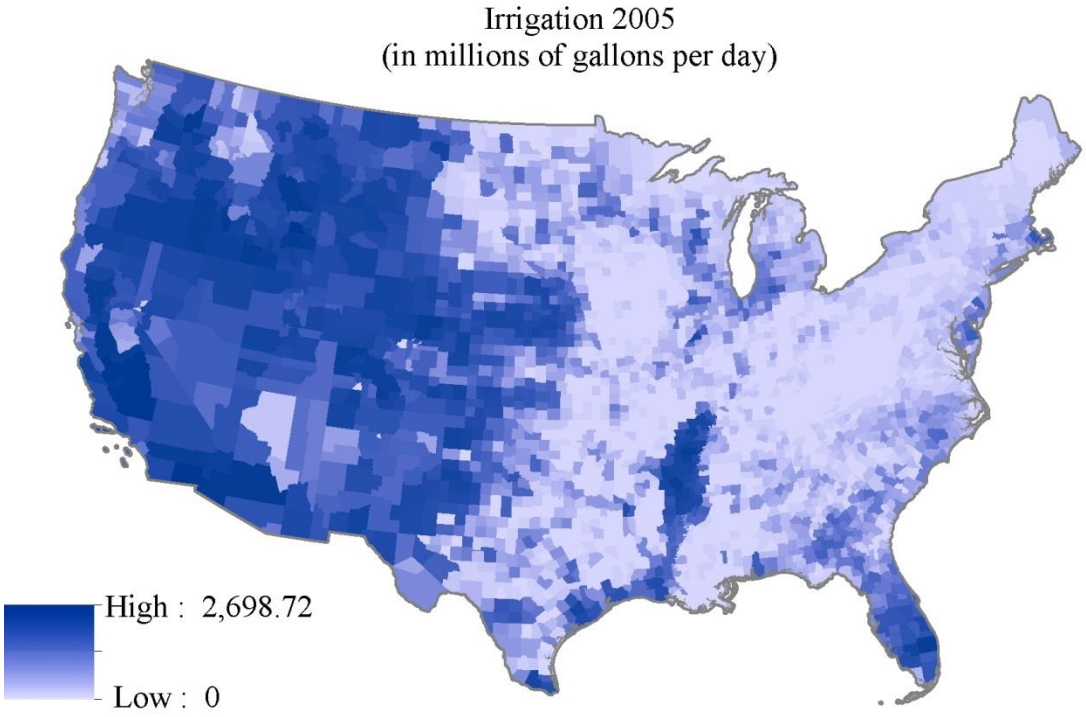

Figure 29 – Estimated fresh water usage for artificial irrigation in 2005, measured in millions of gallons per day. Data from the National Atlas (USGS Water Resources of the United States, 2010).

Similarly it seems evident that some unknown variable or variables are responsible for the major problems modeling the NGP region and the western US in general. One theory is that artificial irrigation, which is much more common in the western half of the country (see Figure 29), increases the amount of habitat available for key mosquito vectors, thereby increasing the likelihood of disease transmission. Irrigation data was not directly included in this study, although it was hoped that the combination of NDVI and land cover information would capture

some of the same patterns. Explicitly including irrigation data in future models may improve results.

It was an assumption of this study (one well supported by the literature) that a national-scale predictive model would invariably perform poorer for specific regions than smaller region-specific models would, owing to the significant regional variations in environmental conditions across the US. The results of the NGP model showed that assumption to be unfounded in this case. The national model R1c, which boasted an overall correlation coefficient of 0.86, dropped to about 0.7 when looking only at the NGP region, while the region-specific NGPa model only mustered a 0.5 correlation coefficient. As much as possible, all other variables were held constant, indicating the region-specific NGP model was a worse predictor of its own region than the national R1c model was. The implication seems to be that this region is subject to some unknown confounding variable(s) that the models were not equipped to predict. Further exploration is needed to determine the exact nature of the interference and what variables might be responsible.

5.1. Summary of Research Questions

This study sought answers to four main research questions: 1) can remotely sensed environmental variables be used to predict WNV incidence rates across the continental US, 2) is prior-year precipitation a better predictor than concurrent year precipitation, 3) is a single national model accurate enough, or is regional variation too strong, requiring smaller region-specific models, and 4) can the small numbers problem be mitigated by estimating WNV incidence from neuroinvasive cases to compensate for underreporting?

With correlation coefficients as high as 0.8, the answer to the first research question is yes, the chosen environmental variables of temperature, precipitation, elevation, NDVI, and land

cover are related to WNV incidence strongly enough to allow predictive modeling. The machine learning techniques employed were able to identify complex relationships between the data, and in the case of the R1c model explained approximately 86% of the observed real-world data.

The second question, as to whether prior year or concurrent year precipitation is a stronger predictor of WNV, cannot be answered at this time. Results were inconclusive and more research will need to be done on this topic to support or refute the claim authoritatively.

The question of whether or not a national model is appropriate across the highly diverse study area of the continental US is harder to answer. Again pointing to models R1a and c, it is tempting to conclude that a national model is effective, however the repeated pattern of model errors clustering spatially in the northern Great Plains region and elsewhere indicates the model is not appropriate for all regions. That said, the follow up model NGP showed that region-specific models may not in fact produce better results than the national model. Owing to this last finding, it was deemed prudent to conclude that a national model is appropriate, as long as it is interpreted with the knowledge of its regional biases and shortcomings.

Finally, as to the novel method of mitigating the small numbers problem with WNV data, this study showed clearly that it was not effective. Not only did the estimated WNV models consistently perform much poorer than their raw WNV counterparts, but their errors were spread over a larger range of county population values, in effect amplifying the small numbers problem instead of mitigating it. The technique here employed is therefore deemed a failure, and its further use is discouraged.

5.2. Limitations and Areas for Future Study

This study was subject to a number of limitations, most notably those associated with the necessity of using counties as the basic study unit and the small numbers problem. Finer

resolution data would be expected to yield better results, but it was not available to the author at the time of this study. The small numbers problem, as discussed previously, is a persistent problem with studies of this nature, and while the mitigation technique here employed was unsuccessful, it is hoped that other techniques might yet be developed to help lessen its impact.

Other areas for improvement and future research include incorporating spatial information explicitly into the predictive model. It has been suggested that simply including latitude and longitude coordinates of county centroids might allow machine learning techniques like Cubist to identify simple spatial patterns in the data. More complicated methods of including topological relationships (perhaps a county-neighbor weights matrix or something similar) might also yield interesting results. It should also be noted that Cubist is only one machine learning program and many other programs and techniques exist, including neural networks, which might be shown to better model the relationships between the environment and WNV risk.

Finally there is the obvious need to identify the confounding variable(s) at work in the NGP region and in the western US in general. Possible culprits include the amount of artificial irrigation (much more common in the western US) which provides excellent mosquito habitat, or perhaps different mosquito vectors or avian hosts which may prefer different environmental conditions. The next major step in this research, once some of the bugs are worked out of the model, would be to incorporate it into a spatial decision support system (SDSS) for use by researchers and public health officials with an interest in early warning detection of areas at high-risk for WNV disease.

REFERENCES

Abler, R. F. (1987). What shall we say? To whom shall we speak? *Annals of the Association of American Geographers*, *77*(4), 511–524.

Allen, T. R., & Wong, D. W. (2006). Exploring GIS, spatial statistics and remote sensing for risk assessment of vector-borne diseases: a West Nile virus example. *International Journal of Risk Assessment and Management*, *6*(4/5/6), 253–275.

Altonen, B. (2002). Historical Disease Maps. *Medical Geography and Disease Surveillance*. Retrieved April 18, 2012, from http://brianaltonenmph.com/gis/historical-disease-maps/

American Geographical Society. (1944). A Proposed Atlas of Diseases. *Geographical Review*, *34*(4), 642–652.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis*, *27*(2), 93–115.

Anyamba, A., Chretien, J. P., Small, J., Tucker, C. J., Formenty, P. B., Richardson, J. H., … Linthicum, K. J. (2009). Prediction of a Rift Valley fever outbreak. *Proceedings of the National Academy of Sciences*, *106*(3), 955–959.

Apperson, C. S., Hassan, H. K., Harrison, B. A., Savage, H. M., Aspen, S. E., Farajollahi, A., … others. (2004). Host feeding patterns of established and potential mosquito vectors of West Nile virus in the eastern United States. *Vector-Borne and Zoonotic Diseases*, *4*(1), 71–82.

Aragon, T. J. (2012). *epitools: Epidemiology Tools.* Retrieved from http://CRAN.R-project.org/package=epitools

ArcGIS Desktop. (2012). Redlands, CA: Environmental Systems Research Institute. Retrieved from http://www.esri.com

Association of American Geographers. (2011). Health and Medical Geography. *AAG Knowledge Communities*. Retrieved April 20, 2012, from http://community.aag.org/AAG/Directory/CommunityDetails/?CommunityKey=740e8665-89fc-4821-bfbc-de983145c0be

Beck, L. R., Rodriguez, M. H., Dister, S. W., Rodriguez, A. D., Rejmankova, E., Ulloa, A., … Legters, L. J. (1994). Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *The American Journal of Tropical Medicine and Hygiene*, *51*(3), 271–280.

Biggerstaff, B. J., & Petersen, L. R. (2003). Estimated risk of transmission of the West Nile virus through blood transfusion in the US, 2002. *Transfusion*, *43*(8), 1007–1017.

Bowden, S. E., Magori, K., & Drake, J. M. (2011). Regional Differences in the Association Between Land Cover and West Nile Virus Disease Incidence in Humans in the United

States. *American Journal of Tropical Medicine and Hygiene*, *84*(2), 234–238. doi:10.4269/ajtmh.2011.10-0134

Brody, H., Rip, M. R., Vinten-Johansen, P., Paneth, N., & Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *Lancet (London, England)*, *356*(9223), 64–68.

Brown, H., Duik-Wasser, M., Andreadis, T., & Fish, D. (2008). Remotely-sensed vegetation indices identify mosquito clusters of West Nile Virus vectors in an urban landscape in the northeastern United States. *Vector-Borne and Zoonotic Diseases*, *8*(1), 197–206. doi:10.1089/vbz.2007.0154

Brown, H. E., Childs, J. E., Diuk-Wasser, M. A., & Fish, D. (2008). Ecological Factors Associated with West Nile Virus Transmission, Northeastern United States. *Emerging Infectious Diseases*, *14*(10), 1539–1545. doi:10.3201/eid1410.071396

Brown, J. A., Factor, D. L., Tkachenko, N., Templeton, S. M., Crall, N. D., Pape, W. J., … Marfin, A. A. (2007). West Nile Viremic Blood Donors and Risk Factors for Subsequent West Nile Fever. *Vector-Borne and Zoonotic Diseases*, *7*(4), 479–488. doi:10.1089/vbz.2006.0611

Brown, T., & Moon, G. (2004). From Siam to New York: Jacques May and the "foundation" of medical geography. *Journal of Historical Geography*, *30*(4), 747–763.

Burton, I. (1963). The quantitative revolution and theoretical geography. *Canadian Geographer*, *7*(4), 151–162. doi:10.1111/j.1541-0064.1963.tb00796.x

Campbell, G. L., Marfin, A. A., Lanciotti, R. S., & Gubler, D. J. (2002). West Nile virus. *The Lancet Infectious Diseases*, *2*(9), 519–529.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An Overview of Machine Learning. In R. S. Michalski, J. Carbonell, & T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach* (pp. 3–23). Palo Alto, California: TIOGA Publishing Co. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.18.5035&type=ab

CDC. (2003). *Epidemic/Epizootic West Nile Virus in the United States: Guidelines for Surveillance, Prevention, and Control* (No. 3rd Revision). Fort Collins, CO: Centers for Disease Control and Prevention.

CDC. (2010). Our History - Our Story. *About CDC*. Retrieved April 18, 2012, from http://www.cdc.gov/about/history/ourstory.htm

CDC. (2012a). *2012 Case Definitions: Nationally Notifiable Conditions Infectious and Non-Infectious Case*. Atlanta, GA.

CDC. (2012b, October). CDC West Nile Virus Homepage. Retrieved October 22, 2012, from http://www.cdc.gov/ncidod/dvbid/westnile/index.htm

Census Bureau Homepage. (2013). *United States Census Bureau*.

Ciota, A. T., Lovelace, A. O., Jia, Y., Davis, L. J., Young, D. S., & Kramer, L. D. (2008). Characterization of mosquito-adapted West Nile virus. *Journal of General Virology*, *89*(7), 1633–1642. doi:10.1099/vir.0.2008/000893-0

Cline, B. L. (1970). New eyes for epidemiologists aerial photography and other remote sensing techniques. *American Journal of Epidemiology*, *92*(2), 85–89.

Congalton, R. G. (2010). Remote Sensing: An Overview. *GIScience & Remote Sensing*, *47*(4), 443–459. doi:10.2747/1548-1603.47.4.443

Cooke, W. H., Grala, K., & Wallis, R. C. (2006). Avian GIS models signal human risk for West Nile virus in Mississippi. *International Journal of Health Geographics*, *5*(1), 36.

CSTE. (2012). *CSTE List of Nationally Notifiable Conditions*. Atlanta, GA: Council of State and Territorial Epidemiologists. Retrieved from http://www.cste.org/webpdfs/CSTENotifiableConditionListAugust2012.pdf

Cubist. (2012). RuleQuest Research. Retrieved from www.rulequest.com

Dambach, P., Machault, V., Lacaux, J. P., Vignolles, C., Sié, A., & Sauerborn, R. (2012). Utilization of combined remote sensing techniques to detect environmental variables influencing malaria vector densities in rural West Africa. *International Journal of Health Geographics*, *11*(1), 8.

DeGroote, J. P., & Sugumaran, R. (2012). National and Regional Associations Between Human West Nile Virus Incidence and Demographic, Landscape, and Land Use Conditions in the Coterminous United States. *Vector-Borne and Zoonotic Diseases*, *12*(8), 657–665. doi:10.1089/vbz.2011.0786

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87.

Eisen, L., & Eisen, R. J. (2011). Using Geographic Information Systems and Decision Support Systems for the Prediction, Prevention, and Control of Vector-Borne Diseases. *Annual Review of Entomology*, *56*(1), 41–61. doi:10.1146/annurev-ento-120709-144847

Eisen, R. J., & Eisen, L. (2008). Spatial Modeling of Human Risk of Exposure to Vector-Borne Pathogens Based on Epidemiological Versus Arthropod Vector Data. *Journal of Medical Entomology*, *45*(2), 181–192. doi:10.1603/0022-2585(2008)45[181:SMOHRO]2.0.CO;2

Foody, G. M. (2006). GIS: health applications. *Progress in Physical Geography*, *30*(5), 691–695. doi:10.1177/0309133306071152

Friis, R. H., & Sellers, T. A. (2009). *Epidemiology for public health practice*. Sudbury, Mass.: Jones and Bartlett Publishers.

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., … Wickham, J. (2011). Completion of the 2006 National Land Cover Database for the Conterminous United States. *PE&RS*, *77*(9), 858–864.

Gates, M. C., & Boston, R. C. (2009). Irrigation linked to a greater incidence of human and veterinary West Nile virus cases in the United States from 2004 to 2006. *Preventive Veterinary Medicine*, *89*(1/2), 134–137.

Gehlke, C. E., & Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, *29*(185), 169–170.

Gibbs, S. E. J., Marlenee, N. L., Romines, J., Kavanaugh, D., Corn, J. L., & Stallknecht, D. E. (2006). Antibodies to West Nile virus in feral swine from Florida, Georgia, and Texas, USA. *Vector-Borne & Zoonotic Diseases*, *6*(3), 261–265.

Glass, G. E., Schwartz, B. S., Morgan III, J. M., Johnson, D. T., Noy, P. M., & Israel, E. (1995). Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health*, *85*(7), 944–948.

Glass, W. G., Lim, J. K., Cholera, R., Pletnev, A. G., Gao, J. L., & Murphy, P. M. (2005). Chemokine receptor CCR5 promotes leukocyte trafficking to the brain and survival in West Nile virus infection. *The Journal of experimental medicine*, *202*(8), 1087–1098.

Goodchild, M. F. (1992). Analysis. In R. F. Abler, M. G. Marcus, & J. M. Olson (Eds.), *Geography's Inner Worlds: Pervasive Themes in Contemporary American Geography* (1st ed., pp. 138–162). New Brunswick, New Jersey: Rutgers University Press.

Grammaticos, P. C., & Diamantis, A. (2008). Useful known and unknown views of the father of modern medicine, Hippocrates and his teacher Democritus. *Hellenic Journal of Nuclear Medicine*, *11*(1), 2–4.

Harrigan, R. J., Thomassen, H. A., Buermann, W., Cummings, R. F., Kahn, M. E., & Smith, T. B. (2010). Economic Conditions Predict Prevalence of West Nile Virus. (W. M. Getz, Ed.)*PLoS ONE*, *5*(11), e15437. doi:10.1371/journal.pone.0015437

Hay, S. I. (2000). An overview of remote sensing and geodesy for epidemiology and public health application. *Advances in Parasitology*, *47*, 1–35.

Hayes, C. G. (2001). West Nile virus: Uganda, 1937, to New York City, 1999. *Annals of the New York Academy of Sciences*, *951*(1), 25–37.

Hayes, E. B., Komar, N., Nasci, R. S., Montgomery, S. P., O'Leary, D. R., & Campbell, G. L. (2005). Epidemiology and transmission dynamics of West Nile virus disease. *Emerging Infectious Diseases*, *11*(8), 1167–1173.

Hayes, R. O., Maxwell, E. L., Mitchell, C. J., & Woodzick, T. L. (1985). Detection, identification, and classification of mosquito larval habitats using remote sensing

scanners in earth-orbiting satellites. *Bulletin of the World Health Organization*, *63*(2), 361–374.

Hippocrates. (1849). *The genuine works of Hippocrates*. (F. Adams, Trans.). New York: Printed for the Sydenham society.

Ho, D. (2012). *Notepad++*. Retrieved from www.notepad-plus-plus.org

Huntington, E., & Cushing, S. W. (1922). *Principles of human geography*. Wiley.

James, P. E., & Jones, C. F. (1954). *American Geography: Inventory & Prospect*. Syracuse University Press.

Janusz, K. B., Lehman, J. A., Panella, A. J., Fischer, M., & Staples, E. (2011). Laboratory Testing Practices for West Nile Virus in the United States. *Vector-Borne and Zoonotic Diseases*, *11*(5), 597–599. doi:10.1089/vbz.2010.0058

Jensen, J. R., Hodgson, M. E., Garcia-Quijano, M., Im, J., & Tullis, J. A. (2009). A remote sensing and GIS-assisted spatial decision support system for hazardous waste site monitoring. *Photogrammetric Engineering and Remote Sensing*, *75*(2), 169–177.

Jensen, John R. (2005). *Introductory digital image processing : a remote sensing perspective*. Upper Saddle River, N.J.: Prentice Hall.

Jensen, John R. (2007). *Remote sensing of the environment : an earth resource perspective*. Upper Saddle River, NJ: Pearson Prentice Hall.

Johnston, A. K. (1856). *Physical Atlas* (2nd ed.). Edinburgh and London: William Blackwood and Sons.

Kilpatrick, A. M., Kramer, L. D., Jones, M. J., Marra, P. P., & Daszak, P. (2006). West Nile Virus Epidemics in North America Are Driven by Shifts in Mosquito Feeding Behavior. *PLoS Biology*, *4*(4), e82. doi:10.1371/journal.pbio.0040082

Kumar, D., Drebot, M. A., Wong, S. J., Lim, G., Artsob, H., Buck, P., & Humar, A. (2004). A seroprevalence study of West Nile virus infection in solid organ transplant recipients. *American Journal of Transplantation*, *4*(11), 1883–1888.

LaDeau, S. L., Calder, C. A., Doran, P. J., & Marra, P. P. (2010). West Nile virus impacts in American crow populations are associated with human land use and climate. *Ecological Research*, *26*, 909–916. doi:10.1007/s11284-010-0725-z

Lanciotti, R. S. (1999). Origin of the West Nile Virus Responsible for an Outbreak of Encephalitis in the Northeastern United States. *Science*, *286*(5448), 2333–2337. doi:10.1126/science.286.5448.2333

Landesman, W. J., Allan, B. F., Langerhans, R. B., Knight, T. M., & Chase, J. M. (2007). Inter-Annual Associations Between Precipitation and Human Incidence of West Nile Virus in

the United States. *Vector-Borne and Zoonotic Diseases*, *7*(3), 337–343. doi:10.1089/vbz.2006.0590

LibreOffice. (2013). Berlin, Germany: The Document Foundation. Retrieved from www.libreoffice.org

Lindsey, N. P., Kuhn, S., Campbell, G. L., & Hayes, E. B. (2008). West Nile Virus Neuroinvasive Disease Incidence in the United States, 2002–2006. *Vector-Borne and Zoonotic Diseases*, *8*(1), 35–40. doi:10.1089/vbz.2007.0137

Liu, H., Weng, Q., & Gaines, D. (2008). Spatio-temporal analysis of the relationship between WNV dissemination and environmental variables in Indianapolis, USA. *International journal of health geographics*, *7*(1), 66.

May, J. M. (1950). Medical Geography: Its Methods and Objectives. *Geographical Review*, *40*(1), 9–41. doi:10.2307/210990

McLean, R. G., Ubico, S. R., Docherty, D. E., Hansen, W. R., Sileo, L., & McNamara, T. S. (2001). West Nile virus transmission and ecology in birds. *Annals of the New York Academy of Sciences*, *951*(1), 54–57.

Meade, M. S., & Emch, M. (2010). *Medical Geography*, *Third Edition*. Guilford Press.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1), 17–23.

Mosby, I. (2009). arbovirus. (M. O'Toole, Ed.)*Mosby's medical dictionary*. St. Louis, Mo.: Mosby.

Mostashari, F., Bunning, M. L., Kitsutani, P. T., Singer, D. A., Nash, D., Cooper, M. J., … Fine, A. D. (2001). Epidemic West Nile encephalitis, New York, 1999: results of a household-based seroepidemiological survey. *The lancet*, *358*(9278), 261–264.

Mostashari, F., Kulldorff, M., Hartman, J. J., Miller, J. R., & Kulasekera, V. (2003). Dead bird clusters as an early warning system for West Nile virus activity. *Emerging infectious diseases*, *9*(6), 641.

NASA. (2001). Landscape Epidemiology and RS/GIS. Retrieved April 18, 2012, from http://geo.arc.nasa.gov/sge/health/landepi.html

NASA Land Processes Distributed Active Archive Center (LP DAAC). (2012). MODIS VI (MOD13A3) Data. USGS/Earth Resources Observation and Science (EROS) Center. Retrieved from https://lpdaac.usgs.gov/get_data

Nasci, R. S., Savage, H. M., White, D. J., Miller, J. R., Cropp, B. C., Godsey, M. S., … Lanciotti, R. S. (2001). West Nile virus in overwintering Culex mosquitoes, New York City, 2000. *Emerging Infectious Diseases*, *7*(4), 742.

Nash, D., Mostashari, F., Fine, A., Miller, J., O'Leary, D., Murray, K., … Layton, M. (2001). The outbreak of West Nile virus infection in the New York City area in 1999. *New England Journal of Medicine*, *344*(24), 1807–1814.

Numbers, R. L. (2000). Medical science before scientific medicine: reflections on the history of medical geography. *Medical History. Supplement*, (20), 217.

O'Leary, D. R., Marfin, A. A., Montgomery, S. P., Kipp, A. M., Lehman, J. A., Biggerstaff, B. J., … Campbell, G. L. (2004). The epidemic of West Nile virus in the United States, 2002. *Vector-borne and Zoonotic Diseases*, *4*(1), 61–70.

Oldstone, M. (1998). *Viruses, plagues, and history* (1st Ed.). New York: Oxford University Press.

Openshaw, S. (1984). Number 38, The Modifiable Areal Unit Problem. In *Concepts and Techniques in Modern Geography* (pp. 1–41). Norwich: Geo Books.

Oregon State University. (2012). PRISM Climate Group. *PRISM Climate Group*. Retrieved November 2, 2012, from http://www.prism.oregonstate.edu/

Ostfeld, R., Glass, G., & Keesing, F. (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in Ecology & Evolution*, *20*(6), 328–336. doi:10.1016/j.tree.2005.03.009

Pavlovsky, E. N. (1965). *Natural nidality of transmissible diseases: with special reference to the landscape epidemiology of zooanthroponoses*. University of Illinois Press.

Petersen, L. R., & Roehrig, J. T. (2001). West Nile virus: a reemerging global pathogen. *Emerging Infectious Diseases*, *7*(4), 611–614.

Peterson, A. T. (2006). Ecologic niche modeling and spatial patterns of disease transmission. *Emerging Infectious Diseases*, *12*(12), 1822.

Peterson, A. Townsend, Pereira, R. S., & De Camargo Neves, V. F. (2004). Using epidemiological survey data to infer geographic distributions of leishmaniasis vector species. *Revista da sociedade Brasileira de Medicina Tropical*, *37*(1), 10–14.

Peterson, A. Townsend, Robbins, A., Restifo, R., Howell, J., & Nasci, R. (2008). Predictable ecology and geography of West Nile virus transmission in the central United States. *Journal of Vector Ecology*, *33*(2), 342–352.

Pringle, D. G. (1995). Mapping Disease Risk Estimates Based on Small Numbers: An Assessment of Empirical Bayes Techniques. *The Economic and Social Review*, *27*(4), 341–363.

Python. (2013). Python Software Foundation. Retrieved from www.python.org

Quattrochi, D. A., Walsh, S. J., Jensen, J. R., & Ridd, M. K. (2004). Remote Sensing. In G. L. Gaile & C. J. Willmott (Eds.), *Geography in America at the Dawn of the 21st Century* (pp. 376–416). New York: Oxford University Press.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81–106.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Machine Learning: Proceedings of the Tenth International Conference* (pp. 236–243). Retrieved from http://cs.ecs.baylor.edu/~hamerly/courses/5325_11s/papers/ibl/quinlan1993combining.pdf

R Development Core Team. (2012). *R*. Vienna, Austria: The R Foundation for Statistical Computing. Retrieved from www.r-project.org

Rappole, J. H., Compton, B. W., Leimgruber, P., Robertson, J., King, D. I., & Renner, S. C. (2006). Modeling movement of West Nile virus in the Western hemisphere. *Vector-borne & zoonotic diseases*, *6*(2), 128–139.

Rappole, J. H., Derrickson, S. R., & Hubálek, Z. (2000). Migratory birds and spread of West Nile virus in the Western Hemisphere. *Emerging Infectious Diseases*, *6*(4), 319–328.

Rappole, J., & Hubalek, Z. (2003). Migratory birds and West Nile virus. *Journal of Applied Microbiology*, *94*, 47–58.

Reisen, W., & Brault, A. C. (2007). West Nile virus in North America: perspectives on epidemiology and intervention. *Pest Management Science*, *63*(7), 641–646. doi:10.1002/ps.1325

Reisen, W.K. (2010). Landscape epidemiology of vector-borne diseases. *Annual Review of Entomology*, *55*, 461–483. doi:10.1146/annurev-ento-112408-085419

Reisen, William K., Fang, Y., & Martinez, V. M. (2006). Effects of Temperature on the Transmission of West Nile Virus by Culex tarsalis (Diptera: Culicidae). *Journal of Medical Entomology*, *43*(2), 309–317. doi:10.1603/0022-2585(2006)043[0309:EOTOTT]2.0.CO;2

Renneboog, N., Gathings, D., Hemmings, S., Makasa, E., Omer, W., Tipre, M., … Luvall, J. C. (2009). *Spatial Analysis of West Nile Virus: Predictive Risk Modeling of a Vector-Borne Infectious Disease in Illinois by means of NASA Earth Observation Systems*. Retrieved from http://science1.nasa.gov/media/medialibrary/2010/03/31/Poster_Spring_2009_Final-DEVELOP_optimized20ppt2_.pdf

Rochlin, I., Turbow, D., Gomez, F., Ninivaggi, D. V., & Campbell, S. R. (2011). Predictive Mapping of Human Risk for West Nile Virus (WNV) Based on Environmental and Socioeconomic Factors. (W. M. Getz, Ed.)*PLoS ONE*, *6*(8), e23280. doi:10.1371/journal.pone.0023280

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). Monitoring vegetation systems in the Great Plains with ERTS (Vol. 1, pp. 309–317). Presented at the 3rd Earth Resource Technology Sattelite (ERTS) Symposium, NASA. Goddard Space Flight Center.

Ruiz, M. O., Chaves, L. F., Hamer, G. L., Sun, T., Brown, W. M., Walker, E. D., … Kitron, U. D. (2010). Local impact of temperature and precipitation on West Nile virus infection in Culex species mosquitoes in northeast Illinois, USA. *Parasites & Vectors*, *3*(1), 19.

Ruiz, M. O., Tedesco, C., McTighe, T. J., Austin, C., & Kitron, U. (2004). Environmental and social determinants of human risk during a West Nile virus outbreak in the greater Chicago area, 2002. *International Journal of Health Geographics*, *3*(1), 8.

Ruiz, M., Walker, E., Foster, E., Haramis, L., & Kitron, U. (2007). Association of West Nile virus illness and urban landscapes in Chicago and Detroit. *International Journal of Health Geographics*, *6*(1), 10–21. doi:10.1186/1476-072X-6-10

RuleQuest Research. (2012). RuleQuest Research Data Mining Tools. Retrieved September 7, 2012, from http://www.rulequest.com/

Rushton, G. (2003). Public Health, GIS, and Spatial Analytic Tools. *Annual Review of Public Health*, *24*(1), 43–56. doi:10.1146/annurev.publhealth.24.012902.140843

Sejvar, J. J., Haddad, M. B., Tierney, B. C., Campbell, G. L., Marfin, A. A., Van Gerpen, J. A., … Petersen, L. R. (2003). Neurologic manifestations and outcome of West Nile virus infection. *JAMA: The Journal of the American Medical Association*, *290*(4), 511–515. doi:10.1001/jama.290.4.511

Sejvar, J. J., Lindsey, N. P., & Campbell, G. L. (2011). Primary causes of death in reported cases of fatal West Nile Fever, United States, 2002–2006. *Vector-Borne and Zoonotic Diseases*, *11*(2), 161–164.

Shaman, J. (2009, February 20). *Climate change and arbovirus disease transmission*. Presented at the Tenth National Conference on West Nile Virus in the United States, Savannah, GA. Retrieved from http://www.cdc.gov/ncidod/dvbid/westnile/conf/February_2009.htm

Shaman, J., Day, J. F., & Stieglitz, M. (2002). Drought-Induced Amplification of Saint Louis encephalitis virus, Florida. *Emerging Infectious Diseases*, *8*(6), 575–580. doi:10.3201/eid0806.010417

Shaman, J., Day, J. F., & Stieglitz, M. (2005). Drought-Induced Amplification and Epidemic Transmission of West Nile Virus in Southern Florida. *Journal of Medical Entomology*, *42*(2), 134–141. doi:10.1603/0022-2585(2005)042[0134:DAAETO]2.0.CO;2

Shuchman, R. A., Malinas, N. P., & Edson, R. (2002). The Role of Remote Sensing and GIS for Impact Modeling and Risk Assessment of Vector Borne Diseases. In *Proceedings for the twenty-ninth International Symposium on Remote Sensing of Environment* (Vol. 29, p. 305). Retrieved from

http://www.ulrmc.org.ua/publication/ecology/GIS%20and%20RS%20in%20Assesment%20of%20Vector%20Borne%20Diseases_eng.pdf

Snow, J. (1855). *On the Mode of Communication of Cholera* (2nd ed.). London: John Churchill.

Solano, R., Didan, K., Jacobson, A., & Huete, D. K. (2010). MODIS Vegetation Index User's Guide. *NASA GSFC, 2.00*. Retrieved from http://nrm.salrm.uaf.edu/~dverbyla/MODIS_Land_Products/pdf_documents/users_guides/snow_users_guide.pdf

Solomonoff, R. J. (1956). *An Inductive Inference Machine* (Privately Circulated). New York City: Technical Research Group.

Soverow, J. E., Wellenius, G. A., Fisman, D. N., & Mittleman, M. A. (2009). Infectious disease in a warming world: how weather influenced West Nile virus in the United States (2001–2005). *Environmental health perspectives*, *117*(7), 1049.

Staples, J. E. (2009, February 19). *WNV surveillance and epidemiology - United States and tropical Americas*. Presented at the Tenth National Conference on West Nile Virus in the United States, Savannah, GA. Retrieved from http://www.cdc.gov/ncidod/dvbid/westnile/conf/February_2009.htm

Sugumaran, R., Larson, S. R., & DeGroote, J. P. (2009). Spatio-temporal cluster analysis of county-based human West Nile virus incidence in the continental United States. *International journal of health geographics*, *8*(1), 43.

Swatantran, A., Dubayah, R., Goetz, S., Hofton, M., Betts, M. G., Sun, M., … Holmes, R. (2012). Mapping Migratory Bird Prevalence Using Remote Sensing Data Fusion. (G. J.-P. Schumann, Ed.)*PLoS ONE*, *7*(1), e28922. doi:10.1371/journal.pone.0028922

Trawinski, P. R., & Mackay, D. S. (2008). Spatial autocorrelation of West Nile virus vector mosquito abundance in a seasonally wet suburban environment. *Journal of Geographical Systems*, *11*, 67–87. doi:10.1007/s10109-008-0070-8

USGS. (2006). Shuttle Radar Topography Mission. *Global Land Cover Facility, University of Maryland*. Retrieved November 28, 2012, from http://glcf.umiacs.umd.edu/data/srtm/

USGS Water Resources of the United States. (2010). Estimated Use of Water in the United States, 2005. Reston, VA: National Atlas of the United States. Retrieved from nationalatlas.gov/atlasftp.html?openChapters=chpwater#chpwater

Wang, F. (2006). *Quantitative methods and applications in GIS*. Boca Raton, Fla.: CRC/Taylor & Francis.

Wang, F., Guo, D., & McLafferty, S. (2012). Constructing geographic areas for cancer data analysis: A case study on late-stage breast cancer risk in Illinois. *Applied Geography*, *35*(1-2), 1–11. doi:10.1016/j.apgeog.2012.04.005

Washino, R. K., & Wood, B. L. (1994). Application of remote sensing to vector arthropod surveillance and control. *American Journal of Tropical Medicine and Hygiene*, *50*(6 suppl), 134–144.

Watson, J. T., Pertel, P. E., Jones, R. C., Siston, A. M., Paul, W. S., Austin, C. C., & Gerber, S. I. (2004). Clinical characteristics and functional outcomes of West Nile fever. *Ann Intern Med*, *141*, 360–365.

Winters, A. M., Eisen, R. J., Lozano-Fuentes, S., Moore, C. G., Pape, W. J., & Eisen, L. (2008). Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado. *The American journal of tropical medicine and hygiene*, *79*(4), 581–590.

World Health Organization. (2007). Global Health Atlas. *World Health Organization*. Retrieved April 23, 2012, from http://apps.who.int/globalatlas/default.asp

World Health Organization. (2012). History of WHO. *World Health Organization*. Retrieved April 20, 2012, from www.who.int/about/history/en/index.html

Young, S. G., & Jensen, R. R. (2012). Statistical and visual analysis of human West Nile virus infection in the United States, 1999–2008. *Applied Geography*, *34*(0), 425–431. doi:10.1016/j.apgeog.2012.01.008

APPENDIX A - PREPROCESSING PYTHON SCRIPTS

As described in section 3.6.1, much of the data preprocessing was accomplished via Python scripts, namely the NDVI data from MODIS and the Temperature data from PRISM. The preprocessing scripts for these datasets are included here in abbreviated form.

A1.NDVI_PREP.PY

```
# ---------------------------------------------------------------------------
# NDVI_Prep.py
# Created on: 2013-01-21 16:21:52.00000
#    (generated by ArcGIS/ModelBuilder)
# Modified by: Sean Young
# Description:
#    Takes a year's worth of MODIS NDVI tiles one month at a time, mosaicks them
#    together, then aggregates the data to the county level using the Zonal
#    Statistics as Table tool from the Spatial Analyst toolbox.
# ---------------------------------------------------------------------------

#-----------------------------------------------------
# Import, Set Product Code, and Check Out Extension
#    Note: glob used for file searching functionality
#-----------------------------------------------------
print "Loading..."
import arceditor
import arcpy
import glob
arcpy.CheckOutExtension("spatial")

#-------------------------
# User-defined Variables:
#-------------------------
yr = "03"
MODIS_tilepath = "D:\\GIS\\sgyoung\\ThesisData\\MODIS\\"
output_gdb = "D:\\GIS\\sgyoung\\ThesisData\\MODISData.gdb"
counties = "D:\\GIS\\sgyoung\\ThesisData\\MyData.gdb\\CONUS_counties_dtl"

#------------------
# Other Variables:
#------------------
mmmyy = "jan"+yr
acqdate = "A20"+yr+"001"
mmmyy2 = "feb"+yr
acqdate2 = "A20"+yr+"032"
mmmyy3 = "mar"+yr
acqdate3 = "A20"+yr+"061"
mmmyy4 = "apr"+yr
acqdate4 = "A20"+yr+"092"
mmmyy5 = "may"+yr
acqdate5 = "A20"+yr+"122"
mmmyy6 = "jun"+yr
```

```
acqdate6 = "A20"+yr+"153"
mmmyy7 = "jul"+yr
acqdate7 = "A20"+yr+"183"
mmmyy8 = "aug"+yr
acqdate8 = "A20"+yr+"214"
mmmyy9 = "sep"+yr
acqdate9 = "A20"+yr+"245"
mmmyy10 = "oct"+yr
acqdate10 = "A20"+yr+"275"
mmmyy11 = "nov"+yr
acqdate11 = "A20"+yr+"306"
mmmyy12 = "dec"+yr
acqdate12 = "A20"+yr+"336"
mosaic_datuminfo =
"GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563
]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]];-400 -400 1000000000;-
100000 10000;-100000 10000;8.98315284119522E-09;0.001;0.001;IsHighPrecision"

#----------------
# Processing Jan
#----------------
print "Starting "+mmmyy

NDVI_Mosaic = "NDVI_" + mmmyy
NDVI_Mosaic_Path = output_gdb+"\\"+NDVI_Mosaic
ndvi_stat = output_gdb+"\\ndvi_"+mmmyy+"_stat"

#----------------------
# Locating MODIS Tiles
#----------------------
print "Finding Tiles..."
h07v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h07v05.005.*.hdf")[0]
h08v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h08v04.005.*.hdf")[0]
h08v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h08v05.005.*.hdf")[0]
h08v06 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h08v06.005.*.hdf")[0]
h09v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h09v04.005.*.hdf")[0]
h09v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h09v05.005.*.hdf")[0]
h09v06 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h09v06.005.*.hdf")[0]
h10v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h10v04.005.*.hdf")[0]
h10v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h10v05.005.*.hdf")[0]
h10v06 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h10v06.005.*.hdf")[0]
h11v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h11v04.005.*.hdf")[0]
h11v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h11v05.005.*.hdf")[0]
h12v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h12v04.005.*.hdf")[0]
h12v05 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h12v05.005.*.hdf")[0]
h13v04 = glob.glob(MODIS_tilepath+"MOD13A3."+acqdate+".h13v04.005.*.hdf")[0]
allTiles =
h07v05+";"+h08v04+";"+h08v05+";"+h08v06+";"+h09v04+";"+h09v05+";"+h09v06+";"+h10v04+"
;"+h10v05+";"+h10v06+";"+h11v04+";"+h11v05+";"+h12v04+";"+h12v05+";"+h13v04

#--------------------------------
# Process: Create Mosaic Dataset
#--------------------------------
print "Creating Mosaic..."
```

```
arcpy.CreateMosaicDataset_management(output_gdb, NDVI_Mosaic, mosaic_datuminfo, "",
"", "NONE", "")

#---------------------------------------
# Process: Add Rasters To Mosaic Dataset
#---------------------------------------
print "Adding tiles to mosaic..."
arcpy.AddRastersToMosaicDataset_management(NDVI_Mosaic_Path, "Raster Dataset",
allTiles, "UPDATE_CELL_SIZES", "UPDATE_BOUNDARY", "NO_OVERVIEWS", "", "0", "1500",
"", "", "SUBFOLDERS", "EXCLUDE_DUPLICATES", "NO_PYRAMIDS", "NO_STATISTICS",
"NO_THUMBNAILS", "", "NO_FORCE_SPATIAL_REFERENCE")

#-----------------------------------
# Process: Zonal Statistics as Table
#-----------------------------------
print "Calculating statistics..."
arcpy.gp.ZonalStatisticsAsTable_sa(counties, "FIPS", NDVI_Mosaic_Path, ndvi_stat,
"DATA", "MEAN")

print "Done with "+mmmyy

#----------------
# Processing Feb
#----------------
print "Starting "+mmmyy2
.
.
.
.
print "Done with "+mmmyy12

print “Success!”
```

A2.NDVI_TABLEMELTER.PY

```
# ---------------------------------------------------------------------------
# NDVI_TableMelter.py
# Created on: 2013-01-22 15:15:13.00000
# Modified by: Sean Young
# Description:
#     Takes one year of NDVI monthly data tables, after mosaicking and aggregation
#     in NDVI_Prep.py and "melts" them together into one table.
# ---------------------------------------------------------------------------

#---------------------
# Import arcpy module
#---------------------
print "Loading..."
import arcpy

#------------------------
# User-defined Variables:
#------------------------
yr = "03"
modispath = "D:\\GIS\\sgyoung\\ThesisData\\MODISData.gdb"
outPath = "D:\\GIS\\sgyoung\\ThesisData\\TestOutput.gdb"

#-----------------
# Other Variables:
#-----------------
deletables = "ZONE_CODE;COUNT;AREA;MIN;MAX"
melters = "MEAN"
mainTable = outPath+"\\ndvi_"+yr
temp02 = modispath+"\\ndvi_feb"+yr+"_stat"
temp03 = modispath+"\\ndvi_mar"+yr+"_stat"
temp04 = modispath+"\\ndvi_apr"+yr+"_stat"
temp05 = modispath+"\\ndvi_may"+yr+"_stat"
temp06 = modispath+"\\ndvi_jun"+yr+"_stat"
temp07 = modispath+"\\ndvi_jul"+yr+"_stat"
temp08 = modispath+"\\ndvi_aug"+yr+"_stat"
temp09 = modispath+"\\ndvi_sep"+yr+"_stat"
temp10 = modispath+"\\ndvi_oct"+yr+"_stat"
temp11 = modispath+"\\ndvi_nov"+yr+"_stat"
temp12 = modispath+"\\ndvi_dec"+yr+"_stat"

#------------------------
# Process: Copy mainTable
#------------------------
print "Copying Jan"+yr+" table to new location"
arcpy.Copy_management(modispath+"\\ndvi_jan"+yr+"_stat",mainTable,"Table")
print "Copied to "+modispath+" successfully."

#------------------------
# Process: Delete Fields
#------------------------
print "Deleting Extra Fields..."
arcpy.DeleteField_management(mainTable, deletables)
```

```
print "Extraneous fields eradicated with extreme prejudice."

#----------------------
# Process: Join Fields
#----------------------
print "Melting tables together..."
arcpy.JoinField_management(mainTable, "FIPS", temp02, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp03, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp04, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp05, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp06, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp07, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp08, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp09, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp10, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp11, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp12, "FIPS", melters)

print "Table melting of "+mainTable+" complete."
```

A3. TEMP_PREP.PY

```
# ---------------------------------------------------------------------------
# Temp_Prep.py
# Created on: 2013-01-21 10:27:48.00000
#     (generated by ArcGIS/ModelBuilder)
# Modified by: Sean Young
# Description:
#     Takes one year of monthly TMax and TMin PRISM Data, projects the data, then
#     aggregates to county level using Zonal Statistics as Table tool from Spatial
#     Analyst toolbox.
# ---------------------------------------------------------------------------

#-------------------------------
# Import and Check Out Extension
#-------------------------------
print “Loading…”
import arcpy
arcpy.CheckOutExtension("spatial")

#-------------------------
# User-defined Variables:
#-------------------------
yr = "03"
inpath = "D:\\GIS\\sgyoung\\ThesisData\\PRISM\\20"+yr
outpath = "D:\\GIS\\sgyoung\\ThesisData\\PRISMData.gdb"
cnty = "D:\\GIS\\sgyoung\\ThesisData\\MyData.gdb\\CONUS_counties_dtl"

#------------------
# Other Variables:
#------------------
tmin_01 = "us_tmin_20"+yr+".01"
tmin_02 = "us_tmin_20"+yr+".02"
tmin_03 = "us_tmin_20"+yr+".03"
tmin_04 = "us_tmin_20"+yr+".04"
tmin_05 = "us_tmin_20"+yr+".05"
tmin_06 = "us_tmin_20"+yr+".06"
tmin_07 = "us_tmin_20"+yr+".07"
tmin_08 = "us_tmin_20"+yr+".08"
tmin_09 = "us_tmin_20"+yr+".09"
tmin_10 = "us_tmin_20"+yr+".10"
tmin_11 = "us_tmin_20"+yr+".11"
tmin_12 = "us_tmin_20"+yr+".12"
tmin_14 = "us_tmin_20"+yr+".14"
tmax_01 = "us_tmax_20"+yr+".01"
tmax_02 = "us_tmax_20"+yr+".02"
tmax_03 = "us_tmax_20"+yr+".03"
tmax_04 = "us_tmax_20"+yr+".04"
tmax_05 = "us_tmax_20"+yr+".05"
tmax_06 = "us_tmax_20"+yr+".06"
tmax_07 = "us_tmax_20"+yr+".07"
tmax_08 = "us_tmax_20"+yr+".08"
tmax_09 = "us_tmax_20"+yr+".09"
tmax_10 = "us_tmax_20"+yr+".10"
```

```
tmax_11 = "us_tmax_20"+yr+".11"
tmax_12 = "us_tmax_20"+yr+".12"
tmax_14 = "us_tmax_20"+yr+".14"
tmin01_stat = outpath + "\\tmin"+yr+"01_stat"
tmin02_stat = outpath + "\\tmin"+yr+"02_stat"
tmin03_stat = outpath + "\\tmin"+yr+"03_stat"
tmin04_stat = outpath + "\\tmin"+yr+"04_stat"
tmin05_stat = outpath + "\\tmin"+yr+"05_stat"
tmin06_stat = outpath + "\\tmin"+yr+"06_stat"
tmin07_stat = outpath + "\\tmin"+yr+"07_stat"
tmin08_stat = outpath + "\\tmin"+yr+"08_stat"
tmin09_stat = outpath + "\\tmin"+yr+"09_stat"
tmin10_stat = outpath + "\\tmin"+yr+"10_stat"
tmin11_stat = outpath + "\\tmin"+yr+"11_stat"
tmin12_stat = outpath + "\\tmin"+yr+"12_stat"
tmin_stat = outpath + "\\tmin"+yr+"_stat"
tmax01_stat = outpath + "\\tmax"+yr+"01_stat"
tmax02_stat = outpath + "\\tmax"+yr+"02_stat"
tmax03_stat = outpath + "\\tmax"+yr+"03_stat"
tmax04_stat = outpath + "\\tmax"+yr+"04_stat"
tmax05_stat = outpath + "\\tmax"+yr+"05_stat"
tmax06_stat = outpath + "\\tmax"+yr+"06_stat"
tmax07_stat = outpath + "\\tmax"+yr+"07_stat"
tmax08_stat = outpath + "\\tmax"+yr+"08_stat"
tmax09_stat = outpath + "\\tmax"+yr+"09_stat"
tmax10_stat = outpath + "\\tmax"+yr+"10_stat"
tmax11_stat = outpath + "\\tmax"+yr+"11_stat"
tmax12_stat = outpath + "\\tmax"+yr+"12_stat"
tmax_stat = outpath + "\\tmax"+yr+"_stat"
projection_info =
"GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563
]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]"
stats = "MIN_MAX_MEAN"

#-----------------------------
# Process: Define Projections
#-----------------------------
print “Defining projections…”
arcpy.DefineProjection_management(tmin_01, projection_info)
arcpy.DefineProjection_management(tmin_02, projection_info)
arcpy.DefineProjection_management(tmin_03, projection_info)
arcpy.DefineProjection_management(tmin_04, projection_info)
arcpy.DefineProjection_management(tmin_05, projection_info)
arcpy.DefineProjection_management(tmin_06, projection_info)
arcpy.DefineProjection_management(tmin_07, projection_info)
arcpy.DefineProjection_management(tmin_08, projection_info)
arcpy.DefineProjection_management(tmin_09, projection_info)
arcpy.DefineProjection_management(tmin_10, projection_info)
arcpy.DefineProjection_management(tmin_11, projection_info)
arcpy.DefineProjection_management(tmin_12, projection_info)
arcpy.DefineProjection_management(tmin_14, projection_info)
arcpy.DefineProjection_management(tmax_01, projection_info)
arcpy.DefineProjection_management(tmax_02, projection_info)
arcpy.DefineProjection_management(tmax_03, projection_info)
arcpy.DefineProjection_management(tmax_04, projection_info)
```

```
arcpy.DefineProjection_management(tmax_05, projection_info)
arcpy.DefineProjection_management(tmax_06, projection_info)
arcpy.DefineProjection_management(tmax_07, projection_info)
arcpy.DefineProjection_management(tmax_08, projection_info)
arcpy.DefineProjection_management(tmax_09, projection_info)
arcpy.DefineProjection_management(tmax_10, projection_info)
arcpy.DefineProjection_management(tmax_11, projection_info)
arcpy.DefineProjection_management(tmax_12, projection_info)
arcpy.DefineProjection_management(tmax_14, projection_info)

#-----------------------------------
# Process: Zonal Statistics as Table
#-----------------------------------
print “Aggregating and calculating statistics…”
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_01, tmin01_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_02, tmin02_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_03, tmin03_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_04, tmin04_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_05, tmin05_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_06, tmin06_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_07, tmin07_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_08, tmin08_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_09, tmin09_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_10, tmin10_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_11, tmin11_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_12, tmin12_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmin_14, tmin_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_01, tmax01_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_02, tmax02_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_03, tmax03_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_04, tmax04_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_05, tmax05_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_06, tmax06_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_07, tmax07_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_08, tmax08_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_09, tmax09_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_10, tmax10_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_11, tmax11_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_12, tmax12_stat, "DATA", stats)
arcpy.gp.ZonalStatisticsAsTable_sa(cnty, "FIPS", tmax_14, tmax_stat, "DATA", stats)

print “Success!”
```

A4.TEMP_TABLEMELTER.PY

```
# ---------------------------------------------------------------------------
# PRISM_TableMelter.py
# Created on: 2013-01-22 15:15:13.00000
# Modified by: Sean Young
# Description:
#    Takes one year of TMax or TMin PRISM tables, after projection and aggregation by
#    Temp_Prep.py and melts them all into one new table.
# ---------------------------------------------------------------------------

#---------------------
# Import arcpy module
#---------------------
print "Loading..."
import arcpy

#-------------------------
# User-defined Variables:
#-------------------------
yr = "08"
minORmax = "MIN"
inpath = "D:\\GIS\\sgyoung\\ThesisData\\PRISMData.gdb"
outPath = "D:\\GIS\\sgyoung\\ThesisData\\TempOutput.gdb"

#------------------
# Other Variables:
#------------------
if minORmax == "MAX":
    pathfill = "tmax"
    deletables1 = "ZONE_CODE;MIN"
    melters = "MAX;MEAN"
if minORmax == "MIN":
    pathfill = "tmin"
    deletables1 = "ZONE_CODE;MAX"
    melters = "MIN;MEAN"
prismpath = inpath+"\\"+pathfill+yr
mainTable = outPath+"\\"+pathfill+yr
temp02 = prismpath+"02_stat"
temp03 = prismpath+"03_stat"
temp04 = prismpath+"04_stat"
temp05 = prismpath+"05_stat"
temp06 = prismpath+"06_stat"
temp07 = prismpath+"07_stat"
temp08 = prismpath+"08_stat"
temp09 = prismpath+"09_stat"
temp10 = prismpath+"10_stat"
temp11 = prismpath+"11_stat"
temp12 = prismpath+"12_stat"
tempYR = prismpath+"_stat"

#-------------------------
# Process: Copy mainTable
#-------------------------
```

```
print "Copying Jan"+yr+" table to new location"
arcpy.Copy_management(prismpath+"01_stat",mainTable,"Table")
print "Copied to new location successfully."

#-----------------------
# Process: Delete Field
#-----------------------
print "Deleting Extra Fields..."
arcpy.DeleteField_management(mainTable, deletables1)
print "Extraneous fields eradicated with extreme prejudice."

#---------------------
# Process: Join Field
#---------------------
print "Melting tables together..."
arcpy.JoinField_management(mainTable, "FIPS", temp02, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp03, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp04, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp05, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp06, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp07, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp08, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp09, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp10, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp11, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", temp12, "FIPS", melters)
arcpy.JoinField_management(mainTable, "FIPS", tempYR, "FIPS", melters)

print "Table melting of "+mainTable+" complete."
```

APPENDIX B - CUBIST FILE FORMATS

Cubist (*Cubist*, 2012) requires a minimum of two input files for model creation, and outputs to at least two others. These files are text and csv formats, but are identified for use in Cubist with special extensions: .names and .data (inputs), .model and .pred (outputs), with two additional optional input files, .test and .cases. Examples of these file formats are presented below, as well as part of the console output from Cubist.

B1. NAMES (*.NAMES) FILE

The names file is the first required input file for Cubist, and it declares the names and types of all the data to be used in the machine learning decision tree generation. The Names file is a plain text (.txt) document, but the extension must be changed to ".names" in order for Cubist to recognize it and read it in properly. The syntax is very simple: comments are delineated with a vertical bar "|" and white space (multiple spaces or tabs) are treated as single spaces, allowing for formatting for user-readability without impacting the syntax of the file.

The target attribute (the variable to be predicted with the model) is always listed first, regardless of its position in the actual data file, but it must be listed again in place along with its type. All variables are then declared in the order in which they occur in the data table, followed by a colon ":", along with a data type, or the instruction for Cubist to ignore that particular attribute. Recognized data types include continuous (used for most numeric attributes), discrete (identified as a list of possible nominal values separated by commas and terminating with a period), date, time, timestamp, and label, and they all must end with a period. The label variable is used to identify cases and is not used in model generation. Cubist can also use implicitly-defined attribute values using simple formulas that can involve any previously defined attribute variable.

Below is an abbreviated version of the names file for R1a:

```
|*************************************************************
| R1a - Averages of all 6 years, used to predict Raw WNV Rates
|*************************************************************

| Target goes first, without declaration of type:
RR38.                     | Target

| List of variables in order comes next, with type declared:
|***County Info***
OID:          ignore.     | Object ID from ArcMap
State:        ignore.     | Two letter state code
County:       ignore.     | County name
FIPS:         label.      | Federal Information Processing Standard Code
|***WNV***
RR38:         continuous. | Raw Rate for 2003-2008 totals.
ER38:         ignore.
|***Precip***
ppt38:        continuous. | Mean Precip 30yr Normal
|***Temp***
tmax38:       continuous. | Mean Max Temp 30yr Normal
tmin38:       continuous. | Mean Min Temp 30yr Normal
|***NDVI***
n0138:        continuous. | Mean NDVI for Jan 2003-08
n0238:        continuous. | Mean NDVI for Feb
n0338:        continuous. | March
n0438:        continuous. | April
n0538:        continuous. | May
.
.
.
.
nlcd95:       continuous. | Area of NLCD06 class 95 – Emergent Herbaceous Wetland
```

B2. DATA (*.DATA), TEST (*.TEST), AND CASES (*.CASES) FILES

The required Data file is a comma-separated, or CSV, text file, but like the Names file it uses a custom extension of ".data". This file must not contain a header row, and the order of the attributes in the data table must match the order listed in the Names file exactly. The optional Test (*.test) file, used only when separate training and test datasets are available, and the optional Cases (*.cases) file, which is only used in the "Sample.c" program, are both the exact same format as the Data file, with the exception that the values for the target variable can be unknown in a Cases file. These files contain only attribute data values separated by commas, with no comments or other special symbols.

Below is a small segment (only the first 8 columns of the first 5 and last rows shown) of the Data file for R1a:

```
1,AL,Autauga County,1001,0,0,137148.4531,2408.917481...
2,AL,Baldwin County,1003,6.083280105,94.89916963,166424.8906,2504.301514...
3,AL,Barbour County,1005,3.564342319,110.4946119,132380.8906,2476.249268...
4,AL,Bibb County,1007,0,0,140364.7656,2388.782959...
5,AL,Blount County,1009,1.812770971,1.812770971,141347.0938,2217.50708...
.
.
.
3105,WY,Weston County,56045,279.8713574,1944.36943,37528.54297,1502.03418...
```

B3. PRED (*.PRED) FILE

The Pred file is a text file output by Cubist with the extension ".pred" that gives predicted values for all test cases used. Cases are identified by the label attribute (which was identified in the Names file), and the Pred file lists both the actual and predicted values for each case, separated by spaces.

An abbreviated example of R1a's Pred file is shown below:

```
(Default value 35.5165825)

   Actual  Predicted    Case
    Value      Value
 --------  ---------    ----
 9.585354  0.0339297    1013
 0.870174  0.9073461    1015
 0.000000  0.4137744    1021
 0.000000  1.0149122    1035
 0.000000  0.3301621    1041
.
.
.
.
33.743671  19.3345871   56029
```

B4. MODEL (*.MODEL) FILE

The Model file output by Cubist is a text file with the extension ".model" that contains the production rules derived from the hierarchical decision tree. While the model file is plain text, every piece of information in the file is tagged so that the model file can be easily parsed and read in by external programs. This makes the file somewhat difficult for humans to read because it appears cluttered, but the tags allow great extensibility when using Cubist models with programs like "Sample.c" or others. See Appendix B5 for a more human-friendly formatting of the model rules, as printed to the console immediately following model generation.

An abbreviated example of the Model file for R1a is provided below (including Rule 1 and Rule 15):

```
id="Cubist 2.07 GPL Edition 2013-02-01"
prec="6" globalmean="35.51658" extrap="0.1" insts="0" ceiling="1276.216" floor="0"
att="RRtot" mean="35.51658" sd="108.1282" min="0" max="1160.2"
att="ppt38" mean="99045.09" sd="35007.9" min="8322.42" max="297105"
att="tmax38" mean="1884.899" sd="463.3917" min="782.126" max="3084.98"
att="tmin38" mean="633.3988" sd="450.516" min="-612.162" max="1975.05"
att="n0138" mean="3620.359" sd="1561.452" min="-125.285" max="8069.57"
att="n0238" mean="3451.231" sd="1544.414" min="-175.993" max="7949.23"
.
.
.
.
att="nlcd95" mean="3.163126e+07" sd="1.260578e+08" min="0" max="2.97298e+09"
sample="0.8" init="12"
entries="1"
rules="15"
conds="2" cover="1723" mean="3.2311018" loval="0" hival="98.8468" esterr="3.3677480"
type="2" att="ppt38" cut="75464.898" result=">"
type="2" att="n1238" cut="3238.8015" result=">"
coeff="5.5157504" att="ppt38" coeff="4e-05" att="tmax38" coeff="0.005" att="tmin38"
coeff="-0.008" att="n0138" coeff="-0.0019" att="n0238" coeff="0.0009" att="n0338"
coeff="0.0024" att="n0438" coeff="-0.001" att="n0638" coeff="0.0032" att="n0738"
coeff="-0.0022" att="n0838" coeff="-0.0017" att="n0938" coeff="0.0022" att="n1038"
coeff="-0.0047" att="n1138" coeff="0.0029" att="n1238" coeff="-0.0017"
.
.
.
.
conds="5" cover="34" mean="381.3927917" loval="24.4569" hival="758.988"
esterr="149.3016052"
type="2" att="nlcd42" cut="61200" result="<="
```

type="2" att="tmax38" cut="1853.483" result="<="
type="2" att="elevMean" cut="464.29883" result=">"
type="2" att="nlcd23" cut="1830600" result="<="
type="2" att="n0338" cut="2303.4329" result=">"
coeff="1590.8806906" att="ppt38" coeff="-0.00938" att="tmax38" coeff="0.061"
att="tmin38" coeff="-0.035" att="n0238" coeff="0.3467" att="n0338" coeff="-0.467"
att="n0438" coeff="-0.0035" att="n0538" coeff="0.0036" att="n0638" coeff="0.0034"
att="n0838" coeff="-0.0018" att="n0938" coeff="0.0019" att="n1038" coeff="0.002"
att="n1138" coeff="-0.0055" att="elevRange" coeff="-0.006" att="elevMean" coeff="-
0.016" att="sloMean" coeff="-0.003305" att="nlcd23" coeff="-1.5e-07" att="nlcd41"
coeff="-7e-09"

B5. CONSOLE OUTPUT

When the Unix version of Cubist is run from the command line, a successful model generation results in detailed output "printed" directly to the console. Most of the content is a human-friendly formatting of the production rules that define the model, and which can be derived with some effort from the Model file. However, Cubist also outputs evaluation data to the console which is not included in the default output files (Pred or Model). This evaluation data is very valuable, including the average error magnitude, the relative error magnitude, and the correlation coefficient as calculated on both the training and the test cases. Cubist also lists how often the various attributes were used in the model, either in rule conditions (marked "Conds") or in the linear equations that define the model's predicted values (marked "Model").

An abbreviated version of the console output for R1a is provided below:

```
Cubist [Release 2.07 GPL Edition]  Fri Feb  1 10:06:24 2013
-------------------------------

    Options:
      Application `R1a'
      Use 80% of data for training
      Random seed 12

    Target attribute `RRtot'

Read 2484 cases (48 attributes) from R1a.data

Model:

  Rule 1: [1723 cases, mean 3.2311018, range 0 to 98.84679, est err 3.3677480]

    if
      ppt38 > 75464.9
      n1238 > 3238.802
    then
      RRtot = 5.5157504 - 0.0047 n1038 + 0.0032 n0638 + 0.0029 n1138
              - 0.008 tmin38 - 0.0022 n0738 + 0.0022 n0938 + 0.0024 n0338
              - 0.0019 n0138 - 0.0017 n0838 + 0.005 tmax38 - 0.0017 n1238
              - 0.001 n0438 + 4e-05 ppt38 + 0.0009 n0238

  Rule 2: [420 cases, mean 4.5232439, range 0 to 157.3587, est err 4.0890737]
.
.
.
```

```
  Rule 15: [34 cases, mean 381.3927917, range 24.45685 to 758.988, est err
149.3016052]

    if
       tmax38 <= 1853.483
       n0338 > 2303.433
       elevMean > 464.2988
       nlcd23 <= 1830600
       nlcd42 <= 61200
    then
       RRtot = 1590.8806906 - 0.003305 sloMean - 0.467 n0338 + 0.3467 n0238
               - 0.00938 ppt38 + 0.061 tmax38 - 0.035 tmin38 - 0.016 elevMean
               - 1.5e-07 nlcd23 - 0.0055 n1138 + 0.0036 n0538 - 0.0035 n0438
               + 0.0034 n0638 - 0.006 elevRange + 0.0019 n0938 - 0.0018 n0838
               + 0.002 n1038 - 7e-09 nlcd41


Evaluation on training data (2484 cases):

    Average  |error|         17.0169770
    Relative |error|               0.31
    Correlation coefficient        0.87


       Attribute usage:
         Conds  Model

          78%    94%    ppt38
          60%    56%    n1238
          34%    82%    tmax38
          32%    35%    elevMean
          24%           nlcd90
          18%    14%    nlcd52
          10%    33%    sloMean
           9%     5%    nlcd42
           7%    12%    nlcd23
           6%           nlcd43
           5%    69%    n0338
                 99%    tmin38
                 95%    n0438
                 78%    n0938
                 77%    n0238
                 76%    n0638
                 76%    n1138
                 75%    n0838
                 74%    n1038
                 72%    n0738
                 67%    n0138
                 25%    n0538
                 20%    elevRange
                 15%    nlcd41
                 12%    asp4
                  9%    nlcd71
                  6%    nlcd82
                  6%    nlcd11
```

```
                5%    nlcd21
                4%    nlcd22
                2%    asp2
                2%    asp8


Evaluation on test data (621 cases):

    Average  |error|         20.7125696
    Relative |error|               0.39
    Correlation coefficient        0.84


Time: 1.6 secs
```