THESIS

# ECOLOGIC FACTORS AND TICK-BORNE RELAPSING FEVER IN THE WESTERN UNITED STATES: COUNTY AND ZIP CODE ANALYSES

Submitted by

Ryan Pappert

Department of Environmental and Radiological Health Sciences

In partial fulfillment of the requirements

For the degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2011

Master's Committee:

Advisor: John Reif

Colleen Duncan
Marie Legare
Paul Mead

ABSTRACT

ECOLOGIC FACTORS AND TICK-BORNE RELAPSING FEVER IN THE WESTERN UNITED STATES: COUNTY AND ZIP CODE ANALYSES

Tick-borne relapsing fever (TBRF) is a rare bacterial disease caused primarily by *Borrelia hermsii* and *Borrelia turicatae* in the western United States and transmitted by *Ornithodoros* species soft ticks. No spatial analyses have been attempted for TBRF, and previous epidemiologic studies were limited to case series and outbreak investigations. This study employed ArcGIS to map counties and zip codes with identified cases of TBRF and neighboring control counties and zip codes. A total of 140 counties with reported cases of TBRF, identified in a previous publication, and 243 counties with no reported cases in 12 states were included in the county level analysis. The zip code level analysis included 60 zip codes with cases of TBRF and 193 control zip codes in California and Washington, using information provided by state health departments. Ecologic factors, including elevation, precipitation, average minimum temperature, average maximum temperature, and land cover, in these areas were compared by frequency analysis and logistic regression analyses. The occurrence of TBRF was associated with elevation, temperature, and evergreen forest land cover in county level analyses, and with elevation and temperature in zip code level analyses. No associations were found with

precipitation or additional land cover variables and TBRF occurrence. Counties ($0.25 \geq p \geq 0.0003$) and zip codes ($0.0007 \geq p \geq 0.03$) with cases were seen in higher proportions at elevations above 500 meters than control counties and zip codes, and elevation was included in logistic regression models at both levels of analysis. A higher proportion of counties with cases were observed in the middle of the range of temperature values, while control counties were evenly distributed ($0.01 \geq p \geq 0.0004$). The association with temperature at the zip code level was less consistent, with higher case zip code proportions observed at lower temperatures ($0.08 \geq p \geq 0.01$). A temperature variable was included in logistic regression analyses at both levels of analysis. Evergreen forest was the majority land cover type in a greater proportion of counties with cases when compared to control counties (total land cover $p = 0.04$) and this variable was only significant in the county level logistic regression analyses. The distribution of land cover variables was not significant at the zip code level ($p = 0.82$) and no zip code level land cover variables were significant in logistic regression analyses.

Similar associations were observed when using logistic regression to analyze high risk counties and control counties ($p = 0.005$), and high risk zip codes and control zip codes ($p = 0.006$). Zip code level analyses of California produced a logistic regression model containing an elevation variable ($p = 0.0002$), while the best model for Washington contained the same variables found in the complete zip code level analysis ($p = 0.07$). These results suggest that ecologic factors including elevation and temperature play a role in areas where TBRF occurs. These factors likely influence the distribution and/or abundance of the tick vectors responsible for this disease or their

preferred hosts. Further refinement of these analyses could lead to the construction of a predictive model that could be used to highlight areas of increased risk of TBRF.

ACKNOWLEDGEMENTS

**Ecologic Data Provided by:**

Precipitation, Average Minimum Temperature, Average Maximum Temperature:

PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created October 31, 2008 (precipitation), October 29, 2008 (average daily maximum temperature and average daily minimum temperature).

Land Cover:

United States Department of Agriculture (USDA), National Resources Conservation Service (NRCS), http://datagateway.nrcs.usda.gov/GDGHome.aspx.

Elevation:

Created by: U.S. Geologic Survey (USGS), EROS Data Center Distributed Active Archive Center

Published by: Environmental Systems Research Institute (ESRI), Redlands, CA, 2000

**County Level TBRF Case Distribution Information from the Publication:**

Dworkin MS, Shoemaker PC, Fritz CL, Dowell ME, Anderson DE Jr., 2002a. The epidemiology of tick-borne relapsing fever in the United States. *American Journal of Tropical Medicine and Hygiene 66*: 753-758.

Full text available at: http://www.ajtmh.org/content/66/6/753.full.pdf+html

**Zip Code Level Information Regarding TBRF Cases Provided by:**

Nicola Marsden-Haug, Washington Department of Health

Curtis L. Fritz, California Department of Health Services

Elisabeth Lawaczeck, Colorado Department of Public Health and Environment (not used in the final analysis)

**Special Thanks to:**

Graduate Committee: John Reif, Colleen Duncan, Marie Legare, and Paul Mead

Colleagues: Kiersten Kugeler (CDC), Katie MacMillan (CDC), Ryan Miller (USDA)

Support: Nina Garbino, BDB Epidemiology group, BDB DRL

TABLE OF CONTENTS

Abstract ..... ii

Acknowledgements ..... iv

Chapter 1: Introduction ..... 1

Chapter 2: Literature Review ..... 4

2.1 Background ..... 4

2.1.1 Organism ..... 4

2.1.2 Vector ..... 6

2.1.3 Pathogenesis ..... 7

2.1.4 Clinical Manifestations ..... 9

2.1.5 Laboratory Diagnostic Techniques ..... 9

2.1.6 Treatment ..... 10

2.1.7 Prevention and Control ..... 11

2.2 Epidemiology ..... 12

2.3 Outbreak Investigations ..... 13

2.4 Spatial Modeling ..... 16

Chapter 3: County Level Analysis ..... 19

3.1 Background ..... 19

3.2 Hypothesis ..... 20

3.3 Specific Aims ..... 20

3.4 Methods ..... 21

3.4.1 Case and Control County Selection ..... 21

3.4.2 Ecologic Data ..... 22

3.4.3 ArcGIS Analysis ..... 23

3.5 Statistical Analysis ..... 24

3.5.1 Variable Definition ..... 24

3.5.2 Frequency Analyses ..... 27

3.5.3 Logistic Regression Analyses ..... 27

3.5.4 Evaluation of Selected Variables and Interaction ..... 29

3.5.5 Comparison of Potential Models ..... 31

3.6 Results ..... 31

3.6.1 Study Population Characteristics ..... 30

3.6.2 Frequency Analysis of Ecologic Variables..... 32

3.6.3 Logistic Regression Analyses of Ecologic Variables ..... 36

3.6.4 Purposeful Variable Selection ..... 37

3.6.5 Assumption of Linearity Evaluation ..... 38

3.6.6 Interaction Terms ..... 40

3.6.7 Comparing Multivariable Models ..... 41

3.6.8 Analysis of High Risk Counties vs. Control Counties ..... 43

3.7 Conclusions ..... 45

Chapter 4: Zip Code Level Analysis ..... 47

4.1 Background ..... 47

4.2 Specific Aims ..... 47

4.3 Methods ..... 48

4.3.1 Case and Control County Selection ..... 48

4.3.2 Ecologic Data ..... 50

4.3.3 ArcGIS Analysis ..... 51

4.4 Statistical Analysis ..... 52

4.4.1 Variable Definition ..... 52

4.4.2 Frequency Analyses ..... 55

4.4.3 Logistic Regression Analyses ..... 55

4.4.4 Evaluation of Selected Variables and Interaction ..... 56

4.4.5 Comparison of Potential Models ..... 56

4.4.6 Model Validation ..... 57

4.4.7 California and Washington Individual Models ..... 57

4.4.8 Predictive Risk Model ..... 57

4.5 Results ..... 59

4.5.1 Study Population Characteristics ..... 59

I. California ..... 59

II. Washington ..... 62

4.5.2 Frequency Analysis of Ecologic Variables ..... 64

4.5.3 Logistic Regression Analyses of Ecologic Variables ..... 68

4.5.4 Purposeful Variable Selection ..... 69

4.5.5 Assumption of Linearity Evaluation ..... 70

4.5.6 Interaction Terms ..... 72

4.5.7 Comparing Multivariable Models ..... 72

4.5.8 Model Validation ..... 74

4.5.9 Predictive Risk Model ..... 75

4.5.10 Analysis of High Risk Zip Codes vs. Control Zip Codes ..... 76

4.5.11 California and Washington Individual Models ..... 79

4.6 Conclusions ..... 84

Chapter 5: Discussion ..... 85

5.1 County Level Analysis ..... 85

5.2 Zip Code Level Analysis ..... 89

5.3 Biological Perspectives ..... 95

5.4 Comparison of Findings with Previous Literature ..... 96

5.5 Study Strengths and Limitations ..... 97

5.5.1 Study Strengths ..... 97

5.5.2 Study Limitations ..... 98

5.6 Recommendations for Future Studies ..... 101

References ..... 103

# CHAPTER 1

Tick-borne relapsing fever (TBRF) is one of eight endemic tick-borne diseases in the United States and is one of the five of these tick-borne diseases that is not nationally notifiable (MMWR, 1997). Cases of TBRF were reported in the United States as early as the beginning of the 20$^{th}$ century. TBRF was first recognized as a public health concern at a symposium held by the American Association for the Advancement of Science in 1942. By that time, cases had been documented from the west coast to Texas and the symposium was held to educate public health officers and physicians regarding the characteristics and spread of the disease (Moursund, 1942). Since that symposium, research on TBRF has increased understanding of the responsible organism and vector, clinical manifestations, pathogenesis, and epidemiology. However, research on this disease is limited, especially compared to Lyme disease, a related tick-borne illness. In fact, the majority of past research on TBRF occurrence consists primarily of case series and outbreak investigations. Further research is needed to gain a more complete picture of this disease and its effect on public health.

Past studies have provided an understanding of the locations and environments that are optimal for transmission of this disease. While observational studies such as

outbreak investigations have provided insight on TBRF occurrence on localized scale, few descriptive studies have examined TBRF distribution patterns on a larger scale, for instance the county level. No research has been published that examines counties or zip codes where TBRF has been reported and compares them to surrounding areas where no cases have been reported using ecologic correlates. Systematic examination of these issues could lead to an increased understanding of the disease and a more targeted approach to disease prevention.

Analyses of the relationship between disease distribution and ecologic correlates have been conducted in numerous publications at varying scales, such as county, parish, zip code and census tract. Such analyses were conducted for other tick-borne diseases, including Lyme disease and tularemia (Eisen et al., 2006, Eisen et al., 2008b), but not for TBRF. This study used methods shown to be effective in analyzing other diseases and applied the same methodology to TBRF in an attempt to approach the disease from a new perspective. Characterizing counties and zip codes where TBRF cases have been reported and comparing them to neighboring areas where no cases were reported will elucidate potential ecologic differences between areas with and without disease. Additionally, comparison between county level and zip code level analysis will give insight into the scale necessary for such associations with disease to be observed. Examining correlations between elevation, habitat type, average temperature, precipitation and location of TBRF cases is a previously unattempted approach to analyzing the factors that affect the occurrence of this disease. The overall goal of this project was to provide insight into ecologic factors that affect TBRF occurrence and

inform future research to increase understanding of this little known and often forgotten disease.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Background

Tick-borne relapsing fever (TBRF) is characterized by recurring febrile episodes accompanied by a variety of nonspecific symptoms including headache, myalgia, arthalgia, shaking chills and abdominal distress. Tick-borne relapsing fever is endemic in the western United States, southern British Columbia, the plateau regions of Mexico, Central and South America, Central Asia, along the Mediterranean, and throughout the majority of Africa (Dworkin et al., 1998). TBRF is a bacterial infection, caused by *Borrelia* species spirochetes. Different species of bacteria that cause TBRF are found worldwide, but in North America infections are caused primarily by *B. hermsii* and *B. turicatae*. The bacteria are transmitted by several species of soft ticks of the genus *Ornithodoros* (family Argasidae), with the species of *Borrelia* named for the species of tick that transmits it (e.g. *Ornithodoros hermsii* transmits *Borrelia hermsii*). Research on TBRF in the U.S. has been limited, especially compared to Lyme disease, and it is believed to be an underreported disease in most areas where it is endemic (Dworkin et al., 2002b).

### 2.1.1 Organism

*Borrelia* spirochetes are actively motile, helical organisms that cause recurring disease cycles through a process in antigenic variation (Stoenner et al., 1982), in which

the surface proteins expressed by the spirochete change over the course of infection (Barbour, 1990). It has been demonstrated that the *Borrelia* serotype (which is based on bacterial surface antigens) does not change while the bacteria are in an infected tick, only during the course of human infection (Schwan et al., 1998). Soft ticks of the family Argasidae (*Ornithodoros* species) transmit the species that cause TBRF, but *Borrelia* infections have occurred via blood transfusion, intravenous drug use, and laboratory accidents (Beck 1942, Favorova et al. 1971, Lopez-Cortez et al. 1989).

There are other spirochetes in the *Borrelia* genus that cause related diseases. *Borrelia burgdorferi* is the causative agent of Lyme disease. *B. burgdorferi* spirochetes are spread through the bite of hard ticks of the *Ixodes* genus and cases occur primarily in northeastern, north-central areas of the United States, with some cases reported in western coastal states as well. Though clinical presentation and endemic regions differ between Lyme disease and TBRF, the two diseases can produce similar results on some diagnostic assays. Two-tiered diagnostic testing for Lyme disease reduces the likelihood of such cross-reactivity. Another related pathogen, *Borrelia recurrentis*, is transmitted by the human body louse and causes a more severe infection known as louse-borne relapsing fever (LBRF). The disease is clinically similar to TBRF, but typically only one relapse is observed. LBRF is prevalent in some African countries and is not endemic in the United States (Murray et al., 2002).

2.1.2 Vector

*Ornithodoros hermsii* and *Ornithodoros turicata*, the two most common vectors of TBRF in the United States, have similar lifestyles but different habitats and hosts. *O. hermsii*, one of the smallest species of *Ornithodoros* ticks, lives in coniferous forests at elevations between 1,500 and 8,000 feet, primarily in western states, including Washington, Idaho, Oregon, California, Nevada, Arizona, New Mexico, Utah, and Colorado (Cooley et al., 1944). Its primary hosts are ground squirrels, tree squirrels and chipmunks (Dworkin et al., 1998). Humans are incidental hosts often exposed at night while staying in structures that have been poorly rodent-proofed, such as rustic cabins.

*O. turicata* is found from Kansas west to California and south to Mexico, but lives in drier habitats at lower elevations (Cooley et al., 1944). It is significantly larger than *O. hermsii,* and unlike *O. hermsii,* secretes large amounts of infectious coxal fluid either during or after feeding. These ticks have often been recovered from underground burrows, and preferential hosts are currently unknown, although coyotes and rodents are among the suspected hosts. Many human exposures to *O. turicata* have occurred in caves in Texas (Rawlings, 1995).

In both species of ticks, *Borrelia* species can be transmitted by any life cycle stage, and tick infection occurs primarily from feeding on the blood of an infected host. Blood meals are brief, usually lasting between 15 and 90 minutes. Persistent infection of tick salivary glands (Schwan et al., 1998) with spirochetes allows for rapid transmission (within a minute) during the short feeding period (Davis, 1955). The

spirochetes can be passed from one stage of the life cycle to another (trans-stadial transmission), and are also vertically transmitted from an infected tick to her offspring. Trans-ovarial (vertical) transmission is possible in both of the discussed *Ornithodoros* species, but is much more common in *O. turicata* than *O. hermsii.* In contrast to hard ticks, the female *Ornithodoros* ticks lay clutches of eggs after each blood meal and can live for years in a favorable environment.

*Ornithodoros parkeri* can be found throughout the west in areas similar to that of *O. turicata* (Thompson et al., 1969). Only one human case of TBRF has been directly linked to an *O. parkeri* bite (Davis, 1955).

2.1.3 Pathogenesis

Following the bite of an infected tick, *Borrelia* spirochetes migrate to the host's blood stream where they begin to multiply, eventually reaching an estimated bacterial blood concentration between $10^5$ and greater than $10^6$ spirochetes per milliliter of blood during symptomatic disease (Stoenner et al., 1982). In contrast to the high level spirochetemia observed during febrile episodes, organisms are microscopically undetectable in the bloodstream during asymptomatic periods. Animal data suggest that during these afebrile intervals the bacteria are sequestered in internal organs such as the liver, spleen, bone marrow and central nervous system. Antigenic variation is responsible for the recurring cycles of spirochetemia and resulting fevers associated with TBRF (Felsenfeld, 1971).

The two groups of outer membrane proteins that participate in antigenic variation are known as “variable small proteins” (vsp) and “variable large proteins” (vlp). Originally, both groups were collectively known as “variable major proteins” or “vmp” (Hinnebusch et al., 1998). These proteins are encoded in DNA sequences on linear plasmids (Barbour, 1990). These proteins are expressed sequentially on the surface of the bacteria, thereby changing the antigenic identity of the bacteria. Up to 40 different serotypes have been identified from the progeny of a single cell of *B.* hermsii, strain HS1 (Restrepo et al., 1994). These multiple alterations prevent the host from eradicating the bacteria, leading to recurrent febrile episodes (Dworkin et al., 2002b).

Information concerning possible complications of TBRF has primarily been observed from experimental animals and autopsies from fatal louse-borne relapsing fever (LBRF) cases. Observed complications associated directly with TBRF include nonspecific dermatologic symptoms (Southern et al., 1969), renal and urologic involvement, as well as thrombocytopenia. Additional complications observed in small numbers of TBRF cases include: hypoxia, elevated liver enzyme levels, arrhythmia, myocarditis, and acute respiratory distress syndrome (ARDS) (MMWR, 2007). Mortality from TBRF in the United States is rare and primarily associated with complications during pregnancy, including spontaneous abortion, premature birth, or neonatal death (Goubau, 1984).

### 2.1.4 Clinical Manifestations

Recurring episodes of fever is the symptom that is most characteristic of TBRF. Several other nonspecific symptoms of TBRF include altered sensorium, headache, myalgia, arthalgia, abdominal pain, and vomiting. Diarrhea can occur in about 25% of cases (Dworkin et al., 1998). The mean incubation period for TBRF is about 7 days, with a range of 4 to more than 18 days (Southern et al., 1969). The average length of the first febrile episode is 3 days with a range between 12 hours to 17 days (Goubau, 1984) and the average time between first episode and first relapse is 7 days. Most cases will experience 2 relapses during the course of infection and around 22% will experience 4 or more relapses (Dworkin et al., 2002a). Mortality associated with TBRF is very low, with only a handful of deaths reported. TBRF infection during pregnancy can result in more severe disease, miscarriage or birth of an infected infant (Dworkin et al., 1998). The Jarisch-Herxheimer Reaction, an increase in symptom severity, is a common complication among cases of TBRF and LBRF that can occur shortly after antibiotic treatment. Diagnosis of TBRF is often difficult because of individual variability between cases and misdiagnosis as other diseases which present with multiple febrile episodes (Dworkin et al., 2002b).

### 2.1.5 Laboratory Diagnostic Techniques

Detection of spirochetes in a patient's blood during a febrile episode, along with a compatible patient history, provides confirmation of a TBRF infection (Burgdorfer, 1976). A drop of blood, stained with Wright's or Giemsa stain, or a wet mount of blood

can be used to detect spirochete motility using bright-field microscopy. A dark field microscope and direct or indirect immunofluorescent staining are also utilized. Spirochetes may be overlooked in the blood for a number of reasons, including lack of suspicion of relapsing fever and examination of blood taken during an asymptomatic interval (Dworkin et al., 2002b). Polymerase chain reaction (PCR) is occasionally used to identify minute quantities of *Borrelia* species DNA which may be present in blood. Serologic confirmation of TBRF requires a four-fold rise in antibody titer between acute and convalescent serum samples, or a single reactive sample if paired sera are unavailable. Commonly used serologic assays include the indirect immunofluorescent antibody test (IFA), the enzyme-linked immunosorbent assay (ELISA), and the immunoblot. The ELISA is used most frequently, and often run in parallel with another test, such as the Western blot, to distinguish between *B. hermsii* and *Borrelia burgdorferi* (Lyme disease) antibodies. Evaluation of sensitivity and specificity has been prevented by small numbers of serum samples from confirmed TBRF patients (Fritz et al., 2004). Additionally, the variability in outer surface proteins expressed over time may lead to reduced reactivity with a positive sample because the antigens expressed in the sample may be different than those used in the assay (Dworkin et al., 2002b).

### 2.1.6 Treatment

TBRF is effectively treated with antibiotics, most commonly penicillin, doxycycline, erythromycin, and tetracycline. A common complication of antibiotic treatment is the Jarisch-Herxheimer reaction (JHR) which may occur on initial treatment

of relapsing fever with an effective antibiotic (Dworkin et al., 2002b). Symptoms of JHR include hypotension, tachycardia, chills, rigors, diaphoresis and elevated body temperature, which usually begin within one to four hours of initial antibiotic dose. In a group of 61 TBRF cases with information available, a JHR was observed in 54% (33 cases) of those treated with antibiotics (Dworkin et al., 1998). This high frequency of patients experiencing a JHR indicates that those being treated for TBRF should be kept under observation for at least two hours after beginning antibiotic treatment. Study in LBRF cases has shown spirochetes disappearing from circulation as large amounts of cytokines are released by the immune system, in addition to altered spirochete morphology with an increased susceptibility to phagocytosis. The stimulus that triggers the massive cytokine release is currently unknown (Griffin, 1998), but it may be related to the death of large numbers of spirochetes in the bloodstream and the release of endotoxin and other immunogenic antigens from these dying spirochetes. No deaths from this reaction have been reported in North America (Dworkin et al., 2002b).

### 2.1.7 Prevention and Control

Control measures for TBRF are difficult to implement because of the longevity of the tick vector and the variety of tick host species that can serve as TBRF reservoirs. Prophylactic antimicrobials can be taken after tick exposure, but this often isn't effective since most tick exposure among cases of TBRF goes unnoticed (Cutler 2010). Prevention of TBRF involves avoiding dwellings and natural areas which may be infested with rodents and ticks. Rodent-proofing homes and rustic cabins, while reducing rodent

habitats around homes, may also reduce the risk of acquiring TBRF (Dworkin et al., 2002b).

**2.2 Epidemiology**

TBRF is endemic in the western United States, southern British Columbia, the plateau regions of Mexico and Central and South America, the Mediterranean, Central Asia and throughout most of Africa (Dworkin et al., 1998). The first reported case of TBRF in the United States occurred in a traveler to Texas in 1905 (Wynns, 1942) and the first documented case in the western U.S. was in 1915 in Jefferson County, Colorado (Meador, 1915). TBRF is not a nationally notifiable disease (MMWR, 1997), meaning that reporting of cases isn't required nationwide, but it is reportable in eleven of the states where it is endemic.

A publication by Mark Dworkin provides the most complete description of the epidemiology of TBRF in the United States currently available (Dworkin et al., 2002a). Records and report forms of TBRF cases were obtained from state health departments in TBRF endemic and neighboring states. The authors identified 450 cases of TBRF (300 confirmed and 150 probable, by their definitions) across 12 western states, many of which had records dating back to the 1970's. The states in which cases of TBRF were reported, from most cases to least, are: California, Colorado, Washington, Idaho, Oregon, Texas, Arizona, Nevada, Utah, New Mexico, Wyoming, and Montana. In these 12 states, 51% of all TBRF cases were reported in only 13 counties which the authors speculated may be because of "better awareness and reporting of TBRF in those

counties, greater popularity of those sites for human visits, a greater density of the tick vector population in those areas, or a combination of these factors." Seasonality was examined for those cases with an available onset date and it was found that the majority of cases occurred in the summer months, especially July and August, when *Borrelia hermsii* was the suspected agent. The causative agent of TBRF in Texas was more likely to be *Borrelia turicatae*, and majority of cases were diagnosed in the early winter months keeping with exposure to the agent during cave exploration. The majority of patients (57%) reported staying in a cabin or rural dwelling and about 40% of cases were traveling outside their state of residence when exposed. For those with information available, the average number of relapses was 2 and 50 cases experienced symptoms of the Jarisch-Herxheimer Reaction (JHR). The major limitation of this study was that information on cases is limited because TBRF is not a nationally notifiable disease and reporting is passive in many states. The authors speculated that this could result in an "underestimation of the distribution and magnitude of TBRF in the United States" (Dworkin et al., 2002a).

**2.3 Outbreak Investigations**

Several outbreaks of tick-borne relapsing fever have been documented in the western United States over the past few decades. These outbreaks have provided valuable information about the distribution of, and risk factors for contracting, TBRF.

One of the first documented outbreaks of TBRF on a large scale occurred on Browne Mountain, near Spokane Washington, in March 1968 (Thompson et al., 1969).

There were 11 cases of TBRF among the 42 boy scouts and scoutmasters camping in the area. Those who spent at least one night in a rodent infested cabin had a much higher chance of contracting TBRF (10 out of 20 became ill) than those who camped only in tents (1 in 22 became ill). Diagnosis of TBRF was based on clinical and epidemiological data, as well as observation of spirochetes in one patient's blood. Additionally, 2 of 18 *O. hermsii* ticks collected from the cabin were shown to be infected with spirochetes. The high attack rate among those staying in the cabin highlights the importance of this environment for the transmission of TBRF.

Two separate outbreaks of TBRF occurred almost two decades apart at the same area of the North Rim of Grand Canyon National Park in Arizona. The first of these outbreaks occurred in the summer of 1973 and included symptoms compatible with TBRF in 27 employees and 35 overnight guests (Boyer et al., 1977). Of these 62 cases, 16 were confirmed by observation of *Borrelia* spirochetes in peripheral blood smears or inoculated Swiss mice. The authors found a significant association between TBRF and sleeping in rustic log cabins, and large amounts of rodent nesting materials were recovered from cabins where patients had stayed. In 1990, in the same area, another smaller outbreak of TBRF occurred. During this outbreak, 15 visitors and 2 employees had illness that met the confirmed or probable case definitions (Paul et al., 2002). Most guests stayed in the same northwest group of cabins in which the outbreak occurred in 1973 (RR=8.2 for northwest vs. southeast cabins) and seven of the patients stayed in the same cabin in the northwest group (RR=98 versus other cabins). All cabins were subsequently evaluated and rodent-proofed as necessary. Nests of pine squirrels, an

important reservoir for TBRF, were observed in cabins where TBRF was likely contracted. It was speculated that an epizootic of plague could have reduced the rodent population, leading the ticks to seek blood meals from humans staying in cabins in the park. In both outbreaks the number of TBRF cases was likely underestimated.

In August and September of 1989, six cases of TBRF were reported among persons who had at different times spent the night in the same cabin at Big Bear Lake, San Bernardino County, California (MMWR, 1990). Of these six cases, TBRF was serologically confirmed in two patients and spirochetes were observed in blood smears of two others. Inhabited ground squirrel burrows were found under the cabin, but no infected ticks were recovered. This outbreak was unusual because the illness was especially severe in one patient who likely suffered from meningeal inflammation. Additionally, four out of six patients had significant gastrointestinal symptoms (nausea and vomiting) and were initially diagnosed with viral gastroenteritis. Gastrointestinal symptoms usually occur in a lower percentage of cases of TBRF.

In late June of 1995, 23 members of a family from Nebraska and Kansas stayed in a rental cabin in Estes Park, Colorado (Trevejo et al., 1998). By late July 1995, 11 (48%) of the 23 family members had become ill. The symptoms of this illness were compatible with TBRF. Additionally, 5 of 30 (17%) other lodgers of this cabin were shown to have symptoms compatible with TBRF. Telephone interviews were conducted to determine behaviors while staying in the cabin. Case-patients were more likely to have slept in the top bunk bed (OR=5.2) or on the floor (OR=28.0) than those that didn't become ill.

Spirochetes were not detected in any patient peripheral blood smears; however, *Borrelia hermsii* was cultured from the blood of one patient. Among 13 convalescent serum samples tested, 3 were positive, 8 were equivocal, and 2 were negative. Four case-patients met the definition of a confirmed case, nine were listed as probable and three were suspected cases. There was evidence of rodent infestation around the cabin, including two rodent carcasses, rodent nesting material and feces. Small mammal trapping yielded multiple rodents, of which two Uinta chipmunks were culture-positive for *B. hermsii*. This outbreak confirmed chipmunks as important reservoirs for TBRF and illustrated the importance of awareness when staying in cabins in endemic areas.

**2.4 Spatial Modeling**

Spatial modeling of disease distribution and ecological factors has been done with other vector-borne diseases, but not with tick-borne relapsing fever. This approach is useful for identifying areas where risk of exposure may be elevated to better target surveillance and control measures (Eisen et al., 2008a).

A model of Lyme disease risk was constructed using information on disease incidence in California (Eisen et al., 2006). A single county, Mendocino County, was used to develop a model incorporating areas with high densities of nymphs of *Ixodes pacificus* and cases of Lyme disease. From these areas of high risk, habitat features were identified that were associated with Lyme disease, and these features were compared to Lyme disease cases occurring in the entire state of California. From this a

statewide predictive model was made to ascertain areas of high risk of exposure to *I. pacificus* nymphs which demonstrated strong associations with elevated Lyme disease risk. Of importance in this paper is the use of zip-code level data in the model, rather than county-level data. The authors conclude that the zip-code scale is useful "to detect small, isolated areas with elevated disease risk that otherwise may go undetected" (Eisen et al., 2006). This is especially useful in the western United States, where counties are often large and ecologically diverse.

Similar processes were used to construct spatial risk models for human plague in both the southwestern United States (Eisen et al., 2007b) and in the West Nile region of Uganda (Winters et al., 2009). In both studies, ecologic correlates of disease incidence were identified via logistic regression and entered in a predictive model that identified areas where cases are likely to occur, including areas where plague may be underreported. Plague is a severe and well reported disease, so the models are fairly complete. The publication by Winters et al. provided the basis for the analysis used in this project.

Mapping and modeling of tularemia, another tick-borne bacterial disease, was completed for a nine-state area in the south-central United States (Eisen et al., 2008b). ArcGIS (ESRI, Redlands, CA) was used to map county-based tularemia incidence from 1990-2003 for nine states, which was analyzed along with data concerning elevation, average climate data, vegetation index and land-cover classifications. Association between habitat type, such as dry forest, grassland, areas near water and tularemia

incidence was determined by ordinal logistic regression, since tularemia incidence data weren't normally distributed. From this regression model, predictive models of risk exposure to tularemia were produced for Arkansas and Missouri, the two states where the majority of cases were reported. This was feasible because reporting for this disease is fairly comprehensive. The model was evaluated by utilizing areas where cases were and were not reported and habitats that were positively and negatively associated with tularemia risk. The study found associations between habitat and disease risk, but the authors suggest that finer-scale models may be more useful for targeting prevention measures and informing local medical personnel.

# CHAPTER 3: COUNTY LEVEL ANALYSIS

## 3.1 Background

Analysis of the relationship between TBRF occurrence and ecologic factors was conducted at two different scales. The initial analysis involved examining the distribution of tick-borne relapsing fever at the county level. Information for this level of analysis was taken from the publication "The Epidemiology of Tick-borne Relapsing Fever in the United States" (Dworkin et al., 2002a). This publication contains the most complete information regarding the distribution of TBRF in the United States. It includes 450 cases from 12 states, reported from January 1977 to January 2000. These cases are grouped at the county level, with county of exposure known for the majority of cases, rather than using county of residence as a surrogate. A summary of the number of counties with TBRF cases in each state is provided in Table 3.1. Comparison of those counties in which TBRF has been documented to neighboring counties where TBRF has not been reported may provide useful information regarding ecologic features that influence transmission of TBRF. Mapping cases for an entire state can provide information on certain habitat types and areas where cases were exposed. Comparing the distribution of TBRF between multiple states can confirm that certain habitats are suitable for transmission. A potential concern with analysis at this level is that many counties in the western United States where TBRF is most common are large, potentially

containing significant variation in habitat types and elevations. Examination of these areas on a finer scale, which was performed in the zip code level analysis, may yield more specific associations between landscape features and TBRF occurrence.

Table 3.1: Counties with cases of tick-borne relapsing fever (TBRF) in the western United States, 1977-2000

| State | Number of Counties with TBRF | Total Number of Counties |
|---|---|---|
| **Arizona** | 3 | 15 |
| **California** | 21 | 57 |
| **Colorado** | 21 | 62 |
| **Idaho** | 8 | 44 |
| **Montana** | 1 | 56 |
| **Nevada** | 4 | 16 |
| **New Mexico** | 33 | 33 |
| **Oregon** | 12 | 36 |
| **Texas** | 16 | 254 |
| **Utah** | 6 | 29 |
| **Washington** | 13 | 39 |
| **Wyoming** | 2 | 23 |

*From Dworkin et al., 2002a

### 3.2 Hypothesis

There is a significant relationship between one or more ecologic factors and the occurrence of cases of TBRF at both the county and zip code levels.

### 3.3 Specific Aims

- Obtain data regarding TBRF cases at the county level for 12 western states in which it is endemic (Dworkin et al., 2002a).
- Identify counties without cases of TBRF adjacent to those counties with reported cases.

- Analyze features of counties with and without TBRF and compare both groups using frequency analysis. Features to be analyzed include: elevation, land cover designations, precipitation, minimum and maximum temperature.
- Extract values for each potential ecologic covariate using ArcGIS.
- Identify features that are statistically associated with the presence and absence of TBRF at the county level using logistic regression models.
- Compare high risk counties with control counties to identify any unique associations not observed in the full case county analysis.
- Discuss the results of the analyses and attempt to draw conclusions from the associations found.

## 3.4 Methods

This research protocol was submitted to the Institutional Review Boards (IRB) of both Colorado State University and the Centers for Disease Control and Prevention and designated as exempt. Approval from the IRB's of the states of California and Washington was not required and therefore not pursued.

### 3.4.1 Case and Control County Selection

Counties with cases of TBRF (Dworkin et al., 2002a) were identified and used to create a layer file in ArcGIS, versions 9.3 and 10, (ESRI, Redlands, CA). These counties with cases of TBRF are henceforth referred to as "case counties." There were 140 counties in 12 states with recognized cases of TBRF. As in the original publication, each case county was categorized according to the total number of cases identified in the county during the original study. The categorization of counties by total number of

TBRF cases was done both for display purposes and an analysis comparing the counties with higher numbers of cases to control counties. Table 3.2 provides more detail as to the number of counties within each category.

Table 3.2: Number of tick-borne relapsing fever (TBRF) cases per county among counties with reported cases in the western United States, 1977-2000

| Case Numbers | Number of Counties |
|---|---|
| 1-5 | 125 |
| 6-10 | 4 |
| 11-15 | 5 |
| > 15 | 6 |

*From Dworkin et al., 2002a

Neighboring counties without TBRF cases that share a contiguous border were selected to serve as "control counties" in the analysis. These counties were selected using the "Selection" menu item and using the "Select By Location" feature in ArcGIS. Any county that touched the boundary of a county in the "case county" layer was selected using this operation. Control counties were only included if they were located in one of the same 12 states as the case counties. The selected counties were then exported as a layer file in ArcGIS. A total of 243 counties were chosen as control counties. Both case and control counties were displayed using the GCS_WGS_1984 projection.

### 3.4.2 Ecologic Data

Elevation was derived from a 1 km resolution digital elevation model (USGS/ESRI, Redlands, CA). The elevation for each county was assessed using the minimum, maximum and average values for that county. This was intended to account for any

variability in elevation that may occur across larger counties with varied topography. Elevation data was displayed in the GCS_WGS_1984 projection. Land cover classification was derived from the National Land Cover Dataset (USDA, National Resources Conservation Service [NRCS]), available for each state individually. The land cover type that was identified as the "majority" for each county was the one associated with that county. Land cover layers were displayed in the projections NAD_1983_10N through NAD_1983_14N, depending on the state being analyzed. Data for precipitation, minimum and maximum temperature were derived from individual models of each using annual averages from 1971 to 2000 (PRISM Climate Group, Oregon State University). These data corresponded well with the time span during which most of the cases included in this study were documented. Minimum, maximum, and mean values were obtained in each county for each variable to account for any variation that might occur across each county for these three potential covariates as well.

### 3.4.3 ArcGIS Analysis

Maps were created using each of the ecologic variables to be analyzed, as well as the map layers for the case and control counties. If the projections were not the same for all layers, the "Project" tool was used to convert layers to the same projection. Once all layers were displayed in the same projection, "Zonal Statistics," were used to extract the data from the ecologic variable layers. Zonal statistics extracts information from the raster data used to represent the ecologic variables and calculates summary statistics (e.g. minimum, maximum, mean, and majority) for each county within the data layers

for case and control counties. This information was then added to a spreadsheet in Microsoft Excel 2007 (Microsoft, Redmond, WA), where information from all counties was consolidated and organized for statistical analysis.

See Figures 3.1 and 3.2 for the distribution case and control counties, respectively.

## 3.5 Statistical Analysis

### 3.5.1 Variable Definition

All counties where TBRF cases were identified were listed as "case." Neighboring counties with no cases of TBRF reported were listed as "control." The minimum, maximum, and mean values of the four continuous ecologic variables (i.e. elevation, average precipitation, average minimum temperature and average maximum temperature) were each treated independently. For example, average precipitation variable was analyzed as three different variables named PPT_MIN, PPT_MAX, and PPT_MEAN, each of which contained the minimum, maximum, and mean average precipitation values for each county, respectively. The land cover variable was categorized numerically, with the number representing the majority land cover or habitat type in each county. The land cover variable was further divided into a series of design or "dummy" variables, with each one representing an individual land cover type. If that specific land cover type was the majority in a county, it was coded "1." If it was not the majority, it was coded "0." The individual habitat types represented by these variables included: open water, developed land, barren land, deciduous forest, evergreen forest, shrub/scrub, grassland/herbaceous, crop/livestock and wetlands.

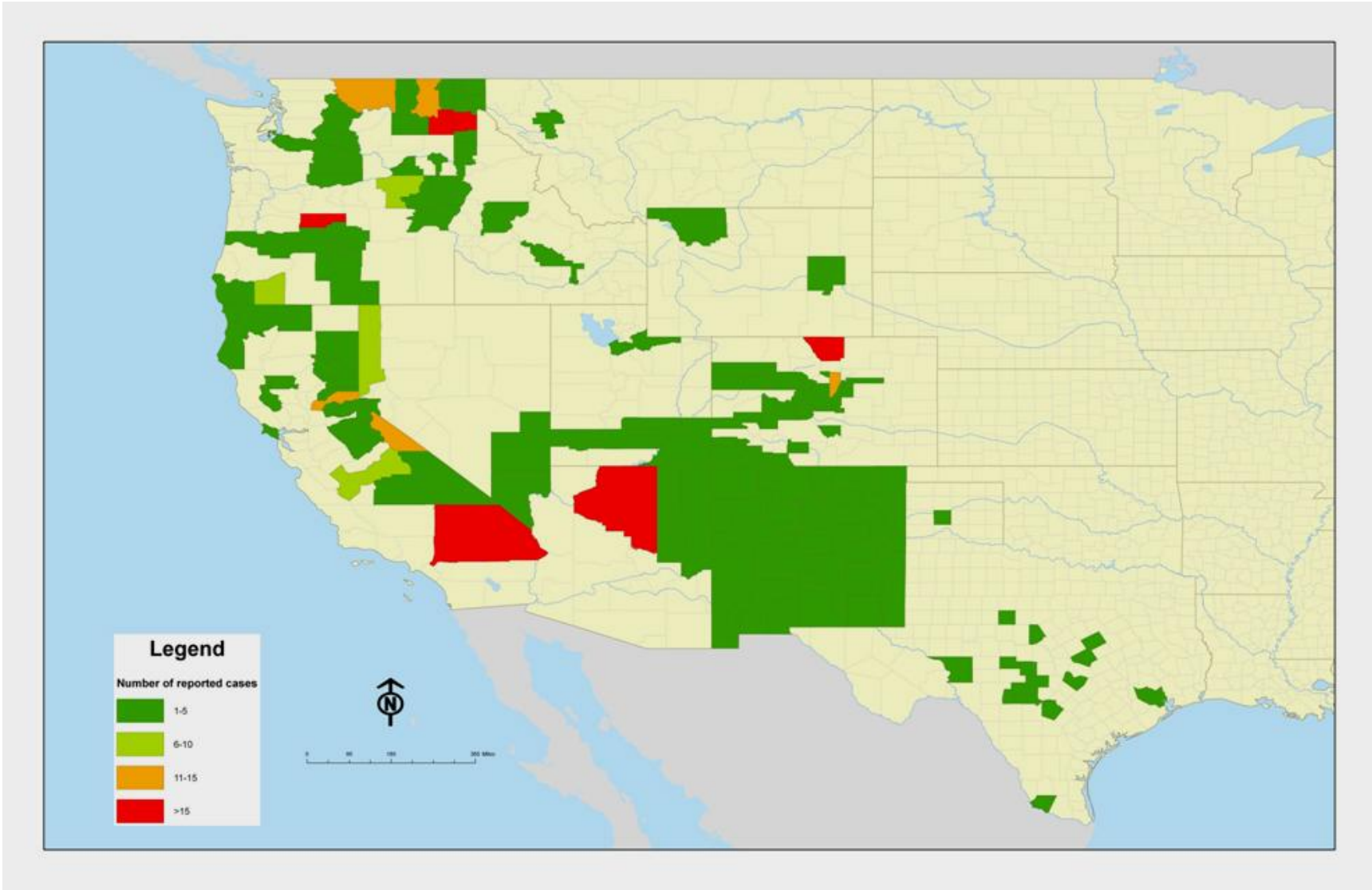


Figure 3.1: Distribution of counties with cases of tick-borne relapsing fever (TBRF) from 1977-2000 (Dworkin et al., 2002a)

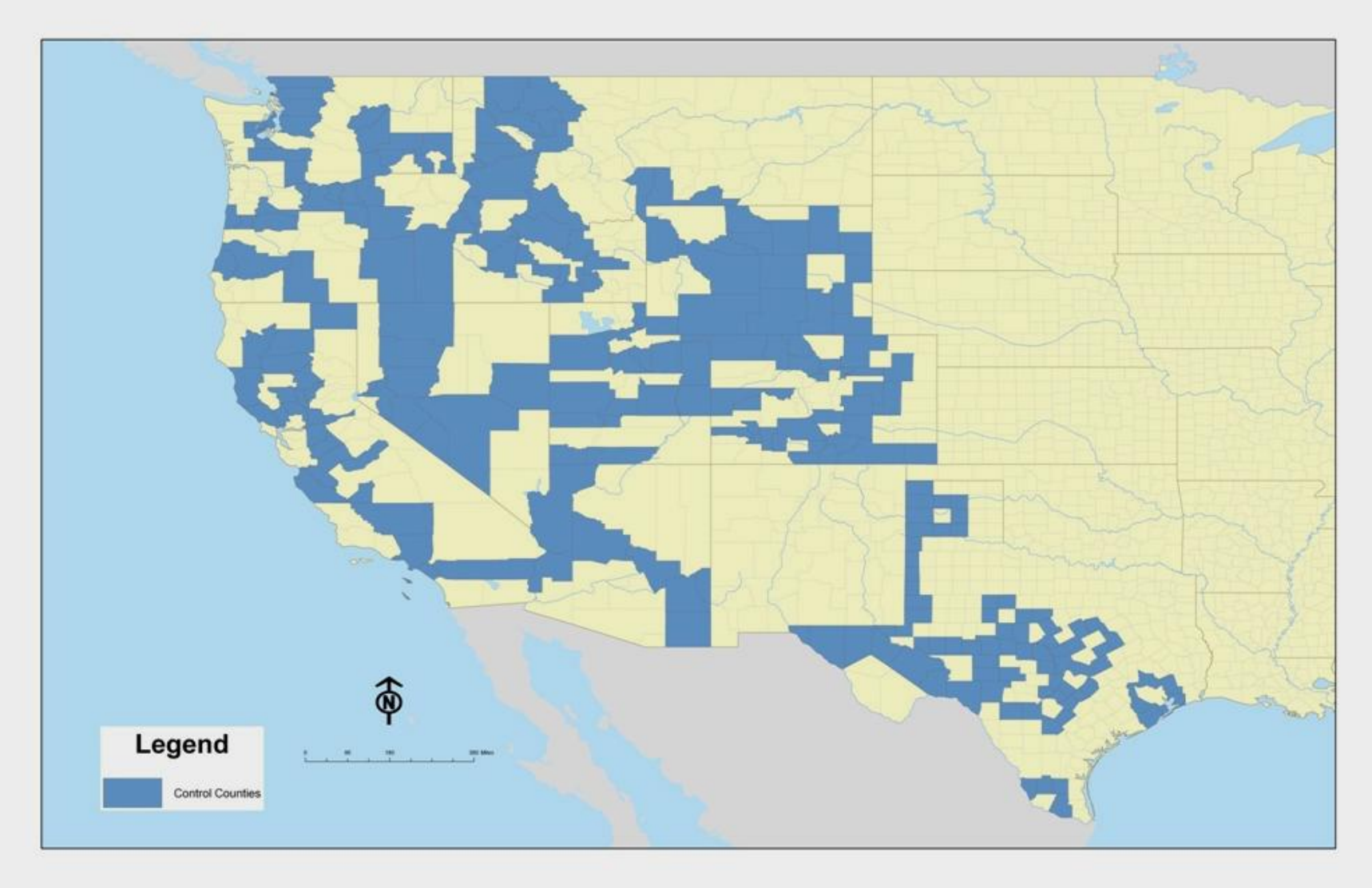


Figure 3.2: Distribution of control counties without reported tick-borne relapsing fever cases

### 3.5.2 Frequency Analyses

The distribution of case and control counties was first compared by frequency analysis. Each ecologic variable was divided into categories based on the range of values observed in that variable. The proportion of case and control counties in each category was calculated by dividing the number of case or control counties in each category by the total number case or control counties. The proportions of case and control counties in each category were compared and differences in the proportions of case and control counties were noted in the results. The overall distribution of the counties in these categories was also noted. Additionally, a chi-square value was calculated for each variable using the distribution of values observed in the frequency tables. A statistically significant chi-square value indicated a difference in the distribution between case and control counties.

### 3.5.3 Logistic Regression Analyses

Binomial logistic regression was conducted on all variables using methods described in *Applied Logistic Regression* (Hosmer and Lemeshow, 2000). All statistical analyses were carried out using the JMP (SAS, Cary, NC) statistical software package, version 9.02.

First, a univariable logistic regression analysis was run for each potential covariate against the dichotomous outcome variable "Case Status," which listed counties as either "case" or "control." A p-value less than 0.25 for the chi-square statistic was required for consideration for the multivariable model. Once the potential

covariates were identified, both forward and backward stepwise logistic regression were run, and the results of each were compared. Similar variables that were significant in the model (e.g. multiple temperature variables) were checked for correlation to each other using a Spearman correlation test. Any two variables with a Spearman's $\rho > 0.8$ could not both be included in the final model. In the case of two correlated variables, the decision on which one to include in the model was based on p-values and performance with other variables in the purposeful variable selection process.

Using the stepwise analyses and Spearman's correlation test as guides, a more purposeful variable selection was conducted without the use of an automated system. After purposeful variable selection was completed, the model coefficients, effect likelihood ratios and Wald statistics were compared to those of the larger models created through stepwise regression to confirm that no drastic changes occurred in the model because of the elimination of variables. The lack of any such changes in these measures confirms that the model with fewer variables was as effective as the larger model. At this stage in the model building process, all variables that were not selected for the multivariable model were added again to identify any that may only have an effect in the presence of other variables. This model, containing all the significant and relevant variables, is referred to as the "preliminary main effects model" (Hosmer and Lemeshow, 2000).

### 3.5.4 Evaluation of Selected Variables and Interaction

After the main variables included in the model were established, the assumption of linearity of all continuous variables was tested. A smoothed scatterplot was created for continuous variables by dividing continuous variables into categories and plotting the proportion of cases in each category against the category itself. From this plot, the shape of the data was noted and variables without obvious linear trends were considered for transformation or categorization. Additionally, a quartile analysis was conducted for some variables. This involved splitting the data into quartiles and constructing three design variables, which were run in a model together and individually. Whether these design variables were an improvement over the continuous variable, coupled with the shape of the scatterplot, lead to a decision regarding possible alteration of the variable. The same methods were applied to variables not included in the multivariate model to confirm that alteration did not produce a variable that made a significant contribution to the model. This model is referred to as the "main effects model" (Hosmer and Lemeshow, 2000).

After the main variables in the model were chosen, interactions between the terms were examined. For consideration as a contributing factor in the model, any interaction term was required to be both statistically significant and biologically plausible.

### 3.5.5 Comparison of Potential Models

Following identification of interaction terms, multiple potential models were compared using various statistical measures, such as Akaike information criterion (AICc) (Akaike, 1974), receiver operator characteristic curves (ROC), and the goodness of fit (or lack of fit) chi-square statistic. The goodness of fit test indicates whether the variables included in the model are sufficient. If the $X^2$ value is not significant ($p > 0.05$), it indicated that no more variables need to be added to the model. The ROC curve is a plot of all sensitivity values on the y-axis against all (1 - sensitivity) values on the x-axis. The area under the curve (AUC) of the ROC curve describes the overall accuracy of the model without the need for a threshold or cutpoint (Fielding & Bell, 1997). The AUC values range from 0.5 to 1.0, with values closer to 0.5 providing poor discrimination between the two groups being classified, and values closer to 1.0 providing excellent discrimination between these two groups. Akaike information criterion is a tool used for model selection only that provides no information regarding the quality of a particular model. The AICc values were used to compare candidate models and choose the most concise or parsimonious model. The model with the lowest AICc value was considered the best model, but models within two AICc units were considered as competing models (Eisen et al, 2010). Finally, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for each model being considered using a probability cut-off value that maximized both sensitivity and specificity. Using the information provided by these tests, the final multivariable model was selected from the competing models by choosing the one with the highest

sensitivity and the best balance among specificity, PPV and NPV (Eisen et al., 2010). This was the model that best represented the relationship between the presence or absence of TBRF cases in a county and the ecologic covariates examined.

## 3.6 Results

### 3.6.1 Study Population Characteristics

All summary information regarding cases is taken from "The Epidemiology of Tick-Borne Relapsing Fever in the United States" (Dworkin et al, 2002a). In this case series, 52% were males, 40% were females, and 8% were missing gender information. The median patient age was 35 years old. There were 300 confirmed cases and 150 probable cases included in this study. A confirmed case was defined as fever and the observation of spirochetes by microscopy. A probable case was defined as relapsing illness with either serologic evidence of infection or epidemiologically appropriate exposure. No further details on method of diagnosis were available. Month of onset of illness was documented for 425 out of 450 cases. The most common months of onset were July (24%) and August (23%), with large numbers of cases documented in June and September as well. Further detail regarding the distribution of TBRF cases by month of illness onset is available in Figure 2 in "The Epidemiology of Tick-Borne Relapsing Fever in the United States" (Dworkin et al, 2002a).

### 3.6.2 Frequency Analysis of Ecologic Variables

Each ecologic variable was divided into categories based on the range of the data displayed by the 140 case counties and 243 control counties being analyzed. In each category, the proportion of case counties was compared to the proportion of control counties. The three majority land cover classifications that were present in the highest number of case counties were: shrub/scrub (40.7%), evergreen forest (37.9%) and grassland/herbaceous (15.7%). Evergreen forest was the only land cover type that had a higher percentage in case counties than control counties, with a difference of 14.4% observed. The variable for majority crop or livestock area was associated with more control counties than case counties, with a difference of about 8%, implying a negative association with TBRF occurrence. The remaining land cover variables were represented in a small number of total counties and had similar frequencies in case and control counties. The overall distribution of the land cover variables was significant, with a $X^2$ = 21.58 ($p = 0.04$). The general pattern of the three elevation variables was lower case county percentages at lower elevations and higher percentages at higher elevations, when compared to control county percentages in the same category. The variables for mean and maximum elevation were statistically significant, while the variable for minimum elevation was not significant ($p = 0.25$). No apparent pattern was observed across the three precipitation variables, with both case and control county frequencies highest among the lower and middle categories; however, the variable for maximum precipitation in a county was statistically significant ($p = 0.003$), with slightly higher proportions of case counties observed at higher precipitation levels. With the three

maximum temperature and the three minimum temperature variables, there was a higher frequency of cases observed in the middle temperature categories, as opposed to the highest and lowest temperature categories. For most temperature variables, control county percentages were generally evenly distributed among all categories and all temperature variables had a statistically significant chi-square value ($p \leq 0.01$ for all variables). Table 3.3 displays the results of the frequency analysis in greater detail.

Table 3.3: Distribution across selected ecologic variables of counties with reported tick-borne relapsing fever (TBRF) cases and neighboring control counties in the western United States, 1977-2000

| Majority Land Cover | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **Open Water** | 0 | 1 | 0.0% | 0.4% |
| **Developed Land** | 2 | 3 | 1.4% | 1.2% |
| **Barren Land** | 0 | 1 | 0.0% | 0.4% |
| **Deciduous Forest** | 0 | 4 | 0.0% | 1.6% |
| **Evergreen Forest** | 53 | 57 | 37.9% | 23.5% |
| **Shrub/Scrub** | 57 | 110 | 40.7% | 45.3% |
| **Grassland/Herbaceous** | 22 | 36 | 15.7% | 14.8% |
| **Crop/Livestock Area** | 5 | 28 | 3.6% | 11.5% |
| **Wetlands** | 0 | 2 | 0.0% | 0.8% |

*Land Cover: $X^2 = 21.58$, $p = 0.04$

| Minimum Elevation (meters) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< 500** | 48 | 103 | 34.3% | 42.4% |
| **501-1000** | 22 | 44 | 15.7% | 18.1% |
| **1001-1500** | 36 | 55 | 25.7% | 22.6% |
| **1501-2000** | 27 | 29 | 19.3% | 11.9% |
| > **2000** | 7 | 12 | 5.0% | 4.9% |
| **Maximum Elevation (meters)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< 1000** | 17 | 69 | 12.1% | 28.4% |
| **1001-2000** | 24 | 53 | 17.1% | 21.8% |
| **2001-3000** | 42 | 61 | 30.0% | 25.1% |
| **3001-4000** | 47 | 54 | 33.6% | 22.2% |
| **> 4000** | 10 | 6 | 7.1% | 2.5% |

| Mean Elevation (meters) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< 500** | 12 | 61 | 8.6% | 25.1% |
| **501-1000** | 29 | 52 | 20.7% | 21.4% |
| **1001-1500** | 32 | 45 | 22.9% | 18.5% |
| **1501-2000** | 32 | 46 | 22.9% | 18.9% |
| **2001-2500** | 21 | 26 | 15.0% | 10.7% |
| **> 2500** | 14 | 12 | 10.0% | 4.9% |

*Min. Elevation: $X^2 = 5.41$, p = 0.25; Max. Elevation: $X^2 = 21.19$, p = 0.0003; Mean Elevation: $X^2 = 19.41$, p = 0.002

| Minimum Precipitation (mm) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< 250** | 50 | 71 | 35.7% | 29.2% |
| **251-500** | 59 | 95 | 42.1% | 39.1% |
| **501-750** | 20 | 41 | 14.3% | 16.9% |
| **751-1000** | 6 | 22 | 4.3% | 9.1% |
| **> 1000** | 5 | 14 | 3.6% | 5.8% |
| **Maximum Precipitation (mm)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< 500** | 7 | 41 | 5.0% | 16.9% |
| **501-1000** | 51 | 102 | 36.4% | 42.0% |
| **1001-1500** | 35 | 51 | 25.0% | 21.0% |
| **1501-2000** | 26 | 17 | 18.6% | 7.0% |
| **> 2000** | 21 | 32 | 15.0% | 13.2% |
| **Mean Precipitation (mm)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< 250** | 3 | 15 | 2.1% | 6.2% |
| **251-500** | 57 | 101 | 40.7% | 41.6% |
| **501-750** | 43 | 61 | 30.7% | 25.1% |
| **751-1000** | 19 | 35 | 13.6% | 14.4% |
| **> 1000** | 18 | 31 | 12.9% | 12.8% |

*Min. Precipitation: $X^2 = 5.39$, p = 0.25; Max. Precipitation: $X^2 = 21.32$, p = 0.003; Mean Precipitation: $X^2 = 4.16$, p=0.38

| Minimum Tmin (Celsius) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< -10** | 17 | 30 | 12.1% | 12.3% |
| **-10 to -5** | 45 | 47 | 32.1% | 19.3% |
| **-5 to 0** | 40 | 48 | 28.6% | 19.8% |
| **0 to 5** | 18 | 35 | 12.9% | 14.4% |
| **5 to 10** | 7 | 38 | 5.0% | 15.6% |
| **> 10** | 13 | 45 | 9.3% | 18.5% |

| Maximum Tmin (Celsius) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< 0** | 6 | 16 | 4.3% | 6.6% |
| **0 to 5** | 57 | 79 | 40.7% | 32.5% |
| **5 to 10** | 51 | 62 | 36.4% | 25.5% |
| **10 to 15** | 21 | 73 | 15.0% | 30.0% |
| **> 15** | 5 | 13 | 3.6% | 5.3% |
| **Mean Tmin (Celsius)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< -5** | 5 | 8 | 3.6% | 3.3% |
| **-5 to 0** | 29 | 52 | 20.7% | 21.4% |
| **0 to 5** | 65 | 73 | 46.4% | 30.0% |
| **5 to 10** | 25 | 55 | 17.9% | 22.6% |
| **> 10** | 16 | 55 | 11.4% | 22.6% |

*Min. Tmin: $X^2 = 22.78$, $p = 0.0004$; Max. Tmin: $X^2 = 16.28$, $p = 0.003$; Mean Tmin: $X^2 = 13.65$, $p = 0.009$

| Minimum Tmax (Celsius) | No. of Counties with TBRF | No. of Control Counties | Percentage of Counties with TBRF | Percentage of Control Counties |
|---|---|---|---|---|
| **< 0** | 4 | 54 | 2.9% | 22.2% |
| **0 to 5** | 35 | 56 | 25.0% | 23.0% |
| **5 to 10** | 45 | 34 | 32.1% | 14.0% |
| **10 to 15** | 26 | 20 | 18.6% | 8.2% |
| **15 to 20** | 9 | 45 | 6.4% | 18.5% |
| **> 20** | 21 | 34 | 15.0% | 14.0% |
| **Maximum Tmax (Celsius)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< 15** | 10 | 18 | 7.1% | 7.4% |
| **15 to 20** | 50 | 102 | 35.7% | 42.0% |
| **20 to 25** | 48 | 43 | 34.3% | 17.7% |
| **25 to 30** | 29 | 77 | 20.7% | 31.7% |
| **> 30** | 3 | 3 | 2.1% | 1.2% |
| **Mean Tmax (Celsius)** | **No. of Counties with TBRF** | **No. of Control Counties** | **Percentage of Counties with TBRF** | **Percentage of Control Counties** |
| **< 10** | 6 | 13 | 4.3% | 5.3% |
| **10 to 15** | 40 | 63 | 28.6% | 25.9% |
| **15 to 20** | 52 | 61 | 37.1% | 25.1% |
| **20 to 25** | 27 | 49 | 19.3% | 20.2% |
| **> 25** | 15 | 57 | 10.7% | 23.5% |

*Min. Tmax: $X^2 = 19.25$, $p = 0.002$; Max. Tmax: $X^2 = 15.51$, $p = 0.004$; Mean Tmax: $X^2 = 12.51$, $p = 0.01$

### 3.6.3 Logistic Regression Analyses of Ecologic Variables

Univariate logistic regression analysis indicated that seven variables chosen for analysis were correlated with the presence or absence of TBRF in a county and 11 were not. Variables that did not have a statistically significant relationship with TBRF occurrence at the county level included: the maximum value for minimum temperature, the maximum value for maximum temperature, the maximum and mean values for precipitation, shrub/scrub, open water, deciduous forest, developed land, grassland/herbaceous, wetlands, and barren land. These variables were not considered for inclusion in the multivariable model. Both forward and backward stepwise regression produced similar results, and the variables selected for the multivariable model included: minimum and mean values for elevation, minimum temperature and maximum temperature, as well as the majority evergreen forest variable. Since multiple elevation and temperature variables are represented in this model, a Spearman correlation test was run to determine if there was any correlation among the variables either within or between groups. The two elevation variables and four temperature variables were highly correlated within their own groups (e.g. minimum elevation was correlated with mean elevation), with $\rho > 0.8$ in every case. However, none of the variables were correlated with another variable outside their specific group, for example no elevation variables were highly correlated with any of the temperature variables. From these analyses it was determined that the final multivariable model would contain variables for elevation, temperature and specifically the land cover variable for the majority of a county being evergreen forest.

### 3.6.4 Purposeful Variable Selection

After these preliminary analyses were complete, a more purposeful variable selection was conducted to identify those to be included in the final model. Elevation and temperature variables were run in multivariable models in different combinations to determine whether any statistically significant relationships existed. Of all the possible combinations of the two elevation and four temperature variables, only the variables for mean elevation and the mean value for maximum temperature produced a model that was statistically significant as a whole, as well as each being statistically significant individually within the model ($p < 0.1$). After these two variables were chosen, the variable representing the majority land cover type for a county being evergreen forest was added to the model. The model containing all three variables was also statistically significant as a whole and for each individual variable. Wald statistics, effect likelihood ratios and variable coefficients were compared between the larger model containing multiple elevation and temperature variables and the smaller model containing only one of each. Only very minor changes were observed in the various measures, indicating that the loss of variables from the larger model did not negatively impact the model in a meaningful way. Therefore, the preliminary main effects model included the variables for mean elevation, the mean value for maximum temperature, and majority evergreen forest.

### 3.6.5 Assumption of Linearity Evaluation

The linearity assumption for the two continuous variables (mean elevation and the mean value for maximum temperature) was tested by examining the smoothed scatterplot and performing quartile analysis on each. For the smoothed scatterplot, the continuous data were divided into categories and plotted against the percentage of case counties in each group. The overall shape of the data was observed, which helped inform decisions about transformation or categorization of the variables. The plot for the mean elevation variable suggested that the linearity assumption was valid.

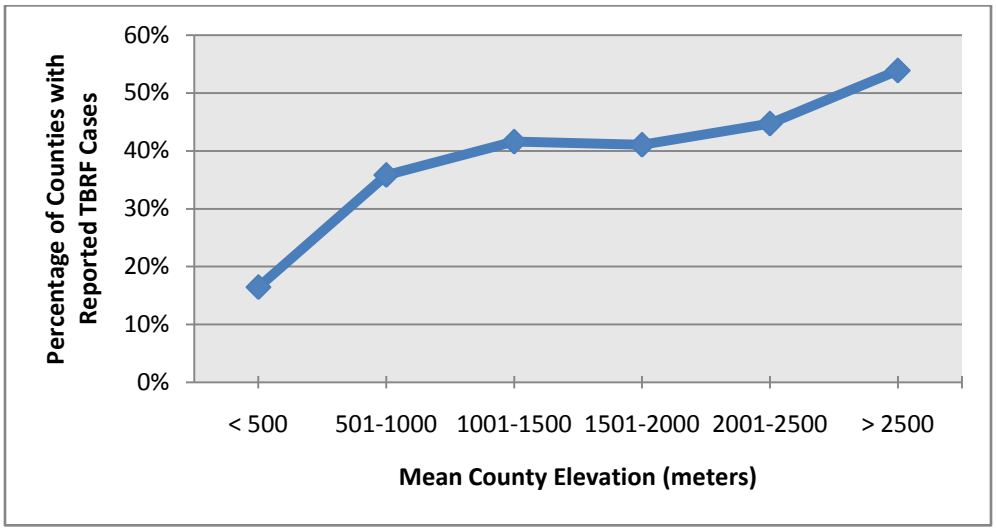


Figure 3.3: Proportion of case counties by mean county elevation, western United States, 1977-2000. A case county is defined as a county with reported tick-borne relapsing fever (TBRF) cases; control counties are neighboring counties without reported TBRF cases. Percentage calculated as the number of case counties divided by total case and control counties within each environmental category.

In order to check the linearity assumption further, the mean elevation variable was divided into quartiles and three design variables were created to represent these quartiles, with the lowest quartile serving as the reference group. These design

variables did nothing to improve the model and only the one comparing the highest quartile of elevation values to the lower three quartiles was statistically significant. Similarly, the shape of the data does not indicate that a transformation is needed and attempted transformations, such as squaring the variable, did nothing to improve its statistical significance. Therefore the mean elevation was left as a continuous variable.

The same methods were used to assess the assumption of linearity in the variable for mean value of maximum temperature in a county. Initial observation of the smoothed scatterplot heavily implied that the variable was not linear, with higher percentages of case counties in the middle temperature values compared to the higher and lower values. Quartile categorical analysis confirmed this, but produced two design variables that were significant. The variable was divided into tertiles and the design variable that was statistically significant compared the middle tertile to the upper and lower ones. This design variable was chosen to be included in the final model because it characterized the pattern observed in the mean maximum temperature variable without violating the assumption of linearity.

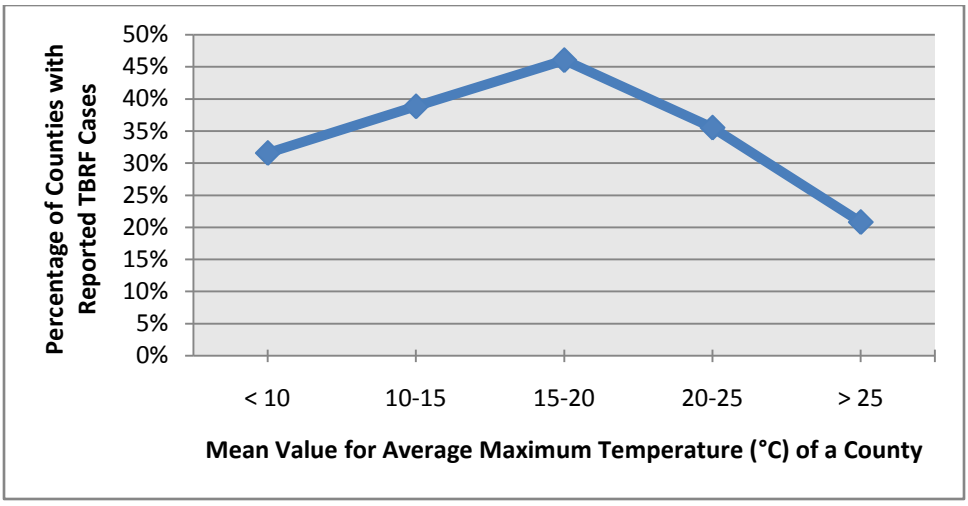


Figure 3.4: Proportion of case counties by mean value for average maximum temperature, western United States, 1977-2000. A case county is defined as a county with reported tick-borne relapsing fever (TBRF) cases; control counties are neighboring counties without reported TBRF cases. Percentage calculated as the number of case counties divided by total case and control counties within each environmental category.

The only precipitation variable that was statistically significant in the univariable analysis, minimum precipitation, was evaluated using the quartile method to confirm that it would not become significant if converted to a categorical variable. Only the design variable for the highest quartile was significant, but it became non-significant once added to the multivariable model ($p = 0.42$). It did nothing to improve the whole model and was therefore excluded. Likewise, transforming variables that were non-significant did not improve their performance in any model constructed.

### 3.6.6 Interaction Terms

The main effects model included the variables for: mean elevation (ELEV_MEAN), the design variable comparing the middle tertile of mean maximum

temperature values to the highest and lowest tertiles (DUMMYT2), and majority evergreen forest (EVGRN-FOR_MAJ). Interactions between all three major variables were evaluated and it was found that the interaction terms between the temperature design variable and the variables for elevation and evergreen forest were statistically significant, while the interaction term for elevation and evergreen forest was highly non-significant ($p = 0.87$). Exclusion of the non-significant interaction term only improved the model as a whole, so it was not considered for inclusion in the final model. Both interaction terms that were statistically significant seemed biologically plausible, and were considered for the final model.

### 3.6.7 Comparing Multivariable Models

The goodness of fit chi-square statistic was compared among all four candidate models being considered, however it was statistically significant for all four models ($p < 0.05$). This implied that additional terms were needed in the final multivariable model, but a model constructed using all potential variables also had a statistically significant chi-square goodness of fit statistic. Likewise, all models with transformed or categorized variables were statistically significant for goodness of fit. The specific goodness of fit chi-square statistic used in the JMP software package could not be identified, but it is possible that the test used was not the most appropriate choice for the data. Alternatively, there could be factors influencing the data that were not considered in this analysis. Regardless, goodness of fit was similar for all models and was not considered in choice of the final multivariable model.

A total of four models were considered as candidates for the final multivariable model. The variables included in each model are listed below.

1. Mean elevation, mean maximum temperature design variable, majority evergreen forest, and interaction terms for (elevation*temperature) and (evergreen forest*temperature)
2. Mean elevation, mean maximum temperature design variable, majority evergreen forest
3. Mean elevation, mean maximum temperature design variable, majority evergreen forest, and the interaction term for (elevation*temperature)
4. Mean elevation, mean maximum temperature design variable, majority evergreen forest, and the interaction term for (evergreen forest*temperature)

Table 3.4: Candidate models for the county level analysis of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence, western United States, 1977-2000

| Mod. ID | Negative log-likelihood | K | AICc | Δ AICc | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 231.61 | 5 | 475.5 | 0 | 0.68 | 62 | 70 | 54 | 76 | ELEV_MEAN, DUMMYT2, EVGRN-FRST_MAJ |
| **2** | 237.67 | 3 | 483.5 | 7.96 | 0.66 | 47 | 79 | 57 | 72 | ELEV_MEAN, DUMMYT2, EVGRN-FRST_MAJ |
| **3** | 234.0 | 4 | 478.1 | 2.65 | 0.68 | 53 | 75 | 55 | 73 | ELEV_MEAN, DUMMYT2, EVGRN-FRST_MAJ |
| **4** | 237.08 | 4 | 484.3 | 8.84 | 0.66 | 49 | 78 | 56 | 72 | ELEV_MEAN, DUMMYT2, EVGRN-FRST_MAJ |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MEAN = mean elevation, DUMMYT2 = mean maximum temperature design variable, EVGRN-FOR_MAJ = majority evergreen forest.

Akaike information criterion (AICc) was examined for each model chosen. The model with the lowest value was considered the best, but models within two AICc units were considered competing models. Model 1 had the lowest AICc value, but Model 3 was almost within two AICc units. The ROC AUC values are the same for the two models, so sensitivity, specificity, PPV and NPV were compared. Model 1 had the higher sensitivity value, with only minor drops in specificity and PPV. Model 1 was chosen as the final multivariable model for the county level analysis. Specifics on the final model parameters can be found in Table 3.5.

Table 3.5: Parameter estimates for the selected multivariate logistic regression model for the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence at the county level, western United States, 1977-2000

| | Parameter Estimates | | | Likelihood ratio test | | |
|---|---|---|---|---|---|---|
| **Model covariates** | **Estimate** | **SE** | **95% C.I.** | $X^2$ | **df** | **p-value** |
| Intercept | -1.59 | 0.34 | (-2.29, -0.96) | 22.37 | 1 | <.0001 |
| ELEV_MEAN | 0.001 | 0.0002 | (0.0006, 0.0015) | 22.05 | 1 | <.0001 |
| DUMMYT2 | -0.38 | 0.13 | (-0.64, -0.11) | 7.88 | 1 | 0.005 |
| EVGRN-FRST_MAJ | -0.64 | 0.17 | (-0.99, -0.32) | 16.76 | 1 | <.0001 |
| EVGRN-FRST_MAJ*DUMMYT2 | 0.35 | 0.17 | (0.03, 0.70) | 4.76 | 1 | 0.03 |
| ELEV_MEAN*DUMMYT2 | -0.001 | 0.0002 | (-0.001, -0.0003) | 10.94 | 1 | 0.0009 |

*df = degrees of freedom; ELEV_MEAN = mean elevation, DUMMYT2 = mean maximum temperature design variable, EVGRN-FOR_MAJ = majority evergreen forest; Whole Model Test $X^2 = 39.68$, df = 5, $p < 0.0001$; goodness of fit $X^2 = 463.22$, $p = 0.002$

### 3.6.8 Analysis of High Risk Counties vs. Control Counties

To determine whether counties with higher case numbers possessed any unique correlations that may have been obscured by the large number of counties with only one case, an analysis was conducted comparing those counties with more than five cases against control counties. This was conducted using the same methods described

in the analysis comparing all case and control counties. The data from the 15 counties with greater than 5 cases reported were first regressed against all 243 control counties included in the full analysis. Following that, they were compared to only the 54 control counties that shared a border with them, to determine if there was a stronger correlation when compared to only neighboring counties.

The analysis comparing the 15 high case counties to all 243 control counties yielded similar results to the full county level analysis. Stepwise regression produced multivariable models containing variables for elevation, evergreen forest, precipitation, and several for temperature. The temperature variables were all correlated, but no other variables were correlated with each other. Purposeful variable selection led to the creation of a model containing variables for elevation (ELEV_MAX) and majority evergreen forest. The interaction term between the two variables was also found to be significant. The two models, with and without the interaction term, were compared to each other as indicated above to determine which best represented the relationship between high case counties and all controls. The model without the interaction term was chosen because of its higher sensitivity value.

Table 3.6: Candidate models for the analysis of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence for high risk counties and all control counties, western United States, 1977-2000

| Mod. ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 51.92 | 2 | 109.94 | 3.24 | 0.76 | 65 | 87 | 99 | 13 | ELEV_MAX, EVGRN-FOR_MAJ |
| **2** | 49.28 | 3 | 106.72 | 0 | 0.78 | 61 | 87 | 99 | 12 | ELEV_MAX, EVGRN-FOR_MAJ |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MAX= mean elevation, EVGRN-FOR_MAJ = majority evergreen forest.

The same procedure was followed to compare the 15 high case counties to only their 54 neighboring control counties. Univariate analysis provided only one elevation and two temperature variables that were significant, with $p < 0.1$. Stepwise regression was performed on all variables with $p < 0.2$, but the only variable that was significant in both stepwise analyses was the variable for minimum value for average maximum temperature in a county (TMAX_MIN). No attempt to combine any two variables in a model was successful and this particular maximum temperature variable was the most statistically significant and had the lowest AICc value, indicating it was the best model for this particular analysis.

**3.7 Conclusions**

Moderate values of maximum temperature (between 0 and 25°C, depending on the specific variable), elevations above 500 meters, and evergreen forest habitat were associated with TBRF occurrence at the county level, based on both the frequency and logistic regression analyses. The percentage of TBRF case counties was higher than control counties in higher elevation categories and lower than control counties in lower elevation categories for the mean and maximum elevation variables ($p \leq 0.002$). There was a higher case county frequency associated with the middle range of average temperatures when compared to the more even distribution of control counties across all temperature categories, with $p \leq 0.01$ for all six temperature variables. Evergreen forest was the majority land cover in a greater proportion of case counties than control counties, with a difference of 14% observed, and the overall distribution of the land

cover variables was significant (p = 0.04). Additionally, interactions between temperature and both elevation and majority evergreen forest were shown to be statistically significant. Precipitation displayed very little association with TBRF occurrence at the county level, with the exception of the statistically significant maximum precipitation variable observed in the frequency analysis (p = 0.003) and the statistically significant univariate logistic regression model for minimum precipitation. The analyses comparing high risk counties and controls yielded similar associations to those found in the full county level analysis. This indicates that the associations observed between ecologic variables and case counties did not differ in areas where more cases were recorded. The consistency among the multiple analyses performed provides confirmation that there was an association between elevation, temperature, and evergreen forest and TBRF occurrence at the county level for these data.

# CHAPTER 4: ZIP CODE LEVEL ANALYSIS

## 4.1 Background

Following the county level analysis, cases of TBRF were analyzed at the zip code level. County level spatial modeling may be acceptable for diseases occurring in the eastern United States, but sub-county scale analyses of disease risk are preferable in the western United States where counties are often large and encompass considerable environmental variability (Eisen and Eisen, 2008). The finer scale of the zip code level analysis could lead to different, and possibly more precise, results than the county level analysis. Data analyzed for the zip code level analysis included only cases with known zip code of exposure located in California and Washington, two of the states with the highest number of reported TBRF cases (Dworkin et al., 2002a). Examination of TBRF distribution and comparison of areas with TBRF to neighboring areas without TBRF has not been attempted previously at the zip code level.

## 4.2 Specific Aims

- Obtain data regarding TBRF cases at the zip code level using information provided by the state health departments of Washington and California.

- Identify zip codes without cases of TBRF adjacent to those zip codes with reported cases.
- Obtain data for ecologic variables in the areas being studied. Variables to be analyzed included: elevation, land cover designations, precipitation, minimum and maximum temperature.
- Extract values for each potential ecologic covariate using ArcGIS.
- Identify features that are statistically associated with the presence of TBRF at the zip code level using logistic regression models.
- Perform the same analyses for each state individually to determine if any differences exist between states.
- Compare high risk zip codes to control zip codes to identify any unique associations not observed in the complete zip code analysis.
- Discuss the results of the analyses and attempt to draw conclusions from the associations found.
- If a model is successfully created, attempt to apply it to zip codes in Oregon and use it to identify areas of potential increased risk for TBRF.

## 4.3 Methods

### 4.3.1 Case and Control Zip Code Selection

Information was requested regarding reported cases of TBRF from the state health departments of California, Colorado, and Washington. The information requested included: zip code of exposure (or zip code of residence if exposure location was unknown), age, gender, month of onset of illness, and the method by which TBRF was diagnosed. No personal identifiers were provided for any of the individual cases used in this study. Upon further examination of the data collected, zip code of

residence was found to be a poor surrogate for zip code of exposure since very few cases were infected in their zip code of residence. Zip code of residence and zip code of exposure were available for 87 cases in the state of California. Among these 87 cases, zip code of exposure was the same as zip code of residence for only 16 cases (about 18%). This information, coupled with the knowledge that cases were often exposed to TBRF while traveling, often outside their state of residence (Dworkin et al, 2002a), led to the decision to include only cases where zip code of exposure was available in the analysis.

The data received from the Colorado Department of Public Health and Environment did not contain sufficient information to identify zip code of exposure for any of the cases listed and was therefore excluded from this analysis.

TBRF cases that met the inclusion criteria were summed based on zip code where exposure to TBRF occurred. Based on the total number of TBRF cases in each zip code, a category was assigned to each zip code. These case number categories were similar to the categories used in the county level analysis (i.e. 1-3, 4-6, 7-9, or 10 cases per zip code). These categories were not used for statistical analysis, but for map display purposes only. A total of 54 cases from Washington and 87 cases from California were included in the analysis. These 141 cases occurred in 60 different zip codes: 29 in Washington and 31 in California.

Table 4.1: Number of zip codes with reported cases of tick-borne relapsing fever (TBRF), California and Washington, 1990-2010

| Case Categories | Number of Zip Codes |
|---|---|
| 1-3 | 52 |
| 4-6 | 4 |
| 7-9 | 3 |
| 10 | 1 |

Neighboring zip codes without TBRF cases that share a contiguous border were selected to serve as "control zip codes" in the analysis. These zip codes were selected using the "Select By Location" feature in ArcGIS (ESRI, Redlands, CA). Any zip code that touched the boundary of a case zip code was selected using this operation. Control zip codes were only included if they were located in either Washington or California. The selected zip codes were then exported as a layer file in ArcGIS. A total of 193 zip codes were chosen as control zip codes. Both case and control zip codes were displayed using the GCS_WGS_1984 projection.

### 4.3.2 Ecologic Data

The ecologic data analyzed in the zip code analysis were the same data used in the county level analysis. Elevation was derived from a 1 km resolution digital elevation model (USGS/ESRI, Redlands, CA). The elevation for each zip code was assessed using the minimum, maximum and average values for that zip code. This was intended to account for any variability in elevation that may occur across larger zip codes with varied topography. Elevation data were displayed in the GCS_WGS_1984 projection. Land cover classification was derived from the National Land Cover Dataset (USDA, NRCS)

available for each individual state. The land cover type that was identified as the "majority" for each zip code was the one assigned to with that zip code. Land cover layers were displayed in the projections NAD_1983_10N through NAD_1983_14N, depending on the state being analyzed. Data for precipitation, minimum and maximum temperature were derived from individual models of each using annual averages from 1971 to 2000 (PRISM Climate Group, Oregon State University). While this time span included several years outside the years when cases were documented (1990-2010), it provided a good representation of an average value for these variables. Minimum, maximum, and mean values were obtained for each variable to account for any variation that might occur across each zip code.

### 4.3.3 ArcGIS Analysis

Maps were created in ArcGIS in the same manner as the county level analysis, using each of the ecologic variables to be analyzed, as well as the map layers for the case and control zip codes. If the projections were not the same for all layers, the "Project" tool was used to convert layers to the same projection. Once all layers were displayed in the same projection, "Zonal Statistics," located in the "Spatial Analyst" toolbox, was used to generate summary statistics for each zip code from the ecologic variable layers. This information was then added to a spreadsheet in Microsoft Excel 2007 (Microsoft, Redmond, WA), where information from all zip codes was consolidated and organized for statistical analysis.

See Figures 4.1 and 4.2 for the distribution case and control zip codes, respectively.

## 4.4 Statistical Analysis

### 4.4.1 Variable Definition

Statistical Analysis was conducted using methods similar to those used in the county level analysis. All zip codes where TBRF cases were identified were listed as "case" zip codes. Neighboring zip codes with no cases of TBRF reported were listed as "control." The minimum, maximum, and mean values of the four continuous ecologic variables (i.e. elevation, average precipitation, average minimum temperature and average maximum temperature) were each treated as their own variable and assigned the same labels used in the county level analysis. The land cover variable was categorized numerically, with the number representing the majority land cover or habitat type in each zip code. The land cover variable was further divided into a series of design or "dummy" variables, with each one representing an individual land cover type, as was done in the county level analysis. The individual habitat types represented by these variables included: open water, developed land, evergreen forest, shrub/scrub, grassland/herbaceous, and crop/livestock area.

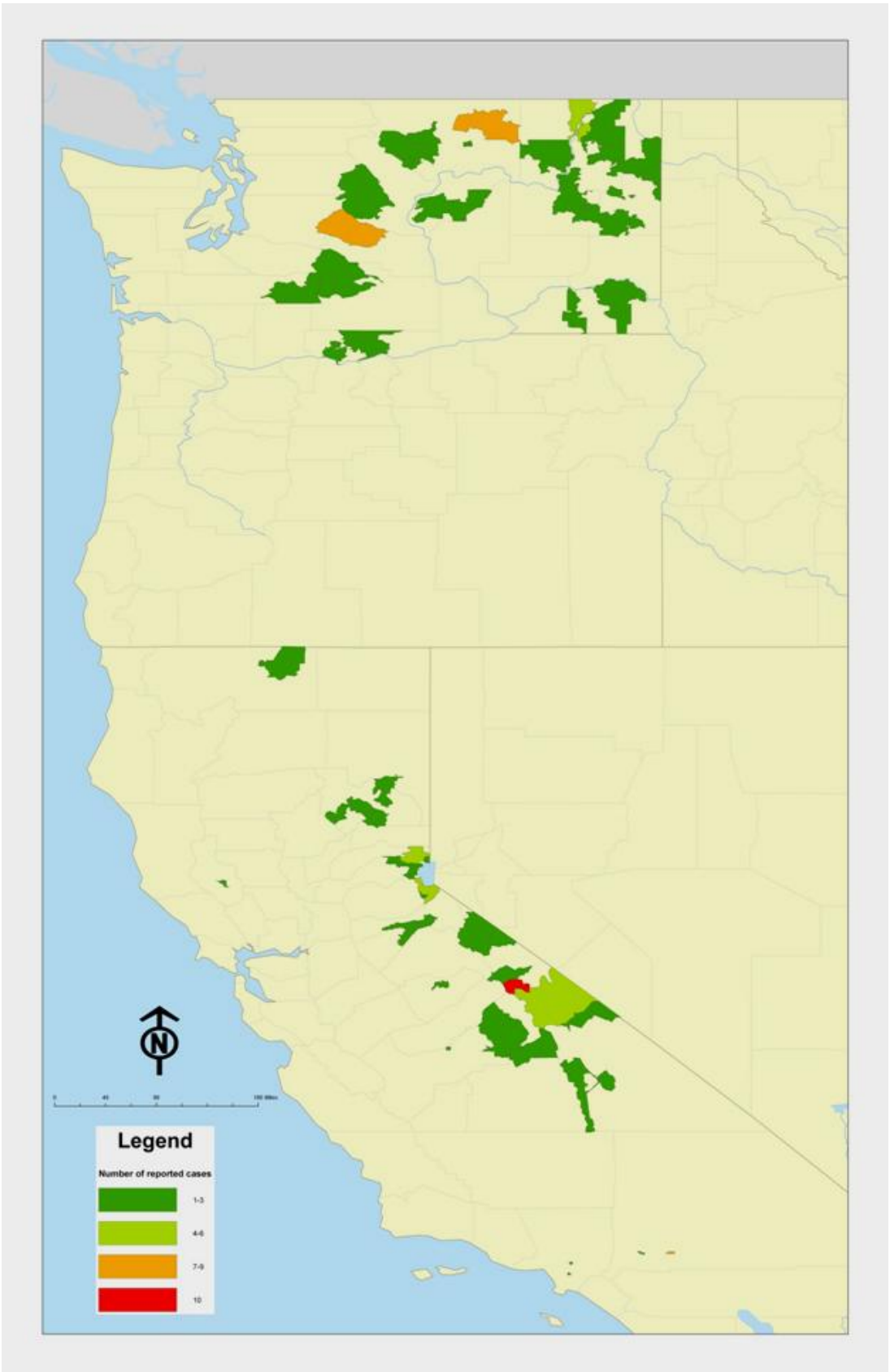


Figure 4.1: Distribution of zip codes with cases of tick-borne relapsing fever (TBRF), 1990-2010


Figure 4.2: Distribution of control zip codes without reported tick-borne relapsing fever cases

### 4.4.2 Frequency Analyses

The distribution of case and control zip codes was compared by frequency analysis as in the county level analysis. Each ecologic variable was divided into categories and the proportion of case and control zip codes in each category was calculated. The proportions of case and control zip codes in each category were compared and a chi-square value was calculated for each variable using the distribution of values observed in the frequency tables. A statistically significant chi-square value indicated a difference in the distribution between case and control zip codes.

### 4.4.3 Logistic Regression Analyses

Binomial logistic regression was conducted on all variables using the methods listed in *Applied Logistic Regression* (Hosmer and Lemeshow, 2000). All statistical analysis was carried out using the JMP (SAS, Cary, NC) statistical software package, version 9.02.

First, a univariable logistic regression analysis was run for each potential covariate against the dichotomous outcome variable "Case Status," which listed zip codes as either "case" or "control." Forward and backward stepwise regression were run as described in the county level analysis, and Spearman correlation coefficients were used to identify correlation among response variables.

Purposeful variable selection was completed, the model coefficients, effect likelihood ratios and Wald statistics were compared to those of the larger models

created through stepwise regression to confirm that no drastic changes occurred in the model because of the elimination of variables. All variables that were not selected for the multivariable model were added again to identify any that may only have an effect in the presence of other variables.

### 4.4.4 Evaluation of Selected Variables and Interaction

After the main variables included in the model were established, the assumption of linearity of all continuous variables was checked. Smoothed scatterplot and quartile analysis were conducted for continuous variables to inform a decision regarding transformation or categorization of the variables. The same methods were applied to variables not included in the multivariable model to confirm that alteration did not produce a variable that made a significant contribution to the model. After the main variables in the model were chosen, interactions between the terms were examined using the methods described in the county level analysis.

### 4.4.5 Comparison of Potential Models

The multiple potential models were compared as before, using various statistical measures, such as Akaike information criterion (AICc) (Akaike, 1974), receiver operator characteristic curves (ROC), and the goodness of fit (or lack of fit) chi-square statistic. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) were calculated for each model and the final multivariable model was selected from the competing models by choosing the one with the highest sensitivity and the best balance among the other three measures.

### 4.4.6 Model Validation

The final multivariable model for the total zip code analysis was validated for use as a predictive model using a leave-one-out method to ensure that the trends observed were not heavily reliant on any one case or control zip code. This method involves removing one zip code from the model, running the model and obtaining the AUC value from the ROC curve. After this is accomplished, the zip code is replaced and the next zip code is removed, the model run again, and the AUC value obtained. This was performed sequentially for every zip code in the model. The average and range of the AUC values were analyzed to confirm that the absence of any single zip code did not drastically affect the overall accuracy of the model that was chosen (Fielding & Bell, 1997).

### 4.4.7 California and Washington Individual Models

The above methods, with the exception of the model validation step, were applied to create individual logistic regression models for California and Washington. Zip codes from each state were separated into two tables and the ecologic data associated with these zip codes were analyzed in the same manner as the data set that included both states.

### 4.4.8 Predictive Risk Model

A predictive model was constructed using the complete zip code level logistic regression model to apply the variables found to be associated with TBRF in Washington and California to the state of Oregon. This was done to highlight areas where TBRF may

be more likely to occur, and where increased surveillance for TBRF may be beneficial. This model may not be completely accurate since the disease is underreported and it would only be based on reported TBRF cases.

The final logistic regression model was entered into the "Raster Calculator" tool in the "Spatial Analyst" toolbox of ArcGIS, version 10. The model is represented by the equation:

$$\text{Logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \text{ [expression 1]}$$

where $P$ is the probability that TBRF is present in a zip code and $\beta_0$ is the intercept. The values $\beta_1$ and $\beta_2$ represent the coefficients of the variables $x_1$ and $x_2$, respectively. This equation produced an output raster layer that was used to define areas as high or low risk in "Raster Calculator" by entering values into the equation:

$$P = e^{\text{Logit}(P)} / (1 + e^{\text{Logit}(P)}) \text{ [expression 2] (Eisen et al., 2010)}$$

The output raster from the above equation corresponded with the probability values observed during the sensitivity and specificity analysis. Using the "Reclassify" tool on the "Spatial Analyst" toolbox, raster values were dichotomized based on the cutoff p-value that maximized sensitivity and specificity. Any values below the cutoff were coded as "0" to signify low risk and any values above the cutoff were coded as "1" to signify high risk. The output from this analysis was a raster layer that displayed areas that were more likely to contain cases of TBRF based on the final model chosen for the total zip code level analysis.

## 4.5 Results

### 4.5.1 Study Population Characteristics

#### I. California

The California Department of Health Services identified and submitted 160 cases of TBRF for analysis between the years of 1990 and 2009. Of these 160 cases, zip code of residence was known for 147 cases and zip code of exposure was known for 87 cases. Cases where exposure to TBRF occurred outside of California were excluded from the analysis. Descriptive summaries included all cases with the appropriate information, not just those used in the final analysis.

The gender distribution of cases in California was heavily skewed towards males, with almost twice as many males infected as females. Of the 156 cases where gender was available, 100 (64%) were male and 56 (36%) were female.

Information regarding age was available for 156 of the cases of TBRF submitted for analysis. The youngest case reported was one year of age; the oldest case was 86 years of age. The mean age was 34 and the median age was 35. A broad spectrum of ages was represented, with most age groups containing 15 to 25 cases. The age group with the greatest number of cases was 40-49. The distribution of case is illustrated in the Figure 4.3, divided into ten year categories.

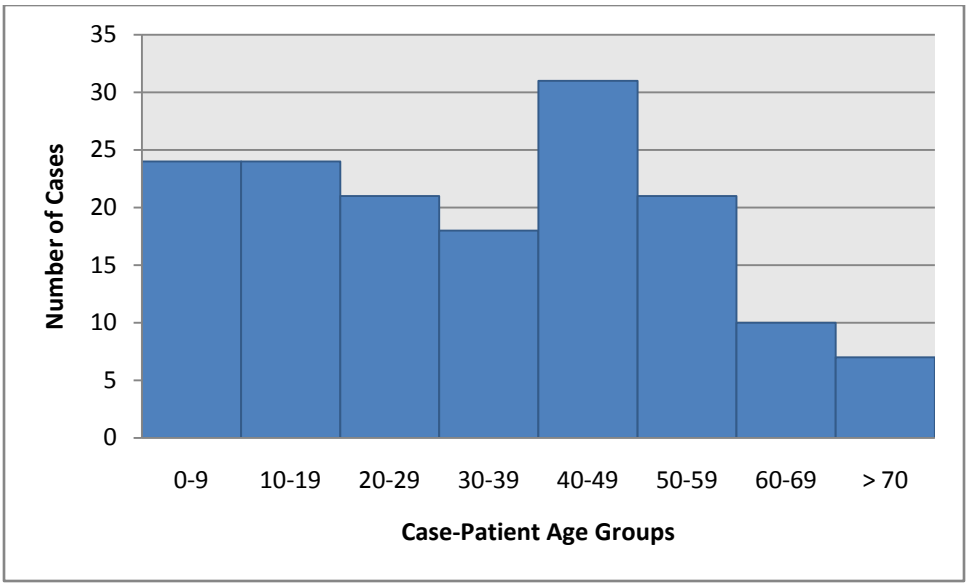


Figure 4.3: Distribution of reported tick-borne relapsing fever (TBRF) cases by patient age, California, 1990-2009

Month of onset of illness was available for 159 cases. Similar to the findings of previous studies (Dworkin et al, 2002a; Dworkin et al, 1998), cases of TBRF present primarily in the summer and early autumn months. Of the 159 cases where month of onset was known: 11% occurred in June, 30% occurred in July, 29% occurred in August, and 11% occurred in September. The distribution of month of onset for all cases in California is available in Figure 4.4.

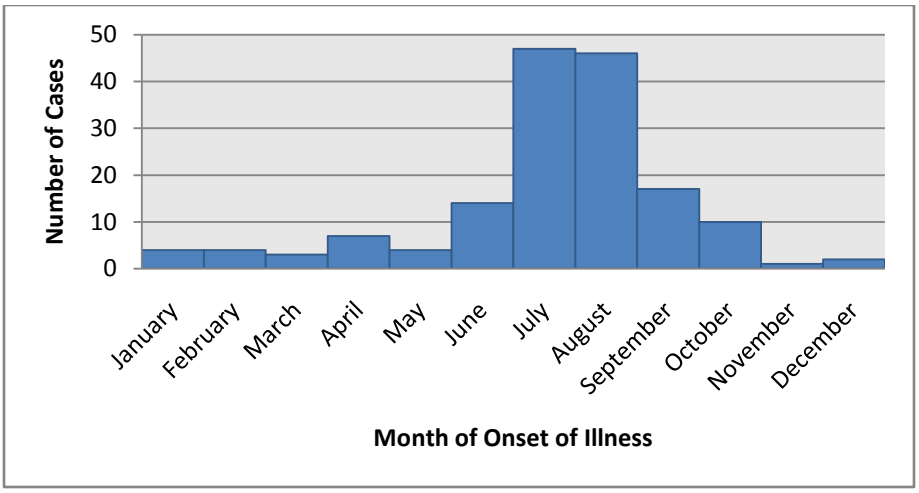


Figure 4.4: Distribution of tick-borne relapsing fever (TBRF) cases by month of onset of illness, California, 1990-2009

The method by which TBRF was diagnosed was available for 149 cases. Observation of spirochetes on a peripheral blood smear was the most common method (77%) and serology was second most common (13%). Cases diagnosed by all methods were included in the analysis to improve statistical power. Only a small number of cases (2%) were not laboratory confirmed by any method. See Figure 4.5 for a summary of diagnostic methods.

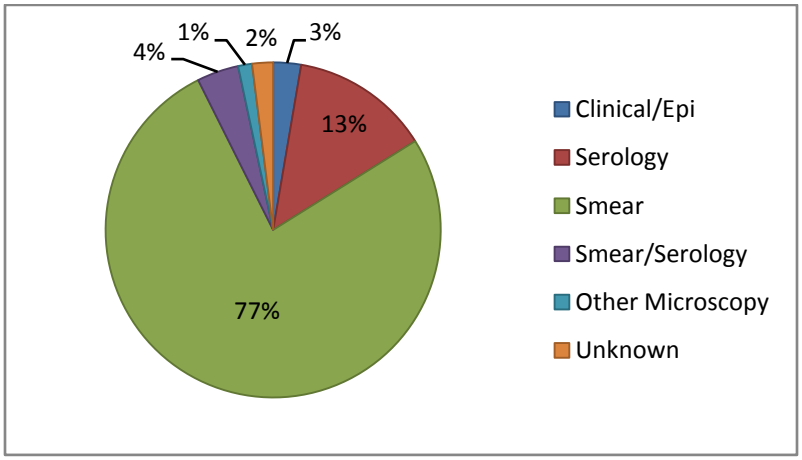


Figure 4.5: Distribution of tick-borne relapsing fever (TBRF) cases by method of diagnosis, California 1990-2009

II. Washington

The Washington State Department of Health identified 115 cases of TBRF that occurred in Washington or were diagnosed in the state of Washington between 1990 and 2010. Zip code of exposure was identified based on the information reported regarding the location of exposure. The majority of the towns in which most cases were exposed have only one zip code, which allowed for precise identification of zip code of exposure. Cases where exposure occurred in a state other than Washington or exposure location information was insufficient for identification of zip code of exposure were excluded from analysis (n = 61). Zip code of residence was provided for only one case, which was excluded from analysis. A total of 54 cases of TBRF where zip code of exposure was identified were included in the analysis.

The gender of TBRF cases was known for all 115 Washington cases. Cases were split almost equally between genders, with 52% male and 48% female. Age data were available for all but one of the cases from Washington. Among cases, the minimum age was less than one year old, the maximum age was 89 years old, the mean age was 35, and the median age was 38. Like the California data, ages were varied with no apparent clustering in any specific age group. The highest number of cases was observed in the 40-49 age group, and similar case numbers were seen in the 0-9 and 10-19 age groups. The distribution of case ages is illustrated in Figure 4.6, with ages divided into 10 year categories.

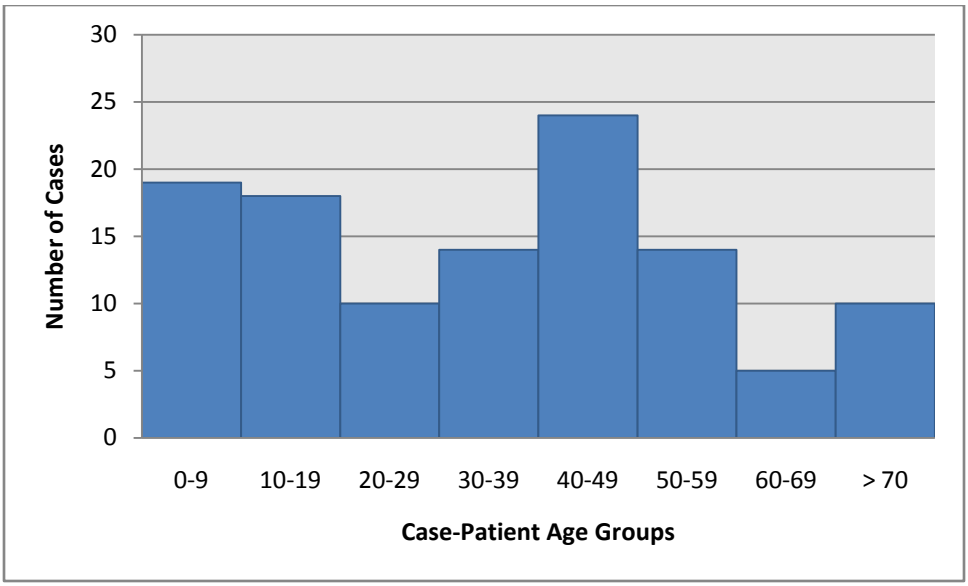


Figure 4.6: Distribution of reported tick-borne relapsing fever (TBRF) cases by patient age, Washington, 1990-2010

Month of onset of illness was available for all 115 cases (Figure 4.7). The seasonal distribution was typical of tick-borne relapsing fever, and other tick-borne diseases: 14% in June, 24% in July, 28% in August, and 11% in September.

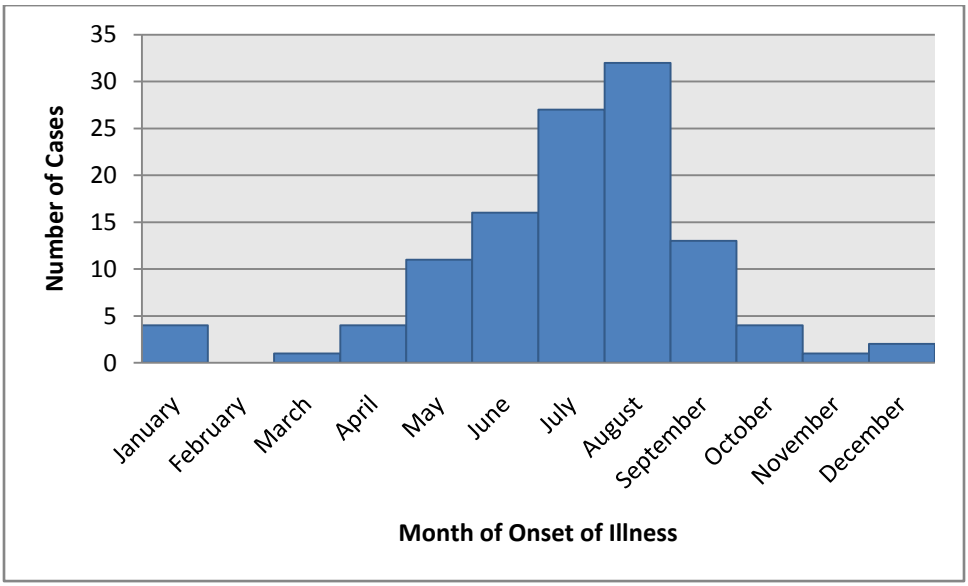


Figure 4.7: Distribution of tick-borne relapsing fever (TBRF) cases by month of onset of illness, Washington, 1990-2010

The method by which TBRF was diagnosed was provided for all 115 Washington cases. As with cases from California, the most common diagnostic methods were blood smear (65%) and serology (11%). For 19 cases, the exact method of diagnosis was unknown but the diagnosis was made through a laboratory test. All cases where zip code of exposure was identified were included in the final analysis to increase statistical power. Cases where diagnosis was based solely on clinical information were not included in the analysis because of the absence of information regarding zip code of exposure. A summary of diagnostic methods is presented in Figure 4.8.

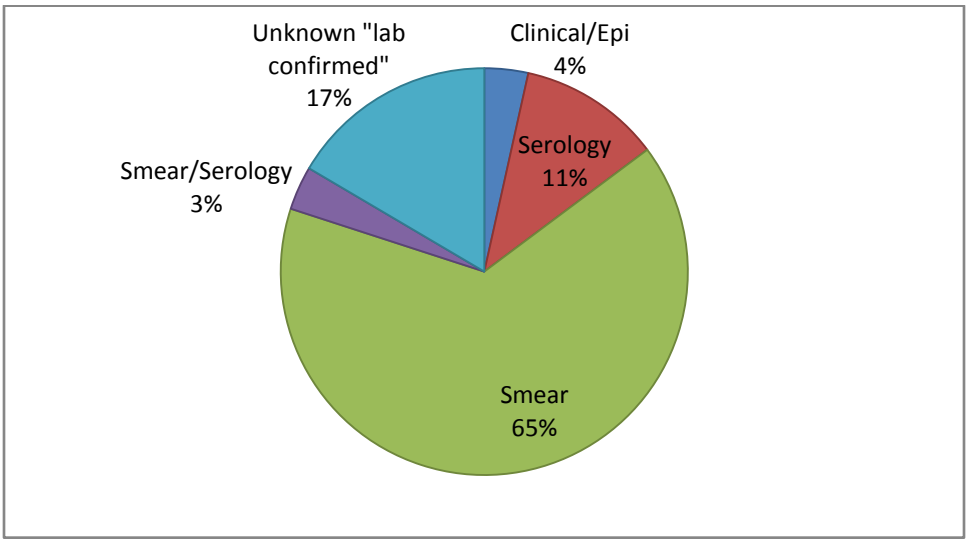


Figure 4.8: Distribution of tick-borne relapsing fever (TBRF) cases by method of diagnosis, 1990-2010

### 4.5.2 Frequency Analysis of Ecologic Variables

Each variable was divided into categories based on the range of the data displayed by the 60 case zip codes and 193 control zip codes being analyzed. In each category, the proportion of case zip codes was compared to the proportion of control zip codes. These

frequency data are displayed in Table 4.2. The three land cover types with the highest number of total zip codes were: evergreen forest (129), shrub/scrub (67), and crop/livestock area (27). Evergreen forest had the highest percentage of case and control zip codes of all land cover types, but a difference of nearly 10% was observed between case zip codes (58.3%) and control zip codes (48.7%). The shrub/scrub land cover classification was the second most common, but it displayed similar proportions among case and control zip codes. Unlike the county level analysis, the overall distribution of the land cover variables was not significant ($p = 0.82$), indicating there was not a significant difference between the distribution of case and control zip codes. As with the county level analysis, all three elevation variables showed case zip codes with lower proportions at lower elevations and higher proportions at higher elevations when compared to the proportions of control zip codes in the same categories ($p \leq 0.03$ for all variables). The precipitation variables were not statistically significant ($p \geq 0.45$ for all variables) and displayed no consistent differences between the proportion of case and control zip codes in each category, with percentages generally higher in the lower and middle categories in both groups. Proportions of case and control zip codes were similar in most categories among the six temperature variables analyzed, with the highest percentages of zip codes observed in the lower and middle temperature categories among both groups; however, all temperature variables were statistically significant ($p \leq 0.08$ for all variables) and higher proportions of case zip codes were observed in some lower temperature categories.

Table 4.2: Distribution across selected ecologic variables of zip codes with reported tick-borne relapsing fever (TBRF) cases and neighboring control zip codes, California and Washington, 1990-2010

| Majority Land Cover | No. of Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **Open Water** | 1 | 1 | 1.7% | 0.5% |
| **Developed Land** | 4 | 18 | 6.7% | 9.3% |
| **Evergreen Forest** | 35 | 94 | 58.3% | 48.7% |
| **Shrub/Scrub** | 14 | 53 | 23.3% | 27.5% |
| **Grassland/Herbaceous** | 0 | 6 | 0.0% | 3.1% |
| **Crop/Livestock Area** | 6 | 21 | 10.0% | 10.9% |

*Land Cover: $X^2 = 5.12$, $p = 0.82$

| Minimum Elevation (meters) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< 500** | 20 | 109 | 33.3% | 56.5% |
| **501-1000** | 20 | 51 | 33.3% | 26.4% |
| **1001-1500** | 8 | 22 | 13.3% | 11.4% |
| **1501-2000** | 5 | 8 | 8.3% | 4.1% |
| **> 2000** | 7 | 3 | 11.7% | 1.6% |
| **Maximum Elevation (meters)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 1000** | 12 | 65 | 20.0% | 33.7% |
| **1001-2000** | 16 | 68 | 26.7% | 35.2% |
| **2001-3000** | 24 | 44 | 40.0% | 22.8% |
| **3001-4000** | 6 | 14 | 10.0% | 7.3% |
| **> 4000** | 2 | 2 | 3.3% | 1.0% |
| **Mean Elevation (meters)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 500** | 5 | 32 | 8.3% | 16.6% |
| **501-1000** | 22 | 90 | 36.7% | 46.6% |
| **1001-1500** | 11 | 33 | 18.3% | 17.1% |
| **1501-2000** | 7 | 21 | 11.7% | 10.9% |
| **2001-2500** | 13 | 12 | 21.7% | 6.2% |
| **> 2500** | 2 | 5 | 3.3% | 2.6% |

*Min. Elevation: $X^2 = 19.13$, $p = 0.0007$; Max. Elevation: $X^2 = 10.83$, $p = 0.03$; Mean Elevation: $X^2 = 14.37$, $p = 0.01$

| Minimum Precipitation (mm) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< 250** | 7 | 32 | 11.7% | 16.6% |
| **250-500** | 29 | 92 | 48.3% | 47.7% |
| **501-750** | 12 | 31 | 20.0% | 16.1% |
| **751-1000** | 9 | 18 | 15.0% | 9.3% |
| **> 1000** | 3 | 20 | 5.0% | 10.4% |

| Maximum Precipitation (mm) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< 500** | 10 | 49 | 16.7% | 25.4% |
| **500-1000** | 18 | 58 | 30.0% | 30.1% |
| **1001-1500** | 15 | 37 | 25.0% | 19.2% |
| **1501-2000** | 10 | 24 | 16.7% | 12.4% |
| **> 2000** | 7 | 25 | 11.7% | 13.0% |
| **Mean Precipitation (mm)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 250** | 1 | 12 | 1.7% | 6.2% |
| **250-500** | 18 | 69 | 30.0% | 35.8% |
| **501-750** | 16 | 39 | 26.7% | 20.2% |
| **751-1000** | 9 | 25 | 15.0% | 13.0% |
| **> 1000** | 16 | 48 | 26.7% | 24.9% |

*Min. Precipitation: $X^2 = 3.71$, $p = 0.45$; Max. Precipitation: $X^2 = 2.92$, $p = 0.57$; Mean Precipitation: $X^2 = 3.37$, $p = 0.50$

| Minimum Tmin (Celsius) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< -5** | 11 | 23 | 18.3% | 11.9% |
| **-5 to 0** | 29 | 55 | 48.3% | 28.5% |
| **0 to 5** | 15 | 83 | 25.0% | 43.0% |
| **5 to 10** | 3 | 21 | 5.0% | 10.9% |
| **> 10** | 2 | 11 | 3.3% | 5.7% |
| **Maximum Tmin (Celsius)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 0** | 6 | 4 | 10.0% | 2.1% |
| **0 to 5** | 36 | 107 | 60.0% | 55.4% |
| **5 to 10** | 13 | 56 | 21.7% | 29.0% |
| > **10** | 5 | 26 | 8.3% | 13.5% |
| **Mean Tmin (Celsius)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 0** | 16 | 28 | 26.7% | 14.5% |
| **0 to 5** | 37 | 116 | 61.7% | 60.1% |
| **5 to 10** | 5 | 37 | 8.3% | 19.2% |
| **> 10** | 2 | 12 | 3.3% | 6.2% |

*Min. Tmin: $X^2 = 12.82$, $p = 0.01$; Max. Tmin: $X^2 = 9.34$, $p = 0.03$; Mean Tmin: $X^2 = 7.84$, $p = 0.05$

| Minimum Tmax (Celsius) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< 5** | 10 | 16 | 16.7% | 8.3% |
| **5 to 10** | 22 | 50 | 36.7% | 25.9% |
| **10 to 15** | 20 | 81 | 33.3% | 42.0% |
| **15 to 20** | 4 | 21 | 6.7% | 10.9% |
| **> 20** | 4 | 25 | 6.7% | 13.0% |

| Maximum Tmax (Celsius) | No. Zip Codes with TBRF | No. of Control Zip Codes | Percentage of Zip Codes with TBRF | Percentage of Control Zip Codes |
|---|---|---|---|---|
| **< 15** | 11 | 14 | 18.3% | 7.3% |
| **15 to 20** | 35 | 117 | 58.3% | 60.6% |
| **20 to 25** | 12 | 44 | 20.0% | 22.8% |
| **> 25** | 2 | 18 | 3.3% | 9.3% |
| **Mean Tmax (Celsius)** | **No. Zip Codes with TBRF** | **No. of Control Zip Codes** | **Percentage of Zip Codes with TBRF** | **Percentage of Control Zip Codes** |
| **< 10** | 1 | 6 | 1.7% | 3.1% |
| **10 to 15** | 40 | 91 | 66.7% | 47.2% |
| **15 to 20** | 14 | 52 | 23.3% | 26.9% |
| **> 20** | 5 | 44 | 8.3% | 22.8% |

*Min. Tmax: $X^2 = 8.24$, $p = 0.08$; Max. Tmax: $X^2 = 7.97$, $p = 0.05$; Mean Tmax: $X^2 = 8.88$, $p = 0.03$

### 4.5.3 Logistic Regression Analyses of Ecologic Variables

Results of the univariate logistic regression analysis were similar to those of the county level analysis. All three elevation variables (all $p < 0.003$) and all six temperature variables (all $p < 0.03$) displayed a statistically significant relationship with the presence of TBRF in a zip code. None of the precipitation variables had a statistically significant relationship with TBRF occurrence. Likewise, the land cover variables, including the variable for majority evergreen forest, were not significant in univariate analyses. Both forward and backward stepwise logistic regression were run on the temperature and elevation variables with the requirement to enter analysis being $p < 0.25$ and the requirement to leave the analysis being $p > 0.1$. The majority land cover variable for evergreen forest was included because of its significance in the county level model, as well as its association with over 50% of case zip codes. Backward stepwise regression yielded a model including only the variable for minimum elevation, while forward stepwise regression yielded a model with variables for minimum elevation, minimum maximum temperature, and majority developed land. The variable for majority

developed land had a p = 0.14, so it was not considered for further analysis. Using these analyses as a guide, a more purposeful variable selection was conducted.

### 4.5.4 Purposeful Variable Selection

Beginning with the minimum elevation variable because of its significance in both stepwise analyses, terms were added to the model individually in order to select potential candidates for the final model. All the variables identified as statistically significant were added to the model containing the minimum elevation variable, and only the variable for minimum value of maximum temperature produced a model in which both terms were significant both individually and as a whole. No other variable added to the model containing these two terms was statistically significant. Likewise, no other combination of variables produced a model that could be considered as an alternative. To be certain no other terms were needed, all variables not included in the two term model were added again and the full model was run. The results of this model indicated that the additional terms were not an improvement over the two term model, with none of them showing statistical significance on an individual level. It was determined that minimum elevation and minimum maximum temperature were the two main terms in the zip code model. A Spearman $\rho = -0.23$ confirmed that the two variables were not correlated with each other and could both be included in the final model.

4.5.5 Assumption of Linearity Evaluation

The assumption of linearity for the two variables included in the model was tested by smoothed scatterplot and quartile analysis, as before. Based on the scatterplot (Figure 4.9), the minimum elevation variable is roughly linear, with percentage of cases increasing with elevation category.

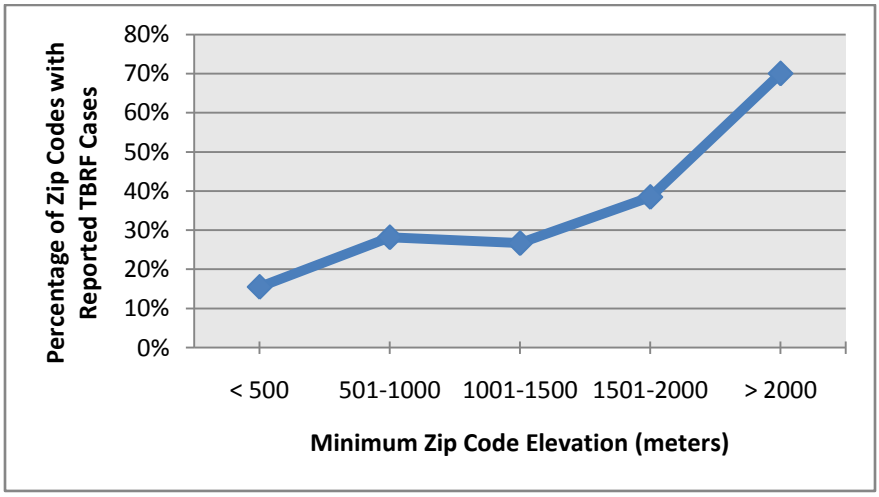


Figure 4.9: Proportion of case zip codes by minimum zip code elevation, California and Washington, 1990-2010. A case zip code is defined as a zip code with reported tick-borne relapsing fever (TBRF) cases; control zip codes are neighboring zip codes without reported TBRF cases. Percentage calculated as the number of case zip codes divided by total case and control zip codes within each environmental category.

Quartile analysis on the minimum elevation variable showed that the three design variables representing the second, third and fourth quartiles were statistically significant when all were included in the same model; however, only the design variable representing the highest quartile was significant individually. There was no improvement to the model when examining these design variables either individually or

in combination with the minimum value for the maximum temperature variable, so it was decided to keep minimum elevation as a continuous variable.

A linear trend was also observed in the scatterplot for minimum value of the maximum temperature variable (Figure 4.10), with the difference being the percentage of case counties decreased with increasing temperature.

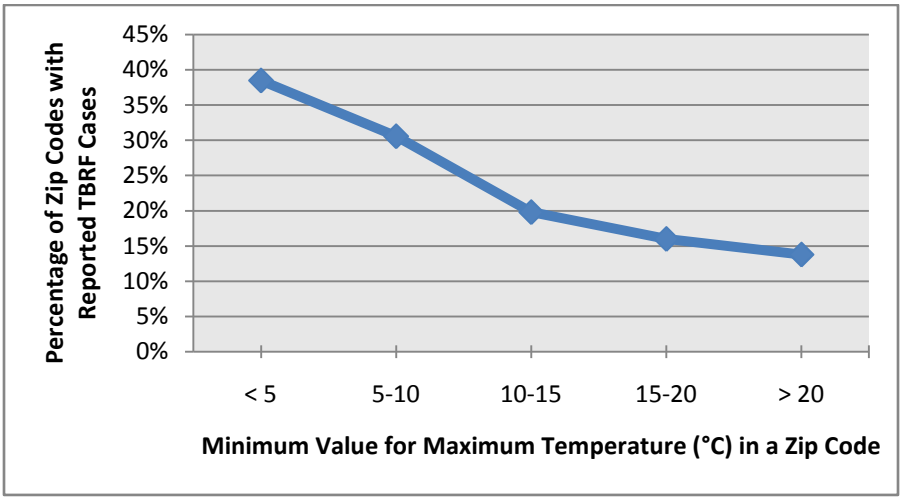


Figure 4.10: Proportion of case zip codes by minimum value for maximum temperature in a zip code, California and Washington, 1990-2010. A case zip code is defined as a zip code with reported tick-borne relapsing fever (TBRF) cases; control zip codes are neighboring zip codes without reported TBRF cases. Percentage calculated as the number of case zip codes divided by total case and control zip codes within each environmental category.

In the quartile analysis, only the design variable for the highest quartile was statistically significant when run with the other two design variables and when run individually. There was no apparent benefit to using the design variable for the highest quartile compared to the original variable so the variable for the minimum value of maximum temperature was left as a continuous variable. A model using the design variables for the highest quartiles of the elevation and temperature variables produced

values similar to those seen in the model with the two continuous variables, but the AICc value was ten units higher than the lowest value; therefore, it was not considered as a candidate for the final multivariable model. Additionally, transformations applied to the variables did not improve their performance in any model.

The precipitation variable that was closest to statistical significance in the univariate analysis (PPT_MAX) was divided into quartiles and design variables created from these categories were analyzed via logistic regression as with other variables. Only the design variable representing the third quartile of the maximum precipitation variable was significant when in a model with the other two design variables, and none of them were significant separately. The significant design variable was no longer significant when included in a model with the elevation and temperature variables in any form, so it was not considered for inclusion in the final multivariable model. All three precipitation variables were correlated with each other, so the results of this analysis were considered representative of all precipitation variables.

### 4.5.6 Interaction Terms

Since only two variables were included in the final model, only the interaction term between the elevation and temperature variables was considered for inclusion in the final model. While inclusion of the interaction term did not affect the whole model test of significance, the interaction term itself was not statistically significant ($p = 0.68$). Since the interaction term did not improve the model, it was not included in the final model.

4.5.7 Comparing Multivariable Models

Goodness of fit chi-square statistics were similar among all four candidate models, with all p-values around 0.3. Since all p-values were not significant, the goodness of fit for each model did not play a role in final model selection. The variables included in the four candidate models are listed below:

1. Minimum elevation
2. Minimum elevation and the minimum value for maximum temperature
3. Minimum elevation, minimum maximum temperature, and interaction term
4. Minimum elevation, minimum maximum temperature, majority developed land

Table 4.3: Candidate models for the zip code level analysis of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence, California and Washington, 1990-2010

| Model ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 131.41 | 1 | 266.87 | 1.99 | 0.65 | 78 | 48 | 32 | 88 | ELEV_MIN |
| **2** | 129.42 | 2 | 264.94 | 0.06 | 0.68 | 67 | 67 | 39 | 87 | ELEV_MIN, TMAX_MIN |
| **3** | 129.34 | 3 | 266.84 | 2.16 | 0.68 | 67 | 66 | 38 | 86 | ELEV_MIN, TMAX_MIN |
| **4** | 128.36 | 3 | 264.88 | 0 | 0.69 | 68 | 66 | 38 | 87 | ELEV_MIN, TMAX_MIN, DEVLP_MAJ |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MIN = minimum elevation; minimum maximum temperature = TMAX_MIN; DEVLP_MAJ = majority developed land

Models must be within two AICc units to be considered competing, which is true of all except Model 3. ROC AUC values were compared and were found to vary no more than 0.04 units between models. Since there was little distinction among ROC AUC values, the model with the highest sensitivity and best balance among specificity, PPV

and NPV was chosen. Model 1 had the highest sensitivity (78%), but was removed from consideration by having the lowest specificity (48%). Models 2 and 4 had nearly identical values among all statistical measures considered. Ultimately, Model 2 was chosen because it behaved similarly to Model 4 while containing one less variable, indicating the variable was superfluous. The most parsimonious model, chosen as the final multivariable model, contains the variables for minimum elevation and the minimum value for maximum temperature in a zip code. Specifics on model parameters can be found in Table 4.4.

Table 4.4: Parameter estimates for the selected multivariate logistic regression model of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence at the zip code level, California and Washington, 1990-2010

| | Parameter Estimates | | | Likelihood ratio test | | |
|---|---|---|---|---|---|---|
| **Model covariates** | **Estimate** | **SE** | **95% C.I.** | **$X^2$** | **df** | **p-value** |
| Intercept | -1.16 | 0.43 | (-2.02, -0.33) | 7.21 | 1 | 0.007 |
| ELEV_MIN | 0.0009 | 0.0003 | (0.0004, 0.0015) | 11.28 | 1 | 0.0008 |
| TMAX_MIN | -0.06 | 0.03 | (-0.12, -0.001) | 3.98 | 1 | 0.046 |

*df = degrees of freedom; ELEV_MIN = minimum elevation; minimum maximum temperature = TMAX_MIN; Whole Model Test $X^2$ = 18.33, df = 2, p = 0.0001; goodness of fit $X^2$ =258.84, p = 0.32

### 4.5.8 Model Validation

The final multivariate model selected above was subjected to a leave-one-out method of model validation to ensure it was not sensitive to any particular case or control zip code (Fielding & Bell, 1997). The method involved removing one zip code, running the model, recording the ROC AUC, replacing the excluded zip code and removing the next zip code. This was done sequentially for all 253 case and control zip codes. The minimum, maximum and mean AUC values are reported below. It was

determined that overall accuracy of the model was not heavily influenced by one specific zip code and none were excluded from the final model.

Table 4.5: ROC AUC values from the leave-one-out validation of the zip code logistic regression model of the relationship between ecologic factors and tick-borne relapsing fever (TBRF) occurrence, California and Washington, 1990-2010

| | Min | Max | Average | Complete Model AUC |
|---|---|---|---|---|
| **ROC AUC** | 0.677 | 0.694 | 0.683 | 0.683 |

*ROC AUC = area under receiver operator characteristic curve

### 4.5.9 Predictive Risk Model

The final model, containing the variables for minimum elevation and the minimum value for maximum temperature, was entered into the Raster Calculator tool in ArcGIS as described in Section 4.5.7. The values for the coefficients and the intercept were entered as they appear in the final equation, but the original raster layers were used in place of the minimum values from each zip code used in the statistical analysis. This was done because the Raster Calculator tool requires a map layer to be used in the preparation of a new raster layer. Since the minimum, maximum and mean temperature and elevation values were correlated with each other, it was considered acceptable to use the source layers from which they were derived as the variables in the regression equation. The output raster layer produced by the first equation was used in place of the "Logit (P)" variable in expression 2, and Raster Calculator used the second equation to produce a second raster layer. Using the "Reclassify" tool in ArcGIS, values in this raster layer were dichotomized into categories based on the cutoff p-value that maximized sensitivity and specificity (p = 0.2341). Any raster values less than

0.2341 were considered low risk and coded as “0” while any values greater than or equal to 0.2341 were considered high risk and coded as “1.” This produced an output raster layer that represented the predictive risk model and displayed the high and low risk areas as two different colors.

Unfortunately, the predictive risk model had very limited ability to distinguish areas of high and low risk. Originally it was planned to create a predictive risk model derived from the zip code model of Washington and California and apply it to Oregon, to determine which areas showed signs of elevated risk. Upon examination of the predictive model layer, the entire state of Oregon was classified as “low risk.” In fact, the layer covered the entirety of the United States and consisted of over 3,000,000 raster units, but only 22 units total were considered “high risk.” All of the 22 units were located in either Washington or California and were located within case zip codes. Since the predictive model provided very little distinction between low and high risk areas, it was not considered for further analysis.

4.5.10 Analysis of High Risk Zip Codes vs. Control Zip Codes

As with the county level analysis, high risk zip codes were examined to determine if they possessed any unique correlations that may have been obscured by the greater number of lower risk zip codes in the total zip code level analysis. The data from the 13 zip codes with more than two cases reported were regressed against both the total number of control zip codes (n= 193) and only the zip codes with which they

share a border (n= 38). The results of these analyses were similar to those observed in the full zip code analysis.

For the analysis which included all controls, all elevation and temperature variables, as well as some precipitation and land cover variables, produced statistically significant univariate regression models. Both stepwise regression models implied that the final model should include an elevation and temperature variable, but only one of each since variables were correlated within groups as before. Purposeful variable selection confirmed that elevation and temperature would be included in the final model; however, several temperature variables produced acceptable models in combination with the same minimum elevation variable. Three candidate models were compared in the same manner described in the complete zip code model. A comparison of the models can be found in Table 4.6. Model 3 was chosen as the final model because it provided the best balance among sensitivity, specificity, PPV and NPV. All three candidate models showed the same correlation between elevation, temperature and TBRF occurrence observed in the total zip code model.

Table 4.6: Candidate models for the analysis of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence for high risk and all control zip codes, California and Washington, 1990-2010

| Mod. ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 33.08 | 2 | 72.3 | 0 | 0.89 | 100 | 63 | 15 | 100 | ELEV_MIN, TMAX_MEAN |
| **2** | 33.49 | 2 | 73.1 | 0.81 | 0.88 | 85 | 73 | 17 | 99 | ELEV_MIN, TMIN_MAX |
| **3** | 33.63 | 2 | 73.4 | 1.09 | 0.88 | 69 | 94 | 45 | 98 | ELEV_MIN, TMAX_MIN |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MIN= minimum elevation, TMAX_MEAN = mean maximum temperature, TMIN_MAX = maximum minimum temperature, TMAX_MIN = minimum maximum temperature.

The analysis of high risk zip codes and only neighboring control zip codes produced fewer significant associations restricted to a few elevation and temperature variables in the univariate analyses. Stepwise regression produced multivariate models that were only significant if they contained multiple temperature variables, which once again were correlated with each other. The only combination of two different variables that produced a viable model was the design variable for the highest quartile of minimum elevation values (DUMMY-ELQ4) and the variable for the maximum value of maximum temperature in each zip code. This multivariate model was compared to the three best univariate models using the measures presented in Table 4.7 to determine which model was the most appropriate choice. Model 1, the only multivariate model, was chosen because of the balance among the values examined. It displayed the same association between elevation, temperature and TBRF occurrence seen in all previous levels of analysis.

Table 4.7: Candidate models for the analysis of the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence for high risk and adjacent control zip codes, California and Washington, 1990-2010

| Mod. ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 23.81 | 2 | 54.1 | 0.33 | 0.76 | 62 | 92 | 73 | 88 | DUMMY-ELQ4, TMAX_MAX |
| **2** | 24.78 | 1 | 53.8 | 0 | 0.75 | 54 | 89 | 64 | 85 | ELEV_MIN |
| **3** | 25.06 | 1 | 54.4 | 0.57 | 0.74 | 69 | 76 | 50 | 88 | TMIN_MAX |
| **4** | 25.12 | 1 | 54.5 | 0.69 | 0.75 | 85 | 66 | 46 | 93 | TMIN_MEAN |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MIN= minimum elevation, DUMMY-ELQ4 = design variable for highest quartile of ELEV_MIN, TMIN_MAX = maximum minimum temperature, TMIN_MEAN = mean minimum temperature.

### 4.5.11 California and Washington Individual Models

The same methods used in the construction and selection of the zip code multivariate model for both states were applied to California and Washington individually. The univariate analysis for California yielded significant models for all three elevation variables and all six temperature variables much like the analysis with both states included. Both forward and backward stepwise regression yielded a model containing only the variable for minimum elevation. A more purposeful variable selection method confirms that no variables are significant when added to the model containing minimum elevation. A smoothed scatterplot revealed a linear distribution for the variable (Figure 4.11), and quartile analysis showed that only the design variable for the highest quartile was statistically significant.

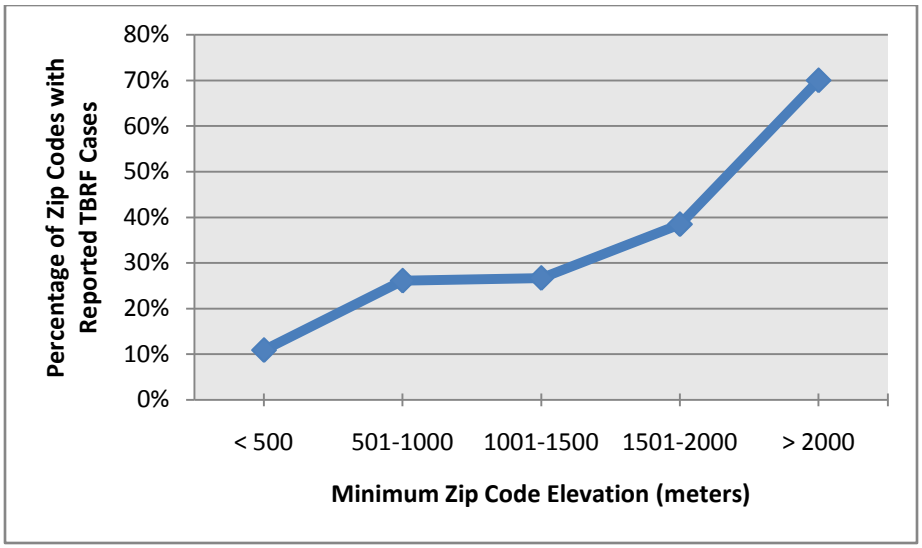


Figure 4.11: Proportion of case zip codes by minimum zip code elevation, California, 1990-2009. A case zip code is defined as a zip code with reported tick-borne relapsing fever (TBRF) cases; control zip codes are neighboring zip codes without reported TBRF cases. Percentage calculated as the number of case zip codes divided by total case and control zip codes within each environmental category.

To further investigate, both the minimum value for maximum temperature and maximum precipitation variables were examined via quartile analysis. None of the precipitation design variables were significant and only the design variable for the top quartile of the temperature variable was significant. No combination of minimum elevation or its design variable and any of these variables produced a model where the second variable added was significant individually. A model containing minimum elevation and the design variable for the highest quartile of the maximum temperature variable was considered because of the AICc value of the model, but it seemed an unlikely choice for the final multivariable model. Interaction terms were considered for the models with more than one variable and the four best models were chosen based on their AICc values. Details are presented in Table 4.8.

Table 4.8: Candidate models for the zip code level analysis of the relationship between ecologic variables and tick-borne relapsing fever occurrence in California, 1990-2009

| Mod. ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 62.06 | 1 | 128.2 | 0 | 0.71 | 71 | 62 | 39 | 86 | ELEV_MIN |
| **2** | 64.45 | 1 | 133.0 | 4.78 | 0.66 | 74 | 57 | 37 | 87 | DUMMY-ELQ4 |
| **3** | 61.87 | 2 | 129.9 | 1.72 | 0.71 | 71 | 64 | 40 | 87 | ELEV_MIN, DUMMY-TX4 |
| **4** | 61.7 | 3 | 131.8 | 3.53 | 0.71 | 74 | 60 | 39 | 87 | ELEV_MIN, DUMMY-TX4 |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MIN= minimum elevation, DUMMY-ELQ4 = design variable for highest quartile of ELEV_MIN, DUMMYTX4 = design variable for highest quartile of minimum maximum temperature variable.

Model 1 was chosen because of its balance of sensitivity, specificity, PPV and NPV. The only model within two AICc units of Model 1 was Model 3, which displayed only very minor changes in the four values being examined despite the inclusion of an additional variable. It was decided that the model containing only the minimum elevation variable was the most parsimonious model for the zip code analysis of California.

Table 4.9: Parameter estimates for the selected logistic regression model for the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence at the zip code level in California, 1990-2009

| | Parameter Estimates | | | Likelihood ratio test | | |
|---|---|---|---|---|---|---|
| **Model covariates** | **Estimate** | **SE** | **95% C.I.** | **$X^2$** | **df** | **p-value** |
| Intercept | -2.34022 | 0.452143 | (-3.30, -1.51) | 26.79 | 1 | <.0001* |
| ELEV_MIN | 0.00126 | 0.000355 | (0.0006, 0.0019) | 14.183 | 1 | 0.0002* |

*df = degrees of freedom; ELEV_MIN = minimum elevation; Whole Model Test $X^2$ = 14.18, df = 1, p = 0.0002; goodness of fit $X^2$ =108.58, p = 0.31

The same methods were applied to zip code data for the state of Washington but the variables behaved differently than the total zip code and California only analyses. None of the variables examined in the univariate analysis were significant.

Since no variables were significant individually, forward and backward stepwise regression were conducted to determine if any variables were significant in combination. Backward stepwise regression produced a model with variables for maximum temperature, and mean and maximum precipitation. All three variables are significant, but the 95% confidence interval for maximum temperature's coefficient contains zero, and the two precipitation variables are highly correlated with each other ($\rho = 0.93$). Attempting to remove either precipitation variable yielded a model that is not statistically significant. Forward stepwise regression produced a model containing two elevation variables, the mean value for minimum temperature and majority crop/livestock land cover. The model as a whole was not significant and none of the terms were significant individually ($p \geq 0.1$). Removing any of these variables did not produce a combination that improved on the model containing all of them. Attempting to create a model that included the same terms as the total zip code level model (minimum elevation and maximum temperature) produced one in which the terms were significant ($p < 0.1$). Adding the interaction term did little to improve the model, but it was considered as a potential candidate. Quartile analysis was performed on minimum elevation, the minimum value for maximum temperature, and maximum precipitation as before, to determine if the variables were statistically significant when examined as categories. Only the design variables for the two middle quartiles of minimum elevation were significant together and none of the design variables for the temperature and precipitation variables were significant in any combination. The two candidate models were compared by AICc, sensitivity, specificity, PPV and NPV values.

Table 4.10: Candidate models for the zip code level analysis of the relationship between ecologic variables and tick-borne relapsing fever occurrence in Washington, 1990-2010

| Mod. ID | Negative log-likelihood | K | AIC | Δ AIC | ROC AUC | Sensitivity | Specificity | PPV | NPV | Independent model variables |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 66.59 | 2 | 139.4 | 0 | 0.66 | 62 | 71 | 38 | 87 | ELEV_MIN, TMAX_MIN |
| **2** | 66.34 | 3 | 141.1 | 1.68 | 0.67 | 69 | 65 | 36 | 88 | ELEV_MIN, TMAX_MIN |

*K = number of estimated parameters in the model; AICc = Akaike information criterion; ROC AUC = area under receiver operator characteristic curve; PPV = positive predictive value; NPV = negative predictive value; ELEV_MIN= minimum elevation, TMAX_MIN = minimum maximum elevation.

Model 1 contained only the two main terms, while Model 2 included the interaction term as well. Sensitivity and ROC AUC improved with the addition of the interaction term, but specificity decreased. Since the improvement in sensitivity was minor, and addition of the interaction term caused the model to have a non-significant p-value (p = 0.12), Model 1 was chosen as the final multivariate model. Additional detail on the model is shown in Table 4.11.

Table 4.11: Parameter estimates for the selected multivariate logistic regression model for the relationship between ecologic variables and tick-borne relapsing fever (TBRF) occurrence at the zip code level in Washington, 1990-2010

| | Parameter Estimates | | | Likelihood ratio test | | |
|---|---|---|---|---|---|---|
| **Model covariates** | **Estimate** | **SE** | **95% C.I.** | $X^2$ | **df** | **p-value** |
| Intercept | -1.244 | 0.716 | (-2.72, 0.12) | 3.02 | 1 | 0.08 |
| ELEV_MIN | 0.002 | 0.001 | (-0.0003, 0.005) | 3.03 | 1 | 0.08 |
| TMAX_MIN | -0.093 | 0.055 | (-0.203, 0.014) | 2.91 | 1 | 0.09 |

*df = degrees of freedom, ELEV_MIN = minimum elevation, TMAX_MIN = minimum maximum elevation; Whole Model Test $X^2$ = 5.32, df = 2, p = 0.07; goodness of fit $X^2$ =133.18, p = 0.36

## 4.6 Conclusions

As with the county level analysis, the zip code level analysis suggests that higher elevations and lower maximum temperature in a zip code were associated with occurrence of TBRF cases. Analysis suggested that increasing elevation was associated with increasing case zip code frequency ($p = 0.0007$), while case frequency was higher than control zip code frequency among lower categories of most temperature variables ($p = 0.08$). No consistent associations were observed among average precipitation variables (all $p \geq 0.45$), further suggesting there was no relationship between these variables and TBRF occurrence at the county or zip code level. While no land cover variable was statistically significant ($p = 0.82$), the majority land cover type was evergreen forest in most of the zip codes analyzed.

Ultimately, the final multivariate model indicated an association between elevation, temperature and occurrence of TBRF cases at the zip code level. This increased the credibility of the county level analysis and reinforced the validity of these findings. Likewise, the results of the high risk zip codes and the individual state analyses were consistent with the full analysis, despite a much smaller sample size. The consistency of the associations observed across all levels of analysis provided confidence in the results of these analyses.

# CHAPTER 5: DISCUSSION

## 5.1 County Level Analysis

For the full county level analysis, 140 counties with cases of TBRF were compared to 243 control counties in order to determine if any associations existed between the TBRF occurrence and variables for the five ecologic factors examined. The majority land cover type was chosen because it was the measure that represented the character of the county and the type of area where individuals would be most likely to be exposed to TBRF. The frequency analysis indicated that, when compared to control counties, a higher proportion of case counties had evergreen forest as the majority land cover and a lower proportion of case counties had crop and livestock areas as the majority land cover ($p = 0.04$). This indicated an association between TBRF and counties which primarily contained evergreen forest, which was consistent with the increased risk of TBRF exposure in rustic cabins (Dworkin et al., 2002a). The distribution of counties observed in the elevation variables implied that there were greater proportions of case counties than control counties at higher elevations (> 500 meters). The minimum elevation variable ($p = 0.25$) was not associated with TBRF occurrence, further reinforcing the idea that TBRF occurrence is associated with higher elevations. The associations between case counties, higher elevations and evergreen forest are consistent with the typical habitat type of *Ornithodoros hermsi*, the primary vector of TBRF in the western United States (Dworkin et al., 2002b). Among the precipitation

variables, no statistically significant difference in proportions was observed between case and control counties, with the exception of the maximum precipitation variable ($p = 0.003$). The significance of this single precipitation variable could indicate a relationship between higher precipitation and TBRF occurrence, possibly related to the tick vector or its preferred rodent host. Conversely, the significance of the maximum precipitation variable could be an errant result due to chance and the lack of significance displayed by the minimum and mean variables could be a more accurate representation of the relationship between TBRF occurrence and precipitation. No association between TBRF occurrence and precipitation has been previously documented in the literature to date and these findings suggest there is no consistent association at the county level. Case proportions were highest among the middle temperature categories, implying that TBRF exposure was more common in areas with a moderate climate (all $p \leq 0.01$). Control counties were evenly distributed among categories in most temperature variables, lending further support for an association between TBRF and temperature at the county level. Associations similar to those found in the frequency analysis were observed in the logistic regression analyses.

The logistic regression model developed to explore the relationship between these ecologic variables and TBRF occurrence at the county level included variables for mean elevation, the mean value for maximum temperature, and majority evergreen forest land cover type. Statistically significant interaction terms between the maximum temperature variable and both the elevation and majority evergreen forest variables were identified and included in the final model. Higher elevations generally have lower

maximum temperatures, and areas with evergreen forest tend to have cooler climates. It was relationships such as these that contributed to the significance of the interaction terms.

All of the logistic regression models produced in this analysis had a goodness of fit statistic that was statistically significant. The significant goodness of fit statistic indicated that it was possible that other factors influence the occurrence of TBRF that were not included in this analysis; however, given the breadth of analyses performed, it was not probable that the inclusion of any of the variables considered would have improved this measure. Further, the exact goodness of fit statistic used by the JMP software package was not identified and may not be the most appropriate test for this particular data set. No other option for goodness of fit tests was available.

While the final model was the best option among the four most likely models in this analysis, its overall accuracy was not very high. The AUC of the ROC was below the range considered acceptable discrimination ($0.7 \leq ROC < 0.8$) and well below the range considered excellent discrimination ($0.8 \leq ROC < 0.9$) (Hosmer & Lemeshow, 2000). The model did discriminate between case and control counties, but with less accuracy than desired. Similarly, sensitivity was lower than ideal, with the model's ability to correctly identify cases only slightly above 60%. Specificity was better, but the overall ability of the model to correctly discriminate between case and control counties was not exceptional. However, the ability to discriminate between case and control counties for a rare disease, based only on ecologic variables, is evidence of the association between these variables and TBRF occurrence.

The analyses that focused on the high risk counties provided results similar to those seen in the analysis of all case counties. When high risk counties were compared to all the control counties variables for elevation and majority evergreen forest were significant in combination; when the high risk counties were compared to only neighboring control counties, a univariate model containing only a maximum temperature variable was judged to be the best model. These were the same associations observed in the final multivariable model for the complete county analysis, with the exception being the maximum elevation variable and the minimum value for the maximum temperature variable were included, rather than the variable for the mean value of each. Since the same associations were detected, the inclusion of fewer variables in the high risk county analyses was a product of reduced sample size. Even with the loss of over 100 counties, these variables were still associated with TBRF occurrence at the county level. The higher case numbers observed in some of these counties can be explained by TBRF outbreaks that occurred at some point during the span of time included in the study. At least four of the high risk counties had a documented TBRF outbreak occur between 1977 and 2000 (Paul et al., 2002; MMWR, 1990; Trevejo et al., 1998; Fritz et al., 2004). The increased case numbers of TBRF seen in these counties could also be related to "better awareness and reporting of TBRF in those counties, greater popularity of those sites for human visits, a greater density of the tick vector population in those areas, or a combination of these factors" (Dworkin et al., 2002a). The logistic regression analyses suggest that the difference between

counties with higher and lower case numbers was not accounted for by the ecologic variables examined in these analyses.

The final model chosen to represent the relationship between TRBF occurrence and the ecologic variables analyzed at the county level showed an association between temperature, elevation and evergreen forest habitat. While similar associations have been observed during outbreak investigations, these associations still existed when analyzing the cases at the scale of the county where the cases were infected. If associations can be detected using only these variables on this scale, it is possible that a model including additional variables on a finer scale may be able to accurately map areas of increased risk of TBRF.

**5.2 Zip Code Level Analysis**

In the zip code analysis, 60 zip codes in California and Washington, identified as exposure locations for cases of TBRF, were compared to 193 surrounding control zip codes to determine if detectable associations existed between TBRF occurrence and the ecologic variables at the zip code level. Similar to the county level frequency analysis, evergreen forest had the highest proportion of case and control zip codes among land cover variables. A difference of 9.6% was observed between case zip codes and control zip codes for the majority evergreen forest variable, indicating that evergreen forest may be associated with TBRF occurrence at the zip code level as well; however, this association was not statistically significant ($p = 0.82$). As with the county level analysis, all three elevation variables were statistically significant (all $p \leq 0.03$), with case zip codes observed in lower proportions at lower elevations and higher proportions at

higher elevations when compared to proportions of control zip codes in the same categories. There was no statistically significant difference between the proportions of case and control counties in the three precipitation variables (all $p \geq 0.45$). This further confirmed that there is no consistent detectable association between TBRF occurrence and precipitation at the zip code or county level. Unlike the county level analysis, case zip code proportions were higher than control zip codes in lower temperature categories of most of the six temperature variables. All temperature variables were statistically significant (all $p \leq 0.08$) and both case and control zip codes were more common in the lower and middle temperature categories. The findings of the frequency analysis did not deviate substantially from those of the county level analysis, with the exception of case proportions being higher among lower values rather than middle values in the temperature variables, the smaller difference between case and control zip codes for the majority evergreen forest variable, and the lack of statistical significance in the distribution of land cover variables. The results of the frequency analysis were largely confirmed in the logistic regression analyses.

As with the county level analysis, associations were found between an elevation and a maximum temperature variable and occurrence of TBRF at the zip code level. For the county level analysis, the mean values for the two variables were included in the final model while the minimum values of the variables were included in the zip code model. All three variables within each group (e.g. minimum, maximum and mean elevation) were correlated with each other at both levels of analysis, so the specific

variables that were chosen (ELEV_MIN and TMAX_MIN) were the most statistically significant within that particular group.

Unlike the county level analysis, none of the zip code level land cover variables showed a significant association with TBRF occurrence in logistic regression analyses. The variable for majority evergreen forest was represented in over 50% of case zip codes, but was also represented in 50% of control zip codes. The lack of significance of land cover variables in the zip code level logistic regression analyses may have been a product of smaller scale, potentially leading to case and control zip codes being over-matched. Many of the counties included in this study were large, increasing the likelihood that neighboring case and control counties would be ecologically dissimilar based on the area included. The smaller area of most zip codes led to more case and control zip codes being characterized by the same variable for majority land cover, simply because smaller neighboring areas are more likely to be ecologically similar.

The goodness of fit chi-square statistic was not significant, indicating sufficient explanatory terms were included in the model. There were almost certainly other factors that influence presence or absence of TBRF in a zip code, but none of the variables excluded from the final model improved it in any way. Why the goodness of fit statistic was significant in the county level analysis but not the zip code level analysis remains unclear.

As with the county analysis, the overall accuracy of the final model was not as high as preferred, but not dismally low. The ROC AUC value was 0.68, which

approached, but was not included in the range of values ($0.7 \leq ROC < 0.8$) that produces "acceptable discrimination" (Hosmer & Lemeshow, 2000). The sensitivity and specificity values were similarly lower than ideal, but close to values that would have provided an acceptable ability to correctly identify true positives and negatives. Compared to the final model that was selected, Model 1 in the total zip code analysis provided an improved estimate of sensitivity (+ 11%) but the drastic decrease in specificity (- 19%) removed it from consideration for the final model. An acceptable balance of all such measures was achieved in the final model, but addition of terms not considered in this analysis may improve the accuracy of this model. The smaller scale of the zip code model led to only a marginal improvement in sensitivity (+5%) and a slight decrease in specificity (-2%) when compared to the county model. Since only two states were included in the zip code analysis compared to the twelve states included in the county analysis, the reduction in scope led to the lack of improvement of overall accuracy for the zip code model.

The results of the analyses comparing high risk and control zip codes did not deviate drastically from the total zip code analysis. The analysis that compared high risk zip codes to all control zip codes yielded several different potential models, all containing the minimum elevation variable and a temperature variable. The model that was chosen contained the same variables as the final model in the complete zip code analysis. The analysis that compared high risk zip codes only to neighboring control zip codes identified correlations with fewer ecologic variables because of the reduced sample size. Ultimately, a final model was chosen that contained the design variable for

the highest quartile of the minimum elevation variable and the maximum value of the maximum temperature variable. The associations between high risk and control zip codes were the same combination of elevation and temperature observed in the full zip code analysis. This implied that zip codes with higher case numbers do not appreciably differ from zip codes with lower case numbers as far as these ecologic variables are concerned. It is likely that the same factors that may have led to increased case numbers at the county level (outbreaks, greater human visitation, etc.) contributed to the higher case numbers observed in these high risk zip codes.

The comparison of TBRF occurrence between California and Washington reinforced conclusions drawn from the total zip code analysis and highlighted differences between the two states. Overall, California had much more variety among zip codes with a greater range of values in most continuous variables. This was easily explained by the state's size and ecologic diversity and the fact that zip codes used in this analysis came from areas across California. The range displayed in the data allowed many variables to be significant individually and led to a univariate model being chosen as the final model. Washington had much less variety in terms of the range of data represented, so differences between case and control zip codes were less pronounced. Like California, case zip codes were distributed across the state; however, Washington is a smaller and less ecologically diverse state than California, which led to the more limited range in data among the continuous variables. This limited data range may have been the cause of no single variable having a statistically significant association with the presence or absence of TBRF among Washington zip codes. Additionally, the only

combination that produced an adequate model was the two variables that comprised the total zip code model. Each state's data contributed an aspect to the total zip code model: California's data showed a strong association between TBRF occurrence and elevation, while Washington's data only produced a significant model with the inclusion of both elevation and temperature. Individual models of states with stark differences in ecologic factors (e.g. Oregon and New Mexico) may provide more insight into different factors that influence TBRF occurrence across the western United States.

The predictive risk model failed to adequately distinguish between areas of high and low risk, or rather areas that are more or less likely to have cases of TBRF. The failure of this model had many contributing factors. The statistical analysis from which the multivariate model was created was conducted using each zip code as one data point, and from that a single value was chosen for each variable to represent that zip code. The coefficients and intercepts from that equation were then applied to the raster layers from which they were derived, which contained a much greater range and amount of data than was used to create the multivariate model. Despite the variables within each variable type being correlated with one another, it is probable that examining the entire range of values for a zip code would produce a different result than a single value representing that entire area. The scale and scope of these analyses also played a role in the failure of the model. For instance, examination of the ecologic variables at the site of an outbreak may produce a predictive risk model that could be applied to surrounding local areas.

Additionally, the overall accuracy of the model used to create it almost certainly played a role in the failure of the predictive risk model. The ROC AUC value indicates that the model was unable to produce acceptable discrimination between case and control zip codes. With a sensitivity of 67% and a specificity of 67%, the model does a poor job of correctly identifying case and control zip codes. The ecologic variables considered in this analysis were only a small portion of factors that may influence TBRF occurrence. Identification and inclusion of additional factors of influence should yield a more accurate predictive risk model. Finally, the data available for this analysis were limited in terms of areas where cases have been reported. Since this is almost certainly not a complete accounting for cases of TBRF in these two states, the quality of the predictive model produced would have been poor even if it had provided better distinction between areas of high and low risk.

**5.3 Biological Perspectives**

The analyses conducted examined only the chosen ecologic variables and the occurrence of TBRF in the western United States. Beyond these ecologic factors, no additional explanatory measures were considered to describe TBRF occurrence in these areas. Since broad ecologic measures were used in analysis, it is almost certain that unaccounted for biological factors are being described by the model. Optimal areas for TBRF transmission contain ecology that is hospitable to the *Ornithodoros* tick vectors and their preferred rodent hosts, such as chipmunks and ground squirrels, and are also commonly visited by their incidental human hosts (Dworkin et al., 2002a). *Ornithodoros*

*hermsi* tick habitat, typically areas of coniferous forest at elevations between 1,500 and 8,000 feet (Dworkin et al., 2002a), is represented in the patterns observed in the land cover and elevation variables. Evergreen forest was the majority land cover in more case than control areas and higher proportions of case areas were observed at higher elevations (> 500 meters) at both the county and zip code levels. It is also possible that areas with reported TBRF cases are more likely contain rustic cabins or vacation homes, or to be popular travel destinations in the summer months. The presence of rustic dwellings or increased human visitation could be features that distinguish counties or zip codes with cases from control counties and zip codes with similar ecologic characteristics. Several TBRF outbreaks were located near or within national parks, such as the north rim of the Grand Canyon (Boyer et al., 1977; Paul et al., 2002) and Estes Park, Colorado (Trevejo et al., 1998). Others occurred in similarly popular outdoor travel destinations, such as Big Bear Lake, California (MMWR, 1998) and Browne Mountain in Washington (Thompson et al., 1969). Future spatial analyses will provide a more complete explanation of factors that influence where the disease occurs if an attempt to account for the peridomestic nature of TBRF transmission is made.

**5.4 Comparison of Findings with Previous Literature**

No publications examining TBRF occurrence and ecologic factors at either the county or zip code level were found in the published literature to date. Results similar to those observed in the zip code analysis regarding seasonality of TBRF were reported in a previous publication (Dworkin et al., 2002a). *Ornithodoros hermsi* habitat was

previously characterized as higher elevation areas consisting of primarily coniferous forest (Dworkin et al., 2002a), which corresponds to the associations found with higher elevations at the county and zip code levels and evergreen forest land cover at the county level. No association between TBRF occurrence and temperature has been published, but the lack of similar ArcGIS analyses for this disease suggests that this association was not previously explored. Similar analyses were used to examine smaller areas in the western United States to explore relationships between ecologic factors and plague (Eisen et al., 2007b) and shared risk of plague and hantavirus (Eisen et al., 2007a). Additionally, relationships between disease occurrence and ecologic factors were successfully identified at both the county level with tularemia (Eisen et al., 2008b) and the zip code level with West Nile virus (Winters et al., 2008).

## 5.5 Study Strengths and Limitations

### 5.5.1 Study Strengths

This study contained the most complete information available in published literature regarding cases of TBRF at the county-level (Dworkin et al., 2002a), in addition to the new zip code level data from state health departments. Analysis conducted at both the county and zip code level, as well as several analyses within these levels, allowed for comparisons to be made between scales and strengthened similar conclusions observed in multiple analyses. Only cases with known zip code of exposure were included in the analysis, eliminating reliance on zip code of residence as a surrogate. The data used in the variables for minimum temperature, maximum

temperature, and precipitation were an average value for that area calculated using annual average values for a 30 year period. This approximated an average value in these ecologic factors for a sizable period of time. Use of minimum, maximum and mean values for temperature and precipitation in each county or zip code increased the likelihood that associations with TBRF occurrence would not be missed solely because of the summary measure chosen to represent that area.

### 5.5.2 Study Limitations

While these were the most complete data available, TBRF is still an underreported disease. Conclusions drawn from these case areas were valid for the case areas, but some associations may be missed because data about TBRF cases was incomplete and not all cases were reported. Cases that were reported as positive based on serological evidence alone could be the result of an earlier infection, making the exposure location incorrect. Likewise, the cases with only a clinical history could have been misdiagnosed, leading to inclusion of a zip code where no actual TBRF cases were documented. Misclassification could also be a concern with the "probable" case group in the publication used for the county level analysis (Dworkin et al., 2002a). If some probable cases were not actual TBRF cases, it may bias the outcome toward the null.

The data analyzed at both the county and zip code levels of analysis were case counts that measured prevalence of reported TBRF in an area over a period of decades. These were not density dependent or population based measures and provided no

measure of incidence for this disease. This issue is further complicated by the fact that that TBRF is often contracting while traveling (Dworkin et al., 2002a), implying that cases exposed in an area may not be a part of the population of that area. A measure accounting for population or human visitation in an area being analyzed should be incorporated into future analyses.

This study was conducted on the county and zip code scales, therefore findings cannot be applied to individual TBRF cases mapped below the zip code level. The associations identified by this study were ecologic in nature, and do not provide a complete description of factors required for TBRF transmission. Given the often peridomestic nature of TBRF exposure, there were likely other factors that affected TBRF transmission on a much smaller scale than was examined during the course of this study. These unaccounted for confounding factors could have influenced the results of these analyses. For instance, the amount of rustic cabins, the primary exposure location for *B. hermsii* infection, in a county or zip code is a probable confounder of the association between TBRF occurrence and the ecologic factors measured because it is related to both the exposure and the disease. A measure accounting for rustic dwellings in an area should be incorporated into any similar analyses conducted in the future.

Neighboring counties and zip codes were chosen as controls because they were likely to be ecologically similar to areas with cases. This choice in control areas could have led to case and control areas being matched on the ecologic factors examined in

these analyses. Since counties in the western United States are large and ecologically diverse, case and control counties were less likely to display ecologic similarities. However, the smaller scale of the zip codes analyzed increased the likelihood that case and control zip codes would be ecologically similar, possibly leading to case and control areas being over-matched at this level of analysis. This could lead to control zip codes being more similar to case zip codes, diminishing the ability to differentiate between case and control zip codes for the ecologic variables analyzed. This ecologic similarity between smaller, neighboring areas may have contributed to the lack of significance of the majority evergreen forest land cover variable in the zip code logistic regression analyses. This could be addressed by the random selection of unaffected counties and zip codes throughout the state being analyzed, rather than the use of neighboring control areas.

Since this study used exposure at the group level for the areas studied, there is potential for ecological fallacy in the interpretation of the results. This could lead to identification of significant associations that were not true, and incorrect conclusions based on these associations. This is especially true of the majority land cover variables. For example, if a county or zip code with TBRF cases is characterized by majority evergreen forest, but contains smaller areas of deciduous forest, it is possible that cases were infected in deciduous forest areas of that county or zip code; however, since the majority land cover variable in that area was evergreen forest, it would be inaccurately associated with TBRF occurrence. The county and zip code level measures of the ecologic variables used in these analyses do not represent the precise locations in which

cases were exposed to TBRF and the results of these analyses should not be applied to individual TBRF cases.

The temperature and precipitation variables used for this analysis provided average values for a 30 year period from 1971 to 2000. These variables did not account for possible substantial changes that may have influenced TBRF occurrence, but provided an average value for temperature and precipitation in the areas studied. Additionally, the years included in the average for these variables do not completely coincide with the years TBRF cases were reported in both levels of analysis, but the average of the variables was considered an acceptable indication of the behavior of the variables over a similar span of time. Likewise, the map layers for land cover of each state were only accurate when they were created. Since both analyses encompassed TBRF cases occurring over a period of 20 years or more, the landscape of the areas where these cases were exposed may not remain constant. Using the majority land cover for each area analyzed was done to capture the overall character of each county or zip code, which may be less likely to change over time than a specific area within a county or zip code.

**5.6 Recommendations for Future Studies**

This study provided evidence of associations between ecologic factors and TBRF occurrence at the county and zip code levels, but the associations observed did not fully characterize areas where TBRF was observed. Analyses could be repeated at either the county or zip code level, but additional factors related to TBRF occurrence should be

investigated. Previous studies have identified a strong association with staying in a rustic cabin and exposure to TBRF. Exploring this association on a county or zip code level by using some measure as a surrogate for the amount of rustic cabins in an area, while accounting for the ecologic variables explored in this analysis, may further refine the model of TBRF occurrence at these levels. Human visitation or tourism revenue is an aspect of an area that should be explored in future analyses. Additionally, analysis of ecologic and additional factors in a smaller area, such as census tract, could provide more specific associations with TBRF occurrence. Performing a spatial analysis in an area where an outbreak occurred by mapping individual case exposure locations would provide much more specific evidence of associations with TBRF cases. The ecologic variables examined could be specific to at least the month in which the outbreak occurred, providing a much more accurate characterization of the conditions under which cases were exposed to TBRF. Finally, mapping the habitat of the *Ornithodoros* species tick vectors responsible for TBRF transmission could lead to a much improved predictive model for areas of greater TBRF risk. This approach has been successfully attempted with *Ixodes pacificus*, (Eisen et al., 2006), the tick species that is the primary vector of Lyme disease in California, and would provide an additional layer of detail to any larger scale future analyses of TBRF.

REFERENCES

Akaike H, 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC-19*: 716-723.

Barbour A, 1990. Antigenic variation of a relapsing fever *Borrelia* species. *Annual Review of Microbiology 44*: 155-171.

Beck MD, 1942. Present distribution of relapsing fever in California. *In* Moulton FR (ed): *A Symposium on Relapsing Fever in the Americas*. Washington, DC, American Association for the Advancement of Science, pp 20-25.

Boyer KM, Munford RS, Maupin GO, Pattinson CP, Fox MD, Barnes AM, Jones WL, Maynard JE, 1977. Tick-borne relapsing fever: an interstate outbreak originating at Grand Canyon National Park. *American Journal of Epidemiology 105*: 469-479.

Burgdorfer W, 1976. The diagnosis of relapsing fevers. *In* Johnson RC (ed): *The Biology of Parasitic Spirochetes*. New York, Academic Press, pp 225-234.

Carlisle RJ, 1906. Two cases of relapsing fever; with notes on the occurrence of this disease throughout the world at the present day. *Journal of Infectious Disease 3*: 233-265.

Centers for Disease Control and Prevention, 1990. Common source outbreak of relapsing fever-California. *MMWR Morbidity and Mortality Weekly Report 39 (34)*: 579-586.

Centers for Disease Control and Prevention, 1997. Case definitions for infectious conditions under public health surveillance. *MMWR Morbidity and Mortality Weekly Report 46 (RR10)*: 1-55.

Centers for Disease Control and Prevention, 2007. Acute respiratory distress syndrome in persons with tick-borne relapsing fever- three states, 2004-2005. *MMWR Morbidity and Mortality Weekly Report 56 (41)*: 1073-1075.

Cooley RA, Kohls GM, 1944. The Argasidae of North America, Central America, and Cuba. *American Midland Naturalist*: Monograph No. 1.

Cutler SJ, 2010. Relapsing fever- a forgotten disease revealed. *Journal of Applied Microbiology 108*: 1115-1122.

Davis GE, 1939. *Ornithodoros parkeri*: Distribution and host data: Spontaneous infection with relapsing fever spirochetes. *Public Health Reports 54*: 1345-1350.

Davis GE, 1955. The endemic relapsing fevers. *In* Hull TG (ed): Diseases Transmitted From Animals to Man. Springfield, IL, Charles C. Thomas, pp 552-565.

Dworkin MS, Anderson DE Jr., Schwan TG, Shoemaker PC, Banerjee SN, Kaasen BO, Burgdorfer W, 1998. Tick-borne relapsing fever in the northwestern United States and southwestern Canada. *Clinical Infectious Diseases 26*: 122-131.

Dworkin MS, Shoemaker PC, Fritz CL, Dowell ME, Anderson DE Jr., 2002a. The epidemiology of tick-borne relapsing fever in the United States. *American Journal of Tropical Medicine and Hygiene 66*: 753-758.

Dworkin MS, Schwan TG, Anderson DE Jr., 2002b. Tick-borne relapsing fever in North America. *Medical Clinics of North America 86*: 417-433.

Eisen RJ, Lane RS, Fritz CL, Eisen L, 2006. Spatial patterns of Lyme disease risk in California based on disease incidence data and modeling of vector-tick exposure. *American Journal of Tropical Medicine and Hygiene 75*: 669-676.

Eisen RJ, Glass GE, Eisen L, Cheek J, Enscore RE, Ettestad P, Gage KL, 2007a. A spatial model of shared risk for plague and hantavirus pulmonary syndrome in the southwestern United States. *American Journal of Tropical Medicine and Hygiene 77*: 999-1004.

Eisen RJ, Enscore RE, Biggerstaff BJ, Reynolds PJ, Ettestad P, Brown T, Pape J, Tanda D, Levy CE, Englethaler DM, Cheek J, Bueno Jr. R, Targhetta J, Montinieri JA, Gage KL, 2007b. Human plague in the southwestern United States, 1957-2004: spatial models of elevated risk of human exposure to *Yersinia pestis*. *Journal of Medical Entomology 44*: 530-537.

Eisen RJ and Eisen L, 2008a. Spatial modeling of human risk exposure to vector-borne pathogens based on epidemiological versus arthropod vector data. *Journal of Medical Entomology 45*: 181-192.

Eisen RJ, Mead PS, Meyer AM, Pfaff LE, Bradley KK, Eisen L, 2008b. Ecoepidemiology of tularemia in the Southcentral United States. *American Journal of Tropical Medicine and Hygiene 78*: 586-594.

Eisen RJ, Griffith KS, Borchert JN, MacMillan K, Apangu T, Owor N, Acayo S, Acidri R, Zielinski-Gutierrez E, Winters AM, Enscore RE, Schriefer ME, Beard CB, Gage KL, Mead PS, 2010. Assessing human risk of exposure to plague bacteria in northwestern Uganda based on remotely sensed predictors. *American Journal of Tropical Medicine and Hygiene 82*: 904-911.

Favorova LA, Chernyshova TF, Mikhailov AK, 1971. Results of inoculation of volunteers with *Borrelia* passaged through lice. *Medical Parazitol. 40*: 443-446.

Felsenfeld O, 1971. Borrelia: Strains, vectors, human and animal borreliosis. St. Louis, Warren H Green, Inc.

Fielding AH, Bell JF, 1997. A review of methods for the assessment of prediction errors in the conservation presence/absence models. *Environmental Conservation 24*: 38-49.

Fritz CL, Bronson LR, Smith CR, Schriefer ME, Tucker JR, Schwan TG, 2004. Isolation and characterization of *Borrelia hermsii* associated with two foci of tick-borne relapsing fever in California. *Journal of Clinical Microbiology 42*: 1123-1128.

Goubau PF, 1984. Relapsing fevers: A review. *Ann Soc Belg Med Trop 64*: 335-364.

Griffin GE, 1998. Cytokines involved in human septic shock- the model of the Jarisch-Herxheimer reaction. *Journal of Antimicrobial Chemotherapy 41 (supplement A)*: 25-29.

Hinnebusch BJ, Barbour AG, Restrepo BI, Schwan TG, 1998. Population structure of the relapsing fever spirochete *Borrelia hermsii* as indicated by polymorphism of two multigene families that encode immunogenic outer surface lipoproteins. *Infection and Immunity 66*: 432-440.

Hosmer DW and Lemeshow S, 2000. Model Building Strategies and Methods for Logistic Regression. *Applied Logistic Regression (2nd Edition)*. New York: John Wiley & Sons, Inc., 91-116.

Lopez-Cortes L, Lozano De Leon F, Gomez-Mateos, et al., 1989. Tick-borne relapsing fever in intravenous drug abusers. *Journal of Infectious Disease 159*: 804.

Meador CN, 1915. Five cases of relapsing fever originating in Colorado, with positive blood findings in two. *Colorado Medicine 12*: 365-368.

Moursund WH, 1942: Historical introduction to the symposium on relapsing fever. *In* Moulton FR (ed): *A Symposium on Relapsing Fever in the Americas*. Washington, DC, American Association for the Advancement of Science, 1-6.

Murray PR, Rosenthal KS, Kobayashi GS, Pfaller MA, 2002. Borrelia. *Medical Microbiology (4th Edition)*. Grigg LL, ed. St. Louis: Mosby Inc., 384-390.

Paul WS, Maupin G, Scott-Wright AO, Craven RB, Dennis DT, 2002. Outbreak of tick-borne relapsing fever at the North Rim of the Grand Canyon: evidence for the effectiveness of preventive measures. *American Journal of Tropical Medicine and Hygiene 66*: 71-75.

Rawlings JA, 1995. An overview of tick-borne relapsing fever with emphasis on outbreaks in Texas. *Texas Medicine 91*: 56-59.

Restrepo BI and Barbour AG, 1994. Antigen diversity in the bacterium *B. hermsii* through "somatic" mutations in the rearranged *vmp* genes. *Cell 78*: 867-876.

Southern PM and Sanford JP, 1969. Relapsing fever: A clinical and microbiological review. *Medicine 48*: 129-149.

Schwan TG and Hinnebusch BJ, 1998. Bloodstream-versus tick-associated variants of a relapsing fever bacterium. *Science 280*: 1938-1940.

Stoenner HG, Dodd T, Larsen C, 1982. Antigenic variation of *Borrelia hermsii*. *Journal of Experimental Medicine 156*: 1297-1311.

Thompson RS, Burgdorfer W, Russell R, Francis BJ, 1969. Outbreak of tick-borne relapsing fever in Spokane County, Washington. *JAMA 210*: 1045-1050.

Trevejo RT, Schriefer ME, Gage KL, Safranek TJ, Orloski KA, Pape WJ, Montinieri JA, Campbell GL, 1998. An interstate outbreak of tick-borne relapsing fever among vacationers at a Rocky Mountain cabin. *American Journal of Tropical Medicine and Hygiene 58*: 743-747.

Winters AM, Eisen RJ, Lozano-Fuentes S, Moore CG, Pape WJ, Eisen L, 2008. Predictive spatial models for risk of West Nile virus exposure in eastern and western Colorado. *American Journal of Tropical Medicine and Hygiene 79*: 581-590.

Winters AM, Staples JE, Oden-Odoi A, Mead PS, Griffith K, Owor N, Babi N, Enscore RE, Eisen L, Gage KL, Eisen RJ, 2009. Spatial risk models for human plague in the West Nile region of Uganda. *American Journal of Tropical Medicine and Hygiene 80*: 1014-1022.

Wynns HL, 1942. The epidemiology of relapsing fever. *In* Moulton FR (ed): *A Symposium on Relapsing Fever in the Americas*. Washington, DC, American Association for the Advancement of Science, 100-105