12-2010

Spatial Epidemiology of Birth Defects in the United States and the State of Utah Using Geographic Information Systems and Spatial Statistics

Samson Y. Gebreab
Utah State University

SPATIAL EPIDEMIOLOGY OF BIRTH DEFECTS IN THE UNITED STATES AND THE STATE OF UTAH USING GEOGRAPHIC INFORMATION SYSTEMS AND SPATIAL STATISTICS

by

Samson Y. Gebreab

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Ecology

Approved:

_____	_____
Robert R. Gillies Major Professor	Ronald G. Munger Committee Member
_____	_____
Jürgen Symanzik Committee Member	Heidi J. Wengreen Committee Member
_____	_____
Christopher Corcoran Committee Member	Byron R. Burnham Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2010

ABSTRACT

Spatial Epidemiology of Birth Defects in the United States and the State of Utah Using Geographic Information Systems and Spatial Statistics

by

Samson Y. Gebreab, Doctor of Philosophy

Utah State University, 2010

Major Professor: Dr. Robert R. Gillies
Department: Watershed Sciences

Oral clefts are the most common form of birth defects in the United States (US) and the State of Utah has among the highest prevalence of oral clefts in the nation. The overall objective of this dissertation was to examine the spatial distribution of oral clefts and their linkage with a broad range of demographic, behavioral, social, economic, and environmental risk factors through the application of Geographic Information Systems (GIS) and spatial statistics. Using innovative linked micromaps plots, we investigated the geographic patterns of oral clefts occurrence from 1998 to 2002 and their relationships with maternal smoking rates and proportion of American Indians and Alaskan Natives (AIAN) at large scales across the US. The findings indicated higher oral clefts occurrence in the southwest and the midwest and lower occurrence in the east. Furthermore, these spatial patterns were significantly related to the smoking rates and AIAN. Then at the small area level, hierarchical Bayesian models were built to examine the spatial variation

in oral clefts risk in the State of Utah from 1995 to 2004 and to assess association with mothers using tobacco, mothers consuming alcohol during pregnancy, and the proportion of mothers with no high school diploma. Next, multi-scalar spatial clustering and cluster techniques were used to test the hypothesis whether there was spatial clustering of oral clefts anywhere in the State of Utah and whether there were statistically significant local clusters with elevated oral cleft cases. Results generally revealed modest spatial variation in oral clefts risk in the State of Utah, with no pronounced spatial clustering, indicating environmental exposures are unlikely plausible cause of oral clefts. However, a few notable areas within Tri-County Local Health District, Provo/Brigham Young University, and North Orem had a tendency toward elevated oral clefts cases. Investigation of the maternal characteristics of these potential clusters supports the hypotheses that maternal smoking, lower education level, and family history are possible causes of oral clefts. Throughout this dissertation, we demonstrated how birth defects data collected by state and local surveillance systems coupled with GIS and spatial statistics methods can be useful in exploratory etiologic research of birth defects.

(186 pages)

ACKNOWLEDGMENTS

First and foremost I would like to thank God, the Almighty, for sustaining and giving me the strength to complete this dissertation work. This dissertation would not have been possible without the assistance, support and guidance of my advisor, Dr. Robert R. Gillies. His insightful comments, discussions, and painstaking reviews and edits of this dissertation were indispensable for completing my PhD work. I will be always grateful for all the friendship, encouragement, and opportunities he has given me over the years.

I would like to express my sincere appreciation to Dr. Ronald R. Munger for allowing me to work on the Utah oral clefts data and financially supporting my research. His invaluable suggestions and comments added greatly to enhance the substance of my dissertation research. I wish to extend special thanks to Dr. Jürgen Symanzik, who provided me with the technical and statistical support on this dissertation. I am truly grateful for his encouragement and mentorship to pursue a dual degree in statistics and to grow as a statistician. I would like to acknowledge Drs. Heidi Wengreen and Chris Corcoran for kindly serving on my committee and reading through my dissertation.

I would like to thank Dr. Nancy Sassano for her assistance in data acquisition and for her generosity and encouragement during my PhD study at USU. I would like to thank the staff in the Department of Watershed Sciences, especially Brian Bailey for his administrative assistance, and John Weeks for his technical support whenever I ran into computer problems. I am also grateful to Sam LeFevre from the Utah Department of

Health, for help with geocoding the oral clefts data for the State of Utah as well as Dr. Marcia Feldkamp and Amy Nance (both from the Utah Birth Defects Network and Utah Department of Health) for their assistance. I also appreciate the helpful comments on Chapter 4 of this dissertation from Dr. Geoff Jacquez, president of BioMedware.

I would also like to express my appreciation to all my friends and colleagues at Utah State University. My appreciation also goes to the entire Eritrean and Ethiopian communities in Utah for their friendship and unforgettable time we had together.

Last but not least, I owe my deepest gratitude to my fiancé, Zeliha Desta, for her unconditional love, understanding, and sacrifice raising our daughter Soliana Samson throughout my PhD work. I dedicate this dissertation to her, my beautiful daughter, and to all the children affected by birth defects.

Samson Y. Gebreab

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Oral clefts are the most common form of birth defects in the United States (US) and constitute a major public health problem in the State of Utah. The prevalence of oral clefts in the State of Utah is among the highest in the nation. Oral clefts have substantial public health impacts in various contexts: namely that of infant mortality, lifelong morbidity and mortality (Czeizel and Sankaranarayanan, 1984; Christensen et al., 2004), and medical cost (Waitzman et al., 1994). In addition, there are extraneous effects that are both psychological and social (Caplan and Weintraub, 1993). Given the burden of oral clefts in the State of Utah, it is imperative that effective prevention strategies and control measures are undertaken, but equally important is the identification of any associated risk factors. Therefore, this dissertation presents a comprehensive analysis of the spatial distribution of oral clefts for the State of Utah.

Currently, there are numerous ongoing efforts to reduce the burden of oral clefts, both at the national and at the state level. To this end, various agencies have established surveillance programs that collect data on the occurrence of birth defects along with other crucial information. At the national level, the National Birth Defects Prevention Network (NBDPN) maintains a network of state and population-based programs for birth defects surveillance and research (NBDPN, 2005). Additionally, many states have their own statewide birth defects surveillance programs that collect and monitor major birth defects outcomes. The Utah Birth Defects Network (UBDN) is one such program that collects, stores, and collates data on major birth defects in Utah and this dataset includes oral

clefts. Birth defects data collected by these programs along with other population studies have been used to monitor trends, examine the causes of birth defects, and subsequently are used to develop policy and prevention measures (NBDPN, 2005).

However, in the case of oral clefts, the etiology remains largely unknown. A significant and growing body of evidence hints that there is an etiologic role at play here that comprises genetic and environmental triggers which manifest both individually and through multiple interactions (Shaw et al., 1996; Romitti et al., 1999; Murray, 2002). Evidence to date suggests that environmental triggers associated with the risk of oral clefts are maternal exposure to smoking (Khoury et al., 1987; Lieff et al., 1999; Chung et al., 2000), alcohol consumption (Munger et al., 1996; Lorente et al., 2000), medication use (Dansky et al., 1991; Parkwyllie et al., 2000), nutritional deficiencies such as multivitamins and folic acid (Shaw et al., 1995; Hayes et al., 1996; Munger et al., 2004), exposure to chemical solvents in the work place or at home (Laumon et al., 1996; Garcia and Fletcher, 1998), drinking water contamination (Bove et al., 1995), environmental lead pollution (Vinceti et al., 2001), ambient air pollution (Ritz et al., 2002; Gilboa et al., 2005), and residing near to hazardous waste sites (Dolk et al., 1998; Orr et al., 2002; Brender et al., 2006).

The majority of epidemiological studies that have examined the risk associated with being born with oral clefts have focused on individual level studies drawn from case-control and cohort studies. Very little attention has been given to the spatial dimensions of the disorder. Geographic variation of oral clefts is commonplace. In fact, studies have shown that there are substantial geographic variations in oral clefts across

various geographical scales in the US and, for that matter, around the world. These variations are ascribed to geographic origin (Vanderas, 1987), racial and ethnic backgrounds (Croen et al., 1998; Tolarova and Cervenka, 1998), socioeconomic status (Murray et al., 1997), lifestyle and nutrition (Munger et al., 2004; Bille et al., 2007), and environmental pollution (Dolk et al., 1998; Ritz et al., 2002; Brender et al., 2006).

Given that so much of the etiology of oral clefts is poorly understood and the fact that there is considerable geographical variation in oral clefts, which is likely due to a broad range of factors (i.e., demographic, genetic, behavioral, social, economic, and environmental), a geographically focused examination of oral clefts may provide a supplementary if not a broader approach towards figuring out the etiology of oral clefts. Moreover, understanding the geographic distribution of oral clefts may assist with hypotheses generation as to the underlying risk factors, that is to say, once potential risk factors are identified, one can further assess them by using more refined epidemiologic studies such as case control studies or cohort studies. A further advantage of performing geographic-type analyses is in the identification of areas of elevated oral clefts risk – this being most advantageous in future planning for health care as well as in the allocation of health resources.

Spatial analysis of disease has played a historical role in the arena of public health, and has long underscored the classical triad in epidemiology as noted by Walter (2000) that is of person, time and place. A famous example, conducted by John Snow, was the mapping of the 1814 London cholera epidemic (Figure 1-1 when etiology of the disease was not scientifically understood.).

Figure 1-1 Snow's map of cholera. The affected well is clearly identified by the concentration of cases in its vicinity (Source: http://archives-fig-st-die.cndp.fr/actes/actes_2000/thouez/t13.gif).

Snow drew dots on a map background which he subsequently used to trace the source of the cholera outbreak, i.e., fecal-contaminated water supplied by the Broad Street pump (Snow, 1855). His direct action in dismantling the pump saved many lives. However, since this investigation, the spatial analysis of disease has been sporadic and

generally limited to simple dot maps and density maps. On the other hand, advances in Geographic Information Systems (GIS), spatial statistical methods, coupled with developments in digital and computing resources, has lead to a very powerful synergistic toolset that now supplies the means with which to effectively analyze any data that is spatially distributed.

As previously noted, during the past two decades the methods for spatial analysis of disease have dramatically improved. Advances in GIS, availability of geo-referenced health and environmental data, along with statistical methodologies, particularly Bayesian methods have allowed investigators to perform routine sophisticated spatial analysis of disease with enhanced precision. So much so that a new sub-field of spatial epidemiology, a hybrid of epidemiology, statistics, and GIS has emerged, and has been increasingly used to assess the spatial variations of chronic and infectious diseases (Elliott and Wartenberg, 2004).

GIS and spatial statistics methods, however, have had limited use in birth defects epidemiology. There is a very limited set of information on the extent and level of geographic variations in oral clefts across the US. Furthermore, little is known about the spatial variations in oral clefts and associated environmental risk factors at small area levels within a state. Given the aforementioned situation, this dissertation, is a step towards the provision of a comprehensive description as to the extent and level of spatial variations in oral clefts at national, regional and small area levels. Further analyses were performed to describe observed spatial variations of oral clefts within the context of demographic, behavioral, social, economic and environmental risk factors. In doing so,

this dissertation draws upon theoretical and methodological approaches from the disciplines of epidemiology, GIS, and spatial statistics. The work outlined in chapters one through four illustrates how GIS and spatial statistic methodologies, coupled with routinely collected oral clefts data from birth defects surveillance systems, can be useful tools to augment oral clefts surveillance and prevention.

RESEARCH PROBLEMS

Oral clefts are one of the most common birth defects in the US. Oral clefts include cleft lip (CL), cleft lip with or without a cleft palate (CL/P), and isolated cleft palate (CP). On average, the occurrence of oral clefts in the US is 1 in 750 (UBDN, 2007). There are, however, marked racial and geographical variations in oral clefts prevalence in the US. In the general population, Asian or Amerindian populations exhibit the highest frequencies, often at 1 in 500 or higher, with Caucasian populations intermediate, and African-American populations the lowest at 1 in 2500 (Murray, 2002).

The impact of oral clefts is profound both socially and economically. Children who are born with an oral cleft require several surgical procedures and complex medical treatments. The estimated lifetime medical cost for each child with an oral cleft is $100,000; this amounts to $750 million if one considers all children born with an oral cleft each year in the United States (Waitzman et al., 1994). In addition to the economic burden, these children and their families often experience profound psychological, behavioral, and physical problems (Caplan and Weintraub, 1993). The defects, therefore,

pose a substantial burden to the individual and their family, and require significant expenditure in terms of health and related services.

In Utah, oral clefts are a major public health problem. The birth prevalence of oral clefts rates in the State is among the highest in the nation. Oral clefts affect 1 in 450 births in Utah compared to 1 in 750 births nationally. Each year more than 100 children with oral clefts are born in Utah (UBDN, 2007). Rates of oral clefts are similar among Utah Caucasians (22.2 per 10,000), Hispanics (22.4 per 10,000), and Asians (18.2 per 10,000). While the rate among Native Americans may be higher (36. 9 per 10,000), the numbers are too small to be certain. The number of affected births among African Americans in Utah is simply too small to draw any conclusions (UBDN, 2007).

The high prevalence of oral clefts in the State of Utah compared to most States in the US is not clearly understood. One approach, to understand why Utah exhibits such a high prevalence, is to investigate the spatial variations in oral clefts at the small area level and is a research topic that has not been given much attention in the literature of the field. Therefore, this dissertation provides a detailed description of the spatial variations in oral clefts in the State of Utah at the small area level for the period from 1995 to 2004. Specifically examined and assessed are the spatial patterns of oral clefts and their associations with certain risk factors; identified also are local clusters with significantly higher number of oral clefts cases than expected.

Spatial Epidemiology of Birth Defects

In recent years, GIS and spatial statistics have been used increasingly in public health and epidemiology. These technologies are useful tools in identifying both the spatial pattern of a disease, areas of excess risks, and in investigating the association between the observed disease incidence and potential risk factors that contribute to a spatial variation in disease risk. As such they are valuable in generating more sophisticated hypotheses for a more in-depth investigation or merely to serve as part of a general health surveillance and monitoring system. However, they have thus far played only a limited role in birth defects research (Siffel et al., 2006). There are only a handful of published studies on birth defects using spatial statistics and GIS. For example, Rushton and Lolonis (1996) used an exploratory spatial analysis approach to find spatial clusters in birth defects rates in Des Moines, Iowa. Forand et al. (2002) used the spatial scan statistic to map elevated and lowered birth defect rates. GIS has been used in the analysis of associations between birth defects and exposures such as hazardous waste sites and air pollution (Orr et al., 2002; Wu et al., 2004; Gilboa et al., 2005; Gilboa et al., 2006), and in the analysis of socioeconomic status and neural tube defects (Wasserman et al., 1998). Recently, Cech et al. (2007) utilized spatial clustering techniques to identify spatial clusters and then evaluate the association between low-level radioactivity in drinking water and rates of oral clefts among residents of Harris County, Texas. Gardner et al. (2007) developed the Automated Spatial Surveillance program (ASSP) to monitor spatial and temporal trends of birth defects. All of these studies illustrate the utility of GIS and spatial statistical methods in birth defects surveillance. However, the full

potential of GIS and spatial statistics capabilities in birth defects research has not yet been fully realized. There are several issues that have slowed down the integration of GIS and spatial statistics into birth defects research; to mention a few – (a) primarily the availability of a high quality geocoded birth defects data, (b) the fact that GIS and spatial statistics techniques are relatively new, and (c) the lack of suitable and accessible software coupled to the shortage of individuals trained in both GIS and spatial statistics techniques (Siffel et al., 2006).

Birth defects surveillance systems typically collect information on the maternal residential address during pregnancy. Prior to any spatial analyses, this residential information generally has to be converted to digital map coordinates (e.g., latitude and longitude) via a technique called geocoding. The geocoding process utilizes both, GIS and specialized geocoding software, after which it is possible to map the coordinates of maternal residential address and subsequently link a location to demographic, socioeconomic, environmental risk factors as well as other potential risk factors that might be associated with birth defects occurrence. For instance, GIS can facilitate the linkage of birth defects data from the UBDN registry with US Census Bureau demographic and socioeconomic data and Environmental Protective Agency (EPA) environmental data.

Several statistical methods have been developed to analyze the spatial patterns of birth defects data. The choice of spatial statistical method depends on the type of geocoded birth defects data as this data is available at either point or area level. Point - level data requires exact coordinates of birth defect records. For spatial analyses point

level is the preferred choice. However, confidentiality and privacy issues of birth defects records more often than not restrict the use of the data at this level. Given such instances, area - level data, where the birth defects cases are aggregated up to some geographical unit (e.g. census tract, county or ZIP code), are more commonly made available for the purposes of research and any privacy concerns are protected (Olson et al., 2006). Hence, the research presented here was performed at the area level, with an emphasis on small area analyses.

The term 'small-area' refers to an area with small 'at-risk' population, but not necessarily small in geographical size/scale (MacNab, 2004). An analysis of birth defects at the small-area level reduces the potential for ecological bias created by the within-area heterogeneity of exposure or other determinants (Lawson et al., 1999; Elliott et al., 2000). Furthermore, at the small-area level statistical tests are able to detect more effectively any local effects that might be linked to environmental circumstances such as industrial pollution in the vicinity. However, when analyzing the spatial variation of birth defects at area level, especially at small-areas, it is important to keep in mind and consider two methodological issues. The first issue is the impact of the modifiable areal unit problem (MAUP) – a phenomenon which occurs when different scales and different zoning methods lead to different statistical results and spatial patterns (Openshaw, 1984). To date there is no solution to the MAUP but one can minimize the negative impact of MAUP through a careful choice of the areal units and zoning designs. In practice, researchers try to select a geographic unit that is as small as possible; however, the selection is often dictated by the availability and depth of the birth defects data. The second issue is the

small-area problem caused by small observed and expected numbers in areas with small ‘at-risk’ population, especially the case for diseases that exhibit low frequency (e.g., birth defects). Such crude birth defect rates for areas with small ‘at-risk’ population produce unstable risk estimates. Problems associated with small-area can be addressed through the use of sophisticated statistical techniques such as Bayesian hierarchical models, which permit to “borrow strength” from neighboring areas and the entire study area to produce more stable risk estimates.

In general, spatial analyses of birth defects at the area level are divided into four categories of study. The first category is explanatory spatial data analysis (ESDA); this endeavors to explore geo-referenced birth defects data for the detection of patterns, isolation of outliers, and cluster identification using graphical plots and figures. For example, the GeoDa spatial analysis toolkit (Anselin, 2003) provides a set of tools for conducing ESDA, (e.g., linking and brushing techniques), all of which help uncover complex relationships between a variable and its covariates. In particular, the multivariate nature of birth defects benefits from using “geographic brushing" techniques for exploring the relations between birth defects and potential risk factors (Monmonier, 1989). In addition, birth defects research can also benefit from the use of innovative linked micromap (LM) visualization techniques that portray the information in a joint geographical and statistical context (Carr and Pierson, 1996). Extension of the techniques towards interactive, web-based mapping is becoming increasingly popular. If implemented, a web-based LM application in the context of birth defects surveillance would provide a set of interactive tools that could be used to dynamically query and sort

the data, and further to compare birth defects rates for different geographic regions, as well as to explore birth defects estimates at different spatial resolutions- all in real time. An example of such an application is the Web-based Atlas of Cancer Mortality in the US (NCI, 2009).

The second category in studying disease mapping deals with the estimation and mapping of birth defects risks across a geographical area. Disease maps provide a visual description of the spatial distribution of birth defects, and are valuable for many reasons the least of which is hypotheses generation and the allocation of health care resources. The standardized morbidity /mortality ratio (SMR) is often used to estimate and map disease risk. However, a particular challenge in estimating disease risk using the SMR at the small-area level, especially for low frequency birth defects, is that there is a tendency for computing unstable risk estimates (Manton et al., 1989; Gelman and Price, 1999). Moreover, this method fails to take into account the presence of spatial dependence in birth defects risks between adjacent areas. One way of handling this problem is to 'borrow strength' by incorporating information from neighboring areas via a smoothing process as is done using Bayesian disease mapping models. Bayesian disease mapping models provide a more robust method of estimation of disease risks that is often more interpretable and informative (Lawson et al., 1999; Elliott et al., 2000) compared to SMR map. Bayesian models, however, require computer intensive iterative procedures, such as Markov Chain Monte Carlo (MCMC) methods for the estimation of model parameters. However, the development of the WinBUGS software (Spiegelhalter et al., 2007) has facilitated the implementation of MCMC. As a result, Bayesian models are now widely

used for the disease mapping of chronic and infectious diseases (Best et al., 2005). Drawing on these techniques, Wu et al. (2004) have recently applied hierarchical Bayesian models in an investigation of the spatial distribution of neural tube birth defects in China.

The third category lies in ecological regression analyses; these techniques have to do with the quantification of any associations that might exist between the spatial variations of birth defects and extraneous risk factors (e.g., environmental). These factors are measured at an aggregated level. Such analyses provide useful in exploratory etiologic research. For example, ecological analysis can generate etiological hypotheses that could set a stage for a comprehensive epidemiologic studies using either case-control or cohort studies. However, it is important to consider three issues when conducting ecological regression analyses. The first is the confounding problem, which is a major problem in the interpretation of the results; for example, it is not uncommon for pregnant mothers residing in a deprived area to reside close to a hazardous environmental source. The second lies in the ecological fallacy; i.e., drawing incorrect individual level inferences. Because of such drawbacks, ecological analyses are not appropriate for assessing causal relationships. Third is that the lack of spatial independence in ecological data, often referred as “spatial dependence”. Spatial dependency can occur because of Tobler's first law of geography summarized as “everything is related to everything else, but near things are more related than distant things” (Tobler, 1970). From a statistical point of view, spatial dependency can lead to the spatial autocorrelation problem in statistics, which violates the assumptions of most standard statistical methods such as

Poisson and logistic regression. So, failure to account for spatial dependency in any ecological analyses can underestimate the standard error and so the significance of any risk factors has a propensity to be overestimated. Instead, it is appropriate to use methods such as hierarchical Bayesian frameworks that incorporate random spatial effects and covariate effects that account for issues that are related to spatial autocorrelation, unmeasured or unknown covariates, and measurement error (Richardson, 1992; Clayton et al., 1993; Best, 1999).

The fourth and final category consists of spatial cluster analyses that deal with conducting formal hypothesis testing in order to determine whether there is spatial clustering of birth defects anywhere in the study area or whether there are local clusters with a high proportion of birth defects that is more than expected. Such investigations serve two fundamental purposes: (a) to alleviate community concern over adverse health outcomes from perceived exposures, and (b) for the purposes of hypotheses generation. In general, there are two types of spatial clustering methods: (i) Spatial global clustering methods and (ii) local cluster detection methods (Besag and Newell, 1991). The former method tests for the presence of spatial clustering in the whole study area but does not provide any information as to location. The later identifies a spatial cluster with excess disease risk without previous knowledge of either how many or where they are located. There are several statistical methods that have been developed for spatial cluster analyses; a detailed review is given in Wakefield et al. (2000) and Waller and Gotway (2004). Among the most significant methodological developments for spatial clustering and cluster analyses at the area level are test for heterogeneity (Potthoff and Whittinghill,

1966), Moran's *I* statistic (Moran, 1950), Tango's excess events test (Tango, 1995), and Tango's maximized excess events test (Tango, 2000) all of which are used for testing global clustering. Methods for the identification of local clusters are local indicators of spatial association (LISA) (Anselin, 1995) and Getis and Ord's local Gi(d) statistic (Getis and Ord, 1992). These are widely used to identify local clusters by measuring spatial autocorrelation and by measuring "concentration" of disease risks between neighboring areas respectively. While the Besag-Newell method (Besag and Newell, 1991) and the spatial scan statistics (Kulldorff, 1997) are widely applied in the detection of local clusters of disease without any pre-selection bias. Several software packages are available for spatial cluster analysis. For example, the opensource DCluster R package that allows testing for both global clustering and local cluster package (Gómez-Rubio et al., 2004; R Development Core Team, 2004) and the opensource SaTScan software that allows detecting purely spatial, purely temporal, or space-time clusters (Kulldorff, 2007). There are also commercially available software such as ClusterSeer (Jacquez et al., 2002), which includes a wide variety of methods for spatial cluster detection and cluster analyses.

OBJECTIVES

The overall objective of this dissertation was to examine the spatial distribution of oral clefts and their spatial linkage with a broad range of demographic, genetic, behavioral, social, economic, and environmental risk factors. The specific objectives of this dissertation are:

1. To explore the geographic patterns of oral clefts rates across states and regions in the US.
2. To determine if the prevalence of oral clefs shows any spatial variability at the small area level in the State of Utah.
3. To identify locations and populations with significantly higher than expected oral clefts risk ("local clusters") in the State of Utah.
4. To investigate whether any spatial patterns observed in objectives 2 and 3 are attributable to spatial differences in demographic, behavioral, socioeconomic, or environmental risk factors.
5. To evaluate the utility of GIS and spatial statistical techniques in spatial surveillance of oral clefts.
6. To furnish reliable risk maps of oral clefts to aid public health officials in intervention programs and allocation of health resources.

RESEARCH HYPOTHESES AND QUESTIONS

The research focuses primarily on the following two hypotheses. The first is that the spatial variation in the prevalence of oral clefts across the State of Utah is hypothesized to be non-random, and the second hypothesis is that spatial associations exist between oral clefts and area-level behavioral and economic characteristics. Specifically, the research aspires to answer the following questions:

1. Are there differences in oral clefts rates across states and regions? How do the rates of oral clefts in the State of Utah compare to the rates in the other states within the US?
2. Is there spatial clustering in the prevalence of oral clefts anywhere in the State of Utah?
3. Do significant local clusters (hotspots) of higher than expected oral clefts cases exist in the State of Utah and, if so, where are the location of these clusters?
4. What specific area-level characteristics are related to oral clefts outcomes and what can this tell us about possible cause of the birth defects?

The research attempts to address the aforementioned questions by utilizing GIS and novel spatial statistical methodologies. Data sources are the US oral clefts data obtained from National Birth Defects Prevention Network (NBDPN, 2005) for the period from 1998 to 2002 and Utah oral clefts data obtained from Center for Epidemiologic Studies, Utah State University, thorough the collaboration of Utah Birth Defects Network (UBDN) for the period from 1995 to 2004.

DISSERTATION STRUCTURE

This dissertation is divided into five chapters.

Chapter 1 this chapter provides an introduction to the research along with outlines covering research problems, study objectives, and hypotheses. Also discussed is the application of geospatial techniques in birth defects research.

Chapter 2 is titled "Visualization and Interpretation of Birth Defects Data Using Linked Micromap Plots." This chapter presents two different templates of LM plots for representing spatially indexed oral clefts rates at two geographical resolutions - at the state level for the US, and at the county level for the State of Utah. The first LM plot displays parallel sequences of micromaps for US states, names of the states, and statistical summaries of selected variables. The LM plot describes spatial patterns of oral clefts and explores the relationships between two variables (a) the rates of oral clefts and the proportion of smoking in pregnant women and, (b) the proportion of American Indian and native Alaskan populations. The second LM plot uses confidence interval statistical plots to represent the uncertainty associated with rate estimates of oral clefts at the county level for the State of Utah.

Chapter 3 is titled "Small Area Mapping and Ecological Analyses of Oral Clefts for the State of Utah for the Period From 1995 to 2004 Using Hierarchical Bayesian Models." Chapter 3 uses hierarchical Bayesian models to explore the spatial variations in oral clefts at the small-area level in Utah by generating stable risk estimates. An assessment as to any relationships between area-specific risk factors and oral clefts is presented and, geographic areas with high-risk of oral clefts are identified. Techniques used comprise (i) a non-spatial model that uses only uncorrelated heterogeneity random effects, and (ii) a spatial model with both uncorrelated heterogeneity and spatially correlated heterogeneity random effects. Models (i) and (ii) were used to generate risk estimates of oral clefts and to describe their spatial variations across small-areas. The two Bayesian models were further extended to incorporate covariate effects measured at the

small-area level for various ecological analyses i.e., oral clefts prevalence associated with socioeconomic status (proportion of mothers with no high school diploma) and behavioral covariates (proportion of mothers using tobacco, and consuming alcohol during pregnancy).

Chapter 4 is titled "A Multi-Scalar Approach to Spatial Clustering and Cluster Analysis of Oral Clefts in the State of Utah: Clues for Etiology." Chapter 4 undertakes analyses to test whether there is disease clustering anywhere in the study area and progresses to identify those local clusters with high concentration of oral clefts cases and to establish if indeed they are statistically significant. Additionally, this chapter investigates whether the maternal characteristics of the cases involved in the detected clusters provide some etiological clues as to their existence. The approach taken was an integrative one to test for global clustering and in the detection of local clusters. Specifically, a heterogeneity test (Potthoff and Whittinghill, 1966), the Moran test (Moran, 1950), and the Tango's maximized excess event test (Tango, 2000) were used to detect overall global clustering in the Utah dataset. The Besag-Newell method (Besag and Newell, 1991) and the spatial scan statistic (Kulldorff, 1997) were used in the detection of local clusters ("hotspots").

Chapter 5 wraps up with a discussion on the findings, and provides conclusions and recommendations for further research directions.

REFERENCES

Anselin L, 2003. GeoDaTM 0.9 User's Guide. Urbana, IL: Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, USA.

Anselin L, 1995. Local indicators of spatial association – LISA. Geogr Anal 27, 93-115.

Besag J, Newell J, 1991. The detection of clusters in rare diseases. J Roy Stat Soc A 154, 143-155.

Best NG, 1999. Bayesian ecological modelling. In: Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R (eds). Disease Mapping and Risk Assessment for Public Health. Wiley, New York, USA, 194–201.

Best N, Richardson S, Thomson A, 2005. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14, 35-59.

Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficy EM, Savrin JE, 1995. Public drinking water contamination and birth outcomes. Am J Epidemiol 141, 850-862.

Brender JD, Zhan FB, Suarez L, Langlois PH, Moody K, 2006. Maternal residential proximity to waste sites and industrial facilities and oral clefts in offspring. J Occup Environ Med 48, 565-572

Caplan DJ, Weintraub JA, 1993. The oral health burden in the United States: a summary of recent epidemiologic studies. J Dent Educ 57, 853-862.

Carr DB, Pierson SM, 1996. Emphasizing statistical summaries and showing spatial context with micromaps. Statistical Computing and Statistical Graphics Newsletter 7, 16-23.

Cech I, Keith BD, Jane W, 2007. Spatial Distribution of Orofacial Cleft Defect Births in Harris County, Texas, 1990 to 1994, and Historical Evidence for the Presence of Low-level Radioactivity in Tap Water. South Med J 100, 560-569.

Christensen K, Juel K, Herskind AM, Murray JC, 2004. Long term follow up study of survival associated with cleft lip and palate at birth. BMJ 328, 1405.

Chung KC, Kowalski CP, Kim HM, Buchman SR, 2000. Maternal cigarette smoking during pregnancy and the risk of having a child with cleft lip/palate. Plast Reconstr Surg 105, 485-491.

Clayton DG, Bernardinelli L, Montomoli C, 1993. Spatial correlation and ecological analysis. Int J Epidemiol 22, 1193–1201.

Croen LA, Shaw GM, Wasserman CR, Tolarova MM, 1998. Racial and ethnic variations in the prevalence of orofacial clefts in California, 1983-1992. Am J Med Genet 79, 42-47.

Czeizel A, Sankaranarayanan K, 1984. The load of genetic and partially genetic disorders in man. 1. Congenital anomalies: estimates of detriment in terms of years of life lost and years of impaired life. Mutat Res 128, 73-103.

Dansky LV, Finnell RH, 1991. Parental epilepsy, anticonvulsivant drugs, and reproductive outcome: Epidemiologic and experimental findings spanning three decades; 2: Human studies. Reprod Tox 5, 301-335.

Dolk H, Vrijheid M, Armstrong B, Abramsky L, Bianchi F, Garne E, Nelen V, Robert E, Scott JE, Stone D, Tenconi R, 1998. Risk of congenital anomalies near hazardous-waste landfill sites in Europe: the EUROHAZCON study. Lancet 352, 423-427.

Elliott P, Wakefield JC, Best NG, Briggs DJ, 2000. Spatial Epidemiology: methods and applications. Oxford University Press, Oxford, UK.

Elliott P, Wartenberg D, 2004. Spatial epidemiology: current approaches and future challenges. Environ Health Perspect 112, 998-1006.

Forand SP, Talbot TO, Druschel C, Cross PK, 2002. Data quality and the spatial analysis of disease rates: congenital malformations in New York State. Health & Place 8, 191–199.

Garcia AM, Fletcher T, 1998. Maternal occupation in the leather industry and selected congenital malformations. J Occup Environ Med 55, 284-286.

Gardner BR, Strickland MJ, Correa A, 2007. Application of the automated spatial surveillance program to birth defects surveillance data. Birth Defects Res A Clin Mol Teratol 79, 559-564

Gelman A, Price PN, 1999. All maps of parameter estimates are misleading. Stat Med 18:3221-3234.

Getis A, Ord JK, 1992. The analysis of spatial association by use of distance statistics. Geogr Anal 24, 189-206.

Gilboa SM, Mendola P, Olshan AF, Langlois PH, Savitz DA, Loomis D, Herring AH, Fixler DE, 2005. Relation between ambient air quality and selected birth defects, seven county study, Texas, 1997-2000. Am J Epidemiol 162, 238-252.

Gilboa SM, Mendola P, Olshan AF, Harness C, Loomis D, Savitz DA, Herring AH, 2006. Comparison of residential geocoding methods in population-based study of air quality and birth defects. Environ Res 101, 256–262.

Gómez-Rubio V, Ferrándiz-Ferragud J, López-Quilez A, 2005. Detecting clusters of disease with R. J Geogr Syst **7**, 189-206.

Hayes C, Werler MM, Willett WC, Michell AA, 1996. Case-control study of periconceptional folic acid supplementation and oral clefts. Am J Epidemiol 143, 1229-1234.

Jacquez GM, Greiling DA, Durbeck H, Estberg L, Do E, Long A, Rommel B, 2002. ClusterSeer User Guide 2: Software for identifying disease clusters. TerraSeer Press, Ann Arbor, USA.

Khoury MJ, Weinstein A, Panny S, Holtzman NA, Lindsay PK, Farrel K, Eisenberg M, 1987. Maternal cigarette smoking and oral clefts: a population-based study. Am J Public Health 77, 623-625.

Kulldorff M, 1997. A spatial scan statistic. Comm in Stat Th & Meth 26, 1481-1496.

Kulldorff M, Information Management Services, Inc., 2007. SaTScan(TM) v7.0: Software for the Spatial and Space-time Scan Statistics.

Laumon B, Martin JL, Collet P, 1996. Exposure to organic solvents during pregnancy and oral clefts: a case-control study. Reprod Tox 10, 15-19.

Lawson AB, Biggeri A, Bohning D, Lesare E, Viel JF, Bertollini R, 1999. Disease mapping and risk assessment for public health. John Wiley & Sons, Chichester, UK.

Lieff S, Olshan AF, Werler M, Strauss RP, Smith J. Michell A, 1999. Maternal cigarette smoking during pregnancy and risk of oral clefts in newborns. Am J Epidemiol 150, 683-94.

Lorente C, Cordier S, Goujard J, Ayme S, Bianchi F, Calzolari E, De Walle HE, Knill-Jones R, 2000. Tobacco and alcohol use during pregnancy and risk of oral clefts. Am J Public Health 90, 415-419.

MacNab YC, 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accid Anal Prev 36, 1019-1028.

Manton KG, Woodbury MA, Stallard E, Riggan WB, Creason JP, Pellom AC, 1989. Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. J Am Stat Assoc 84, 637-650.

Monmonier M, 1989. Geographic brushing: enhancing exploratory analysis of the scatterplot matrix. Geogr Anal 21, 81-84.

Moran PAP, 1950. Notes on continuous stochastic phenomena. Biometrika, 37, 17-23.

Munger R, Romitti P, Daack-Hirsch S, Burns T, Murray J and Hanson J, 1996. Maternal alcohol use and risk of orofacial cleft birth defects. Teratol 54, 27-33.

Munger RG, Sauberlich HE, Corcoran C, Nepomuceno B, Daack-Hirsch S, Solon FS, 2004. Maternal vitamin B-6 and folate status and risk of oral cleft birth defects in the Philippines. Birth Defects Res A Clin Mol Teratol 70, 464-71.

Murray JC, 2002. Gene/environment causes of cleft lip and/or palate. Clin Genet 61, 248-256.

Murray JC, Daack-Hirsch S, Buetow KH, Munger R, Espina L, Paglianawan N, Villanueva E, Rary J, Magee K, Magee W, 1997. Clinical and epidemiologic studies of cleft lip and palate in the Philippines. Cleft Palate Craniofac J 34, 7-10

National Birth Defects Prevention Network (NBDPN), 2005. Birth defects surveillance data from selected states, 1998-2002. Birth Defects Res A Clin Mol Teratol 73, 758-853.

National Cancer Institute (NCI), 2009. State cancer profiles [http://www.statecancerprofiles.cancer.gov/]

Olson KL, Grannis SJ, Mandl KD, 2006. Privacy protection versus cluster detection in spatial epidemiology. Am J Public Health 96, 2002–2008.

Openshaw S, 1984. The modifiable areal unit problem. Concepts and techniques in modern geography 38. Geo Books, Nowich, UK.

Orr M, Bove F, Kaye W, Stone M, 2002. Elevated birth defects in racial or ethnic minority children of women living near hazardous waste sites. Int J Hyg Environ Health 205, 19-27.

Parkwyllie L, Mazotta P, Pastuszak A, Moretti ME, Beique L, Hunnisett L, Friesen HH, Jacobson S, Kasapinovic S, Chang D, Diavcitrin O, Chitayat D, Nulman I, Einarson TR, Koren G, 2000. Birth defects after maternal exposure to corticosteroids: prospective cohort study and meta-analysis of epidemiological studies. Teratol 62, 385–392.

Potthoff RF, Whittinghill M, 1966. Testing for homogeneity in the Poisson distribution. Biometrika 53, 183–190.

R Development Core Team, 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.r-project.org.

Richardson S, 1992. Statistical methods for geographical correlation studies. In: Elliott P, Cuzick J, English D, Stern R (eds). Geographical and environmental epidemiology: methods for small-area studies. Oxford University Press, Oxford, UK, 181–204.

Ritz B, Yu F, Fruin S, Chapa G, Shaw GM, Harris JA, 2002. Ambient air pollution and risk of birth defects in Southern California. Am J Epidemiol 155, 17–25.

Romitti P, Lidral A, Munger R, Daack-Hirsch S, Burns T, Murray J, 1999. Candidate genes for nonsydromic cleft lip and palate and maternal cigarette smoking and alcohol consumption; evaluation of genotype-environment interactions from a population-based case-control study of orofacial clefts. Teratol 59, 39-50.

Rushton G, Lolonis P, 1996. Exploratory spatial analysis of birth defect rates in an urban population. Stat Med 15, 717-726.

Shaw GM, Lammer EJ, Wasserman CR, O'Malley CD, Tolarova MM, 1995. Risk of orofacial clefts in children born to women using multivitamins containing folic acid periconceptionally. Lancet 346, 393-396.

Shaw GM, Wasserman CR, Lammer EJ, O'Malley CD, Murray JC, Basart AM, Tolarova MM, 1996. Orofacial clefts, parental cigarette smoking, and transforming growth factor-alpha gene variants. Am J Med Genet 58, 551-561

Siffel C, Strickland MJ, Gardner BR, Kirby RS, Correa A, 2006. Role of geographic information systems in birth defects surveillance and research. Birth Defects Res A Clin Mol Teratol 76, 825–833

Snow J, 1855. On the mode of communication of cholera (2nd Edition). Churchill, London, UK.

Spiegelhalter D, Thomas A, Best N, Lunn D, 2007. WinBUGS user manual version 1.4. Medical Research Council Biostatistics Unit. http://www.mrc-bsu.cam.ac.uk/bugs.

Tango T, 1995. A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 14, 2323–2334.

Tango T, 2000. A test for spatial disease clustering adjusted for multiple testing. Stat Med 19, 191-204.

Tobler W, 1970. A computer movie simulating urban growth in the Detroit region. Econ Geogr 46, 234-240.

Tolorova M, Cervenka J, 1988. Classification and birth prevalence of orofacial clefts. Am J Med Genet 75, 126-137.

Utah Birth Defect Network (UBDN), 2007. Orofacial Clefts at a glance. http://health.utah.gov/birthdefect/defects/orofacial.html. Accessed August 2007

Vanderas AP, 1987. Incidence of cleft lip, cleft palate, and cleft lip and palate among races: a review. Cleft Palate J 24, 216-225.

Vinceti M, Rovesti S, Bergomi M, Calzolari E, Candela S, Campagna A, Milan M, Vivoli G, 2001. Risk of birth defects in a population exposed to environmental lead pollution. Sci Total Environ 278, 23-30.

Waitzman NJ, Romano PS, Scheffler RM, 1994. Estimates of the economic costs of birth defects. Inquiry 31, 188-205

Wakefield JC, Kelsall JE, Morris SE, 2000. Clustering, cluster detection, and spatial variation in risk. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (eds). Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, 128-152.

Waller LA, Gotway CA, 2004. Applied Spatial Statistics for Public Health Data. John Wiley & Sons, New York, USA.

Walter SD, 2000. Disease mapping: a historical perspective. In: Elliot P, Wakefield JC, Best NG, Briggs DJ (eds). Spatial epidemiology: methods and application. Oxford University Press, Oxford, UK, 223-239.

Wasserman CR, Shaw GM, Selvin S, Gould JB, Syme SL, 1998. Socioeconomic status, neighborhood social conditions, and neural tube defects. Am J Public Health 88, 1674-1680

Wu JL, Wang JF, Meng B, Chen G, Pang LH, Song XM, Zhang KL, Zhang T, Zheng XY, 2004. Exploratory spatial data analysis for the identification of risk factors to birth defects. BMC Public Health 4, 23-33.

CHAPTER 2

VISUALIZATION AND INTERPRETATION OF BIRTH DEFECTS DATA USING LINKED MICROMAP PLOTS[1]

ABSTRACT

Many states have implemented birth defects surveillance systems to monitor and disseminate information regarding birth defects. However, many of these states rely on tabular methods to disseminate statistical birth defects summaries. An innovative presentation technique for birth defect data that portrays the information in a joint geographical and statistical context is the linked micromap (LM) plot. LM plots were generated for oral cleft data at two geographical resolutions—USA states and counties of Utah. The LM plots also included demographic and behavioral risk data. A LM plot for the USA reveals spatial patterns indicating higher oral cleft occurrence in the southwest and the midwest and lower occurrence in the east. The LM plot also indicates relationships between oral cleft occurrence and maternal smoking rates and the proportion of American Indians and Alaskan Natives. In particular, the five states with the highest oral cleft occurrence had a higher proportion of American Indians and Alaskan Natives. Among the 15 states with the highest oral cleft occurrence, nine had a smoking

---

[1] *The material for this chapter was previously published as:* Gebreab, S. Y., Gillies, R. R., Munger, R. G., Symanzik, J. (2008). Visualization and interpretation of birth defects data using linked micromap plots. Birth Defects Research Part A Clinical and Molecular Teratology, 82 (2):110 – 119

rate of 16% or higher while among the 15 states with the lowest oral cleft occurrence only one state had a smoking rate greater than 16%. The LM plot for the state of Utah shows no clear geographic pattern, due perhaps to a relatively small number of cases in a limited geographic area. LM plots are effective in representing complex and large volume birth defects data. Integration to birth defects surveillance systems will improve both presentation and interpretation.

BACKGROUND

Birth defects are one of the leading causes of infant mortality and childhood morbidity in the US; the statistics for the US hold that birth defects alone account for 21% of all infant mortality (CDC, 1998). Most of these birth defects result in a range of disabilities where the economic cost of medical treatment is substantial: according to Waitzman *et al*. (1994), the estimated lifetime cost of care for the number of US children born with the 18 most common birth defects exceeds $8 billion per year. In addition to the economic effects, children who are born with such birth defects often experience long-lasting psychological and physical burdens. For many years now, there has been a continuous and concentrated effort to monitor birth defects through data collection and surveillance systems—with the ultimate goal of establishing prevention strategies. As a result, many states have developed monitoring systems that collect data on the occurrence of birth defects along with other crucial information—the underlying objectives being to catalogue and disseminate information regarding the prevalence of birth defects (Sever, 2004). These data are very important in providing information as to the monitoring of

such fundamentals as the occurrences and trends of birth defects. Moreover, historical characteristics of the data are particularly useful in public health planning services, implementing prevention strategies such as allocating finite resources to the most affected areas, and improving health care access to affected children and families. Furthermore, these data are essential in a scientific sense as they are often used to generate hypotheses that are used to research the risk factors that may be associated with birth defects.

As a component of public health surveillance, states collect data on 45 major birth defects and related information (National Birth Defects Prevention Network [NBDPN], 2005). In addition, these states and many US public health agencies (e.g., CDC and NBDPN) play an important role in making birth defects data accessible to the public through differing media. However, they tend to depend on tables to disseminate the birth defects in-formation. For example, in its role to inform the public, the NBDPN published birth defects data for the period of 1998–2002 (NBDPN, 2005). The published report contains a myriad of data that consist of estimates for each birth defect by state, race/ethnicity, and, for some birth defects, by age of mother–all of which are in tabular form that constitute multiple rows and columns that run through many pages.

Publishing large statistical datasets in tabular form is an important way of managing data but is not particularly informative from an interpretative standpoint. It may be difficult and frustrating for a reader to observe trends, relationships, and anomalies that may be present in the data. A user is forced to scan through many pages of tables, and tries to build a visual picture that permits an integrative understanding of the numbers, for example, which state has the maximum number of cases in a particular year.

Equally, it can be argued that tabular data are especially useful to researchers who are interested in utilizing the raw data to conduct research; however, researchers likewise require a conversion of bulk tabular data into a visual framework to help not only in understanding the structure of the data, but further to facilitate the analysis of the data. Furthermore, there is value in reporting to the public in an informative way while at the same time facilitating the presentation of data for policy makers to enable them to make informed and timely decisions. These aforementioned circumstances suggest that the conversion of tabular data into a visual and ordered context can illustrate patterns and relationships and so forth in the data to an observer that would erstwhile be elusive and moreover, in the most practical sense, be an efficient vehicle for disseminating information to the public and decision makers.

Visualization techniques offer a set of tools that can be used to simplify large and complex datasets into more comprehensible forms. They offer the ability to transform large public health datasets such as birth defects data into a more meaningful representation of the underlying epidemiological information in a revealing way without overwhelming the reader. Using visualization, trends, relationships, and anomalies that were not at first obvious in the tables can be revealed quickly. Moreover, visualization increases the effectiveness of communicating information to the public and further enables users to do a critical evaluation of the data while at the same time likely reducing errors in its interpretation but maintaining consistency.

Over many years, many visualization tools have been developed to convert tabular information into visual graphs or plots (e.g., Carr, 1994), but a fairly recent

development in the field highlights the use of linked micromap (LM) plots (Carr and Pierson, 1996) as a way of displaying geographically indexed data. LM plots use multiple small maps (called micromaps) to visualize complex data structures in a geographical context. LM plots have already been used in many fields, including environmental science (Carr and Pierson, 1996; Carr et al., 1998; Symanzik *et al*., 1999), ecology (Carr *et al*., 2000a), epidemiology (Carr *et al*., 2000b; Symanzik *et al*., 2003), and in the case of federal statistical summaries (Hurst et al., 2003). However, LM plots have not been specifically applied to birth defects data. The purpose of this article is to highlight and examine the use of LM plots in presenting geographically indexed birth defects data. Specifically, it will demonstrate the use of LM plots to graphically represent statistical summaries and their associated uncertainties for oral cleft occurrences (oral clefts are defined as a cleft lip and cleft palate birth defects, where occurrence of oral clefts observed is prevalence at birth). This is done at two geographical resolutions: (1) at the state level for the US, and (2) at the county level for the state of Utah. Furthermore, LM plots are used to graphically relate oral cleft occurrence estimates with associated demographics and behavioral data collected at the same geographical resolutions.

A final important point is that of ensuring confidentiality. All the data used in the construction of the LM plots were aggregated values and so an individual's information is kept strictly confidential. In fact, LM plots are not designed to show specific data at a particular location but more to group information into manageable units such as a statistical summary that by its very nature removes the individual from the picture.

MATERIALS AND METHODS

Data Sources, Breakdown, and Aggregation

Birth defects and other variables of interest (including data on demographics and behavioral risk factors) were obtained from various sources. National data for oral cleft occurrence and livebirths for the period of 1998– 2002 were obtained from the NBDPN (2005) as issued in Birth Defects Research (Part A). Thirty-five states participated in reporting up to 45 major birth defects and, of these, 31 states contributed oral cleft occurrence. The relevant data for oral cleft occurrence were compiled for each state. Next, occurrence of oral cleft in each state was computed per 10,000 livebirths (NBDPN, 2005) for the same period.

Oral cleft occurrence for the state of Utah was obtained from a case-control study of oral cleft occurrence under-taken by the Center for Epidemiologic Studies, Utah State University that covered the period of 1995–2004. The cases used in the study were originally obtained from the Utah Birth Defect Network, a statewide surveillance program that monitors and detects birth defects in Utah. All cases had street address information of the mother's residence at the time of birth. The street address information was transformed (geocoded) to map coordinates and then aggregated to the county level. The live births at the county level for the same period (1995–2004) were obtained from US census data (http://quickfacts.census. gov/qfd/index.html) after which the oral cleft occurrence in each county was computed per 10,000 livebirths for the period of 1995–2004.

Details of the geocoding process are as follows. Case addresses were geocoded using the ArcView geocoding utility and Dynamap/2000 (Version 14.3). Street File Network information for the state of Utah was obtained from Geographic Data Technology (GDT), Inc. (GDT, 2004). Of the total cases, 96.6% of them were automatically geocoded or interactively geocoded with minor editing for spelling, street aliases, and acronyms. Certain addresses (0.5%) were unmatched and geocoded manually with the assistance of internet mapping services such as Yahoo Maps, MapQuest, and Google Maps. A number of the cases (2.7%) did not have a geocodable address but geo-coded either to the city or zip code centroid. The geographic centroids were obtained from a 2004 Municipalities shapefile or a zipcode shapefile available from the Utah Automated Geographic Reference Center (AGRC) (Utah AGRC, 2006). The remaining cases (0.2%) were excluded from the analysis because no address was resolved or the location resided outside the state of Utah.

Maternal smoking is a well-established risk factor for oral clefting (Khoury *et al*., 1989; Little *et al*., 2004). Therefore, data on the percentage of maternal smoking during pregnancy were obtained for the purpose of relating this particular risk factor with the oral cleft occurrence; this was done at the state level. The data on the percentage of maternal smoking during pregnancy for 2002 were obtained from Mathews and Rivera (2004). The data were collected from birth certificates and reported by 49 states (including the District of Columbia and New York City) to CDC's National Vital Statistics System, operated by the National Center for Health Statistics. Of the 31 states that reported oral clefts, only California had not collected data on maternal smoking using

the same protocol as the rest (data collected as to maternal smoking at time of pregnancy was through birth certificates; California was an exception, as it does not conform to the standard format used by the other states, hence it was coded as data not available), but instead of excluding it from the analysis, it was included as missing data. As to the reliability of the maternal smoking data collection, Mathews and Rivera (2004) note:

> Second, prenatal smoking is underreported on birth certificates. Underreporting might be related to the wording of the smoking question, the timing of the data collection (e.g., during prenatal care versus after the live birth), and the stigma associated with smoking during pregnancy, particularly in cases of poor birth outcome. However, despite underreporting, the trends and variations in smoking derived from birth certificate data have been confirmed with data from other sources (e.g., National Survey of Family Growth and Pregnancy Risk Assessment Monitoring System). (p. 913)

In adition, demographic factors, that is, race and ethnicity, are also understood to be risk factors in oral cleft occurrence. For example, the risk is particularly high in the American Indian and Alaskan Native (AIAN) population (Coddington and Hisnanick, 1996). Therefore, data on the proportion of AIAN in the population for the year 2000 was obtained from the U.S. Census Bureau (2000), accessible at http://www.census.gov/prod/2002pubs/ c2kbr01-15.pdf.

Visualization Technique

The graphical visualization technique presented in this article is referred to as LM plots. LM plots provide an alternate way (compared to traditional choropleth maps—for a comparative discussion on the relative merits of choropleth maps (see Symanzik and Carr, 2008) of displaying geographically indexed statistical summaries (e.g., oral cleft occurrence for each state or counties within a state) in a corresponding spatial context

(Carr and Pierson, 1996; Carr et al., 1998). LM plots combine both an exploratory analysis capability together with traditional statistical graphics while maintaining the geographical context.

Before LM plots are programmed and subsequently displayed (using the statistical software package S-plus or on the web), LM plots require a generalized map to work from, that is, a smoothed or simplified boundary defining a geographical region. However, such boundaries (e.g., state or county) that exist as Geographic Information System (GIS) data layers often consist of a large number of vertices that are considerably more than that required for micromap depiction on the display. Therefore, it is necessary to reduce redundant vertices in a polygon but only to the point of maintaining the essential shape and neighborhood relationship of the polygons that comprise the micromap. A generalized map for the US is available online at ftp://galaxy.gmu.edu/pub/dcarr/newsletter/micromap/. To produce a generalized map for the state of Utah, a boundary shape file was obtained from the (Utah AGRC, 2006). Using ArcGIS, a desktop GIS package, the simplified boundaries were generalized. The generalization routine applied is based on the Douglas-Peucker line simplification algorithm (Douglas and Peucker, 1973). Finally, after generalizing the boundaries, LM plots for the US and the state of Utah were created using the S-plus statistical software package. The sample S-plus code for creating LM plots is also obtainable from Dan Carr's ftp site at ftp://galaxy.gmu.edu/ pub/dcarr/newsletter/micromap/.

RESULTS

Template for LM Plots

A typical template of a LM plot consists of four key features (Carr and Pierson, 1996). Figure 2-1 shows a hypothetical LM plot. The first feature is three or more sequence panels in parallel linked by location. In the hypothetical case, Figure 2-1 shows five parallel sequences of panels. The first (leftmost) sequence of panels is the micromap panel itself that typically contains small caricatures of map outlines of a region. The caricature map maintains the shape and neighborhood relationship while making the small subregions more visible. The second (from the left) sequence of panels is the label panel that provides the names of the geographical subregions (here, Region 1 through Region 10). The third through the fifth (from the left) sequence of panels display the statistical summaries. These panels may represent many forms of statistical summaries including box-plots, dot-plots (as shown in Fig. 2-1), time series plots, CIs, and so forth. Sorting the geographic subregions based on the statistical variable(s) of interest is the second feature. Sorting improves perception between consecutive panels from the top to the bottom of the display. The third feature is the partitioning of the regions into perceptual groups of size five or less to allow the viewer's attention to focus on explicit areas at a time. The fourth feature is color and location that links corresponding elements within the parallel sequence panels, that is, the color red in the top-most panels relates to the geographic subregion in the northeast of the map, the subregion name (Region 5), and a red dot in each of the three statistical panels. The color red is reused in the next

consecutive set of panels for Region 2, but there is no relationship between Region 5 and Region 2 as one might at first assume. Simply, there do not exist enough distinguishable colors to populate an entire display (with, say, 50 different subregions) such that colors have to be reused in different panels.

In the hypothetical Figure 2-1, the rows are sorted by decreasing values with respect to the statistical panel 2. The statistical data displayed in the statistical panels 1 and 2 show a strong positive association (the correlation *r* calculated as 0.99), expressed in the almost parallel behavior of the dots and lines representing the values for these two variables. In contrast, the statistical data in panel 3 and 1 (as well as 3 and 2) show a strong negative association (the correlation r calculated as 20.94 for 3 and 1 and as 20.92 for 3 and 2). This negative association is seen in the movement of the dots and lines in opposite directions for these variables. Moreover, the data in panel 3 show an unusual outlier, the value for Region 1. It is this outlier that considerably reduces the almost perfect negative association otherwise present in this data. Just a simple numerical calculation of r might not be able to reveal the influence of a single subregion on the overall relationship.

The map panels of the LM plot in Figure 2-1 exhibit a strong geographic pattern: highest occurrences with respect to the statistical panels 1 and 2 can be found in the north and in the east; lowest occurrences can be found in the west and in the south. Additional features of LM plots exist and are described in more detail in Symanzik and Carr (2008).

US Level LM Plots

Figure 2-2 shows the LM plot for the 31 US states that reported on oral cleft occurrence. The figure shows five vertical columns that are linked by geographic location. The first column shows the generalized outline of the US wherein are drawn the map caricatures for the states. In particular, Alaska and Hawaii are modified in size and shifted towards the 48 contiguous states. Otherwise, the island to the east of Virginia represents Washington, D.C. that otherwise would not be visible. Note that redundant details of a state's boundaries are left out; however, the essential fraction that designates the boundary shape and neighborhood relationships is preserved (other than Washington, D.C.), while at the same time small states such as Rhode Island are magnified such that their assigned color is evident on the map. The second column shows the state names along with a dot in the linking color. The last three columns illustrate three statistical variables. In this particular example, dot-plots represent the three variables oral cleft occurrence, maternal smoking rate, and the AIAN proportion in the population. All the corresponding micromaps, labels, and statistical panels are linked through the same color designation. Note that five distinct colors are used to distinguish the states within a particular micromap frame.

The data in Figure 2-2 are sorted by oral cleft occurrence from largest to smallest. The micromaps are further divided into two main blocks with Texas in the middle—Texas defines the median occurrence and is plotted (and identified) in black between those states that lie above and below this median. The data are further partitioned into six micromaps each containing a grouping of five states. Such sorting (here descending) and

breaking of a long list of states into smaller groups highlights the data from a discrete visual perspective and so draws the viewer's attention to a few subregions at a time. Further-more, it also provides a viewer with additional visual perspective, that is, by sorting and breaking the data apart into, in this case, six micromaps. These LM plots provide a viewer with considerably more information than what would otherwise have been provided by a series of tables or an overall map representation (e.g., a chrolopleth map) alone. Viewers can now easily navigate through the LM plot to a place of interest in order to review oral cleft occurrence and related statistics without having to leaf backward and forward through a collection of tables or, for that matter, a series of maps. Moreover, viewers can compare the oral cleft occurrence of a particular state with a benchmark oral cleft occurrence or other states in an easier fashion. For example, it is immediately clear from the LM plot that Alaska (ranked 1st) exhibits a much higher oral cleft occurrence compared to Utah (ranked 2nd). The LM plot also reveals states that had oral cleft occurrence above, below, or equal to the median and shows states that surpassed the national average (which is 17.7 per 10,000 occurrences, i.e., 1.25 on a log10 scale). The national average is indicated with a vertical red line.

Figure 2-2 also provides a viewer with a quick overview of any spatial patterns present in oral cleft occurrence. The LM plot is very effective in revealing spatial trends. The immediate impression about spatial patterns observed in Figure 2-2 is of a few small groups of states that certainly raise questions about oral cleft occurrence similarities. For example, there is a noticeable elevation in the west (including Alaska) as compared to an observable low occurrence in the east-northeast.

However, a glance at the series of micromaps in Figure 2-2 reveals further details in spatial patterns. For example, light gray shading is used as a foreground to distinguish states above the median occurrence (i.e., in Texas) from those states below the median occurrence. The light gray shading draws attention to higher oral cleft occurrence in the upper half of the plot and lower oral cleft occurrence in the lower half of the plot. The state with the median occurrence (Texas) is shaded in all individual micromaps. The use of such shading provides additional spatial detail. As one can see in Figure 2-2, high oral cleft occurrence is primarily to be found in the west and the midwest with the exception of California, while the east coast states show up as a broad area of lower oral cleft occurrence.

LM plots can also display multiple variables simultaneously and this allows the viewer to explore the relationships between these variables. As shown in Figure 2-2, viewers can observe the relationship between oral cleft occurrence and maternal smoking—as mentioned earlier no data on maternal smoking rates were collated for California. The map shows that 9 of the 15 states that are above the median oral cleft occurrence have smoking rates above 16% (1.2 on a log10 scale) compared to only 1 of the 14 states that are below the median oral cleft occurrence. This difference is statistically significant (p=0.0052) as tested through a two-tailed Fisher's exact test. This is consistent with the smoking-cleft association that is well established and noted previously.

The rightmost statistical panel reveals a positive relationship between oral cleft occurrence and the proportion of AIAN in the population. In fact, 7 of the 15 states with

above the median oral cleft occurrence have an AIAN population equal to or above 1.3% (0.114 on a log10 scale), while none of the 15 states with below the median oral cleft occurrence exceeds the same AIAN population level. This difference is also statistically significant (p=0.00632, two-tailed Fisher's exact test).

Utah Level Analysis

Figure 2-3 illustrates a LM plot for oral cleft occurrence by county for the state of Utah. The overall design of the LM plot in Figure 3 follows the LM template: it shows five sequence columns, the first column being a map demarking the counties of Utah, while the second column contains the county name with associated color labels. The next three columns show the statistical panels for three variables for each county respectively oral cleft occurrence (counts divided by number of live births) for each county. The counties are ranked according to the oral cleft occurrence from highest to lowest and are partitioned into seven micromaps. The number of counties in Utah is 29 and therefore it is not evenly divisible by five. Symanzik and Carr (2009) provide suggestions of how to partition subregions into the micromaps when the number of geographic units (in this case the counties of Utah) within a LM plot are not evenly divisible by the number of geographic units to be displayed in a single micromap. Here, the first three micromaps and the last three micro-maps each display four counties while the fourth (middle) micromap displays five counties. Note that in this representation of the LM plot the median is not explicably drawn but the first three micromaps outline counties above the median while the last three micro-maps outline counties below the median. The county

with the median occurrence (Garfield) is shaded in all individual micromaps. No additional counties are out-lined in the fourth (middle) micromap (other than the five counties that constitute this micromap).

One supplementary statistical representation included in Figure 2-3 is the addition of CIs as part of the statistical oral cleft occurrence panel. The CIs indicated by connected small dots correspond to the 95% lower and upper confidence limits. The larger colored dots refer to, as before, the oral cleft occurrence in each county. The 95% CI was calculated for each occurrence using an exact Poisson distribution (Leslie, 1992). A viewer can now appreciate the fact that the oral cleft occurrence of each county is not quite the ‘‘true’’ (actual) oral cleft occurrence and that the CIs describe the uncertainties of the occurrence estimates, that is, the true value of the occurrence falls most likely somewhere between the limits of the CIs. Moreover, readers can also observe that counties where the occurrence is calculated from limited data (i.e., are more uncertain) have wider CIs and vice versa. As an example, consider how Daggett County (ranked 1st) with an oral cleft occurrence of 102.5 per 10,000 (resulting in a value of 2.01 on a log10 scale) compares to Salt Lake County (ranked 18th) that has an oral cleft occurrence of 12.7 per 10,000 (resulting in a value of 1.1 on a log10 scale). Upon initial examination of the occurrence information alone, one might be tempted to infer that Daggett County has a higher oral cleft occurrence when compared to Salt Lake County. However, the conclusion is somewhat different when one takes the CI information of both counties into account: it is evident that Daggett County has a wide 95% CI, compared to Salt Lake County, which has a narrow 95% CI. The implication that one should take from the

additional information is that the oral cleft occurrence for Daggett County is less reliable, while one may consider the occurrence for Salt Lake County to be more representative/reliable. This is justified by the data displayed in the counts and live-births statistical panels.

The addition of counts and livebirths into the statistical panels in Figure 2-3 provides a viewer with a more complete picture of the statistical assessment of oral cleft data at the county level. Certainly, viewers can appreciate the importance of these two variables by just comparing the oral clefts occurrence and counts in the statistical panels. As indicated in Figure 3, counties such as Salt Lake, Utah, Cache, Davis, Weber, and Box Elder have sizeable numbers in the counts and livebirths categories (a direct result of these counties being more heavily populated). This translates to narrow CIs. In contrast, counties such as Daggett, Garfield, Kane, Millard, and Sanpete correspondingly exhibit wide CIs—a direct result of a sparser population base. Overall, this demonstrates the interdependence of occurrence, counts, and livebirth numbers and implies that both the number of counts and livebirths determine the reliability of the oral cleft occurrence.

DISCUSSION

This article demonstrates the use of LM plots for the display of geographically indexed oral cleft occurrence at two geographical levels—the state and the county level. It is important to note that there are inherent limitations in the data used in this article. To begin with, all birth defects data (including oral clefts) were collected at the state level as compiled by the NBDPN—that is, the NBDPN only maintains the network of state and

population-based programs for birth defects. Thus, there may be differences in the standards used when gathering birth defects data and level of ascertainments among states; this may result in certain extremes of the variability of oral cleft occurrence among states that may obscure the true difference of the oral cleft occurrence among states. Maternal smoking and AIAN data are also not without limitations as they were only available for a single year, that is, 2002 and 2000, respectively, and do not cover the same period as the oral cleft data. Despite these limitations, we respect the differences in the state and census data and surmise that the limitations in the data are not so extreme that they may preclude the visualization and analysis presented here. The data can still provide us with important insights as to patterns and relationships in birth defects. However, the readers should be alert to these limitations and use caution when they interpret the results derived from these data. Hence, our intent is not to draw definitive conclusions from the LM plots but rather to show how the visualizations can order the data such that an easier interpretation is possible. From experience in the use of micromapping, the authors believe that LM plots may well have an important role to play in birth defect surveillance because of the many advantages a LM plot offers over tabular or other graphical methods of representation and elucidate this further with the following statements.

The first advantage is that LM plots provide an improved way of viewing and communicating information about birth defects. By sorting and breaking the datasets into a series of micromaps, LM plots simplify the visual appearance by encouraging selective focus. Viewers can immediately spot their home state or county for review of the status of

birth defects, and in this manner, they can engage in meaningful discussion. Moreover, LM plots allow viewers to make rapid and meaningful comparisons between different regions. Viewers can compare the rate of a particular state of interest with bench-mark values (median or national average) or with other states in a stratified environment. This kind of profiling of states or counties (above or below a central tendency) is valuable information for planning public health services and their subsequent decision criteria like that of resource allocation.

The second advantage of LM plots is that they present statistical summaries and estimates of birth defects in a spatial context. Unlike traditional statistical graphical methods, LM plots combine both exploratory analysis and traditional statistical graphics while maintaining the spatial context; this is very important in birth defects epidemiology because of the intrinsic spatial nature of the events. LM plots are also very effective at describing the spatial elements of the oral cleft occurrence, that is, the varied geographical distribution of the oral clefts as well as their spatial clustering. LM plots are particularly effective in highlighting subregions in a series of micro-maps, and in doing so, they reveal detailed spatial pat-terns that otherwise might not have been detected from data tables alone. As was illustrated in Figure 2-2, as one moved from the high to median oral cleft occurrence and from the median to the low oral cleft occurrence, a spatial pattern emerged. High oral cleft occurrence tended to be in the western and midwestern states, while the east coast (especially the northeast) revealed a region of low oral cleft occurrence. Such insights are as valuable for hypothesis generation as for identifying areas of high or low risk.

A third advantage of LM plots is associated with the efficacy of the technique of micromapping in handling multiple variables. It is well known that causes for birth defects are, by nature, multivariate, which advocates the linking of birth defects data with potential risk factors in order that one may investigate underlying patterns and relationships. LM plots effectively facilitate this by displaying multiple variables alongside oneanother. This capability allows readers to quickly view associations between variables and further pinpoint any anomalous relationship(s) that may exist between variables. Figure 2-2 illustrates this by displaying maternal smoking and AIAN alongside the oral cleft occurrence. In particular, the association observed between oral cleft occurrence and AIAN was immediately evident for the 15 states in the top three micromaps when compared with the remaining states.

Also shown was the capability of LM plots to display uncertainties of the oral cleft occurrence estimates. Reporting uncertainties along with the occurrence are particularly helpful to the viewer as this permits an assessment as to the reliability of the data. Viewers are able to appreciate that the big dots (Fig. 2-3) are not representative of the true value but the fact that CIs indicate that there is a range into which the true occurrence falls. The viewer can also note that states or counties with small count values and livebirths produce less reliable in-formation on the occurrence as exhibited by wider CIs, while states or counties with a large number of counts and livebirths create an occurrence that is more reliable and is evidenced by narrower CIs.

In addition to the earlier description of LM plot templates, an ample set of templates are available that offer readers considerable flexibility in visualizing their data

via LM plots. For example, the statistical panels of LM plots can take many different forms such as box-plots, bar-plots, histograms, or time series plots. These alternate statistical plots offer additional avenues for one to query the underlying structure of the data and to examine pat-terns and relationships in the data. For example, Carr *et al.* (1998) used LM plots to effectively depict time series data for per capita carbon dioxide emissions. One could imagine a similar time series LM plot that would examine the trend of NTDs before and after mandatory fortification of cereal grain products with folic acid. One can also manipulate the colors by using a different set of colors or hues. Furthermore, the beauty of LM plots is that they are not limited to static representations of summary statistics; web-based LM plots can provide users with real-time data to interactively and dynamically query, sort, and compare different regions over different resolutions, for example, at the state or county level. Such web-based LM plots also permit dynamic links between databases and automatic updates of data. In this capacity, Symanzik *et al*. (1999) developed web-based interactive LM plots for the US Environmental Protection Agency, and in a similar fashion, Wang *et al*. (2002) developed web-based LM plots for the National Cancer Institute, micromap website (National Cancer Institute, 2003) accessible at http://statecancerprofiles.cancer.gov/ micromaps/.

A final interesting aspect of the national cleft data that pertains to the eastern states lies in the fact that the oral cleft occurrences in these states all fall below the median occurrence. This is notable because the northeastern states are generally high in cancer rates (Hao *et al.*, 2006) and many (Zhu *et al*., 2002; Mili *et al.,* 1993a,b; Windham

et al., 1985) have suggested that cancer and birth defects may share common causes linked to location—these data, at least for clefts, do not support that notion.

In conclusion, LM plots provide a constructive geographic representation coupled to a statistical visualization tool, which also have an exploratory capability. In the context of the integration of LM plots towards the monitoring of birth defects, there is certainly provision, if not tremendous advantage, in the utilization of LM plots to augment the presentation of birth defect data. Further, the application of LM plots has distinct merit in the enhancement of data analysis, the generation of scientific hypotheses, as well as in the integration of data of various forms (e.g., census, environment, etc.). These aforementioned aspects, when linked together, can facilitate planning of public health services towards such aims as targeting limited resources to places with the greatest need.

REFERENCES

Carr DB. 1994. Converting tables to plots. Technical Report 101, Center for Computational Statistics, Fairfax, VA: George Mason University.

Carr DB, Olsen AR, Courbois JP, et al. 1998. Linked micromap plots: named and described. Statist Comput Stat Graph Newslet 9:24–32.

Carr DB, Olsen AR, Pierson SM, et al. 2000a. Using linked micromap plots to characterize Omernik ecoregions. Data Mining and Knowl-edge Discovery 4:43–67.

Carr DB, Pierson SM. 1996. Emphasizing statistical summaries and show-ing spatial context with micromaps. Stat Comput Stat Graph Newslet 7:16–23.

Carr DB, Wallin JF, Carr DA. 2000b. Two new templates for epidemiology applications: linked micromap plots and conditioned choropleth maps. Stat Med 19:2521–2538.

Centers for Disease Control and Prevention (CDC). 1998. Tends in infant mortality attributed to birth defects—United States, 1980–1995. MMWR 47:773–778.

Coddington DA, Hisnanick JJ. 1996. Midline congenital anomalies: the estimated occurrence among American Indian and Alaska Native infants. Clin Genet 50:74–77.

Douglas D, Peucker T. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Can Cartographer 10:112–122.

Geographic Data Technology (GDT). 2004. Dynamap/2000 v14.3 street net work file in ArcView shapefile format, county tile for the state of Utah, on digital optical disk (CD).

Hao Y, Ward EM, Jemal A, et al. 2006. U.S. congressional district cancer death rates. Int J Health Geog 5:28.

Hurst J, Symanzik J, Gunter L. 2003. Interactive federal statistical data on the web using ViZn. Comput Sci Stat 35:CD.

Khoury MJ. 1989. Does maternal cigarette smoking during pregnancy cause cleft lip and palate offspring? Am J Diseases Children 143:333– 337.

Leslie D. 1992. Simple SAS macros for the calculation of exact binomial and Poisson confidence limits. Comput Biol Med 22:351–361.

Little J, Cardy A, Munger RG. 2004. Tobacco smoking and oral clefts: a meta-analysis. Bull WHO 82:213–218.

Mathews TJ, Rivera CC. 2004. Smoking during pregnancy: United States, 1990–2002. MMWR 53:911–915.

Mili F, Khoury MJ, Flanders WD, et al. 1993a. Risk of childhood cancer for infants with birth defects. I. A record-linkage study, Atlanta, Georgia, 1968–1988. Am J Epidemiol 137:629–638.

Mili F, Lynch CF, Khoury MJ, et al. 1993b. Risk of childhood cancer for infants with birth defects. II. A record-linkage study, Iowa, 1983– 1989. Am J Epidemiol 137:639–644.

National Birth Defects Prevention Network (NBDPN). 2005. Birth defects surveillance data from selected states, 1998–2002. Birth Defects Res A Clin Mol Teratol 73:758–853.

National Cancer Institute (NCI). 2003. State Cancer Profiles. Dynamic views of cancer statistics for prioritizing cancer control efforts in the nation, states, and counties. http://statecancerprofiles.cancer.gov/ micromaps/.

Sever LE. 2004. Guidelines for conducting birth defect surveillance. 2nd ed. Atlanta, GA: National Birth Defects Prevention Network, Inc.

Symanzik J, Axelrad DA, Carr DB, et al. 1999. HAPs, micromaps and GPL—visualization of geographically referenced statistical summaries on the world wide web. In: Annual Proceedings (ACSM-WFPS-PLSO-LSAW 1999 Conference CD). American Congress on Surveying and Mapping, Bethesda, MD.

Symanzik J, Carr DB. 2008. Interactive linked micromap plots for the display of geographically referenced statistical data. In: Chen C-H, Härdle W, Unwin A, editors. Handbook of Data Visualization, New York: Springer. (in press).

Symanzik J, Gebreab S, Gillies R, et al. 2003. Visualizing the spread of West Nile Virus, 2003 Proceedings. Alexandria, VA: American Statis-tical Association, CD.

U.S. Census Bureau. 2000. The American Indian and Alaska Native popu-lation (AIAN). Census 2000 Brief. http://www.census.gov/prod/ 2002pubs/c2kbr01-15.pdf

Utah Automated Geographic Reference Center (Utah AGRC). 2006. http:// agrc.utah.gov/agrc_gisservices/gisservicesintro.html

Waitzman NJ, Romano PS, Scheffler RM. 1994. Estimates of the economic costs of birth defects. Inquiry 31:188–205.

Wang X, Chen JX, Carr DB, et al. 2002. Geographic statistics visualization: web-based linked micromap plots. Comput Sci Eng 4:90–94.

Windham GC, Bjerkedal T, Langmark F. 1985. A population-based study of cancer incidence in twins and in children with congenital malfor-mations or low birth weight, Norway, 1967–1980. Am J Epidemiol 121:49–56.

Zhu JL, Basso O, Hasle H, et al. 2002. Do parents of children with con-genital malformations have a higher cancer risk? A nationwide study in Denmark. Br J Cancer 87:524–528.

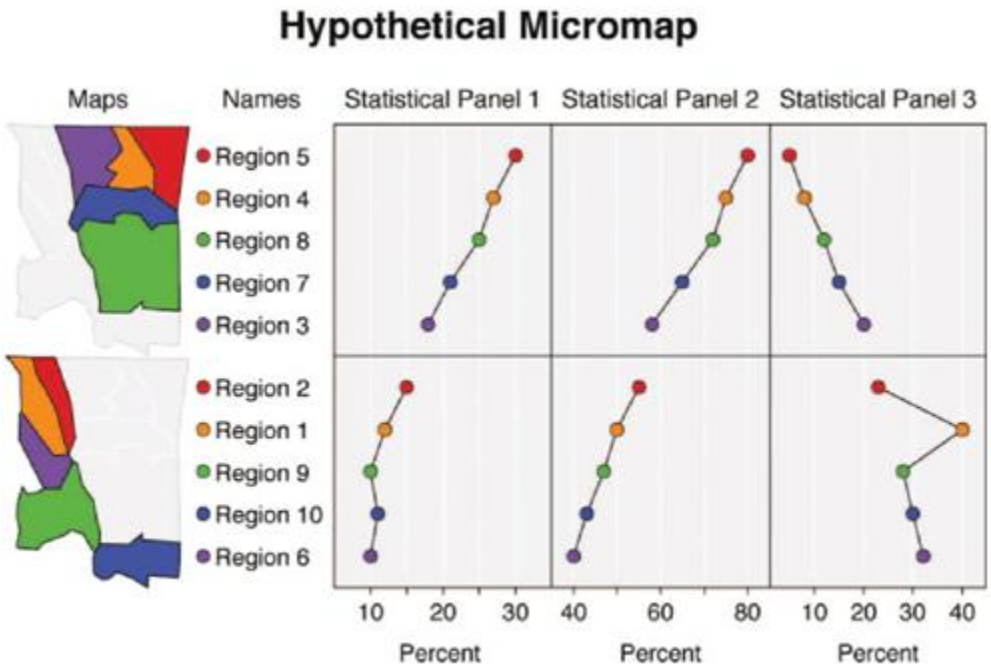


Figure 2-1 Hypothetical LM plot illustrating the main features of such plots: the leftmost sequence of map panels, the second (from the left) sequence of label panels, and the third through the fifth (from the left) sequence of statistical panels.

Figure 2-2 LM plot showing oral cleft occurrence by state for the period of 1998–2002. Only oral cleft occurrence for 31 out of the 50 US states was available and displayed here. Smoking rates for California were not available. The red lines show the national average (i.e., mean) of oral cleft occurrence (17.7 per 10,000), smoking rate 16%, and AIAN proportion of 1.3%. Note that Texas had the median oral cleft occurrence among the 31 states for which data were available.

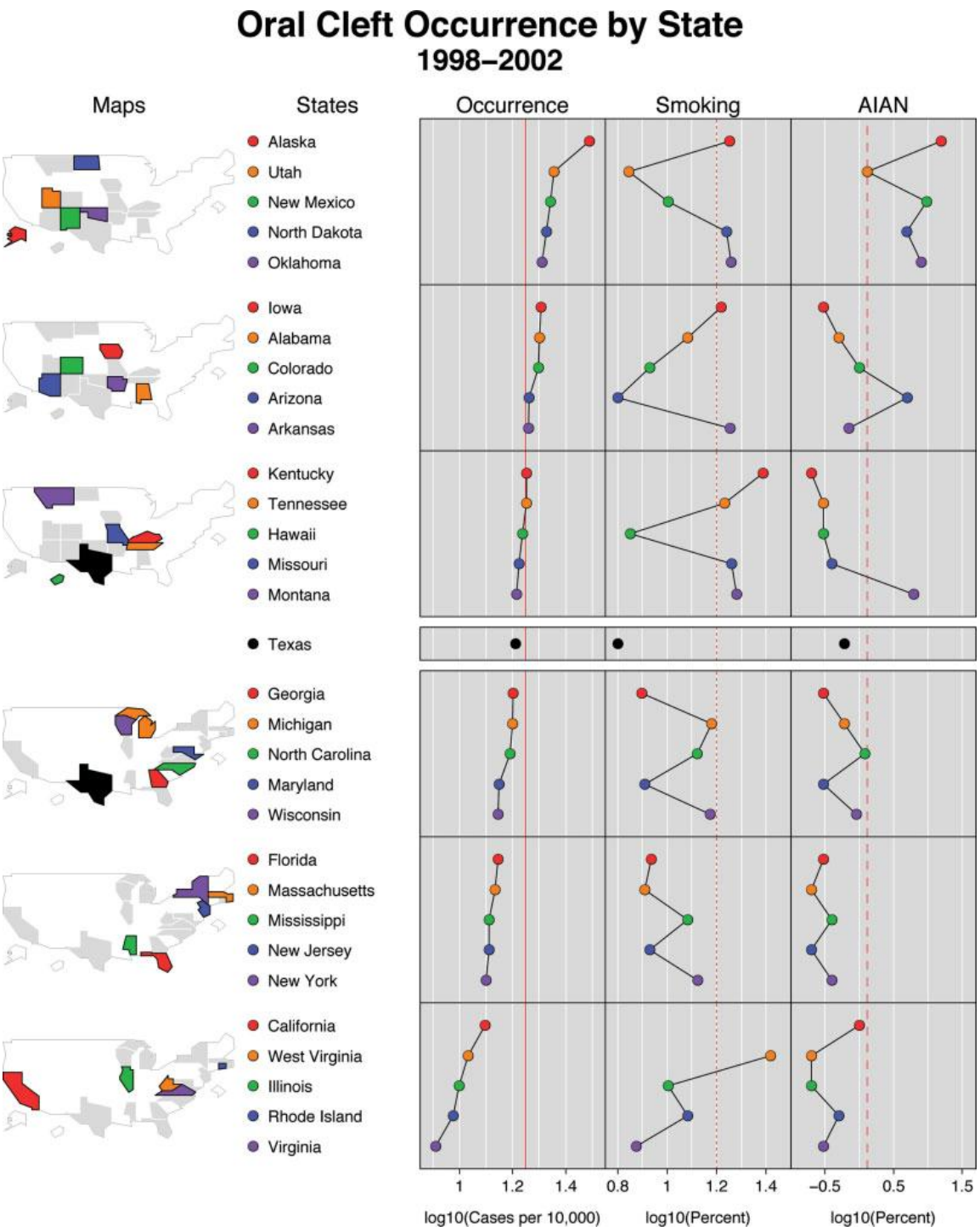
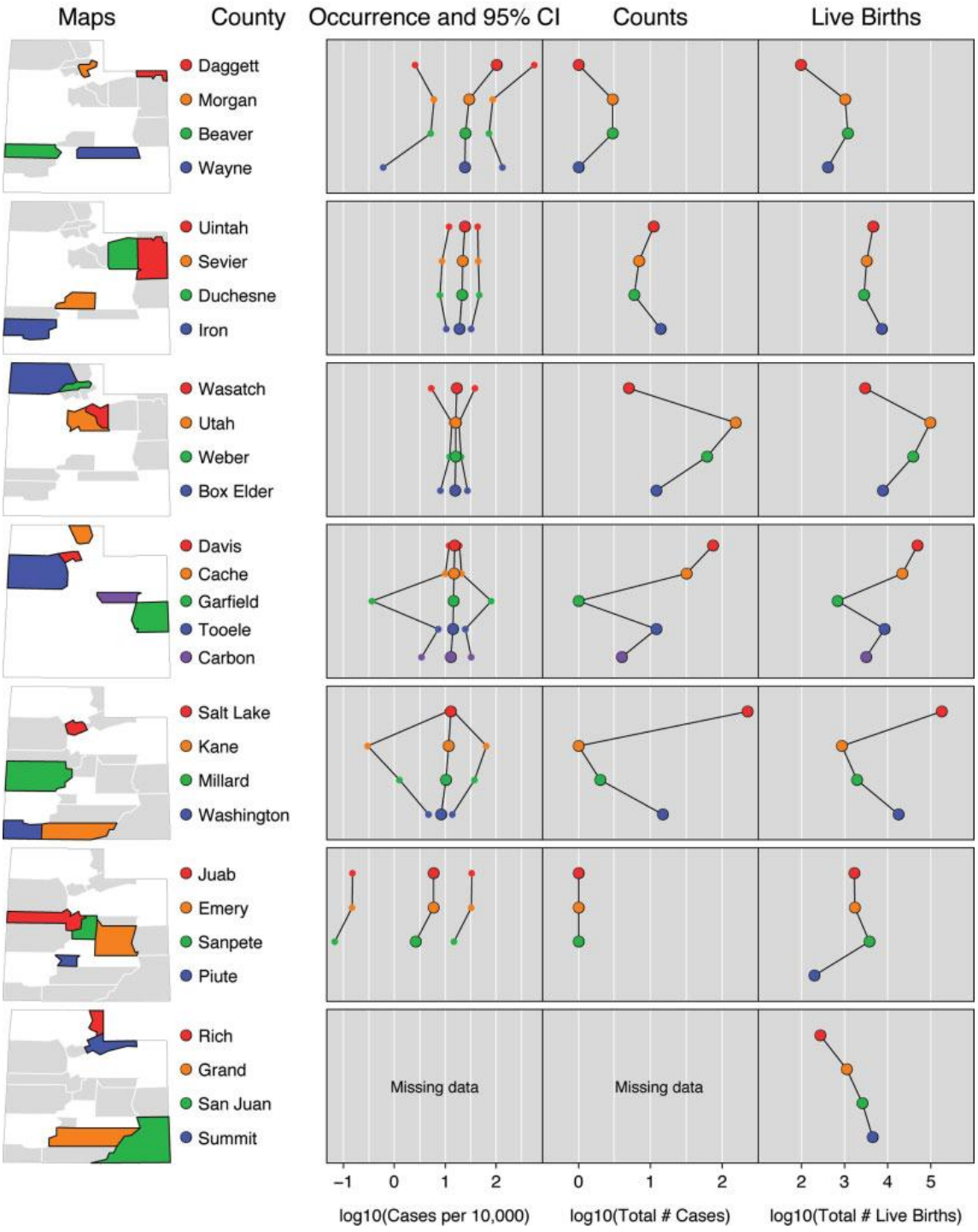

Oral Cleft Occurrence by State
1998–2002
Maps
States
Occurrence
Smoking
AIAN
Alaska
Utah
New Mexico
North Dakota
Oklahoma
Iowa
Alabama
Colorado
Arizona
Arkansas
Kentucky
Tennessee
Hawaii
Missouri
Montana
Texas
Georgia
Michigan
North Carolina
Maryland
Wisconsin
Florida
Massachusetts
Mississippi
New Jersey
New York
California
West Virginia
Illinois
Rhode Island
Virginia
1
1.2
1.4
0.8
1
1.2
1.4
−0.5
0.5
1.5
log10(Cases per 10,000)
log10(Percent)
log10(Percent)

Figure 2-3 LM plot of oral clefts occurrence for the State of Utah by county for the period of 1995-2004. Only oral clefts occurrence for 24 out of the 29 counties in Utah were available.

Oral Cleft Occurrence by County for the State of Utah
1995–2004



CHAPTER 3

SMALL AREA MAPPING AND ECOLOGICAL ANALYSES OF ORAL CLEFTS IN THE STATE OF UTAH USING HIERARCHICAL BAYESIAN MODELS[2]

ABSTRACT

The objective of this paper was to assess the spatial variation in oral clefts risk and to investigate the associations between oral clefts risk and three area-level covariates in the State of Utah using four hierarchical Bayesian models. Oral clefts data aggregated at sixty-one Utah small areas for the period from 1995 to 2004 were used. The four fitted hierarchical models were: i) a non-spatial model with uncorrelated heterogeneity random effects only; ii) a spatial model with both uncorrelated heterogeneity and spatially correlated random effects, and iii) a non-spatial regression with covariates and uncorrelated heterogeneity random effects and iv) a spatial regression model with covariates and both uncorrelated heterogeneity and spatially correlated random effects. The main findings were that the models produced more smoothed and interpretable maps of oral clefts risk compared to the SMR map. However, the models detected little evidence of spatial variation in oral clefts risk, with no pronounced evidence of spatial clustering. A few areas in Tri-county Local Health District, Provo/Brigham Young University, and North Orem areas were highlighted with a tendency of high relative risks

[2] *Part of the material for this chapter was presented* at the Urban and Regional Information Systems Association's (URISA) GIS in Public Health Conference, May 20-23, 2007, New Orleans, Louisiana.

indicating potential "local clusters" of oral clefts. The ecological analysis confirmed an association between mother using tobacco and the risk for oral clefts as has been reported in the literature while the other factors considered, mothers consuming alcohol and mothers with no high school diploma were not statistically significant. The smoothed maps of oral clefts risk produced here along with the highlighted areas of excess oral clefts can assist decision makers in initializing prevention measures and prioritizing health resources for these areas.

INTRODUCTION

Oral clefts are one of the most common birth defects in the United States (US). They are major cause of infant mortality, and lifelong morbidity and mortality (Czeizel and Sankaranarayanan, 1984; Christensen et al., 2004). Furthermore, they are associated with considerable economic (Waitzman et al., 1994), psychological and social problems (Caplan and Weintraub, 1993). Oral clefts also constitute a major public health issue in the State of Utah evidenced by the prevalence of oral clefts which lies around 1 in 450 births as compared to 1 in 750 births nationally (UBDN, 2007). Of significance, however, is the fact that this figure is the second highest in the US next to the State of Alaska (Gebreab et al., 2008). Yet it is still unclear why Utah exhibits higher oral clefts prevalence as compared to other states in the US and for that matter other countries around the world.

While the etiology of oral clefts remains largely unknown, it is hypothesized that there exists an interaction between the genetic disposition and environmental triggers

likely enhance the risk of having a child with oral clefts (Mossey et al., 2009; Romitti et al., 1999; Murray, 2002). When the etiology is largely unknown or poorly characterized, one approach to investigate the role that the environment might play in triggering oral clefting is to carry out a geographical or ecological analysis of the birth defect. It is well documented the prevalence of oral clefts varies considerably at large scales across the US and around the world. These variations have been ascribed to geographic origin (Vanderas, 1987), racial and ethnic backgrounds (Croen et al., 1998; Tolarova and Cervenka, 1998), socioeconomic (Clark et al., 2003), lifestyle and nutritional status (Bille et al., 2007), and through genetic variation (Murray, 2002). However, what is less understood is the extent to which these large-scale variations in oral cleft rates reproduce at smaller scales. The extent of geographic variation in oral clefts at small area levels has not been detailed in published literature. Therefore, we examined the spatial variations in oral cleft rates at the so-called small area level in Utah and explored their spatial link with risk factors for oral clefts measured at area-level. Spatial analysis of oral clefts at small-area level allowed us to investigate the spatial patterns of oral clefts adequately while protecting the birth defects records' privacy. Furthermore, this approach offered greater flexibility to detect localized effects of environmental factors while minimizing ecological bias (Elliott et al., 2000).

In recent years, small-area disease mapping and ecological analyses have gained increasing popularity and recognition; this is due in part to enhanced geographic information systems (GIS) capabilities, which have fueled the availability of geo-referenced data, that when coupled together have created a powerful tool and have

allowed public health professionals to carry out more high-level disease mapping and ecological analyses at the small-area level (Elliott et al., 2000; Walter, 2000). Disease maps, in the form of mortality and morbidity atlases, are now routinely produced by health agencies and such maps have improved our knowledge concerning the factors that result in particular spatial patterns of disease (Waller and Gotway, 2004). Moreover, disease maps are increasingly utilized as a basis in the generation of hypotheses in search of etiological evidence (e.g. demographic, behavioral factors, socio-economic variables, environmental hazards, or genetic disposition) that might act as a precursor for the development of an ailment. Furthermore, disease maps are particularly useful in disease surveillance where they can be applied to identify high risk areas for the purposes of preventive medicine and effective resource allocation (Wakefield et al., 2000).

The most widespread statistical methods for disease mapping and ecological analyses have been based on the standardized mortality/morbidity ratio (SMR) and Poisson regression. While effective in many contexts, they are not applicable when computed from sparsely populated area as they tend toward unstable risk estimates, especially for a disease with low frequency of occurrence (Manton et al., 1989; Gelman and Price, 1999). What is more these methods are not appropriate when the data exhibit more over-dispersion than assumed by the Poisson model (Breslow, 1984; Elliott et al., 1995). One source of over-dispersion is when the disease risk is not constant within areas or depends on unknown or unmeasured covariates. A more complex source of over-dispersion is due to the failure to account for the presence of spatial dependence of disease risks in nearby areas, such spatial dependence may arise due to one or more

spatially distributed covariates that have not been observed and/or measured (Clayton et al., 1993).

To address these problems, researchers advocate the use of Bayesian models, which offer a flexible framework for disease mapping and ecological analysis at small area level (Besag et al., 1991; Richardson, 1992; MacNab, 2004; Wakefield, 2007). One of the advantages of Bayesian models is that they provide a more stable risk estimates by "borrowing strength" from neighboring areas or entire geographical areas via the inclusion of random effects (MacNab, 2004). What is more, Bayesian models can handle over-dispersion through these random effects. In essence, the random effects can be thought of as latent variables which capture the effects of unmeasured and /or unobserved covariates that are both randomly distributed or spatially structured covariates (Richardson, 1992; Best, 1999; Clayton et al., 1993). In addition, Bayesian models offer a richer set of inferential outcomes while, at the same time, accounting for the uncertainty in the parameter of interest, including local estimates of oral cleft rates, and associated probabilistic summaries(e.g., probability of exceedance threshold and credible intervals) to quantify uncertainty.

The use of Bayesian models for disease mapping was first introduced by Clayton and Kaldor (1987), and was further developed into a fully Bayesian settings by Besag et al. (1991). However, full Bayesian inference of parameters is computationally intractable and requires computer intensive simulations, such as Markov Chain Monte Carlo (MCMC) methods. Fortunately, these issues have been addressed by the development of a statistical software package WinBUGS (Spiegelhalter et al., 2003), which facilitated the

implementations of MCMC simulations, such as Gibbs sampler (Gilks et al., 1996). As the result, full Bayesian models are increasingly being used for disease mapping and ecological analyses of chronic and infectious diseases (Best et al., 2005). Examples of recent applications of Bayesian modeling include, investigation of the spatial distribution of prostate cancer incidence in UK (Jarup et al., 2002), malaria in South Africa (Kleinschmidt et al., 2002), and neural tube birth defects in China (Wu et al., 2004).

In this study, we used hierarchical Bayesian models to conduct spatial and ecological analysis of oral clefts in the State of Utah at the small-area level for the period from 1995 to 2004. The objectives of this paper were threefold (i) to examine the spatial variation in oral cleft risks by generating statistically stable risk estimates (ii) to assess the associations between oral cleft risks and potential risk factors measured at area-level, and (iii) to accurately identify geographic areas with significantly elevated oral cleft risks. The specifics will follow but the aim is to first fit two types of Bayesian models to estimate the relative risks. They are the non-spatial model with single uncorrelated heterogeneity random effects, and the spatial model that partitions the random effects into uncorrelated heterogeneity and spatially correlated random effects, thus adjusting for the presence of spatial autocorrelation in the oral clefts data. The next step involves extending the two Bayesian models to incorporate covariates effects measured at area-level for ecological analyses, identified later as non-spatial regression and spatial regression models. The covariates included in the regression models are (a) the proportion of mothers using tobacco during pregnancy, (b) the proportion of mothers consuming alcohol during pregnancy and, (c) the proportion of mothers with no high

school diploma. The choice of covariates is based on documented major risk factors for oral clefts (Lieff et al., 1999; Munger et al., 1996; Krapels et al., 2004; Mossey et al., 2009).

The paper is organized in the following way. The next section (section 2) describes the data and data sources. Following, the statistical notations and maximum likelihood method are introduced in section 3. Section 4 provides a brief overview of the hierarchical Bayesian models, including the non-spatial, spatial and regression models. Here further discussion is provided on hyperprior specification, model implementation, convergence and selection. Section 5 presents the results of the analyses and is followed by some discussion in section 6. Section 7 consists of some conclusions and, given the analyses reflections and perspectives for future works.

MATERIALS AND METHODS

Data and Data Sources

The oral clefts data used in this study was provided through a case-control study set up to investigate Utah child and family health carried out by the Center for Epidemiologic studies at Utah State University. Data were originally obtained from Utah Birth Defects Network (UBDN), a statewide surveillance program that monitors and detects birth defects in Utah. The UBDN began monitoring efforts in 1994 by collecting data on neural tube defects. In 1995 monitoring efforts were extended to include oral clefts. Four years later in 1999, monitoring all major birth defects was undertaken (UBDN, 2007). For the purposes of this study, the data covers the period from1995 to

2004; over this period a total of 894 oral clefts cases and 458,593 live births were utilized yielding a crude rate of oral clefts of 19.5 per 10,000.

All individual cases had street address or zip code information on the mother's residence at the time of birth. A first step involved geocoding street addresses into a map coordinate system. In order to protect confidentiality, the data were then aggregated to "small geographic areas" (referred hereafter as the small areas) for subsequent analyses. Utah is divided into 61 small areas (Figure 3-1) for the purposes of public health assessment (Haggard et al., 1998). As the number of oral clefts per year aggregated at the small area level was quite low, mapping and ecological analyses were carried out using the aggregated number of oral cleft cases over the ten-year period, i.e., 1995 to 2004. The numbers of oral cleft cases within any small area ranged from 4 to 36 and the live births ranged from 2,998 to 18,177. Data on the live births (the population at risk) for each of the 61 small areas for the study period were obtained from the Utah Department of Health (UDOH, 2007).

For the purpose of the ecological analyses, as mentioned in the introduction, three covariates were considered: (a) the proportion of mothers using tobacco during pregnancy, (b) the proportion of mothers consuming alcohol during pregnancy and, (c) the proportion of mothers with no high school diploma. Here, maternal education is considered as an indication for socioeconomic status. The data for these covariates were obtained from UDOH (2007) for the study period from 1995 to 2004. These were publicly available at the UDOH web site (http://ibis.health.utah.gov/).

Maximum Likelihood Method

The study area for the state of Utah is divided into *i*=1, . . . , 61 disjoint small areas. We denote y_i to be the observed count of oral clefts cases within each area *i* and N_i the number of live births within each area *i*. Given that we are dealing with small areas and a relatively rare-disease, the observed count of oral clefts cases in each area was assumed to follow an independent Poisson distribution with mean $E_i\theta_i$:

$$y_i \sim Poisson\ (E_i\theta_i) \tag{1}$$

where E_i is the expected count of oral clefts cases in each area *i*. Typically, the expected count of cases is standardized for the age of the mother and race of the infant; this is especially so for oral clefts data. However, these data were not available; hence, it was not possible to standardize for such confounding variables. Therefore, E_i was calculated as $E_i = \sum y_i / \sum N_i$.

θ_i is the area-specific relative risk, the parameter of interest. First, we estimated θ_i using the maximum likelihood method, which is commonly referred as SMR. The SMR was computed as:

$$\hat{\theta} = SMR_i = y_i \big/ E_i \tag{2}$$

And the standard error of SMR was defined accordingly as $\sqrt{y_i}\big/E_i$. The SMR is the most common measure for disease mapping. However, as described in the introduction, the SMR is not always an appropriate measure for disease mapping because

it provides a less reliable risk estimate, especially for rare disease or sparsely populated areas. In addition, this approach does not take into account the presence of spatial autocorrelation of disease risks in nearby areas. Such correlation may be due to spatially correlated covariates which are not included in the model. To circumvent these problems, we employed hierarchical Bayesian models that permit "borrowing strength" from data across areas to achieve stable risk estimates (Besag et al., 2005).

Hierarchical Bayesian models

Within the Hierarchical Bayesian framework, our parameter of interest (e.g., the relative risks) within each small area was treated as random and assumed to have underlying probability distributions called prior distributions. The prior distributions were parameterized by hyperparameters and were, in turn, defined by hyperprior distributions leading to a hierarchical model structure. Using Bayes's theorem, these prior distributions and the likelihood of the data were combined to obtain posterior distribution of the relative risks.

In the ensuing analyses, four types of Bayesian hierarchical models of increasing complexity that took into account for uncorrelated heterogeneity, spatial correlated random effects, and covariate effects were fitted. The first model (Model 1) was non-spatial model incorporated only uncorrelated heterogeneity random effects while the second spatial model (Model 2) incorporated both spatially correlated and uncorrelated heterogeneity random effects. Models 1 & 2 were further broadened to include covariate

effects and were referred as non-spatial regression model (Model 3) and spatial regression model (Model 4). Here, we briefly outlined the models.

Model 1: Non-spatial model

The first fit was the non-spatial model (also called exchangeable model) and included only uncorrelated heterogeneity random effects in the model. These random effects was used to account for extra-Poisson variation in the oral clefts data due to important unobserved risk factors that do not have systematic spatial patterns and shrinks the relative risks to a global smoothing, and is written as

$$\log(\theta_i) = \alpha + v_i \tag{3}$$

where α is the intercept or the overall mean relative risk in the study area and v_i represents area-specific uncorrelated heterogeneity random effects and assumed to have an independent Gaussian distribution of zero mean and precision hyperparameter τ_v^2 as shown below.

$$v_i \sim N(0, \tau_v^2)$$

Model 2: Spatial model (Conditional Autoregressive Model - CAR model)

The second fit was the spatial model based on the assumption that oral clefts risk in nearby areas are likely to be similar to each other (i.e., the presence of spatial autocorrelations in the oral cleft data) in part because they reflect the underlying spatially varying risk factors. Therefore, Model 2 was augmented with spatially correlated random

effects (u_i) that accounted for the spatial clustering of oral clefts risk between neighboring areas. The formulation of the model is specified as

$$\log(\theta_i) = \alpha + v_i + u_i \tag{4}$$

where u_i represents area-specific spatially correlated random effects and follow an intrinsic Conditional autoregressive (CAR) model (Besag et al., 1991).

$$u_i | u_{j,j\neq i} \sim N\left(\frac{\sum_j w_{ij} u_j}{\sum_j w_{ij}}, \frac{\tau_u^2}{\sum_j w_{ij}}\right)$$

The parameter w_{ij} is weight matrix defining the relationship between areas *i* and its neighbor *j*. Here, the definition adopted was w_{ij} =1 if areas *i* and *j* were adjacent (i.e., share a common border) and w_{ij} = 0 if not. The hyperparmeter τ_u^2 represents the precision that controls the amount of variability in u_i and is conditional on the neighborhood structure defined by the weights w_{ij}.

Model 2 is more often referred to as a convolution model (Mollie, 1996) or BYM model after Besag, York and Mollie (1991) and is essentially the sum of two independent v_i and u_i random effects. This model allows us to determine the extent and the amount of spatial clustering in the oral clefts data, and the relative importance of spatially correlated random effects compared to uncorrelated heterogeneity random effects. In the case of u_i, the proportion of total variation due to spatial clustering was computed from the posterior distribution of ψ following (Best et al., 1999; Eberly and Carlin, 2000) as shown below.

$$\psi = \frac{sd(u)}{sd(v) + sd(u)}$$

Where $sd(v)$ and $sd(u)$ are the empirical marginal standard deviations of v_i and u_i respectively. The parameter ψ ranges between 0 and 1. If the estimate of ψ is close to 1 then the total variation is dominated by the spatial clustering while a value close to zero indicates that the spatial clustering is negligible.

Ecological analyses

In addition to modeling the relative risks of oral clefts in Utah, also investigated was the association between oral cleft risks and the following three area-level covariates (a) the proportion of mothers using tobacco during pregnancy, (b) the proportion of mothers consuming alcohol during pregnancy and, (c) the proportion of mothers with no high school diploma. The prevalence maps for the covariates are given in Figures 3-2, 3-3, and 3-4, respectively for (a) - (c). The covariates effects were modeled in the presence of uncorrelated heterogeneity and/or spatially correlated random effects and were added linearly into model 1 and model 2 as prior distributions. Consequently, the regression models take the form:

$$\log(\theta_i) = \alpha + \sum_{j=1}^{p} \beta_j x_j + v_i \qquad (5)$$

$$\log(\theta_i) = \alpha + \sum_{j=1}^{p} \beta_j x_j + v_i + u_i \qquad (6)$$

In equations 5 and 6, x_j was the value of the j^{th} area-level covariate and β_j was the corresponding regression coefficient for the j^{th} area-level covariate. Therefore,

equations 5 and 6 respectively comprise models 3 (non-spatial regression model) and 4 (spatial regression model).

Hyperprior specification

In order to perform full hierarchical Bayesian analysis, hyperprior distributions were assigned for the hyperparameters α, β, τ_v^2 and τ_u^2. Since we had no prior knowledge, we assumed independent and vague hyperpriors for all these hyperparameters. Bernardinelli et al. (1995) contains more detailed information on issues related to the selection and interpretation of the various priors. For α and β, we assigned uniform prior distributions as representative of vague beliefs. Although these are improper priors, it has been shown that these assumptions do not lead to an improper posterior distribution (Mollie, 1996). For precision hyperparameters of τ_v^2 and τ_u^2, we assigned non-informative Gamma (0.001, 0.001) hyperprior distributions. The non-informative hyperprior specifications for the precision hyperparameters allow the likelihood data to dominate the prior information; hence, it will have minimum effect on the inference of relative risks and regression coefficients.

Model implementation and convergence

The aforementioned four models were fitted to the oral clefts data using WinBUGS version 1.4 (Spiegelhalter et al., 2003). All models were simulated with two independent chains starting with dispersed initial values. Convergence of the four models were monitored by visual inspection of (a) trace plots (in this cases, a sample

should resemble a random scatter about a stable mean value), (b) autocorrelation graphs (here, a high autocorrelation graph, near to one, indicates slow mixing), and (c) Gelman-Rubin (GR) convergence statistics which is based on upon the ratio of between and within chain variance (this ratio should converge to 1.0 Best et al., 1999).

The four models described above had different "burn-in" iterations, with the slower convergence for the regression models. Convergence was detected at 40,000 iterations for the non-spatial and spatial models and at 60,000 for the regression models. Depending on the complexity of the models, the first 40,000 - 60,000 iterations were discarded as “burn-in” and each model was run for a further 20,000 iterations, giving 40,000 (2 chains x 20,000) samples with acceptable Monte Carlo (MC) errors of <5% of the sample posterior standard deviation. These pooled samples were, then, used to obtain the posterior means and credible intervals of the parameter interests of relative risks, regression coefficients, heterogeneity and, spatial autocorrelation, spatial fraction and probability of exceedance.

Model Comparison

The deviance information criterion (DIC) proposed by (Spiegelhalter et al., 2002) was used to compare the performance of the four models. The DIC is similar to the Akaike Information Criterion (AIC) (Akaike, 1973) that combines model fit and complexity. The DIC is calculated by adding the effective number of parameters (complexity) to the posterior mean deviance (goodness-fit) of a model. The effective number of parameters is estimated by the difference between the posterior mean of the

deviance and the deviance at the posterior estimates of the parameters of interest (see Appendix B for detailed information on DIC).

The 'best fit' model is the one with the smallest DIC value. Differences in DIC of around three or more are considered 'significant' while differences of 10 or more provide evidence of substantial difference in a model fit (Kelsall and Wakefield, 1999). The DIC value for each model was computed at the same time as the MCMC simulation using WinBUGS.

RESULTS

Figure 3-5 displays observed SMR oral clefts in the State of Utah from 1995 to 2004. The SMR map showed modest variability in SMR, with very few areas exhibiting extreme SMR values. The SMR values ranged from 0.43 to 1.83, with a mean value of 0.31. The map also revealed areas with high anomalies of oral clefts risk (shaded as dark grey); these were located in areas of Tri-County Local Health District (LHD) (53), Box Elder County (2), East Orem (46), and Wasatch County (52). However, the last three small areas might be outliers due to random variability. On the whole, though, the SMR map showed no evidence of an apparent spatial pattern.

Figures 3-6 to 3-9 show maps of the Bayesian estimates of relative risks (RR) derived from the non-spatial, spatial models, non-spatial regression, and spatial regression models, i.e., models 1 to 4, respectively. The Bayesian based estimates of RR were considerably smoothed, and had a narrower set of ranges in comparison to SMR values (Figure 3-5). This is clearly illustrated by the plot in Figure 3-10 where it has

shown that the Bayesian estimates of RR for models 1 to 4 shrunk substantially towards one. For example, the non-spatial RR ranged from 0.90 to 1.12, while the spatial RR ranged from 0.88 to 1.13. The effect of smoothing was particularly obvious for sparsely populated areas, for example, the non-spatial model (model 1) substantially smoothed the SMR values for the Box Elder County (2) from 1.60 to 1.06 and for the East Orem (46) from 1.72 to 1.08. On the other hand, the SMR values for areas that had large populations were preserved.

Overall, the four models produced broadly similar patterns in oral clefts risk. In fact, a comparison of the DIC values (see Table 3-1) among the models showed that there were almost no discrepancies among the models, i.e., the DIC values were essentially similar. Therefore, it is apparent that the addition of the spatially random effects and / or covariates to the models did not improve the reduced non-spatial model; indicating lack of presence of any "significant" spatial clustering in the oral clefts data.

Table 3-2 reports the variance of the uncorrelated heterogeneity $(sd(v))$ and the spatially correlated $(sd(u))$ random effects along with the spatial fraction (ψ) for the four models. As shown in Table 3-2, the variance of $sd(v)$ using non-spatial model was 0.094 (95% CI: 0.026, 0.199). This corresponds to 1.37-fold variation (95% CI: 1.09-fold to 1.70-fold variation) in risk of oral clefts between highest and lowest 5% of areas, indicating modest heterogeneity in the risk for oral clefts across 90% of the small areas in the State of Utah. The inclusion of the spatially correlated random effects in Model 2 decreased the variance of the uncorrelated heterogeneity random effects to 0.085 (95% CI: 0.024, 0.189). Furthermore, the inclusion of the covariates to models 3 and 4 further

reduced the variance of the uncorrelated heterogeneity random effects to 0.089 (95% CI: 0.024, 0.196) and 0.080 (95% CI: 0.023, 0.181) respectively. The small change in the variance of the uncorrelated heterogeneity random effects with the inclusion of the covariates suggests that there is little evidence in the data to determine conclusively the effects of these covariates on the spatial variation of oral clefts. The proportion of total variation captured by the spatially correlated random effects, ψ was around 42% (95% CI: 11% to 79%), indicating the uncorrelated heterogeneity random effects dominates over the spatially correlated random effects. In other words, there is little evidence of spatial clustering in the oral clefts data.

The parameter estimates for the associations between oral clefts and three area-level covariates are reported in Table 3-3 and the corresponding maps are shown in Figures 3-8 & 3-9. Note that the coefficient estimates (β_{1-3}) were virtually similar under both models 3 and 4. For this discussion, we use the estimates from the non-spatial regression model and the corresponding map is shown in Figure 3-8. From model 3, there was marginally significant association between oral clefts risk and mothers using tobacco, showing a relative risk of 1.025 and 95% CI of (1.000, 1.051). The association between the oral clefts risk and mothers consuming alcohol was found to be statistically not significant, indicated by a relative risk of 0.958 and 95% CI of (0.874, 1.044). Similarly, the association between oral clefts risk and mothers with no high school diploma was found to be statistically non significant, shown by a relative risk of 0.995 and, with 95% CI of (0.984 to 1.005).

The maps of high risk areas using model 1 and model 3 are presented in Figures 3-11 & 3-12 respectively. In addition, Table 3-4 shows the exceedance probability for selected areas for the non-spatial model. To identify areas with high risk of oral clefts, the posterior probability of the RR was computed in each area that exceeded a threshold of one, i.e., Pr (RR > 1.00). The high risk map using model 1 (Figure 3-11) identified three areas Tri-County local health district (LHD) (53), Provo/Brigham Young University (BYU) (44), and North Orem (47) as having at least 75% probability of excess relative risk greater than one, but none of them were statistically significant different from one. The high risk map after adjusting for the covariates using model 3 (Figure 3-12) shows that the probability for area Tri-County LHD dropped from 80% to 73%, suggesting the high risk of oral clefts in this area could be due to high proportion of mothers using tobacco (see Figure 3-2). On the other hand, the exceedance probability for the other two areas 44 and 47 was unchanged, an indication that there are no clear reasons why these areas should show a tendency of excess risk. Although none of these areas are statistically significant high risk areas, they should be closely monitored for potential risk factors.

DISCUSSION

Four Bayesian models of increasing complexity were applied to investigate and understand the risk of oral clefting in Utah, the use of which was to generate and map accurate risk maps, to identify possible clusters of oral clefts and further to assess the relationships between oral clefts and covariates measured at the small area level. The results shown here illustrate there is potential benefit in using hierarchical Bayesian

models to obtain viable assessments of underlying risk in oral clefts, if simply by overcoming some of the problems associated with the conventional SMR method. Of note was the fact that the SMR method had a tendency to misinterpret the true underlying patterns of oral clefts in Utah. In essence unreliable high risk areas were mistakenly defined as "true" high risk due to a small population size by the SMR method while the converse was true where genuinely high risk areas masked by the random noises were not identified.

Using Bayesian models, however, it was possible to achieve robust estimates of relative risks by "borrowing strength" from the neighboring areas or entire areas. It is quite clear from the comparison of the SMR map versus the Bayesian smoothed maps (Figures 3-6 to 3-9) that the Bayesian models produced more homogenous and interpretable maps by eliminating the random variation. Furthermore, the Bayesian models reduced the chance of obtaining "false" clusters while providing a much clearer picture of the "true" high risk areas. However, a limitation of the Bayesian models were excessive shrinkage of the estimates, which may mask the detection of areas with low to moderate excess risk of oral clefts (Richardson et al., 2004; Bergamaschi et al., 2006; Goovaerts and Gebreab, 2008).

Findings revealed modest evidence of heterogeneity in oral clefts risk across the small areas of the State of Utah, with no pronounced evidence of spatial clustering. This suggests that the occurrence of oral clefts is mainly driven by risk factors that are randomly distributed across the State. However, the study identified three areas (Figure 3-11 & 3-12) with a tendency of oral cleft cases as shown by their probability values

being exceeding one (Table 3-4); these areas were Tri-County LHD (53), Provo/BYU (44) and North Orem (47). Although none of these areas was statistically significant, they cannot be regarded as that of random noise and should be closely monitored for potential risk factors (the one of tobacco use is suggested) using epidemiological studies.

The Bayesian framework also enabled the exploration of any associations between oral clefts risks and area-level risk factors. Our study found a positive but marginally significant association between the mothers using tobacco and the risk for oral clefts. This result is in concurrence with previous reports that mother's tobacco use is an important risk factor for cleft lip and/or cleft palate (Khoury et al., 1987; Lieff et al., 1999; Chung et al., 2000). Other studies (e.g., Munger et al., 1996; Lorente et al., 2000) have noted that mother's alcohol consumption during pregnancy is also a risk factor for oral clefts. However, in our study the association between oral clefts risk and mother consuming alcohol was not statistically significant; this inconsistency can be attributed to the use of ecological data, i.e., through aggregation of data direct relationships between oral clefts risk and the mother's alcohol consumption may simply be obscured.

Further investigation the association between oral clefts and mothers with no high school diploma was found to be statistically non-significant; this is in contrast to the positive and significant association found in a case-control study conducted using the data from Utah child and family study (Moss, 2006). The lack of association with the mothers education could be due again to the fact that this study is based on aggregated data and not at individual level, which can mask the true association that is occurring at the individual level. Another possible reason may be a mother's education by itself might

not be a sufficient indicator, but simply a complex variable that is a combination of other variables such as maternal racial/ethnicity or mother's habits (i.e., tobacco use, alcohol consumption, drug abuse, intake of multivitamin supplements during pregnancy, etc.). However, regardless of its direct association with oral clefts, it is important that a mother's education is considered as a surrogate for socioeconomic status or as a proxy for covariates that are unmeasured or difficult to measure. For example, women with lower education are more likely to smoke likely to take folic acid during pregnancy (Clark et al., 2003; Krapels et al., 2004), which several studies have shown multivitamin and folic acid supplementation reduce the risk for oral clefting (van Rooji et al., 2004, Wilcox et al., 2007).

There are a number of limitations that could have affected our results. A major limitation of was that we could not adjust for age and race, because these data were available. The results potentially could be affected by the difference in race and age between areas; for example, the excess oral clefts cases in the Tri-County area might be due to high number of the American Indian population in that area. The converse might also be true that important spatial patterns or areas of excess oral clefts cases have not been accounted for. However, it is surmised that adjustment for age and race would have had a negligible effect on the results because most of the cases and live births are Caucasian followed closely by Hispanic, and they have virtually the same rates (UBDN, 2007).

Another concern about the data used in this study is that there a number of unaccounted oral clefts cases that have not been included. For example, the oral clefts

rate as reported by the UBDN is 22.2 per 10,000 for the period from 1995 to 2003 compared to this study's rate of 19.5 per 10,000 for the period from 1995 to 2004. Although the difference between these two rates is small and the percentage of missing cases very low; there might exist a small bias which might have slightly affected our results. However, as long as they are distributed randomly, any bias should be negligible.

The results here are based on so-called ecological analysis that assesses the association between aggregated covariates and the risk for oral clefts. The inference based on these analyses cannot be directly transferred to the individual level, because of the ecological bias known as ecological fallacy. Nevertheless, ecological analysis such as is presented here can be useful for generating testable hypotheses using epidemiological studies such as case-control or cohort studies (Best, 1999). Furthermore, they are particularly useful because the data are freely available and the fact that cost and confidentiality reasons can limit the amount of research conducted at the individual level and likely means that such analyses, such as was conducted here, will likely continue to provide either etiological clues or monitoring guidance in any spatially distributed disease or ailment.

SUMMARY AND CONCLUSIONS

Demonstrated here was the benefit of hierarchical Bayesian models through the development and provision of reliable maps of oral clefts risk maps at small area level. The overall analysis revealed modest spatial variations in oral clefts risk but no pronounced spatial pattern; this implies that the occurrence of oral clefts in the State of

Utah appears to be not driven by spatially distributed risk factors. The study also identified areas of potential "hotspots" for oral clefting, particularly Tri-county LHD, North Orem, and Provo/BYU, which warrant further investigation. The ecological analysis did however confirm an association between mothers using tobacco and the risk for oral clefts as has been reported in the literature in the subject area. The other factors considered, a mothers consuming alcohol and mothers with no high school diploma were not statistically significant in this specific study.

While our study is based on ecological analysis, studies such as this are useful in exploratory etiologic research. The results from this study are potentially valuable to epidemiologists and public health authorities in Utah towards developing preventive and control measures that can be applied to reduce the burden of oral clefts in the State. The smoothed maps of oral clefts risk produced here along with the highlighted areas of excess oral clefts can be useful in further investigating etiological factors, initializing prevention measures and prioritizing health resources for affected areas. For instance, a risk factor like mother's tobacco use can be prevented through smoking cession intervention programs. On a more esoteric front, such studies also prove to be valuable in generating hypotheses and oftentimes set the stage for more comprehensive epidemiological studies using either cohort or case-control studies.

However, there are several issues that could lead to improvements of this ecological study. For example, future work should account for known potential confounding factors, and would certainly expand the models to account for measurement errors (Best, 1999; Bernardinelli et al., 2000). In the future, the Bayesian models should

extend to include temporal or seasonal variations in oral clefts. Furthermore, an opportunity exists in a future to explore other choices of neighborhood weight matrix that reflect better the irregular size and shape of the small-areas of the study area, such as the State of Utah (Goovaerts and Gebreab, 2008). Certainly the techniques applied here should be extended in a categorization of oral clefts that reflects cleft lip with or without cleft palate (CL/P) or cleft palate only (CP) separately, as they appear to have discrete etiologies (Harville et al., 2005), this may shed light on whether they have similar spatial patterns and if they share common risk factors. Finally, it should be noted that our ecological study of oral clefts is only preliminary. For refined understanding of the etiology and development of intervention for oral clefts, more investigation of the case-control data collected by Utah child and family study will be required.

REFERENCES

Akaike H, 1973. Information theory and an extension of the maximum likelihood principle. In, Petrov B, Caáki F (eds). 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.

Bernardinelli L, Clayton D, Montomoli C, 1995. Bayesian estimates of disease maps, How important are priors? Stat Med 14, 2411-2431.

Bernardinelli L, Pascutto C, Montomoli C, Gilks W, 2000. Investigating the genetic association between diabetes and malaria, an application of Bayesian ecological regression models with errors in covariates. In: Elliott P, Wakefield JC, Best NG,

Briggs DJ (eds). Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, 286–301.

Bergamaschi R, Montomoli C, Candeloro E, Monti M, Cioccale R, Bernardinelli L, Fratino P, Cosi V, 2006. Bayesian mapping of multiple sclerosis prevalence in the province of Pavia, northern Italy. J Neurol Sci 244, 127-131.

Besag J, York JC, Mollié A, 1991. Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann Inst Stat Math 43, 1–59.

Best NG, Arnold RA, Thomas A, Waller LA, Conlon EM, 1999. Bayesian models for spatially correlated disease and exposure data (with discussion). In: Bernardo JM, Berger JO, Dawid AP, Smith AFM (eds). Bayesian Statistics 6, Oxford University, Press, Oxford, UK, 131-156.

Best NG, 1999. Bayesian ecological modelling. In: Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel JF, Bertollini R (eds). Disease Mapping and Risk Assessment for Public Health. John Wiley and Sons, Chichester, UK, 193–201.

Best NG, Richardson S, Thomson A, 2005. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res 14, 35–59.

Bille C, Olsen J, Knudsen VK, Olsen SF, Rasmussen K, Murray JC, Andersen MA, Christensen K, 2007. Oral clefts and life style factors – A case–cohort study based on prospective Danish data. Eur J Epidemiol 22, 173-181.

Breslow NE, 1984. Extra-Poisson variation in log-linear models. Applied Statistics 33, 38-44.

Caplan DJ, Weintraub JA, 1993. The oral health burden in the United States, a summary of recent epidemiologic studies. J Dent Edu 57, 853-62.

Christensen K, Juel K, Herskind AM, Murray JC, 2004. Long term follow up study of survival associated with cleft lip and palate at birth. BMJ 328, 1405

Chung KC, Kowalski CP, Kim HM, Buchman SR, 2000. Maternal cigarette smoking during pregnancy and the risk of having a child with cleft lip/palate. Plast and Reconstr Surg 105, 485-491.

Clark JD, Orth D, Mossey PA, Orth M, Sharp L and Little J, 2003. Socioeconomic status and orofacial clefts in Scotland, 1989 to 1998. Cleft Palate Craniofac J 40, 481-485

Clayton DG, Kaldor J, 1987. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43, 671-681.

Clayton DG, Bernardinelli L, Montomoli C, 1993. Spatial correlation and ecological analysis. Int J Epidemiol 22, 1193–1201.

Croen LA, Shaw GM, Wasserman CR, Tolarova MM, 1998. Racial and ethnic variations in the prevalence of orofacial clefts in California, 1983-1992. Am J Med Genet 79, 42-47.

Czeizel A, Sankaranarayanan K, 1984. The load of genetic and partially genetic disorders in man. 1. Congenital anomalies, estimates of detriment in terms of years of life lost and years of impaired life. Mutat Res 128, 73-103.

Eberly LE, Carlin BP, 2000. Identifiability and Convergence Issues for Markov Chain Monte Carlo Fitting of Spatial Models. Stat Med 19, 2279-2294.

Elliott P, Martuzzi M, Shaddick G, 1995. Spatial statistical methods in environmental epidemiology, a critique. Stat Methods Med Res 4, 137-159.

Elliott P, Wakefield J, Best N, Briggs DJ, 2000. Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, 3–14.

Gebreab SY, Gillies RR, Munger RG, Symanzik J, 2008. Visualization and interpretation of birth defects data using linked micromap plots**.** Birth Defects Research Part A, Clinical and Molecular Teratol 82, 110-119.

Gelman A, Price PN, 1999. All maps of parameter estimates are misleading. Stat Med 18, 3221-3234.

Gilks WR, Richardson S, Spiegelhalter DJ, 1996. Monte Carlo Markov Chain in practice. Chapman and Hall, London, UK.

Goovaerts P, Gebreab S, 2008. How does Poisson kriging compare to the popular BYM model for mapping disease risks. Int J Health Geogr 7, 6.

Haggard LM, Shah GH, Rolfs RT, 1998. Assessing health status, establishing geographic areas for small area analysis in Utah. Utah's Health: An Annual Review Vol V., 1997-1998.

Harville EW, Wilcox AJ, Lie RT, Vindenes H, Abyholm F, 2005. Cleft lip and palate versus cleft lip only, are they distinct defects? Am J Epidemiol 162, 448–453.

Jarup L, Best NG, Toledano MB, Wakefield J, Elliott P, 2002. Geographical epidemiology of prostate cancer in Great Britain. Int J Cancer 97, 695–699.

Kelsall JE, Wakefield, 1999. Discussion of Bayesian models for spatially correlated disease and exposure data, by Best NG, Arnold RA, Thomas A, Conlon E, Waller

LA Thomas A, Conlon EM, Arnold R. In: Bernado JM, Berger JO, Dawid AP, Smith AFM (eds). Bayesian Statistics 6, Oxford University Press, Oxford, UK, pp. 151.

Khoury MJ, Weinstein A, Panny S, Holtzman NA, Lindsay PK, Farrel K, Eisenberg M, 1987. Maternal cigarette smoking and oral clefts, a population-based study. Am J public Health 77, 623-625.

Kleinschmidt I, Sharp B, Mueller I, Vounatsou P, 2002. Rise in malaria incidence rates in South Africa, a small area spatial analysis of variation in time trends. Am J Epidemiol 155, 257–264.

Krapels IP, van Rooij IA, Ocke MC, van Cleef BA, Kuijpers-Jagtman AM, Steegers-Theunissen RP, 2004. Maternal dietary B vitamin intake, other than folate, and the association with orofacial cleft in the offspring. Eur J Nutr 43,7-14.

Lieff S, Olshan AF, Werler M, Strauss RP, Smith J. Michell A, 1999. Maternal cigarette smoking during pregnancy and risk of oral clefts in newborns. Am J Epidemiol 150, 683-94.

Lorente C, Cordier S, Goujard J, Ayme S, Bianchi F, Calzolari E, De Walle HE, Knill-Jones R, 2000. Tobacco and alcohol use during pregnancy and risk of oral clefts. Am J Public Health 90, 415-419.

MacNab YC, 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Acc Anal Prev 36, 1019-1028.

Manton KG, Woodbury MA, Stallard E, Riggan WB, Creason JP, Pellom AC, 1989. Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. J Am Stat Ass 84, 637-650.

Mollie A, 1996. Bayesian mapping of disease. In: Gilks WR, Richardson S, Spiegelhalter DJ (eds). Markov Chain Monte Carlo in Practice. Chapman & Hall, New York, USA, 359–79.

Moss MM, 2006. Smoking, Anemia, and risk of oral clefts in Utah. MS thesis, Utah State University, Logan, USA.

Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC, 2009. Cleft lip and palate. Lancet 374, 1773-85.

Munger R, Romitti P, Daack-Hirsch S, Burns T, Murray J, Hanson J, 1996. Maternal alcohol use and risk of orofacial cleft birth defects. Teratol 54, 27-33.

Murray JC, 2002. Gene/environment causes of cleft lip and/or palate. Clin Genet 61, 248-256.

Richardson S, 1992. Statistical methods for geographical correlation studies. In: Elliott P, Cuzick J, English D, Stern R (eds). Geographical and Environmental Epidemiology, Methods for Small-Area Studies. Oxford University Press, Oxford, UK, 181–204.

Richardson S, Thomson A, Best N, Elliott P, 2004. Interpreting posterior relative risk estimates in disease mapping studies. Environmental Health Perspect 112, 1016-1025.

Romitti P, Lidral A, Munger R, Daack-Hirsch S, Burns T, Murray J, 1999. Candidate genes for nonsydromic cleft lip and palate and maternal cigarette smoking and alcohol consumption; evaluation of genotype-environment interactions from a population-based case-control study of orofacial clefts. Teratol 59, 39-50.

Spiegelhalter DJ, Best NG, Carlin BP, van-der Linde A, 2002. Bayesian measures of model complexity and fit (with discussion). J R Stat Soc B 64, 583–640.

Spiegelhalter DJ, Thomas A, Best N, Lunn D, 2003. WinBUGS user manual version 1.4. Medical Research Council Biostatistics Unit. http,//www.mrc-bsu.cam.ac.uk/bugs.

Tolorova M, Cervenka J, 1998. Classification and birth prevalence of orofacial clefts. Am J Med Genet 75, 126-137.

Utah Birth Defect Network (UBDN), 2007. Orofacial Clefts at a glance. Information available at http,//health.utah.gov/birthdefect/defects/orofacial.html. Accessed August 2007.

Utah Department of Health (UDOH), 2007. Utah's indicator based-information system for public. Information available at http,//ibis.health.utah.gov/ Accessed August 2007.

Vanderas AP, 1987. Incidence of cleft lip, cleft palate, and cleft lip and palate among races, a review. Cleft Palate J 24, 216-225.

van Rooij IA, Ocke MC, Straatman H, Zielhuis GA, Merkus HM, Steegers-Theunissen RP, 2004. Periconceptional folate intake by supplement and food reduces the risk of nonsyndromic cleft lip with or without cleft palate. Prev Med 39, 689–694.

Waitzman NJ, Romano PS, Scheffler RM, 1994. Estimates of the economic costs of birth defects. Inquiry 31, 188-205.

Wakefield J, 2007. Disease mapping and spatial regression with count data. Biostatistics 8, 158-183.

Wakefield JC, Best NG, Waller L, 2000. Bayesian approaches to disease mapping. In: Elliott P, Wakefield JC, Best NG, Briggs DJ (eds). Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, 104–127.

Waller LA, Gotway CA, 2004. Applied Spatial Statistics for Public Health Data. John Wiley & Sons, New York, USA.

Walter SD, 2000. Disease mapping, a historical perspective. In: Elliott P, Wakefield JC, Best NG, Briggs DJ (eds). Spatial epidemiology: methods and applications. Oxford University Press, Oxford, UK, 223-239.

Wilcox AJ, Lie RT, Solvoll K,Taylor J, McConnaughey DR, Abyholm F, Vindenes H, Vollset SE, Drevon CA, 2007. Folic acid supplements and risk of facial clefts: national population based case-control study. BMJ 334, 464

Wu JL, Wang JF, Meng B, Chen G, Pang LH, Song XM, Zhang KL, Zhang T, Zheng XY, 2004. Exploratory spatial data analysis for the identification of risk factors to birth defects. BMC Public Health 4, 23–33.

Table 3-1 Posterior mean deviance ($\overline{D}$), effective number of parameters (p_D), and model comparison criterion (DIC) for each model.

Model	$\overline{D}$	$D(\bar{\theta})$	p_D	DIC
Non-Spatial model (Model 1)	331.11	322.08	9.03	340.14
Spatial Model (Model 2)	329.31	317.94	11.37	340.68
Non-spatial regression (Model 3)	330.3	319.04	11.26	341.56
Spatial regression (Model 3)	327.82	314.27	13.55	341.36

Table 3-2 Posterior means and 95% credible intervals (CIs) of the standard deviations of the random effects for the four Bayesian models.

Model	$sd(v)$ (95% CI)	$sd(u)$ (95% CI)	ψ (95% CI)
Non-Spatial model (Model 1)	0.094 (0.026, 0.199)	–	–
Spatial model (Model 2)	0.085 (0.024, 0.189)	0.06 (0.015, 0.150)	0.422 (0.114, 0.790)
Non-spatial regression (Model 3)	0.089 (0.024, 0.196)	–	–
Spatial regression (Model 4)	0.08 (0.023, 0.181)	0.072 (0.017, 0.170)	0.472 (0.135, 0.816)

Table 3-3 Posterior coefficients and 95% CIs of area-level covariates associated with oral clefts for the non-spatial and spatial regression models.

Covariates	Posterior mean (β^*)	$SD(\beta)$	MC error	RR	95% CI (RR)
Non-Spatial Regression (Model 3)					
Intercept (β_0)	-0.0732	0.0861	1.39E-03	0.934	(0.786, 1.097)
Smoking (β_1)	**0.0249**	**0.0125**	**2.60E-04**	**1.025**	**(1.000, 1.051)**
Alcohol (β_2)	-0.0434	0.0452	5.24E-04	0.958	(0.874, 1.044)
Education (β_3)	-0.0055	0.0056	1.04E-04	0.995	(0.984, 1051)
Spatial Regression (Model 4)					
Intercept (β_0)	-0.0893	0.0905	1.78E-03	0.918	(0.7618, 1.090)
Smoking (β_1)	**0.0261**	**0.0134**	**2.94E-04**	**1.026**	**(1.000, 1.053)**
Alcohol (β_2)	-0.042	0.0473	6.01E-04	0.96	(0.8727, 1.050)
Education (β_3)	-0.0053	0.0058	1.11E-04	0.995	(0.9833, 1.006)

Table 3-4 Posterior means of RR and probability of exceedance for selected areas using the non- spatial model.

Area	SMR	RR	SD	MC error	95% CI	Pr (RR > 1)
Tri-County HDL(53)	**1.834**	**1.135**	**0.155**	**0.00326**	**(0.916, 1.521)**	**0.83**
Wasatch Co.(52)	1.711	1.059	0.132	0.0016	(0.848, 1.374)	0.658
Provo/BYU (47)	**1.458**	**1.107**	**0.136**	**0.00234**	**(0.898, 1.431)**	**0.792**
East Orem (46)	1.721	1.08	0.14	0.00211	(0.861, 1.419)	0.709
North Orem (44)	**1.428**	**1.101**	**0.132**	**0.00198**	**(0.895, 1.417)**	**0.785**
Elder Box (2)	1.595	1.064	0.13	0.00155	(0.854, 1.372)	0.677

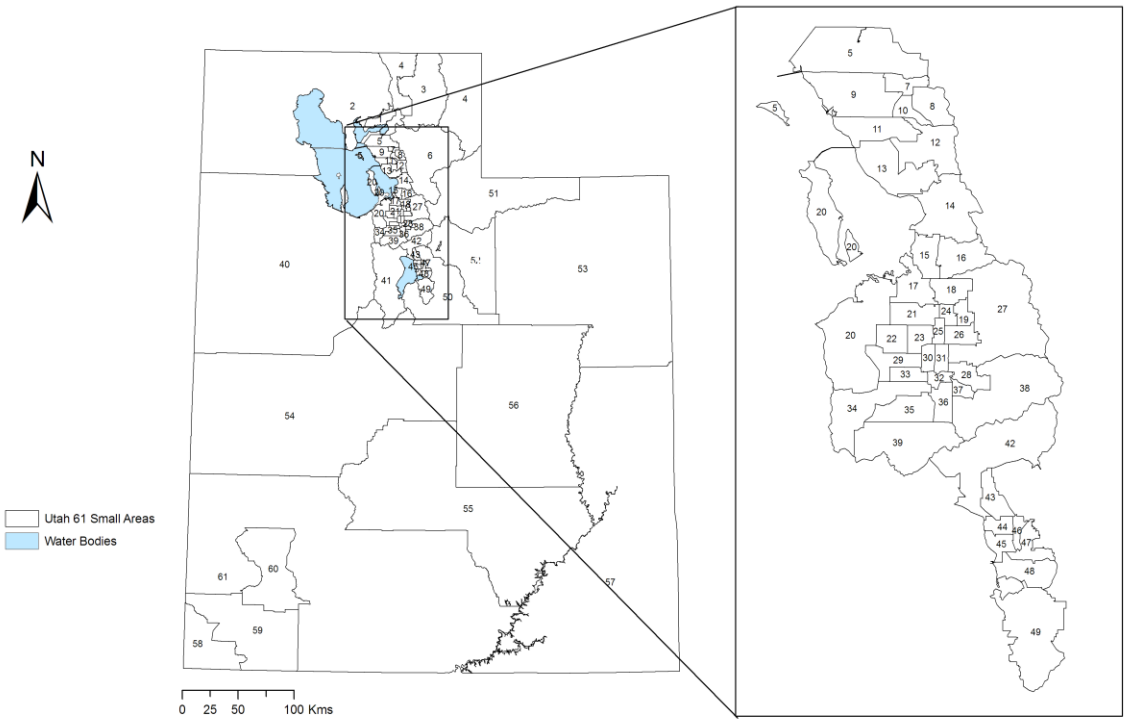


Figure 3-1 The study area for the State of Utah showing 61 small-geographic areas with their corresponding area ID.

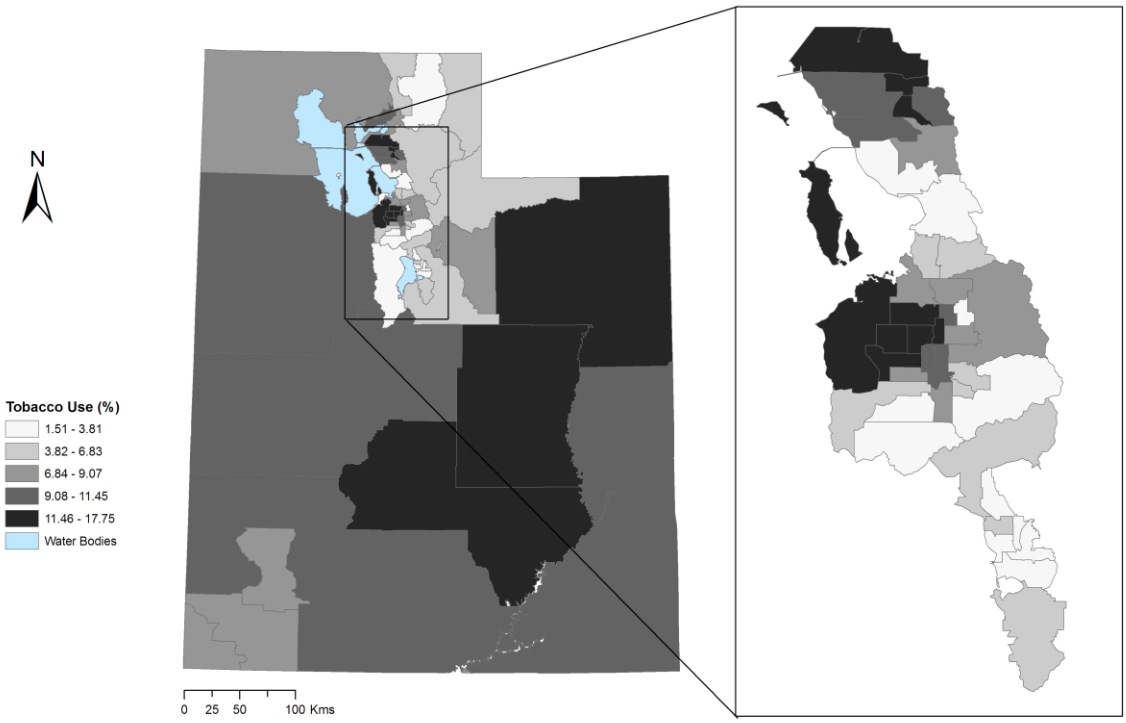


Figure 3-2 The proportion of mothers using tobacco during pregnancy from 1995 to 2004 across Utah small areas. The color scheme goes from a light grey (low smoking percentage) to a dark grey (high smoking percentage).

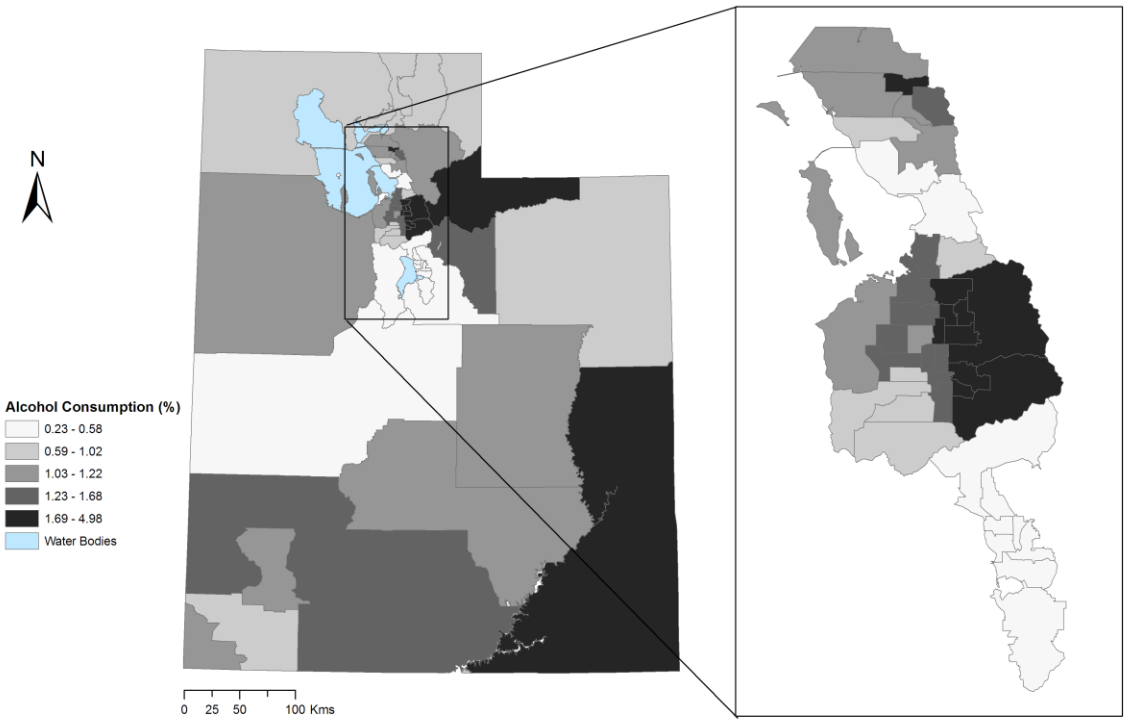


Figure 3-3 The proportion of mothers consuming alcohol during pregnancy from 1995 to 2004 across Utah small areas. The color scheme goes from a light grey (low alcohol consumption percentage) to a dark grey (high alcohol consumption percentage).

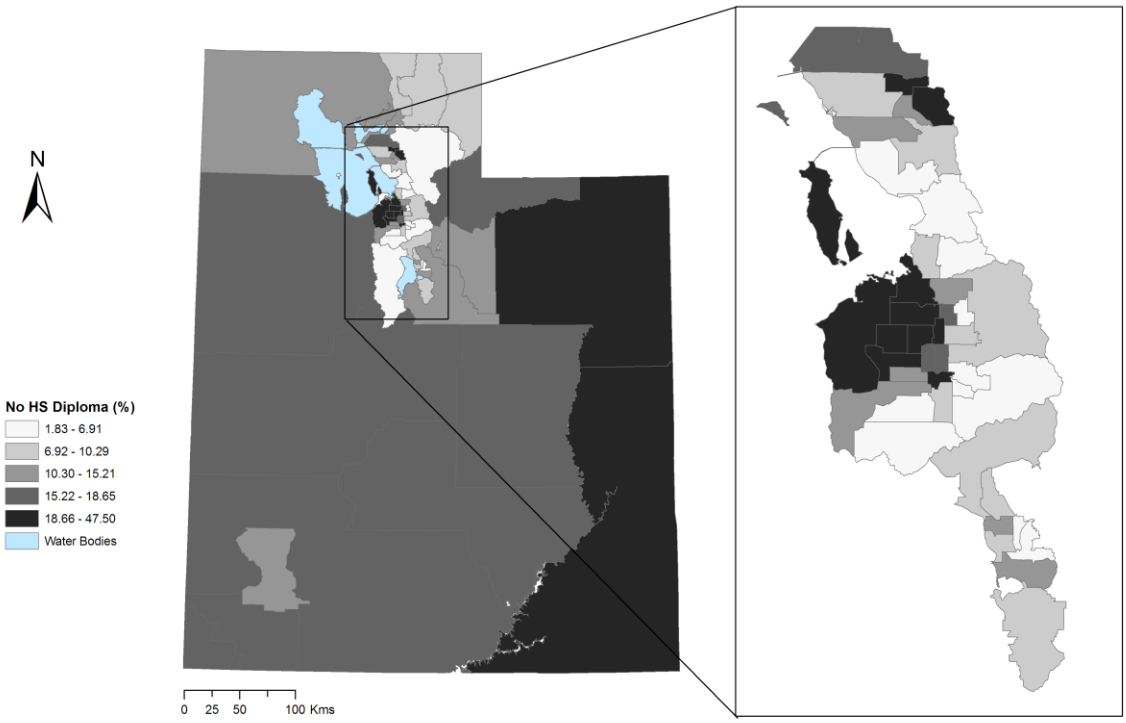


Figure 3-4 The proportion of mothers with no high school (HS) diploma during pregnancy from 1995 to 2004 across Utah small area. The color scheme goes from a light grey (low percentage of mothers with no HS diploma) to a dark grey (high percentage of mothers with no HS diploma).

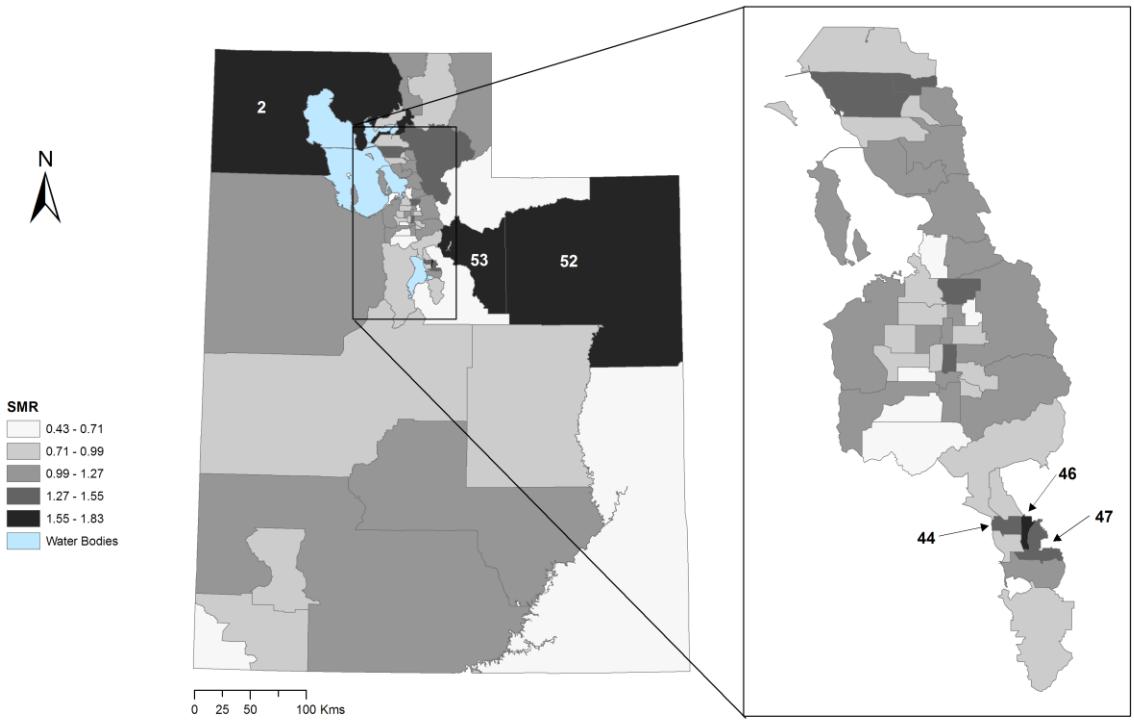


Figure 3-5 Raw Standardized Morbidity Ratio (SMR) values for oral clefts from 1995 to 2004 across Utah small areas. The color scheme goes from a light grey (low SMR value) to a dark grey (high SMR value).

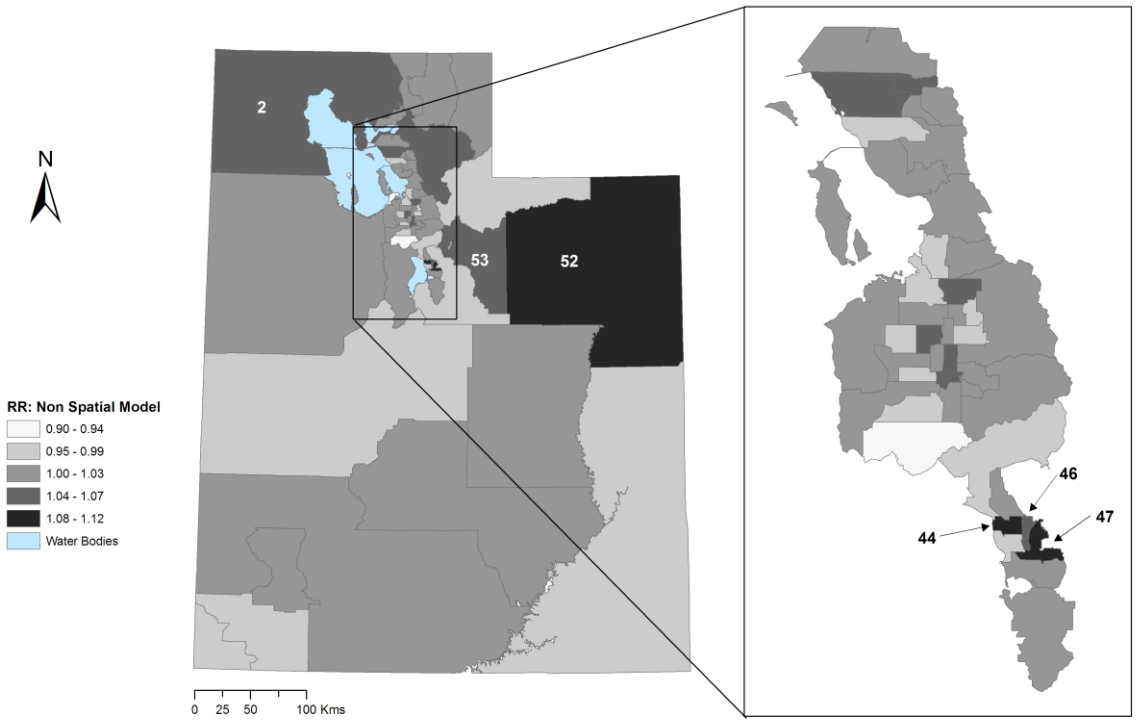


Figure 3-6 Bayesian smoothed relative risks (RR) for oral clefts from 1995 to 2004 across Utah small areas using non-spatial model (Model 1). The color scheme goes from a light grey (low RR value) to a dark grey (high RR value).

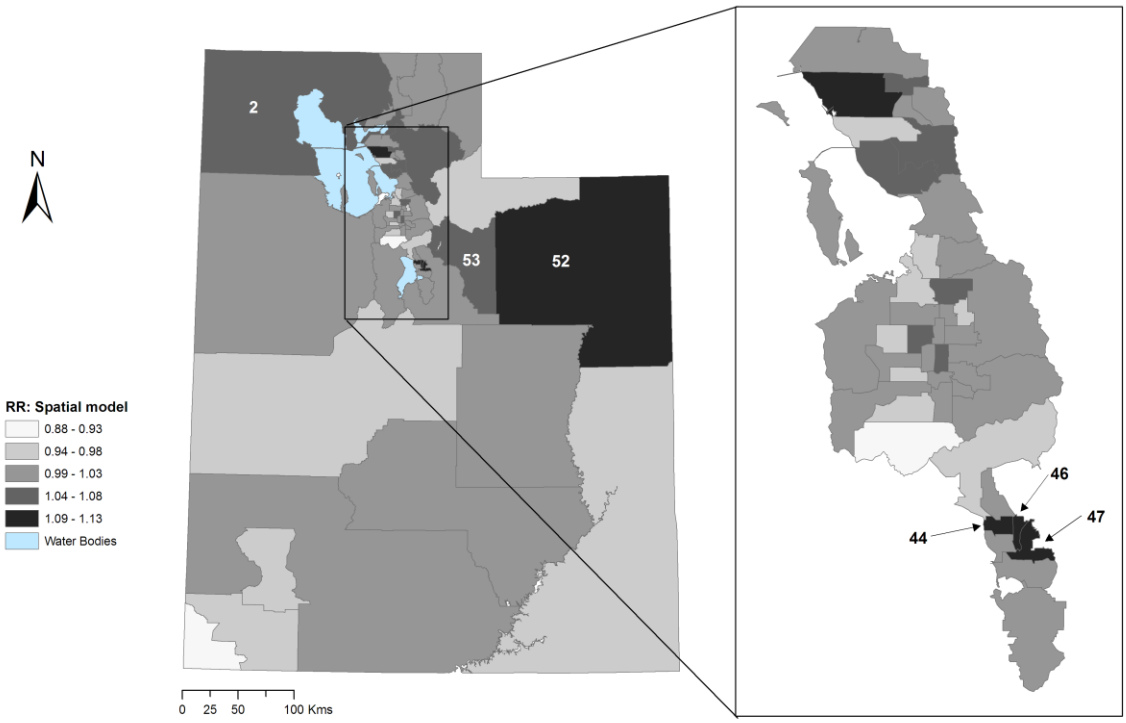


Figure 3-7 Bayesian smoothed relative risks (RR) for oral clefts from 1995 to 2004 across Utah small areas using spatial model (Model 2). The color scheme goes from a light grey (low RR value) to a dark grey (high RR value).

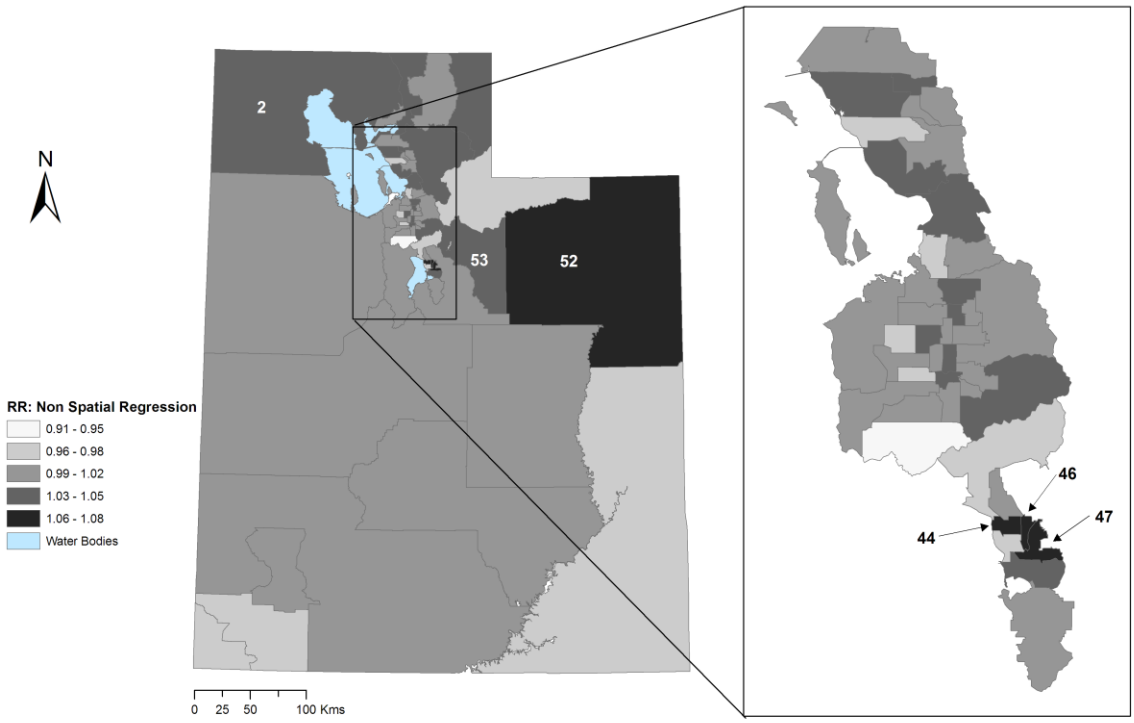


Figure 3-8 Bayesian smoothed relative risks (RR) for oral clefts from 1995 to 2004 across Utah small areas using non-spatial regression model (Model 3). The color scheme goes from a light grey (low RR value) to a dark grey (high RR value).

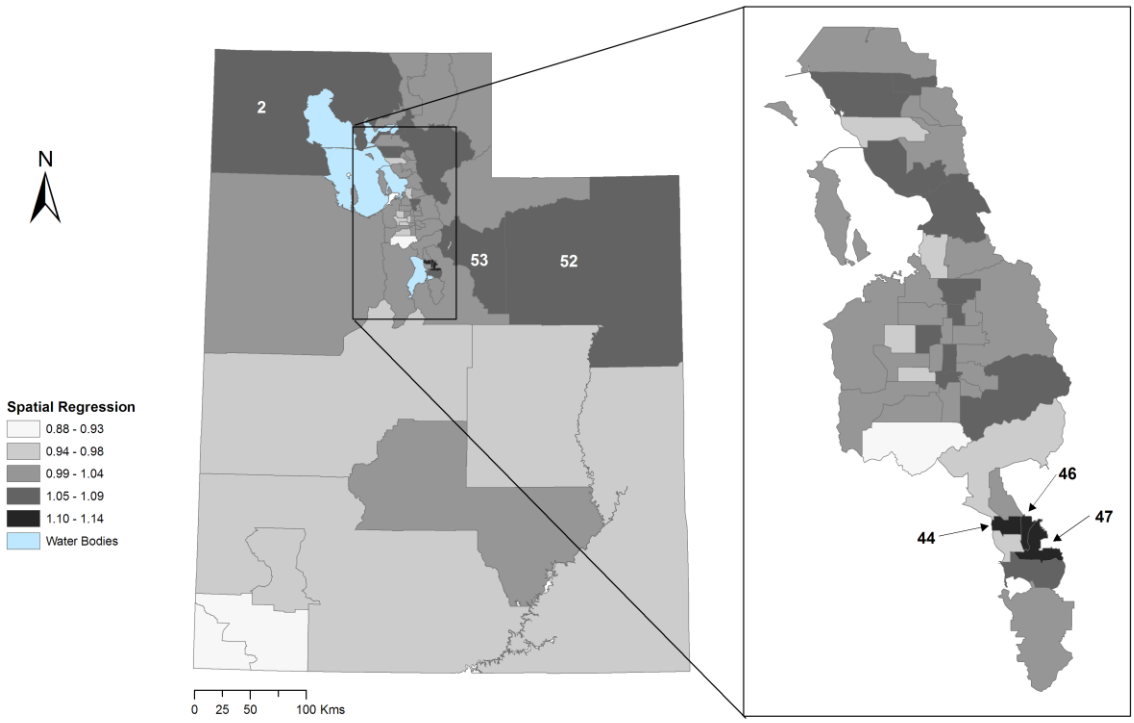


Figure 3-9 Bayesian smoothed relative risks (RR) for oral clefts from 1995 to 2004 across Utah small areas using spatial regression model (Model 4). The color scheme goes from a light grey (low RR value) to a dark grey (high RR value).

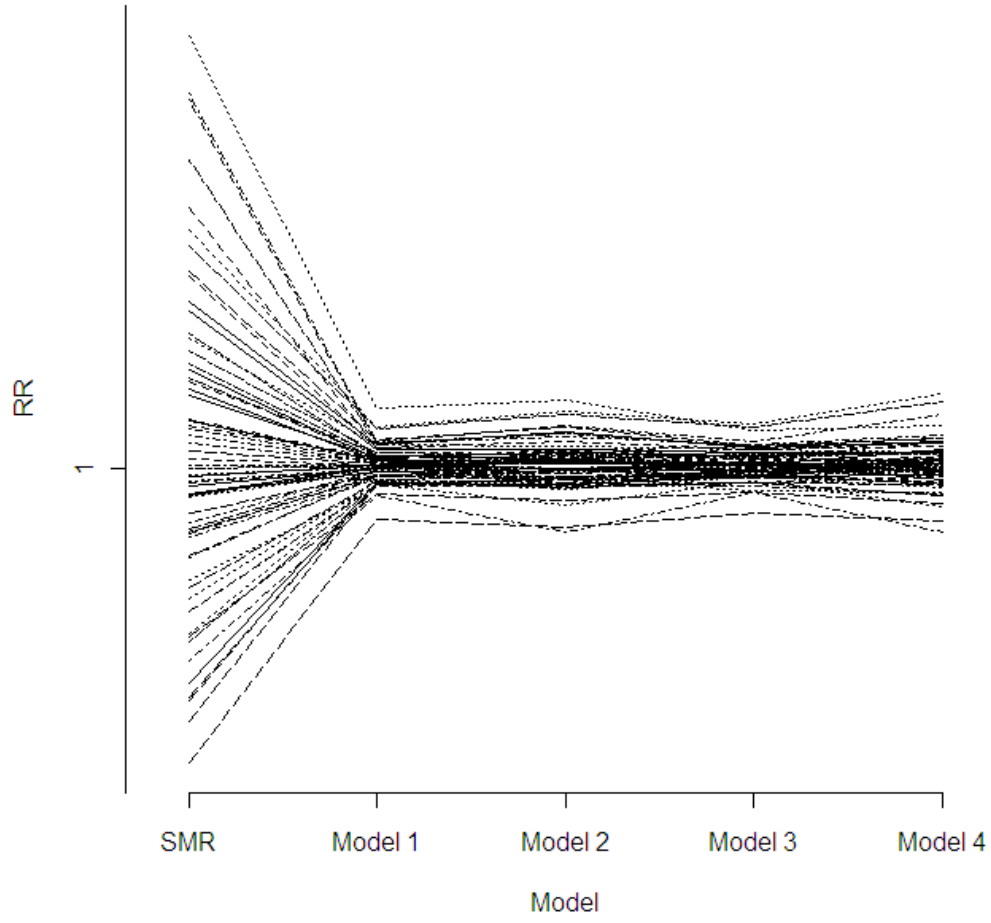


Figure 3-10 Plot showing the degree of smoothing of the raw SMR value using the different Bayesian models – SMR (Raw SMR value), Model 1 (Non-Spatial model), Model 2 (Spatial model), Model 3 (Non-Spatial Regression model), and Model 4 (Spatial Regression model). The Bayesian models shrunk the raw SMR towards to one.

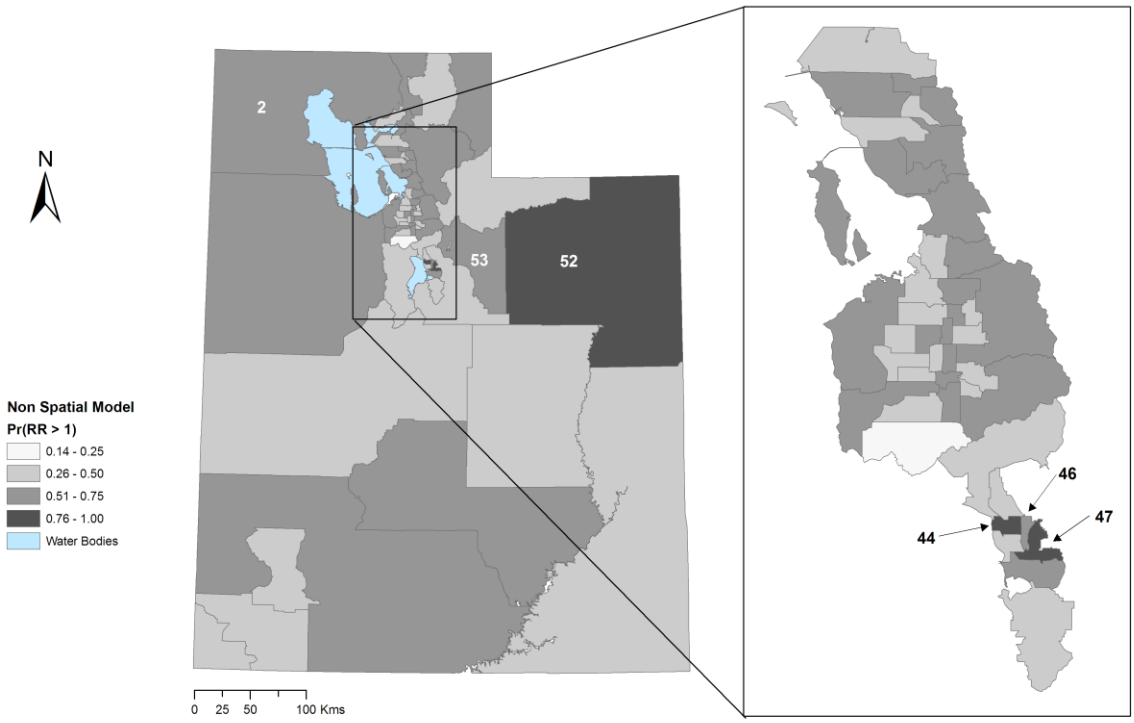


Figure 3-11 Thematic map of exceedance probability of oral clefts risk greater than 1, i.e., Pr (RR > 1) using non-spatial model (model 1). The color scheme goes from a light grey (low probability being the RR value greater than 1) to a dark grey (higher probability the RR value being greater than 1).

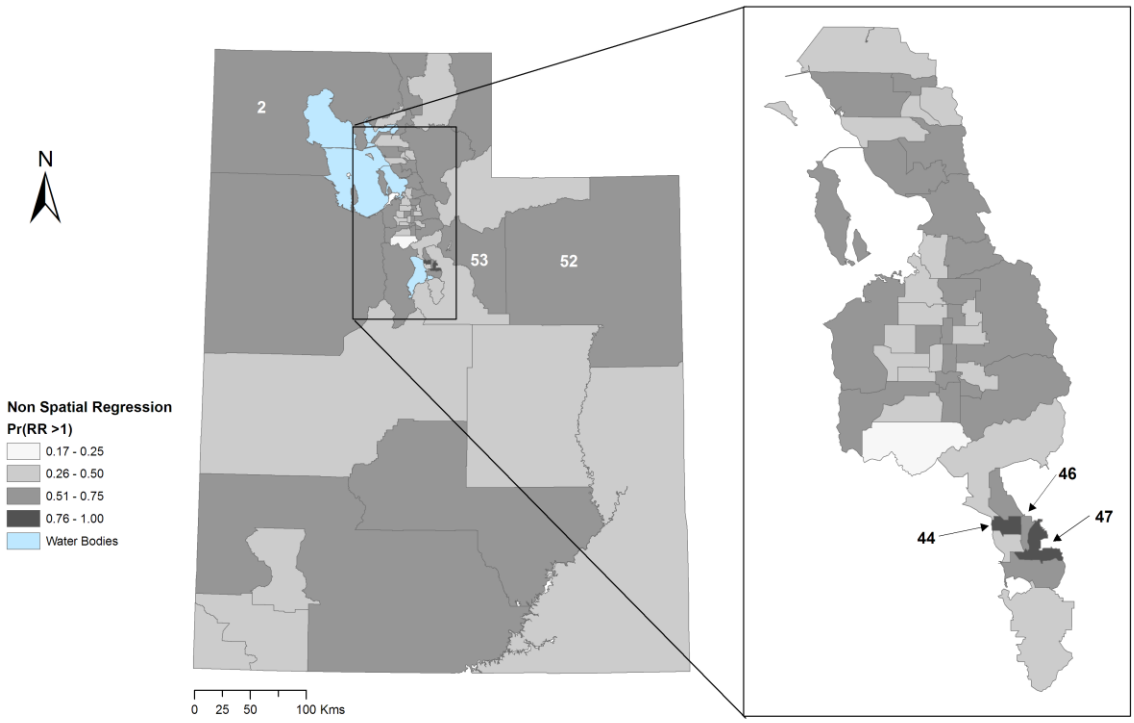


Figure 3-12 Thematic map of exceedance probability of oral clefts risk greater than 1, i.e., Pr (RR > 1) using non-spatial regression model (model 3). The color scheme goes from a light grey (low probability being the RR value greater than 1) to a dark grey (higher probability the RR value being greater than 1).

CHAPTER 4

A MULTI-SCALAR APPROACH TO THE SPATIAL CLUSTER ANALYSIS OF ORAL CLEFTS IN THE STATE OF UTAH FROM 1995 TO 2004: WITH CLUES FOR ETIOLOGY[3]

ABSTRACT

The state of Utah has among the highest prevalence of oral clefts in the United States (US). Yet the reasons are unclear. Spatial cluster analysis was conducted to determine whether or not global clustering was present, to identify clusters of excess oral cleft cases, and to examine the maternal characteristics of the cases involved within the clusters. For the spatial cluster analysis, a total of 894 cases of oral cleft were identified by the state of Utah Birth Defects Network from 1995 to 2004. Tests for global clustering were performed using the Potthoff-Whittinghill (PW), Moran's *I*, and Tango's Maximized Excess Events Test (MEET) statistics. To detect local clusters of excess oral clefts, the Besag-Newell (BN) method and the spatial scan statistic were used. Finally, utilizing a subset of the oral cleft database, a descriptive analysis was carried out in order to reveal characteristics of the cases involved in the identified clusters. No evidence of spatial heterogeneity was found using the PW statistic and there was no evidence of spatial clustering using either Moran's *I* or Tango's MEET. However, the spatial scan statistic

---

[3] *The material for this chapter will be submitted as:* Gebreab SY, Gillies RR, Munger GM and Symanzik J to Environmental Health.

and BN method detected a most likely cluster of high oral cleft cases within the Tri-County Local Health District (LHD) and Wasatch County, with borderline statistical significance. A secondary cluster within Provo/Brigham Young University (BYU), North Orem, and East Orem areas was detected, but was not statistically significant. Our analysis revealed a higher number of cases with the characteristics of maternal smoking, lower education level and family history in the most likely cluster. Conversely, in the secondary cluster, there were modest number of cases with maternal lower education level and family history accompanied with only one case of maternal smoking. The Utah Study, using a variety of spatial techniques, revealed little evidence to support the existence of a single point source of environmental exposure causing oral clefts. However, our study revealed a tendency of excess oral cleft cases in some areas that are related to maternal smoking, lower education level, and family history. These geographic areas may warrant further investigation using epidemiological methods designed to account for the risk factors and covariates revealed in this study.

INTRODUCTION

Oral clefts are one of the most common birth defects in the United States (US) and the State of Utah has among the highest prevalence of oral clefts in the US [1]. The geographical distribution oral clefts birth defect via micromaps is shown in Gebreab *et al*. [2]. Oral clefts include a cleft lip with or without a cleft palate (CL/P), isolated cleft palate (CP) and isolated cleft lip (CL). Oral clefts affect 1 in 450 births in Utah compared to 1 in 750 births nationally [3]. There is, however, marked variation by ethnic groups

[4,5]; for example, oral clefts prevalence rates are considerably higher in American Indian and Native Alaskan populations, intermediate in Caucasians, and lowest in African Americans. In addition, there is a substantial geographic variation in the prevalence of oral clefts related to geographic origin [6,7] and socioeconomic status [8].

The etiology of oral clefts still remains elusive, but it is hypothesized to be caused by a combination or interaction of genetic and environmental factors [9,10]. Oral clefts have been linked to a variety of environmental risk factors, including maternal smoking [7,11,12,13], alcohol consumption [14,15], maternal nutrition [16] as well as instances of medication use [17,18]. Additionally, exposure to chemical solvents in the work place or at home [19,20], to contaminated drinking water [21], environmental lead pollution [22], ambient air pollution [23], or residing near to hazardous waste sites [24,25] have attracted considerable attention as potential risk factors for the development of oral clefts.

Risk factors for oral clefts may be unevenly and unequally distributed geographically, and thus may result in spatial patterning of oral clefts [7]. Therefore, investigation of the spatial patterns of oral clefts using techniques that involve spatial cluster methods can be of great public health importance for identifying areas of elevated risk and for advancing our understanding of the etiology of the birth defect. A spatial "cluster" can be defined as an area with unusual concentration of cases of a birth defect in a small area [26]. Identification of spatial clusters with unusual concentration of oral cleft cases at small area level may uncover potential causes such as environmental exposures, communal behavioral risk factors, and socioeconomic determinants, or perhaps, a shared genetic susceptibility. Furthermore, spatial cluster analysis is relevant for planning and

delivery of health services such as targeting potential intervention programs and resources allocation to the affected children.

Several statistical methods have been developed for spatial cluster analysis of a disease [27-31]. In general, Besag and Newell [32] distinguished the methods into two groups including tests for global clustering and tests for detecting local clusters. The former are used to determine whether there is spatial clustering present throughout the study area, without identifying any specific clusters that may exist. Whereas, the latter are designed to identify local clusters (often referred as hotspots) with no prior knowledge of their number, size, and location and subsequently determine whether they are statistically significant or not [33].

The purpose of this paper is to investigate the spatial patterns of oral clefts in the State of Utah during the period from 1995 to 2004. Despite high oral clefts rates in the State, it is still unclear why Utah exhibits high oral clefts prevalence when compared to most other states in the US. Moreover, there is very little information available in the published literature about the nature and the extent of spatial patterns of oral clefts in the State of Utah at small area level. Therefore, three-step spatial analysis was performed first to assess whether there was a general tendency of global clustering in the oral cleft data, second to identify local clusters ("hotspots") and determine the statistical significance of the clusters. Finally, to examine the characteristics of the cases involved in these clusters in order to gain insight about the etiology of oral clefts.

In the analysis outlined in this paper, we utilize a battery of tests for global clustering and local clusters in tandem to provide a complete description of the different

aspects of spatial patterns that may be present in the oral clefts data. This stems from the fact that first, we have no knowledge *a priori* as to the "true" underlying spatial patterns of oral clefts in the State of Utah, i.e., it is unclear as to the nature and the magnitude of clustering to be tested. Second, each method differs in their definition of clustering test, hence, in their ability to capture different aspects of spatial patterns even though they all similarly assume a null hypothesis of spatial randomness [27,30,34,35]. For example, some measure over-dispersion using goodness-of-fit statistics [36]. Some measure spatial autocorrelation, i.e., the tendency for higher (or lower) values to cluster more closely in space (e.g., [37]). While others scan for potential clusters with elevated incidence of disease by delineating using a circular shape [38] or using an irregular shape [39]. Third, each spatial cluster method has associated strengths and weaknesses, and there is no single best method that captures all the different aspects of spatial patterns with sufficient statistical power [40,41]. Fourth, by using a combination of methods, our goal is also to corroborate the consistency and validity of the results to each other. For example, if the different methods produce similar clusters of size, shape, and location in a logically consistent manner, then our confidence in the results are improved.

Finally, our study is motivated by the recent emphasis on the need for a multi-scalar approach that is epitomized in such studies as the investigation of the spatial patterns of Creutzfeldt-Jakob disease in France [42], breast, lung and colorectal cancer in Long Island, New York [43], brain cancer in the US [44], low birth weight in Shelby County, Tennessee [45], breast cancer in Upper Cape Cod, Massachusetts [46], childhood acute leukemia in France [47] and different cancer data types in Connecticut and the US

[48]. All these studies underscored the need for a multi-scalar approach to the spatial cluster analysis of a disease in order to provide a more complete description of the spatial patterns of a disease.

Given the aforementioned, in the analysis that follows, we explore Utah oral clefts data using a combination of methods that include tests for global clustering and local clustering. We tested global clustering in the Utah oral clefts data by using the Potthoff - Whittinghill (PW) statistic [36], Moran's *I* statistic [37], and Tango's maximized excess event test (MEET) [49,50]. To detect the presence of local clusters ("hotspots") of oral clefts, we employed the Besag-Newell (BN) method [32] and the spatial scan statistic [38].

METHODS

Datasets and Data Sources

The oral cleft data used in this study were obtained from the case-control study of the Utah oral clefts study carried out at the Center for Epidemiologic Studies at Utah State University [51]. A total of 894 cases were identified from 1995 to 2004 by the Utah Birth Defect Network (UBDN), a State–wide birth defects surveillance program that began monitoring oral cleft birth defects in 1995 and added all major birth defects in 1999 [3].

All of the individual cases had street address, or zip code information or both of the mother's residence at the time of birth. We first geocoded the street address to a map coordinate system. The geocoding details are reported in [2]. To protect confidentiality,

the data were subsequently aggregated to "small geographic areas" (referred to as small areas) for analysis. Utah is divided into 61 small areas (Figure 4-1) for the purposes of public health assessment [52]. Data on the live births (population at risk) for each of the 61 small areas for the study period were obtained from the Utah Department of Health (UDOH) [53]. A total of 458,593 live births were identified during the study period. The number of oral cleft cases within any small area ranged from 4 to 36 while the number of live births ranged from 2,998 to 18,177.

In addition, we obtained data on maternal characteristics during pregnancy for 560 (63%) of the cases that participated and completed interviews as part of the Utah oral clefts study. The variables included in our analysis were family history (parental and relative history) of congenital malformations, maternal education, maternal smoking status (active and passive smoking), and maternal alcohol consumption during pregnancy.

Statistical Framework

The study area for the state of Utah was divided into i=1, . . . , m(61) disjoint small areas. Let O_i be the observed number of oral cleft cases within each area i, and let N_i be the corresponding number of live births in area i. The total number of cases (O_+) and live births (N_+) for the entire study area are given by $O_+ = \sum_{i=1}^{m} O_i$ and $N_+ = \sum_{i=1}^{m} N_i$, respectively. The statewide oral clefts rate is expressed as $R = O_+ / N_+$.

Under the null hypothesis of "no clustering" and/or "no cluster," the oral cleft cases are randomly distributed across the small areas and we expect that the number of

cases is proportional to the number of births in each area [30]. So, the expected number of oral cleft cases (E_i) in each area i is computed as the statewide oral clefts rate (R) multiplied by the number of live births in each area i, i.e., $E_i = N_i R$. Given the definitions for O_i and E_i, the standardized morbidity ratio (SMR) is computed as the ratio of the observed to the expected number of oral cleft cases, i.e., $SMR_i = O_i / E_i$ for each area i.

We want to mention here that we did not adjust the expected values for the age of the mother and the race of the infant which are commonly done. This is because the data on the age of the mother and the race of the infant for some of the cases and all live births were not available. Hence, it was impossible to adjust for these confounding factors. However, we surmised that this does not limit the validity of our analyses, because age and race distribution of the cases tend to closely follow the age and race distribution of the live births at risk. Moreover, with over 90% of the cases and live births being Caucasian and Hispanic and the rates for the Caucasian and Hispanic being similar, any adjustment for age and race would have a negligible effect on the expected values.

Test for Global Clustering

The presence of global clustering of oral clefts in the State of Utah was tested by using three methods. The PW statistic [36] to assess the spatial heterogeneity, Moran's I statistic [37] to measure the spatial autocorrelation, and Tango's MEET statistic [49, 50] to establish evidence of overall global clustering in the oral clefts data. The general procedure in the application of these tests was essentially the same, i.e., we tested the null

hypothesis of "no clustering" against an alternative hypothesis of "there is clustering" and summarized the evidence of clustering across the study area using a single *p*-value [33]. For all three methods, the *p*-values of the clustering were obtained using Monte Carlo simulations i.e., by comparing the observed statistics from the real oral clefts dataset to the test statistics generated from 999 random replicas of the dataset under the null hypothesis. The analyses for PW and Moran's *I* were implemented in the R statistical software [54]. Tango's MEET was implemented using S-plus code obtained from Dr. Toshiro Tango which is publicly available at http://www.niph.go.jp/soshiki/gijutsu/download/meet/index.html. A summary of each method follows.

1. ***Potthoff - Whittinghill's Statistic:*** The PW statistic [36] is the uniformly most powerful (UMP) test of random pattern against spatial heterogeneity. Under the null hypothesis of no spatial heterogeneity, the variance of the observed number of oral cleft cases would equal the expected number of oral cleft cases. The PW alternative hypothesis is that the ratio of the variance to the expected is greater than one. If the ratio is greater than one, then there is over-dispersion relative to the Poisson distribution and relatively large numbers of cases would arise in some areas, i.e., more than predicted under the Poisson distribution. The PW statistic is defined as [29].

$$PW = E_+ \sum_{i=1}^{m} \frac{O_i(O_i - 1)}{E_i}, \qquad (1)$$

where O_i and E_i are used as defined above, and $E_+ = \sum_{i=1}^{m} E_i$.

2. ***Moran's I Statistic:*** Moran's *I* statistic is a global measure of spatial autocorrelation that quantifies the tendency of high (or low) oral clefts risk areas to be clustered together [37]. Moran's *I* values range from -1 to 1. A zero Moran's *I* indicates absence of spatial autocorrelation (null hypothesis of no clustering), a positive Moran's *I* indicates positive autocorrelation, i.e., similar values of oral clefts risks tend to cluster together, whereas a negative Moran's *I* indicates negative spatial autocorrelation, i.e., high values of oral clefts risk tend to be located next to low values of oral clefts risk. Moran's *I* statistic is calculated as

$$I = \frac{n\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(Z_i - \overline{Z})(Z_j - \overline{Z})}{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}\sum_{i=1}^{m}(Z_i - \overline{Z})^2} \qquad (2)$$

where, $Z_i = SMR_i$, $\overline{Z} = 1/m \sum_{i=1}^{m} Z_i$, and w_{ij} is the m x m weight matrix defining the "closeness" between area i and its neighbor j. There are various ways of defining weights. The most commonly used are adjacency and distance-based measures of weight. The adjacency weight is defined as

$$w_{ij=}\begin{cases} 1 & \text{if areas } i \text{ and } j \text{ are adjacent} \\ 0 & \text{if areas } i \text{ and } j \text{ are not adjacent or if } i = j \end{cases}$$

The distance weight is defined as

$$w_{ij=}\begin{cases} 1 & \text{if } d_{ij} \le d \text{ for some fixed distance } d \\ 0 & \text{if } d_{ij} > d \text{ or if } d_{ii} = 0 \end{cases}$$

The parameter d_{ij} is the Euclidean distance in kilometers between the centroids of areas *i* and *j*, and *d* is a user specified spatial autocorrelation scale in kilometers. In practice, we have no *a priori* knowledge of the scale of spatial autocorrelation, thus we specify different values of *d*. For the purposes of this study, we used both adjacency and distance-based weights to measure spatial autocorrelation in the oral clefts data. For the distance-based weights, we set several successive values of *d* = 5, 10, 15, 20, 25, 30, 35, 40, 50 *km.*

A limitation of Moran's *I* statistic is that it requires a constant variance assumption. It is difficult to meet this assumption for our data because of varying population sizes across Utah's small areas. Therefore, we also considered the empirical Bayesian index (EBI) proposed by Assunção and Reis [55]. The EBI is a population-based adjusted Moran's *I*, which is robust in detecting spatial autocorrelation in the presence of population heterogeneity.

3. ***Tango's Statistic****:* The Tango statistic is used to establish evidence of overall clustering in the oral clefts data. Tango [49] first proposed the excess events test statistic (EET), which is a weighted sum of the excess number of cases (observed minus expected) in area *i* times the excess number of events in area *j*, then weights the difference by a measure of the distance between the areas, with a higher weighting given when the two area are close. For a given parameter λ, the statistic is defined as

$$EET(\lambda) = \sum_{i=1}^{m}\sum_{j=1}^{m} e^{-4\left(d_{ij}/\lambda\right)^2}(O_i - E_i)(O_j - E_j) \qquad (3)$$

where, d_{ij} is used as defined above, and λ is a clustering scale parameter in kilometers chosen by the user. A large λ is sensitive to large clusters, while a small λ is sensitive towards small clusters. Since we have no *a priori* knowledge about the clustering scale parameter, in practice we evaluate the method with a range of λ values. However, this creates multiple testing problems. To overcome this problem and to be able to detect clustering irrespective of its clustering scale, Tango [50] proposed the Maximized Excess Events Test (MEET).

$$MEET = \min_{0 \le \lambda \le U} P\{EET(\lambda) > eet(\lambda) \mid H_0, \lambda\} \quad (5)$$

where, *eet* (λ) is the observed value of the *EET* statistic as a function of λ, and *U* is an upper limit on λ specified by the user; this usually varies continuously from a small value near zero upwards until λ reaches about half of the size of the whole study area. $P(\min)$ is the minimum of the profile of *p*-value of EET for λ.

In this study, we had no *a priori* knowledge on the spatial clustering scale of oral clefts data, so we set several values of λ = 5, 10, 15, 20, 25, 30, 35, 40, 50 *km*. We obtained *p*-values for each λ value and adjusted *p*-value over all λ values.

Tests for Detection of Local Clusters

To detect the presence of any local clusters (hotspots) of excess oral cleft cases and to evaluate their statistical significance, we employed the Besag and Newell [32] method and the spatial scan statistic [38]. Both methods tested a null hypothesis of "no cluster" against an alternative hypothesis that "there is at least one cluster" and computed *p*-values associated with each candidate cluster. As before, the *p*-values of candidate

clusters were obtained using Monte Carlo simulations i.e., by comparing the observed statistics from the real oral clefts dataset to the test statistics generated from 999 random replicas of the dataset under the null hypothesis. Analysis for the BN method was implemented in the R statistical software package [54] and the spatial scan statistic was implemented using the SaTScan software [56]. A brief description of these two methods follows.

1. ***The Besag and Newell Method:*** The BN method was originally developed to improve the limitations of Openshaw's Geographical Analysis Method (GAM) [57]. The BN method was first used in the detection of childhood leukemia clusters in northern England [32]. The basis of this method is relatively simple. First a user chooses a parameter *k*, a number of oral cleft cases representing the size of a cluster to be detected. For each area *i*, the remaining areas are ordered according to their increased distance from the centroid of area *i*. Then, circles are drawn centered on each area *i* to include the minimum number of neighboring areas needed to accumulate at least *k* cases, i.e., the radius of the circle increases until it contains *k* or more cases. If it contains *k* or more cases, then the procedure stops. The final step then involves the calculation of associated statistics that encompass the last circle formed around the centroid of area *i.*

Let $O_{j(i)}$ be the observed number of cases in area *i* and its *j* closest neighbors and let $N_{j(i)}$ be the observed number of live births in area *i* and its *j* closest neighbors. Let M_i be the random variable containing the minimum number of nearest areas around area *i* that are needed to accumulate at least *k* cases, where m_i is an observed value of M_i, i.e., $m_i = \min\{j : (O_{j(i)} + 1) \geq k\}$. The significance level of the test is obtained by using a

Poisson distribution under the null hypothesis that there is no cluster at the centroid of area *i*. The statistics is defined as

$$P(M_i \le m_i \mid k) = 1 - P(M_i > m_i \mid k) = 1 - \sum_{j=0}^{k-1} \frac{(N_{j(i)}R)^j}{j!} \exp(-N_{j(i)}R) \qquad (6)$$

A critical issue in the BN method is the choice of *k*, the cluster size. The choice of *k* is arbitrary, especially since we have no prior knowledge on cluster sizes, if present, in the oral clefts data. Besag and Newell [32] recommend repeating the test using several values of *k*, and mapping all the clusters that attain statistically significant (at the 5% level) for different values of *k*. For our study, we evaluated the BN method for several values of cluster sizes (k = 22 to 38). We report clusters as "consistent" if their statistical significance persisted over three values of *k* and "less clear" if their statistical significance appeared only on fewer than three *k* values as was suggested by Newell and Besag [58].

2. ***Spatial Scan Statistic****:* The spatial scan statistic tests for clusters of any size and at any location without a pre-selection bias, by using circular windows with a continuously variable radius [38]. The method accounts for multiple testing and inhomogeneous population density. It was first applied to leukemia disease in Sweden [59] and breast cancer in the northeastern United States [60]. This method imposes a circular scanning window on the map and lets the center of the circle move over the study area so that at different positions the window includes different sets of neighboring areas. An area is included if its centroid lies within the circle. For each circle centroid, the radius varies continuously from zero to a user-defined maximum population size (usually

set not to exceed 50% of the underlying population). In this way, the circular window is flexible both in location and in size. In total, the method creates an infinite number of distinct circular windows, each with a different set of neighboring areas within it, and each is a potential cluster that may consist of a single area or a large number of neighboring small areas. For each circle, a likelihood ratio is computed for the alternative hypothesis that there is a higher rate of oral clefts than expected inside the circle against the null hypothesis that the oral clefts rates inside and outside the circle are the same. Let $L_{i(j)}$ be the likelihood under the alternative hypothesis that there is a cluster in area *i* and its *j* closest neighbors, and let L_0 be the likelihood under the null hypothesis. The likelihood ratio statistic is defined as

$$\frac{L_{j(i)}}{L_0} = \left(\frac{O_{j(i)}}{N_{j(i)}R} \right)^{O_{j(i)}} \left(\frac{O_+ - O_{j(i)}}{O_+ - N_{j(i)}R} \right)^{O_+ - O_{j(i)}} \tag{7}$$

The test statistics is

$$LR = \max_{i,j} \frac{L_{j(i)}}{L_0} I(O_{j(i)} > N_{j(i)}R) \tag{8}$$

where *LR* is the maximum likelihood ratio, and *I( )* is the indicator function, that is equal to '1' when the observed number of oral cleft cases inside the circle is more than expected, i.e., $O_{j(i)} > N_{j(i)}R$ otherwise it is equal to '0'. The circle with the maximum likelihood ratio among all radius sizes at all possible locations is considered as the most likely cluster (also known as the *primary cluster*) and is followed by other non-overlapping secondary clusters according to the order of the likelihood ratios.

The spatial scan statistic requires specifying the maximum population size; however, there are no clear guidelines for the selection of the appropriate maximum population size. For our study, we set different values of maximum population size at 10%, 25%, and 50 % of the total population.

Maternal Characteristics of Local Clusters

The maternal characteristics of the local clusters identified by the spatial scan statistic were examined using a descriptive analysis. Summary statistics (number and percentage) describing some of the maternal characteristics of the cases involved in the local clusters were constructed. The descriptive analysis was restricted only to those cases that had detailed maternal characteristics information. The maternal characteristics included in this study were the number and percentage of mothers with lower education level (without some college level), with family history (if at least one of the parents or blood- relatives had congenital malformation), with active and passive maternal smoking history, and with a history of maternal alcohol consumption during pregnancy.

RESULTS

Figure 4-2 shows a map of SMR and indicates the distribution of oral clefts in the State of Utah. The SMRs range from 0.43 to 1.83 around an overall mean of 1.03 and standard deviation of 0.31. A visual inspection of Figure 2 reveals that high anomalies of oral cleft cases are located in the following areas, Tri-County Local Health Department (LHD) (53), East Orem (46), Wasatch County (52), and Box Elder County (2) - all are

shaded as dark grey. However, only the SMR for Tri-County LHD was statistically significant at the 5% level. On the whole, the SMR map shows no evidence of an apparent spatial pattern.

Table 4-1 shows the results from the rest of the formal tests performed for clustering. The test for heterogeneity using the PW statistic resulted in a *p*-value of 0.237 (ref., Figure 4-3), indicating no evidence of spatial heterogeneity in the risk of oral clefts across the study area. Moran's *I* statistic was evaluated using distance and adjacency-based weights. A statistically significant spatial autocorrelation was evidenced at $d = 5$ *km*. However, no evidence of spatial autocorrelation for $d = 10$ *km* or greater was observed. Moran's *I* statistic based on adjacency weight showed no evidence of statistically significant spatial autocorrelation, with Moran's *I* value = -0.147 and *p*-value=0.956. Similarly, the result from the EBI (population-adjusted Moran's *I* statistic) showed no evidence of spatial autocorrelation (Moran's $I = -0.115$, *p*-value= 0.905) in the oral clefts data. The results for Tango's EET and MEET, as summarized in Table 4-1 are also visualized in Figure 4-4 both for the unadjusted *p*-value (EET) for different values of λ and the adjusted *p*-value (MEET) for overall clustering. As seen in Table 4-1 and Figure 4-4, Tango's EET indicates no evidence of global clustering in the range of 5 to 50 *km*, with an overall MEET adjusted *p*-value of 0.229.

Table 4-2 shows the summary information for the clusters detected by both the BN and the spatial scan statistic. First we applied BN with differing cluster sizes (k) ranging from 22 to 38. This method detected three possible clusters as identified in Table 4-2 and shown in Figure 4-5. The first cluster was detected at $k = 22$ and was centered at

the Tri-County LHD (53) area. This cluster was later expanded to include Wasatch County (52) over three values of k =30, 32, 36 (Figure 4-5), with p-values <0.05 as revealed in Table 4-2. According to the definition of Newell and Besag [55], this cluster is classified as "consistent" since it was statistically significant (at the 5% of level) over three values of k. Of note in Table 4-2, this cluster had a relative risk of 1.799 and contained 37 observed cases compared to 20.57 expected. The method also detected two "less clear" clusters that were statistically significant (p-values <0.05), but only at two k values (k =36 and 38). One of the clusters was centered at Provo/Brigham Young University (BYU) (47) with a radius of 5 *km* and included East Orem (46). This cluster had a relative risk of 1.523 and contained 39 observed cases compared to 25.60 expected. The third cluster (k=36, 38) was centered at East Orem (46) with a radius of 4 *km* and included North Orem (44). This cluster had a relative risk of 1.498 and contained 40 observed cases compared to 26.70 expected.

The spatial scan statistic detected two clusters of excess oral cleft cases, although not strictly significant at the 5 % threshold at all maximum population sizes of ≤ 10%, ≤ 25% and ≤ 50%. Figure 4-6 displays the primary cluster (red circle, covering areas 52 & 53) and a secondary cluster that was detected (green circle, covering areas 44, 46, & 47). Summary information is provided in Table 4-2. The primary cluster, with the largest likelihood ratio, comprised the Tri-County LHD (53) and Wasatch County (52) areas with a radius of 111 *km*. This cluster contained 37 observed oral cleft cases compared to 20.57 expected and had the largest relative risk of 1.833. This cluster was borderline significant (p=0.063, p=0.089, and p=0.100) at population sizes ≤ 10%, ≤ 25% and ≤ 50%,

respectively. This method also detected a non-overlapping secondary cluster comprising North Orem (44), East Orem (46) and Provo/BYU (47) with a radius of 4.6 *km*. This cluster contained 68 observed cases compared to 45.91 expected cases and had a relative risk of 1.521, but it was not statistically significant (*p*-values were in the range of 0.111 to 0.165) within their different maximum population sizes.

We examined the maternal characteristics of the oral cleft cases involved in the clusters detected by the spatial scan statistic. Some of these maternal characteristics are summarized in Table 4-3. In the primary cluster, Tri-County LHD (53) had twenty seven cases observed during the study period, detailed information on maternal characteristics were available for only fifteen cases. Among these, twelve (80%) cases had maternal lower education level and five (33.33%) had an active maternal smoking history during pregnancy. Furthermore, out of the fifteen cases, six (40%) cases had a family history of congenital malformation in this area as shown in Table 4-3. Wasatch County (52) had ten observed cases during the study period, but only three cases had information on maternal characteristics. Of these, one case had a family history of congenital malformation, but none of the cases had maternal lower education level, maternal smoking or alcohol consumption history during pregnancy. In the secondary cluster, which comprises the North Orem (44), East Orem (46), and Provo/BYU (47) areas, the total observed number of cases during the study period was sixty eight cases, but detailed information on maternal characteristics were only available for forty nine cases. Among these, the number of cases with lower maternal education level was thirteen cases (26.5%) and

fourteen cases (28.6%) had a family history of congenital malformation, but only one case had a history of maternal smoking which belonged to North Orem (46).

DISCUSSION

We have presented a statewide spatial cluster analysis of oral clefts in the State of Utah at the small area level. Our objectives were three-fold: a) to investigate whether clustering of oral cleft cases was present anywhere in Utah, b) to detect specific clusters that manifest a significant excess of oral cleft cases, and c) to examine the maternal characteristics of the oral cleft cases involved in the clusters identified under item b).

Our study used a multi-scalar approach to capture the different types of spatial patterns present in the Utah oral clefts data. First we preformed test for spatial heterogeneity using the PW statistic. This provided no evidence of spatial heterogeneity in oral clefts risk across the small areas of the State of Utah. However, this method does not provide information about the spatial pattern of the deviations i.e., whether areas of high (or low) deviations are spatially correlated or widely separated from each other. Thus, we used Moran's *I* statistic with the goal of measuring spatial autocorrelation of oral clefts rates between small areas.

The results obtained from Moran's *I* statistic showed no evidence of spatial autocorrelation in the oral clefts data despite multiple testing at different scales of distance values (ref., Table 4-1). One of the criticisms of Moran's *I* statistic is that it assumes constant population size, which was difficult to meet for our data. To overcome this problem and further confirm the results, we evaluated EBI, which is effective in

accounting for differing population sizes in the presence of population heterogeneity. The result from EBI was still statistically non-significant which likewise confirmed the absence of spatial autocorrelation in the oral clefts data.

Next, we applied Tango's MEET statistic to asses for evidence of overall clustering. This method has several appealing features in comparison to the PW and Moran's *I* statistics. First, it serves as a general purpose test for evidence of general clustering by incorporating aspects of goodness – of – fit tests and spatial autocorrelation [30,61]. Second, it accounts for heterogeneous population sizes and multiple testing problems. Third, MEET has been shown to have a higher statistical power in detecting overall clustering in comparison to Moran's *I* statistic [40,41]. The results obtained by Tango's MEET reflected those findings obtained by the PW and Moran's *I* statistics. Tango's MEET showed no evidence of general clustering within the 5 to 50 *km* range with an overall adjusted p=0.229, which suggests the absence of global clustering in the oral clefts data.

Our finding of lack of global clustering by the three methods suggests a common environmental risk factor is unlikely the plausible cause of oral clefts. In other words, it eliminates the existence of a single and strong point-source of environmental exposure such as air pollution, contaminated water, hazardous waste, factory emissions, or maternal infection causing the high prevalence of oral clefts in the State of Utah. Despite the lack of evidence for global clustering in our study, it is possible to have local clusters with an excess of oral cleft cases [33]. Local clusters, if present, may reflect an aggregation of mothers at high risk of giving a child with oral clefting due to some non-

environmental factor such as socioeconomic factors, demographic composition, and genetic susceptibility.

We tested for any presence of local clusters by applying the BN and spatial scan statistic. Both methods identified two non-overlapping clusters in the same locations of similar sizes (ref., Table 4-2, Figures 5 and 6). However, their main point of departure was in the estimation of statistical significances (*p*-values) as was indicated in Table 4-2. Much of the variation is due the fact that the BN statistic provides *p*-values for each cluster location, but does not adjust for multiple testing at different *k* values, so it is more likely to detect false positive clusters [32]. To circumvent this problem, Newell and Besag [58] suggested classifying clusters into "consistent" or "less clear" on the basis of their statistical significance over three values of *k*. Following their classification, only the cluster found within the Tri-County LHD and Wasatch County was "consistent" since it was statistically significant ($p < 0.05$) over three values of *k*, whereas the cluster within the North Orem, East Orem, and Provo/BYU areas would be defined as "less clear."

An alternative approach to that of BN method is the spatial scan statistic. Unlike the BN method, the spatial scan statistic adjusts the *p*-values for multiple testing inherent in the many potential cluster locations and sizes. The primary cluster of oral clefts was found in the Tri-County LHD and Wasatch County but reported borderline significant ($0.05 < p < 0.1$) at the different maximum population sizes. Given that this cluster was identified by BN method as a "consistent" cluster, we can reasonably conclude that there is a tendency toward excess oral cleft cases in the Tri-County LHD and Wasatch areas. The spatial scan statistic found a secondary cluster in North Orem, East Orem, and

Provo/BYU areas but was not statistically significant (*p*-values were in the range 0.111 to 0.165) regardless of maximum population size. This is not surprising since the BN method identified the same cluster as a "less clear" cluster.

Another important difference between the spatial scan statistic and the BN method in searching for local clusters is the cluster size. The spatial scan statistic does not require pre-specification of cluster size; it searches for a cluster at any location and of any size up to a maximum population size without pre-selection bias [38]. In contrast, the BN method requires specifying the cluster size (*k*) a *priori* and looks specifically for that cluster size. Thus, the size of clusters to be detected is highly dependent on the choice of *k*. If too small, large clusters cannot be detected, and if too large, spurious clusters may be produced [62]. Therefore, the BN method is a good choice when the size and the scale of a cluster are erstwhile known. Another advantage of the spatial scan statistic is that it has higher statistical power for detecting the most likely cluster compared to the BN method [40,41], especially when the most likely cluster is circular in shape.

The spatial scan statistic does, however, have some drawbacks. First, the method uses circular windows to detect clusters, which may inadvertently include surrounding areas with non-elevated risk. Its design is therefore not directly applicable for the detection of non-circular clusters such as clusters that tends to follow linear features like rivers or overhead power lines. To avoid such limitations and to identify non-circular clusters, irregular-shaped [39] or elliptical-shaped [63] scans have been proposed. Second, the spatial scan statistic tends to produce conservative *p*-values for the secondary clusters [38], thus it may underestimate the statistical significance of secondary clusters.

Third, the choice of maximum population size is somewhat arbitrary, and there are no clear guidelines for an appropriate choice of maximum population size. For example, Hjalmars *et al.* [59] used 10% of the total population to define the windows while Kulldorff *et al.* [60] used 50%.

Although the etiology of oral clefts is poorly understood, oral clefts are related to such factors as maternal low socioeconomic status, maternal smoking, nutrition intake, alcohol consumption, and medication use. Investigation of the maternal characteristics of the primary and secondary clusters revealed maternal smoking history, maternal lower education level, and family history of congenital malformations were dominant characteristics in the primary cluster, particularly in the Tri-County LHD area where the percentage of cases with maternal smoking history during pregnancy (33%), maternal lower education level (80%), and family history (40%) were higher compared to the other areas in the study (ref., Table 3). This observation was also consistent with UDOH report that reported that this area had the highest proportion (17%) of pregnant women smoking during pregnancy in comparison to the other areas over the period 1995-2004 [53]. Although a higher smoking history, lower education level, and family history might explain excessive oral cleft cases in this area, racial/ethnicity composition might be also a factor; this is because this area overlaps with American Indian reservations where there is a higher proportion of Native Americans in the population. For the secondary cluster (North Orem, East Orem, and Provo/BYU), we found that the percentage of cases with maternal lower education level and family history was modest and was accompanied by only one case with maternal smoking history (i.e., North Orem). Of the areas in the

secondary cluster, North Orem exhibited a higher number of cases with associated family history as a compounding factor as well as maternal lower education.

Our association study was essentially descriptive because of small sample size and therefore prevents us from drawing a definitive conclusion with regard to this association; however, this study provides further evidence in support of a causal role of maternal smoking during pregnancy, maternal lower education level, and family history in oral clefts. Our findings are consistent with the Utah case-control study of oral clefts by Moss [64] that maternal smoking history, maternal lower education level, and family history were significantly associated with oral clefts. Furthermore, previous studies have showed a strong association between oral clefts and maternal smoking [11,12,13], maternal lower education level [65], and family history [10]. However, further investigations designed to account for the risk factors and covariates revealed in our study would be necessary. Future research should also examine whether there is a particular genetic marker related to oral clefts and explore the etiology of oral clefts in terms of gene-environment interactions.

Limitations and Strengths

This study is not without limitations and, like many that involve statistical analyses, care must be taken when interpreting the results. First, this study did not adjust for confounding variables such as the mother's age and race of the infant because these data were not available to us. It is feasible that the detected clusters might be attributed due to uneven distributions of the race of the infant and/or age of the mother. For

example, it is possible that the excess oral cleft cases in Tri-County LHD might be confounded due to the high proportion of Native Americans in that area. Conversely, lack of adjustment can also have the opposite effect, that other potential clusters might have been obscured from being detected. Obviously, future follow-up studies should include information on the mother's age and race of the infant to account for possible confounding effects in the study of the spatial patterns of oral clefts in the State of Utah.

Second, the total cases of oral clefts used in this study were not the complete record as was compiled by the UBDN. The oral clefts rate reported by UBDN [3] for the period 1995 to 2003 was 22.2 per 10,000 compared to this study's rate of 19.5 per 10,000 for the period from 1995 to 2004. Despite the fact that the difference between the rates is small, such differences may lead to a spurious spatial variation between small areas. Third, as mentioned, the descriptive analysis of the maternal characteristics of the local clusters was restricted to those cases that had maternal information, which limited our ability to generalize the results. Nevertheless, we believe that the subset information was helpful in shedding some light on the etiology of oral clefts.

Fourth, we considered adjusting for the repeated tests required in the sensitivity analysis to parameters (such as spatial scale parameters) but do not believe this necessary for two reasons: (a), evaluating sensitivity to scale parameters was undertaken to identify the underlying spatial scale of the process, and not for the purposes of statistical inference. (b), the spatial clusters that were identified are suggestive in that they have borderline statistical significance at the alpha level of 0.05. Adjustment for multiple tests

would increase the *p*-values but should not alter qualitatively the concordance of results across the multiple methods.

Despite its limitations, this study had several strengths. A major strength of this study was that it demonstrated the significance of spatial cluster analysis in characterizing the spatial patterns of oral clefts at the small area level. The role of spatial statistics techniques and Geographic Information Systems (GIS) in birth defects surveillance is rarely explored. Clearly, this study illustrated how spatial cluster techniques coupled with data summarized at the small area level can be a powerful tool in birth defects surveillance in a preliminary fashion which is arguably cost-effective. This study was useful in providing valuable etiological clues such as behavioral, genetic and environmental causes of oral clefts and in identifying high-risk populations and locations that could be utilized for further epidemiological studies as well as for health service planning and delivery.

A second strength of this study was that a multi-scalar approach was used to better understand the different aspects of the spatial patterns present in the oral clefts. By using different combinations of spatial cluster methods, we not only provided a complete picture of the spatial patterns of oral clefts in the State of Utah, but also highlighted the strengths and weaknesses of each method as well as corroborating the consistency and validity of the results produced by the different methods. Finally, the third strength of this study was that we had the opportunity to examine the maternal characteristics of the local clusters based on a subset of oral cleft cases available to us, which confirmed the

role of established risk factors including maternal smoking, maternal education level, and family history in causing oral cleft birth defects.

CONCLUSIONS

In conclusion, our analysis did not reveal any evidence of global clustering of oral clefts in the State of Utah suggesting little evidence to support the existence of a single and strong point-source of environmental exposure that might cause oral clefts. Alternatively, the most likely explanation for the high rate of oral cleft cases in the State of Utah is attributed to demographic characteristics, maternal behavioral factors and family history of the population at risk. In particular, the high oral cleft cases found within Tri-County LHD support the role of established risk factors including maternal smoking during pregnancy, maternal lower education level, and family history in oral clefts etiology. Although epidemiologic studies such as case-control or cohort studies would be needed to draw more firm conclusions on the causal factors of oral clefts, a study such as this has demonstrated the usefulness of spatial cluster analysis in generating etiological hypotheses and identifying local clusters of excess oral cleft cases for further epidemiological studies, and for health service planning and delivery.

REFERENCES

[1] National Birth Defects Prevention Network (NBDPN): **Birth defects surveillance data from selected states,** 1998-2002. *Birth Defects Research (Part A): Clinical and Molecular Teratology* 2005, **73:**758-853.

[2] Gebreab SY, Gillies RR, Munger RG, Symanzik J: **Visualization and interpretation of birth defects data using linked micromap plots**. *Birth Defects Research* Part A: *Clinical and Molecular Teratology* 2008, **82:**110-119.

[3] Utah Birth Defect Network (UBDN): **Orofacial Clefts at a glance**. [http://health.utah.gov/birthdefect/defects/orofacial.html]. August 2007.

[4] Croen LA, Shaw GM, Wasserman CR, Tolarova MM: **Racial and ethnic variations in the prevalence of orofacial clefts in California, 1983-1992.** *American Journal of Genetics* 1998, **79:**42-47.

[5] Tolarova MM, Cervenka J: **Classification and birth prevalence of orofacial clefts**. *American Journal of Medical Genetics* 1998, **75:**126-137.

[6] Vanderas AP: **Incidence of cleft lip, cleft palate, and cleft lip and palate among races: a review.** *Cleft Palate Journal* 1987, **24:**216-225.

[7] Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC: **Cleft lip and palate**. *Lancet* 2009, **374:**1773-85.

[8] Murray JC, Daack-Hirsch S, Buetow KH, Munger R, Espina L, Paglianawan N, Villanueva E, Rary J, Magee K, Magee W: **Clinical and epidemiologic studies of cleft lip and palate in the Philippines**. *Cleft Palate Craniofacial Journal* 1997, **34:**7-10.

[9] Shaw GM, Wasserman CR, Lammer EJ, O'Malley CD, Murray JC, Basart AM, Tolarova MM: **Orofacial clefts, parental cigarette smoking, and transforming growth factor-alpha gene variants**. *American Journal of Medical Genetics* 1996, **58:**551-561.

[10] Murray JC: **Gene/environment causes of cleft lip and/or palate**: *Clinical Genetics 2002*, **61:**248-256.

[11] Khoury MJ, Weinstein A, Panny S, Holtzman NA, Lindsay PK, Farrel K, Eisenberg M: **Maternal cigarette smoking and oral clefts: a population-based study**. *American Journal of Public Health* 1987, 77:623–625

[12] Lieff S, Olshan AF, Werler M, Strauss RP, Smith J, Mitchell A: **Maternal cigarette smoking during pregnancy and risk of oral clefts in newborns.** *American Journal of Epidemiology* 1999, **150:**683-694.

[13] Chung KC, Kowalski CP, Kim HM and Buchman SR: **Maternal cigarette smoking during pregnancy and the risk of having a child with cleft lip/palate.** *Plastic and Reconstructive Surgery* 2000, **105:**485-491.

[14] Munger R, Romitti P, Daack-Hirsch S, Burns T, Murray J, Hanson J: **Maternal alcohol use and risk of orofacial cleft birth defects**. *Teratology* 1996, **54:**27-33.

[15] Lorente C, Cordier S, Goujard J, Ayme S, Bianchi F, Calzolari E, De Walle HE, Knill-Jones R: **Tobacco and alcohol use during pregnancy and risk of oral clefts.** *American Journal of Public Health* 2000, **90:**415-419.

[16] Munger R: **Maternal nutrition and oral clefts. In *Cleft Lip and Palate: From Origin to Treatment*.** Edited by Wyzsynski D. New York: Oxford University Press; 2002:170-192.

[17] Dansky LV and Finnell RH**: Parental epilepsy, anticonvulsivant drugs, and reproductive outcome: Epidemiologic and experimental findings spanning three decades; 2: Human studies.** *Reproductive Toxicology* 1991, **5:**301-335.

[18] Parkwyllie L, Mazotta P, Pastuszak A, Moretti ME, Beique L, Hunnisett L, Friesen HH, Jacobson S, Kasapinovic S, Chang D, Diavcitrin O, Chitayat D, Nulman I, Einarson TR, Koren G: **Birth defects after maternal exposure to corticosteroids: prospective cohort study and meta-analysis of epidemiological studies.** *Teratology* 2000, **62:**385–392.

[19] Laumon B, Martin JL, Collet P: **Exposure to organic solvents during pregnancy and oral clefts: A case-control study.** *Reproductive Toxicology* 1996, **10:**15-19.

[20] Garcia AM and Fletcher T: **Maternal occupation in the leather industry and selected congenital malformations.** *Occupational and Environmental Medicine* 1998, **55:**284-286.

[21] Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficy EM, Savrin JE: **Public drinking water contamination and birth outcomes.** *American Journal of Epidemiology* 1995, **141**:850-62.

[22] Vinceti M, Rovesti S, Bergomi M, Calzolari E, Candela S, Campagna A, MilanM, Vivoli G: **Risk of birth defects in a population exposed to environmental lead pollution**. *Science of the Total Environment* 2001, **278:**23-30.

[23] Ritz B, Yu F, Fruin S, Chapa G, Shaw GM, Harris JA: **Ambient air pollution and risk of birth defects in southern California** *American Journal of Epidemiology* 2002, **155:**17–25.

[24] Brender J.D., Zhan F.B., Suarez L., Langlois P.H., Moody K: **Maternal residential proximity to waste sites and industrial facilities and oral clefts in**

offspring. *Journal of Occupational and Environmental Medicine 2006*, **48:**565-572.

[25] Dolk H, Vrijheid M, Armstrong B, Abramsky L, Bianchi F, Garne E, Nelen V, Robert E, Scott JE, Stone D, Tenconi R: **Risk of congenital anomalies near hazardous-waste landfill sites in Europe: the EUROHAZCON study.** *Lancet* 1998, **352:**423-427.

[26] Dolk H: **The role of the assessment of spatial variation and clustering in environmental surveillance of birth defects.** *European Journal of Epidemiology* 1999, **15:** 839-845.

[27] Kulldorff M: **Statistical methods for spatial epidemiology: tests for randomness.** In *GIS and Health*. Edited by Loytonen M, Gatrell A. London: Taylor & Francis; 1998:49-62.

[28] Lawson AB, Kulldorff M: **A review of cluster detection methods.** In *Disease mapping and risk assessment for public health decision-making*. Edited by Lawson AB, Biggeri A, Bohning D, Lesaffre E, Veil J, Bertollini R. London: Wiley; 1999:99-110.

[29] Wakefield JC, Kelsall JE, Morris SE: **Clustering, cluster detection, and spatial variation in risk.** In *Spatial epidemiology: Methods and Applications*. Edited by Elliot P, Wakefield JC, Best NG, Briggs DJ. New York: Oxford University Press; 2000:128-152.

[30] Waller LA, Gotway CA: *Applied Spatial Statistics for Public Health Data*. New York: John Wiley and Sons; 2004.

[31] Gómez-Rubio V, Ferrándiz-Ferragud J, López-Quilez A: **Detecting clusters of disease with R.** *Journal of Geographic Systems* 2005, **7:**189-206.

[32] Besag J, Newell J: **The detection of clusters in rare diseases.** *Journal of the Royal Statistical Society, Series A* 1991, **154:**143-155.

[33] Waller LA, Hill EG, Rudd RA: **The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations.** *Statistics in Medicine* 2006, **25:**853-865.

[34] Waller LA, Jacquez GM: **Disease models implicit in statistical tests of disease clustering.** *Epidemiology* 1995, **6:**584-590.

[35] Jacquez, GM: **Cluster Morphology Analysis.** *Journal of Spatial and Spatio-Temporal Epidemiology* 2009, **1:**19-29.

[36] Potthoff RF,Whittinghill M: **Testing for homogeneity in the Poisson distribution.** *Biometrika* 1966, **53:**183–190.

[37] Moran PAP: **Notes on continuous stochastic phenomena.** *Biometrika* 1950, **37:**17-23.

[38] Kulldorff M: **A spatial scan statistic.** *Communication Statistics Theory Methods* 1997, **26:**1481-1496.

[39] Duczmal L, Assuncao R: **A simulated annealing strategy for the detection of arbitrary shaped spatial clusters.** *Computational Statistics and Data Analysis* 2004, **45:**269-286.

[40] Kulldorff M, Tango T, Park PJ: **Power comparisons for disease clustering tests**. *Computational Statistics and Data Analysis* 2003, **42:**665–684.

[41] Song C, Kulldorff M: **Power evaluation of disease clustering tests.** *International Journal of Health Geographics* 2003, **2:**9.

[42] Huillard d'Aignaux J, Cousens SN, Delasnerie-Lauprête N, Brandel JP, Salomon D, Laplanche JL, Hauw JJ, Alpêrovitch A: **Analysis of the geographical distribution of sporadic Creutzfeldt-Jakob disease in France between 1992 and 1998.** *International Journal of Epidemiology* 2002, **31:**490-495.

[43] Jacquez GM, Greiling DA: **Local clustering in breast, lung and colorectal cancer in Long Island, New York.** *International Journal Health Geographics*, 2003, 2:3

[44] Fang Z, Kulldorff M, Gregorio DI: **Brain Cancer in the United States, 1986–95: A geographic analysis.** *Neuro-Oncology* 2004, **6:**78-82.

[45] Ozdenerol E, Williams BL, Kang SY, Magsumbol MS: **Comparison of spatial scan statistic and spatial filtering in estimating low birth weight clusters.** *International Journal of Health Geographics* 2005, **4:**19.

[46] Ozonoff A, Bonetti M, Forsberg L, Pagano M: **Power comparisons for an improved disease clustering test.** *Computational Statistics and Data Analysis.* 2005, **48:**679–684.

[47] Bellec S, Hemon D, Rudant J, Goubin A, Clavel J: **Spatial and space-time clustering of childhood acute leukaemia in France from 1990 to 2000: a nationwide study**. *British Journal of Cancer* 2006, **94:**763–770.

[48] Kulldorff M, Song C, Gregorio D, Samociuk H, DeChello L: **Cancer map patterns: are they random or not?** *American Journal of Preventive Medicine* 2006, **30:**S37-S49.

[49] Tango T: **A class of tests for detecting 'general' and 'focused' clustering of rare diseases**. *Statistics in Medicine* 1995, **14:**2323–2334.

[50] Tango T: **A test for spatial disease clustering adjusted for multiple testing.** *Statistics in Medicine* 2000, **19:**191-204.

[51] Munger RG, Tamura T, Johnston KE, Feldkamp ML, Pfister R, Carey JC: **Plasma zinc concentrations of mothers and the risk of oral clefts in their children in Utah.** *Birth Defects Research (Part A): Clinical and Molecular Teratology* 2009, **85:**151-155.

[52] Haggard LM, Shah GH, Rolfs RT: **Assessing health status: establishing geographic areas for small area analysis in Utah.** *Utah's Health:* An Annual Review, Vol V., 1997-1998.

[53] Utah Department of Health (UDOH): **Utah's indicator based-information system for public.** [http://ibis.health.utah.gov/.]. August 2007.

[54] R Development Core Team: *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2007. http://www.r-project.org.

[55] Assuncao RM, Reis EA: **A new proposal to adjust Moran's I for population density**. *Statistics in Medicine* 1999, **18:**2147-2162.

[56] Kulldorff M and Information Management Services, Inc: **SaTScan(TM) v7.0: Software for the Spatial and Space-time Scan Statistics** 2007.

[57] Openshaw S, Charlton M, Wymer C, Craft A: **A mark i geographical analysis machine for the automated analysis of point data sets.** *International Journal of Geographic Information Systems* 1987, **1:**335–358.

[58] Newell JN, Besag JE: **The detection of small-area database anomalies.** In *Methods for investigating localized clustering of disease*. 135th edition. Edited by Alexander FE, Boyle P: Lyon: International Agency for Research on Cancer; 1996: 88–100.

[59] Hjalmars U, Kulldorff M, Gustafsson G, Nagarwalla N: **Childhood leukemia in Sweden: using GIS and a spatial scan statistic for cluster detection.** *Statistics in Medicine* 1996, **15:**707-715.

[60] Kulldorff M, Feuer EJ, Miller BA, Freedman LS: **Breast cancer clusters in Northeastern United States: a geographic analysis.** *American Journal of Epidemiology* 1997, **146:**161-70

[61] Rogerson PA: **The detection of clusters using a spatial version of the chi-square goodness of-fit statistic.** *Geographical Analysis* 1999, **31:**130-147.

[62] Le ND, Petkau AJ, Rosychuk R: **Surveillance of clustering near point sources.** *Statistics in Medicine* 1996, **15:**727-740.

[63] Kulldorff M, Huang L, Pickle L, Duczmal L: **An elliptic spatial scan statistic.** *Statistics in Medicine* 2006, **25:**3929-3943.

[64] Moss MM: **Smoking, Anemia, and risk of Oral Clefts in Utah.** *MS thesis*, Utah State University, Nutrition and Sciences; 2006.

[65] Krapels IP, van Rooij IA, Ocke MC, van Cleef BA, Kuijpers-Jagtman AM, Steegers-Theunissen RP: **Maternal dietary B vitamin intake, other than folate, and the association with orofacial cleft in the offspring.** *European Journal of Nutrition* 2004, **43:**7-14.

Table 4-1 Summary results for global clustering tests for oral clefts data in the State of Utah, 1995-2004 (MC p-values refers to Monte Carlo p-values).

Test statistic	**Parameter**	**Value**	**MC* *p*-values**
PW statistic		805000	0.237
Moran's *I*	adjacency	-0.147	0.956
d(km)			
	5.0	0.548	0.028
	10	-0.021	0.523
	15	-0.034	0.571
	20	0.002	0.363
	30	0.009	0.36
	40	-0.024	0.556
	50	-0.012	0.354
EBI	adjacency	-0.115	0.905
Tango's EET (unadjusted *p*-value)			
λ (km)	5.0	0.00130	0.171
	10	0.00133	0.142
	15	0.00134	0.147
	20	0.00132	0.167
	30	0.00129	0.180
	40	0.00126	0.186
	50	0.00121	0.201
Tango's MEET (adjusted *p*-value)	5 - 50		0.229

Table 4-2 Summary results for cluster detection tests for the oral clefts data in the State of Utah, 1995-2004.

Test statistic	Parameter	Area ID	Radius (*km*)	Observed	Expected	Relative Risk	MC *p*-value
Besag-Newell statistic							
	$k = 22$	53	0.00	27	14.72	1.834	0.045
	$k = 30$	53, 52	100	37	20.57	1.799	0.030
	$k = 32$	53, 52	100	37	20.57	1.799	0.012
	$k = 36$	47, 46	5.00	39	25.60	1.523	0.030
		53, 52	100	37	20.57	1.799	0.001
		46, 44	4.00	40	26.70	1.498	0.049
	$k = 38$	47, 46	5.00	39	25.60	1.523	0.013
		46, 44	4.00	40	26.70	1.498	0.023
Spatial scan statistic	$\leq$ *10%*	53, 52	111	37	20.57	1.833	0.063
		46, 44, 47	4.60	68	45.91	1.521	0.111
	$\leq 25\%$	53, 52	111	37	20.57	1.833	0.089
		46, 44, 47	4.60	68	45.91	1.521	0.144
	$\leq 50\%$	53, 52	111	37	20.57	1.833	0.100
		46, 44, 47	4.60	68	45.91	1.521	0.165

Table 4-3 Maternal characteristics by cluster and area. Shown are counts (percentages) within each cluster and area.

Cluster	Lower Education	Active Smoking	Passive Smoking	Alcohol Use	Paternal History	Relative History	Family History
Primary Cluster (N = 18)	12 (66.7%)	5 (27.8%)	4 (22.2%)	1(5.6%)	3 (16.7%)	5 (27.8%)	7 (38.9%)
52 (N* = 3)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0(0.0%)	1 (33.3%)	0 (0.0%)	1 (33.3%)
53 (N = 15)	12 (80.0%)	5 (33.3%)	4 (26.7%)	1(6.7%)	2 (13.3%)	5 (33.3%)	6 (40.0%)
Secondary Cluster (N = 49)	13 (26.5%)	1 (2.0%)	1 (2.0%)	1(2.0%)	8 (16.3%)	6 (12.2%)	14 (28.6%)
44 (N = 20)	8 (40.0%)	1 (5.0%)	1 (5.0%)	1(5.0%)	5 (25.0%)	4 (20.0%)	9 (45.0%)
46 (N = 9)	2 (22.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (11.1%)	1 (11.1%)
47 (N = 20)	3 (15.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	3 (15.0%)	1 (5.0%)	4 (20.0%)

*N represents the number of cases that had maternal characteristics information.

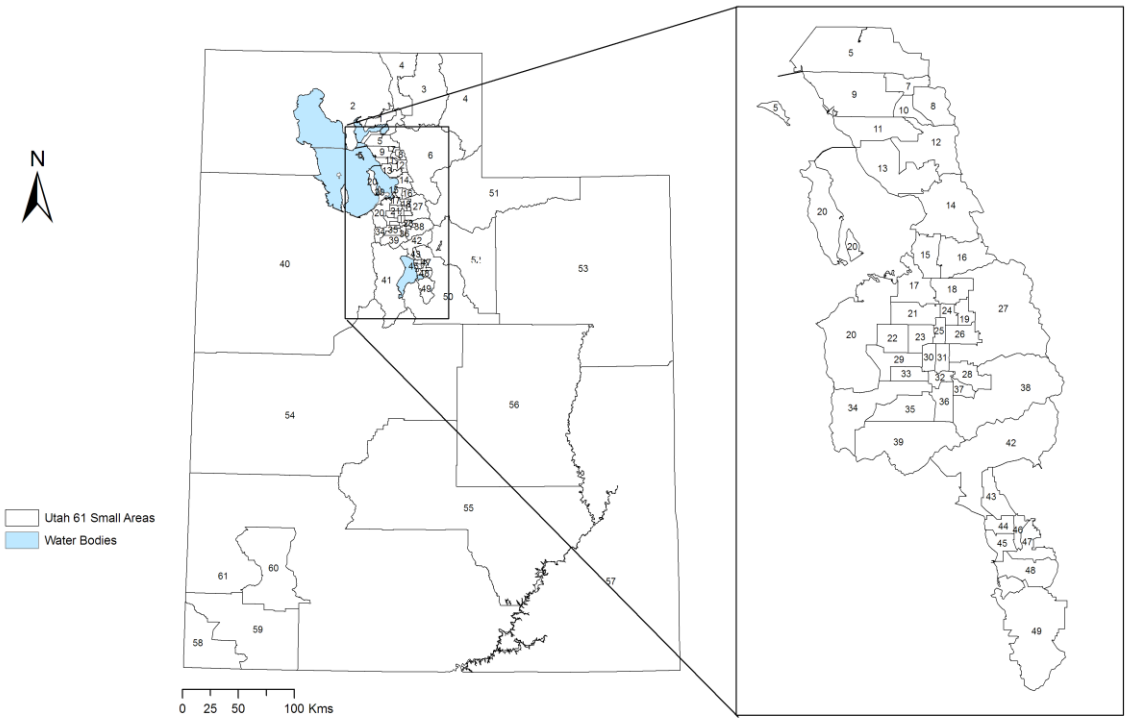


Figure 4-1 The study area for the State of Utah showing 61 small-geographic areas with their corresponding area ID

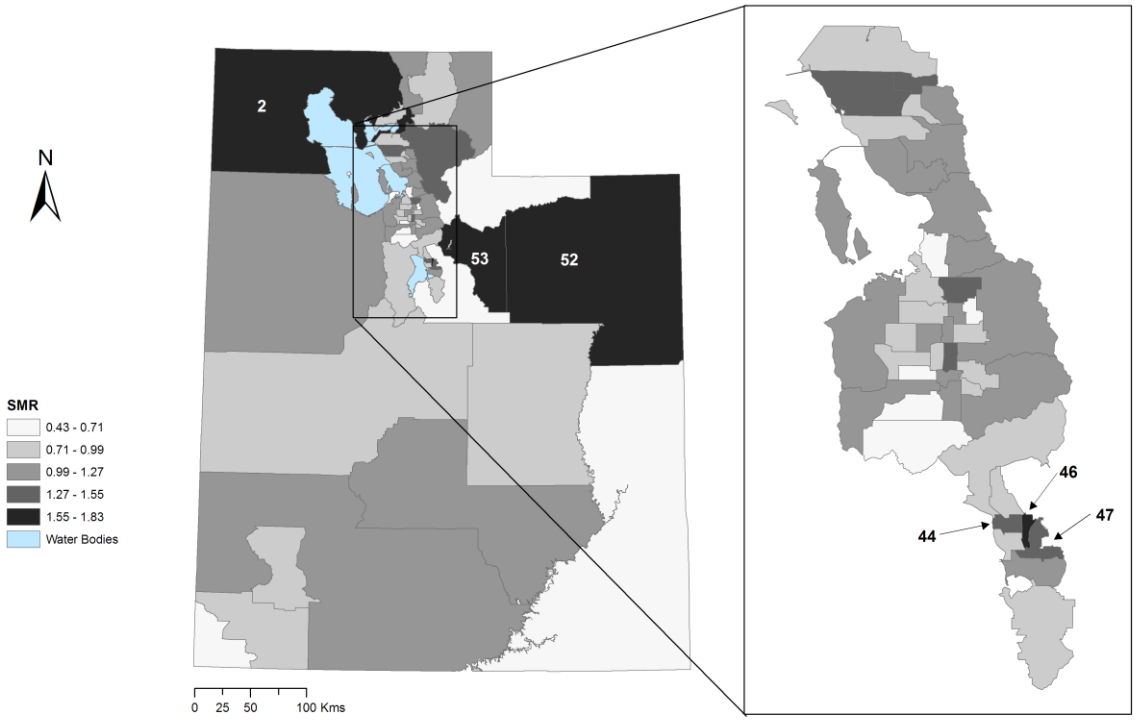


Figure 4-2 SMR of oral clefts by small-geographic areas for the State of Utah, 1995 – 2004. The color scheme goes from a light grey (low SMR value) to a dark grey (high SMR value).

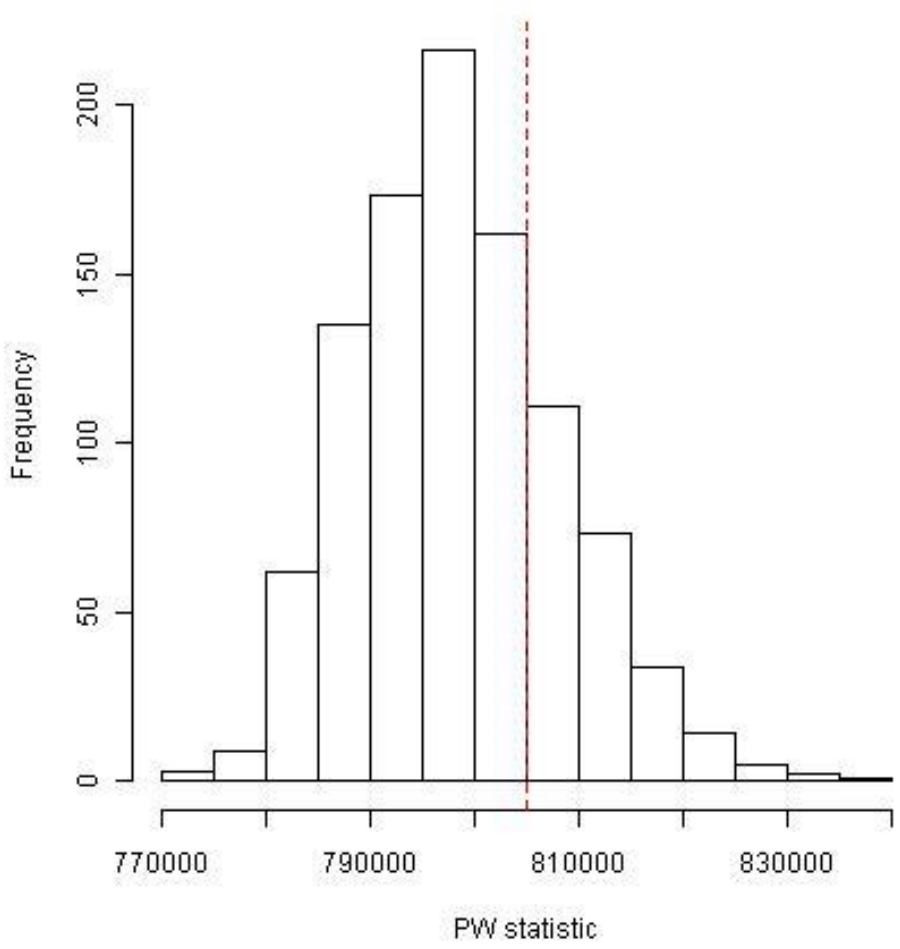


Figure 4-3 Histogram of the PW statistic for the oral clefts data. The vertical line denotes the PW statistic (80500) and the corresponding p-value (0.237).

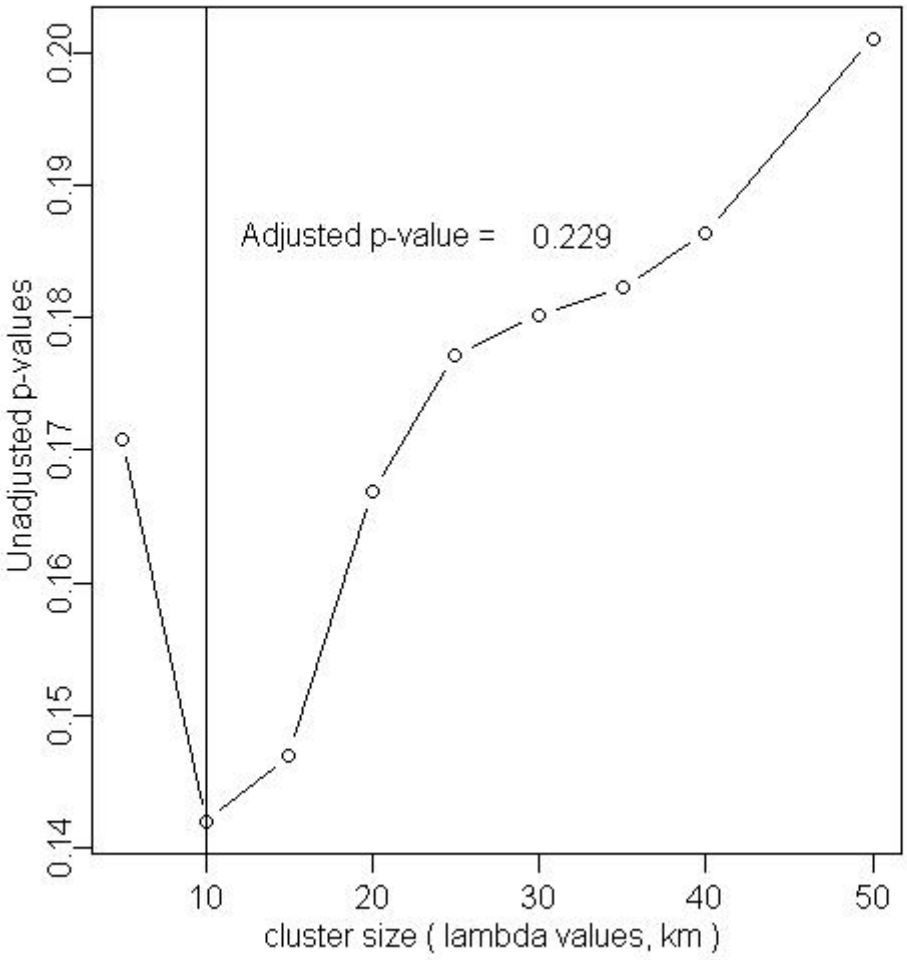


Figure 4-4 The Profile p-value of EET statistic for the oral clefts data. The vertical line denotes the optimal which attains the minimum of the profile p-value.

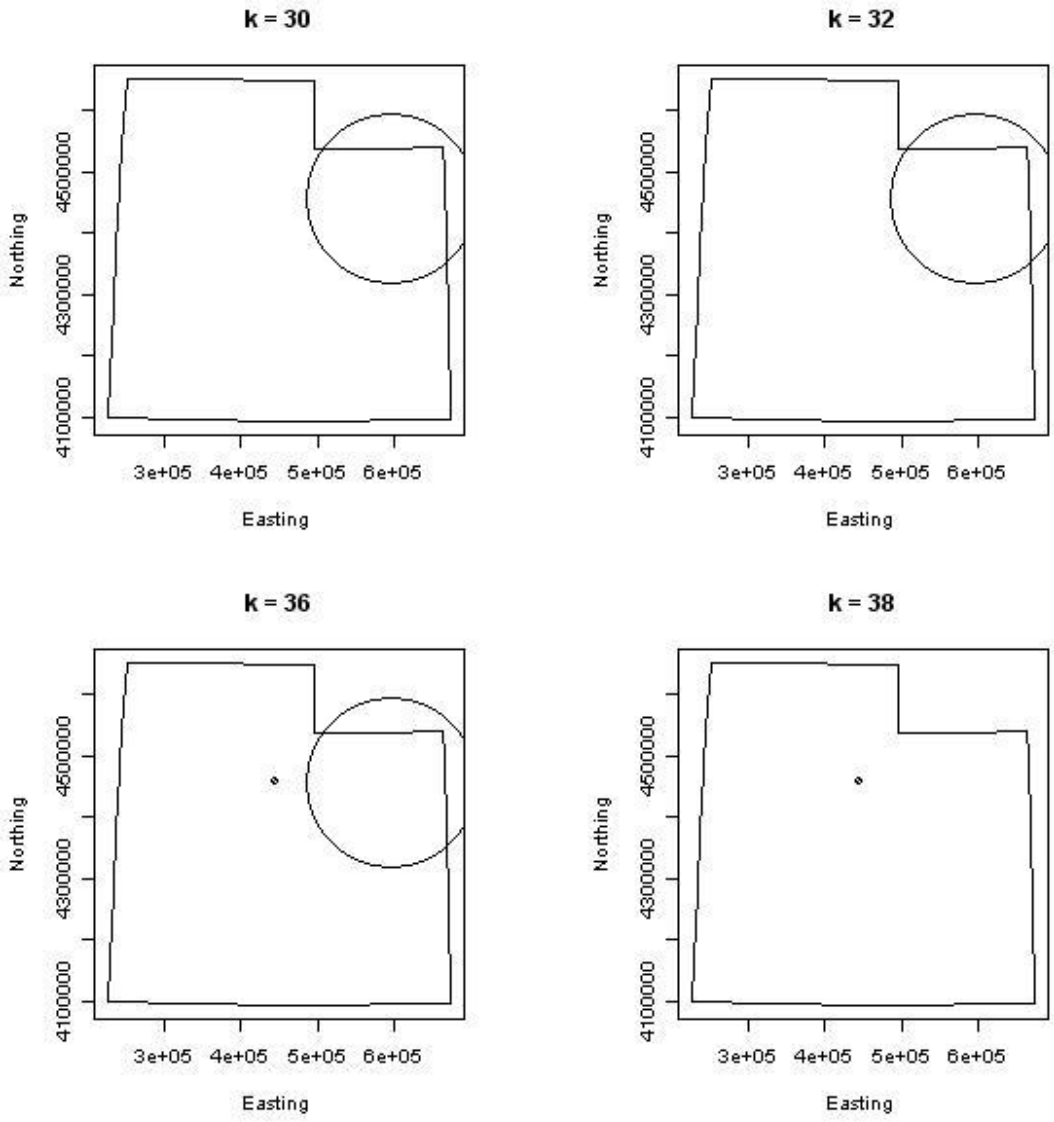


Figure 4-5 The most significant clusters of oral clefts using the BN statistic for the cluster sizes of k =30, 32, 36, and 38. Easting refers *x*-coordinate and Northing refers y-coordinate.

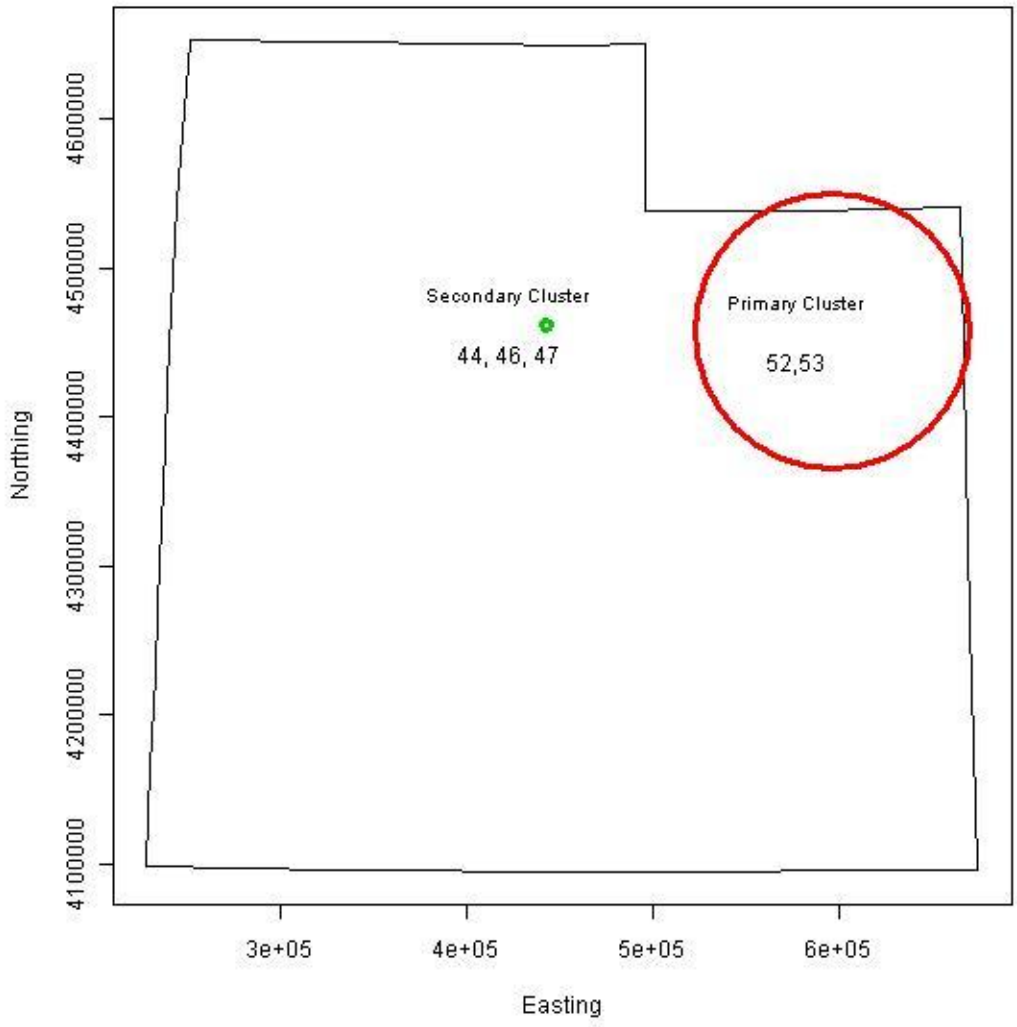


Figure 4-6 The oral clefts clusters detected using the spatial scan statistic. The red circle corresponds to the most likely cluster (primary cluster) and the green circle corresponds to the secondary cluster.

CHAPTER 5

SUMMARY, CONCLUSIONS, AND FUTURE DIRECTIONS

SUMMARY

The overall objective of the research conducted here was to undertake a comprehensive evaluation of the spatial distribution of oral clefts and establish any linkages that might exist with a broad range of demographic, behavior, social, economic, and environmental risk factors, through the application of Geographic Information Systems (GIS) and spatial statistics methodologies. Major research themes addressed were i) to explore differences in oral clefts rates across states and regions in the United States (U.S). and within Utah counties, and for the State of Utah, ii) to examine the spatial variations in the prevalence of oral clefts at the small-area level; iii) to identify high-risk populations and locations of oral clefts and iv) to assess the extent to which specific individual level and area-level risk factors explain the spatial variations and the local clusters identified in ii) and iii).

From a methodological point of view, the research themes were conducive to the application of GIS technologies and several innovative statistical methodologies, those being 1) the visualization of geographically indexed oral clefts rates using linked micromap (LM) plots; 2) Bayesian spatial and ecological analyses of small-scale spatial patterns of oral clefts, and 3) identification of high-risk populations and locations of oral clefts using a numerous spatial clustering techniques. Each chapter exemplifies detailed

discussion. In this chapter, a summary of the salient findings for each chapter is provided along with conclusions and directions for future research.

In Chapter two, an innovative visualization technique called LM plotting was used to represent geographically indexed oral clefts at two geographical resolutions – at the state level for the U.S. and at County level for the State of Utah using data obtained from the National Birth Defects Prevention Network (NBDPN) and the Utah Birth Defects Network (UBDN). Many states and local agencies have implemented birth defects surveillance systems to monitor and disseminate information regarding birth defects. However, many of these agencies rely solely on tabular methods to disseminate statistical birth defects summaries, which force readers and public health officials to construct their own visualization in order to uncover trends, relationships, and anomalies that may be present in the data. The LM representation, and particularly so the web-based application, provides an alternative presentation technique for birth defect data that perhaps can move the field away from just tabular formats alone but goes further in portraying the information in a joint geographical and statistical context which is unique in itself

Two template (US state and Utah County) LM plots were used to represent statistical summaries of oral clefts and their spatial indices. At the state level, the LM plot displayed five parallel sequences of panels: US and state micromaps, state names along with three statistical summary panels, including oral clefts rates, proportion of maternal smoking during pregnancy, and proportion of American Indian and Alaskan Native (AIAN). The purpose of this LM plot was to reveal overall spatial trends and explore the

relationships between the statistical summary panels. At the county level, the LM plot displayed four parallel sequences: State of Utah and counties micromaps, names of counties and statistical summary panels of oral clefts rates, count of oral clefts cases and total number of live births, with one of the statistical panels showing confidence intervals (to indicate uncertainty) of the oral clefts estimates in each county.

The main epidemiological result that came from the state level LM plot was that oral clefts were a major public health issue in the State of Utah, i.e., the state with the next to one the highest prevalence of oral clefts in the U.S after Alaska. Moreover, the LM plot revealed spatial patterns indicating that higher oral cleft occurrences were observed to be in the southwest and the Midwest and that lower occurrences were found in the East of the country. The plot also revealed significant and positive associations between oral cleft occurrence and maternal smoking rates and the proportion of American Indians and Alaskan Natives (States with a high percentage of AIAN population exhibited high oral clefts rates. In particular, there were five states (Alaska, Utah, New Mexico, North Dakota, and Oklahoma) in this category that scored highest in oral clefts rates. Among the 15 states with the highest oral cleft occurrence, nine had a smoking rate of 16% or higher while among the 15 states with the lowest oral cleft occurrence only one state had a smoking rate greater than 16%.

At the state level, the LM plot showed counties with reliable rate estimates such as Salt Lake, Cache, Weber and Box Elder resulted in narrow confidence intervals a direct result of these counties being more heavily populated. In contrast counties like Daggett, Garfield, Kane, Millard and Sanpete resulted in wide confidence intervals a

direct result of these counties being more sparsely populated. Such a representation means that any readers can directly appreciate the uncertainty associated with rate estimations. Furthermore, graphical plots that include uncertainty estimates mean that public health officials are more informed and this purports a more objective decision-making process in relation to targeted disease control.

In conclusion, LM plots offer many advantages over traditional choropleth map and tabular methods of data presentation as they are more effective at representing oral clefts data both at the state and county levels. The integration of micromaps into birth defects surveillance will enhance data collection, data analysis, and hypothesis generation but will also aid in any planning of public health services.

In Chapter three, an investigation of small scale spatial patterns of oral clefts in the State of Utah was carried out using Bayesian spatial and ecological models. The research goals were to demonstrate how Bayesian modeling techniques might be used to provide reliable oral clefts risk estimates and to highlight areas of high prevalence, and then explain the results, in terms of potential risk factors measured at area-level. The Bayesian analysis produced various unique model -based maps of oral clefts for the State of Utah. Such maps of oral clefts distribution are important tools for guiding surveillance and effective control of oral clefts because they (a) provide useful information on areas at high risk of oral clefts and, (b) help to optimize the allocation of resources. In addition, the oral clefts maps can be used to assess the effectiveness of intervention programs.

The Bayesian analysis comprised four hierarchical models specifically; a non-spatial model and a spatial model partitioned accordingly to account for random effects

and spatial autocorrelation in the data. The resulting risk estimates for all models were similar if not identical. All models produced smooth and interpretable maps of oral clefts risk. Moreover, all models filtered out that attributed to random noise from the "true" high oral clefts risk areas. The models detected modest small scale variations in oral clefts risks in the State of Utah. A few areas in Tri-county LHD, Provo/BYU, and North Orem areas were highlighted with high relative risks indicating possible "local clusters" of oral clefts. Comparison of the Deviance Information Criterion (DIC) of the two models indicated that the addition of spatially correlated random effects did not markedly the outcome.

Furthermore, by extending the Bayesian models to include covariates, it was possible to assess the associations between oral clefts prevalence and ecological risk factors. The results point towards a statistically significant positive association between mother's tobacco use and the risk for oral clefts; this finding is consistent with several previous studies (Khoury et al., 1987; Lieff et al., 1999; Wyszynski et al., 1997; Chung et al., 2000). On the other hand, there was a positive but not statistically significant association between a mother's alcohol consumption and the risk for oral clefts. Other studies of a mother's alcohol use during pregnancy (e.g., Munger et al., 1996) have documented this as statistically significant. The inconsistency here is likely problems associated with using ecological data that is aggregated. In such circumstances, direct relationships between oral clefts risk and mother's alcohol use can be obscured and difficult to establish.

There was no strong association between education background and oral clefting. Although the association was not significant, it is important to note that the direction of the relationship was positive. The lack of association between oral clefts and mothers with no high school diploma could be due to the fact that it was based on aggregated data, which could obscure the true association at the individual level. However, regardless of its direct association with oral clefts, it is important to consider maternal education as a proxy for covariates that are unmeasured or difficult to measure; for example, women with less education are less likely to take folic acid during pregnancy, it has been shown in several studies that multivitamin and folic acid supplementation reduces the risk of oral clefting (van Rooji et al., 2004; Wilcox et al., 2007).

In conclusion, by accounting for spatial dependency and area-level covariates in small-area data, Bayesian hierarchical methods provide more reliable estimate of oral clefts risk. In addition, genuine areas of elevated oral clefts were highlighted and area-level characteristics associated with the risk patterns of oral clefts were identified. The small scale oral clefts maps produced by using Bayesian methods can have an important role in planning and intervention programs. Specifically, the maps can be used by the Utah Department of Health (UDOH) or UBDN to direct surveillance and channel resources or intervention strategies by virtue of more robust accurate estimates of oral clefts.

Chapter four approaches the problem from another perspective where multi-scalar approaches to the spatial clustering and cluster analysis were introduced. Here the objectives were to test whether clustering of oral clefts cases were present anywhere in

Utah, to identify any local clusters of excess oral clefts that might be found, and to undertake an examination of the maternal characteristics involved with the clusters. Global clustering and local cluster testes were used in tandem to detect the different aspects of spatial patterns present in the Utah oral clefts data. Specifically, Potthoff - Whittinghill, Moran's *I* and Tango's MEET statistics were used to test for presence of global clustering and Besag - Newell and the spatial scan statistics were applied to detect for local clusters. Each of these methods is sensitive to different aspects of spatial patterns and so, they complement each other.

The results of the multi-scalar approach using the Potthoff – Whittinghill, Moran's *I* and Tango's MEET methods indicated no evidence of spatial clustering of oral clefts rates across the study area. This finding suggests that there is little evidence to support the existence of a strong source of environmental exposure or maternal infection affecting oral cleft outcome. However, a tendency of excess oral clefts cases was identified in the Tri-county LHD, Wasatch county, Provo/BYU, East Orem and North Orem areas using spatial scan statistic and Besag and Newell method. Subsequent investigation of the maternal characteristics involved in these areas showed that maternal smoking use, maternal lower education level, and family history of congenital malformation were high within the Tri-County LHD, whereas the Provo/BYU and North Orem areas showed only a modest number of cases with maternal lower education level and family history of congenital malformation but very low maternal smoking.

In conclusion, using a multi-scalar approach, it was possible to probe different aspects of spatial patterns in oral cleft prevalence over Utah and so provide a more

critical assessment of the processes that might underlie their distribution. Despite finding no evidence of spatial clustering, a few areas with excess oral clefs cases were signaled out. Furthermore, these areas were associated with established risk factors of oral clefts.

LIMITATIONS AND FUTURE DIRECTIONS

Although the research just detailed does provide a comprehensive evaluation of the spatial distribution of oral clefts, there are several limitations that must be discussed and addressed in the future work. One limitation was that the NBDPN oral clefts data used were collected from different state birth defects surveillance programs. The variation in the rates of oral clefts across states and regions may reflect differences in the collection and ascertainment methods of state-based birth defect surveillance systems (e.g., difference in case ascertainments, case inclusion criteria, and inclusion of elective terminations and still births) rather than the true difference among states and regions. Therefore, for a more meaningful characterization of the spatial variations in oral clefts across states and regions, guidelines are needed on standardized methods of birth defects collection; cases case ascertainments, and inclusion criteria.

As mentioned, this research study did not adjust for confounding factors (e.g., age of the mother and race of infant) for the Bayesian spatial cluster analysis and mapping because these data were not made available to us by the UDOH despite an IRB request. These confounding factors might influence the results of the spatial analysis, but it is doubtful, for reasons stated earlier, that the results were overly affected by a lack of these

adjustments. Regardless, future studies of spatial analysis of oral clefts should control, wherever possible, for relevant demographic and socioeconomic confounders.

The Utah oral clefts data used in this research study might not be complete record. For example, the oral clefts rates reported by UDOH are 22.2 per 10,000 for the period 1995-2003 compared to this research study's rates of 19.5 per 10,000 for period 1995-2004. Although the difference between these two rates is small, an incomplete dataset could result in spurious geographical variation. However, again it is unlikely that the results would change substantively since we found only modest geographical variation or clustering.

The research was based on aggregated data; essentially it was an ecological study. Although there is a valuable contribution in terms of helpful information garnered for pre-epidemiologic studies of oral clefts, a limitation of such a study is that it is difficult to establish casual relationships between oral clefts and any potential risk factors. Future studies should involve a more detailed investigation using either case-control or cohort studies. Especially important for future research initiatives should be a focus on those areas identified as hotspots in order to pinpoint specific individual risk factors that may be unique to these areas. In addition, future studies should move to identify any genes that are risk factors for oral cleft development and examine the interactions between those genes and environmental factors.

The effects of ambient air and environmental pollutants on oral clefts have not been directly addressed in this research, however other studies have examined the effects of exposure to ambient air pollution (Ritz et al., 2002; Gilboa et al., 2005) and

environmental hazardous waste sites (Brender et al., 2006) during pregnancy. Future studies should consider the contribution of these pollutants in Utah oral cleft outcome at refined scale. For instance, one can examine whether oral clefts cases tend to cluster nearby to so-called "superfund" sites or under elevated air pollution levels (e.g., strong subsidence inversion events where particulate matter, nitrogen dioxide, carbon monoxide and ozone can be concentrated and exceed "safe" levels); in such scenarios the use of focused testes such as Lawson and Waller (1996) might be very insightful.

The current study was based on the collective types of oral clefts grouped together. However, a growing body of evidence suggests that there exists etiologic heterogeneity between different types of oral clefts. To better understand the underlying etiology of oral clefts and whether common or different genes and environmental risk factors play a causal role for the different types of oral clefts, future studies should break down the analysis by type.

From a methodological point of view, some of the methods for global clustering (e.g., Tango's MEET) and local cluster testes (e.g., Spatial Scan Statistics) formally accounted for multiple testing, while others such as Moran's *I* and BN method did not account for multiple testing. Future studies should consider a standardized approach to address the multiple testing problems that will allow a researcher to first maintain statistical rigor but second and more importantly, allow inter-comparison of results between different methods to be the case.

Further spatial analyses of oral clefts at a finer scale level (individual level) should be conducted to confirm or capture spatial patterns that will be missed by using

aggregated data. An important expansion would be to consider the inclusion of temporal data to allow for a better understanding of possible temporal patterns. Furthermore, to better understand the etiology of oral clefts and explore how physical and social environment of neighborhood might influence the risk of oral clefts, future studies should consider using multilevel modeling. Multilevel modeling is a powerful construct that allows one to estimate the contribution to the outcome of both individual level and neighborhood level to the total variation.

CONCLUSIONS

In conclusion, the research presented here presents additional insights into oral clefts epidemiology through a geographical perspective. The findings indicate that higher oral clefts occurrence in the southwest and the midwest and lower occurrence in the east, with the patterns of oral clefts occurrence significantly related to smoking rates and American Indians and Alaskan Natives. The results also indicate that the State of Utah has the second highest prevalence of oral clefts in the U.S. However, small-area analysis of Utah oral clefts data revealed modest spatial variation in oral clefts risks in the State of Utah, with no pronounced spatial clustering. Our finding of lack of global clustering suggests a common environmental risk factor is unlikely to be a plausible cause of oral clefts. In other words, it eliminates the existence of a single and strong source of environmental exposure such as air pollution, contaminated water, hazardous waste, factory emissions, or maternal infection causing high prevalence of oral clefts in the State of Utah. Despite the lack of evidence for spatial clustering in a study, a few notable areas

(i.e. within the Tri-County LHD, Wasatch County, Provo/BYU, North Orem and East Orem) there was a tendency towards high aggregation values of oral clefts cases, indicating possible local clustering of oral clefts.

Furthermore, the results support the hypotheses that maternal smoking, family history, and maternal education background are significant risk factors of oral clefts for the state. However, further evaluation of the role of these factors is required through the application of individual data; especially in areas that exhibit high oral clefts cases. In addition, throughout the research venture effort was made to reveal the usefulness of GIS and novel spatial statistical methodologies for birth defects surveillance. It was indeed demonstrated how birth defects data collected by state and local monitoring systems coupled with GIS and spatial statistics methods could be useful as a preliminary and cost-effective method of characterizing the epidemiology of birth defects.

Finally, the findings and the methodological applications demonstrated here can pose an important role for guiding further epidemiological studies and for aiding public health officials in surveillance and control activities. Specifically, governmental agencies such as UDOH or UBDN can use the information to optimize the allocation of health resources for oral clefts control or intervention activities such as smoking cessation programs, and / or nutritional and multivitamin supplement use as a preventative measure.

REFERENCES

Brender JD, Zhan FB, Suarez L, Langlois PH, Moody K, 2006. Maternal residential proximity to waste sites and industrial facilities and oral clefts in offspring. J Occup Environ Med 48, 565-572.

Chung KC, Kowalski CP, Kim HM and Buchman SR, 2000. Maternal cigarette smoking during pregnancy and the risk of having a child with cleft lip/palate. Plast and Reconstr Surg 105, 485-491.

Gilboa SM, Mendola P, Olshan AF, Langlois PH, Savitz DA, Loomis D, Herring AH, Fixler DE, 2005. Relation between ambient air quality and selected birth defects, seven county study, Texas, 1997–2000. Am J Epidemiol 162, 238–252.

Khoury MJ, Weinstein A, Panny S, Holtzman NA, Lindsay PK, Farrel K, Eisenberg M, 1987 Maternal cigarette smoking and oral clefts: a population-based study. Am J Public Health 77, 623-625.

Lawson AB, Waller L, 1996. A Review of Point Pattern Methods for Spatial Modelling of Events around Sources of Pollution. Environmetrics **7**, 471-488.

Lieff S, Olshan AF, Werler M, Strauss RP, Smith J, Michell A, 1999. Maternal cigarette smoking during pregnancy and risk of oral clefts in newborns. Am J Epidemiol 150, 683-694.

Munger R, Romitti P, Daack-Hirsch S, Burns T, Murray J, Hanson J, 1996. Maternal alcohol use and risk of orofacial cleft birth defects. Teratol 54, 27-33.

Ritz B, Yu F, Fruin S, Chapa G, Shaw GM, Harris JA, 2002. Ambient air pollution and risk of birth defects in Southern California. Am J Epidemiol 155, 17–25.

van Rooij IA, Ocke MC, Straatman H, Zielhuis GA, Merkus HM, Steegers-Theunissen RP, 2004. Periconceptional folate intake by supplement and food reduces the risk of nonsyndromic cleft lip with or without cleft palate. Prev Med 39, 689–694.

Wilcox AJ, Lie RT, Solvoll K, Taylor J, McConnaughey DR, Abyholm F, Vindenes H, Vollset SE, Drevon CA, 2007. Folic acid supplements and risk of facial clefts: national population based case-control study. BMJ 334, 464

Wyszynski DF, Duffy DL, Beaty TH, 1997. Maternal cigarette smoking and oral clefts; a meta-analysis. Cleft Palate Craniofac J 34, 206-210.

APPENDICES

APPENDIX A

Permission and Release Letters

From: Goldweber, Paulette - Hoboken <pgoldweb@wiley.com>
Date: Mon, Dec 20, 2010 at 9:12 AM
Subject: RE: Permission
To: Samson Gebreab <samygeb@gmail.com>

Dear Samson:

Thank you for your request. John Wiley & Sons, Inc. has no objections to your proposed reuse of this material.

Permission is hereby granted for the use requested subject to the usual acknowledgements (title, volume number, issue number, year, page numbers. Copyright [year and owner]. And the statement "This material is reproduced with permission of John Wiley & Sons, Inc."). Any third party material is expressly excluded from this permission. If any of the material you wish to use appears within our work with credit to another source, authorization from that source must be obtained. This permission does not include the right to grant others permission to photocopy or otherwise reproduce this material except for versions made by non-profit organizations for use by the blind or handicapped persons.

Best wishes,

Paulette Goldweber
Associate Manager, Permissions
Global Rights
John Wiley & Sons, Inc.
ph: 201-748-8765
f: 201-748-6008
pgoldweb@wiley.com

-----Original Message-----

From: Samson Gebreab [mailto:samygeb@gmail.com]
Sent: Monday, December 20, 2010 9:03 AM
To: Goldweber, Paulette - Hoboken
Subject: Permission

Permissions Department
John Wiley & Sons, Inc.
111 River St. MS 4-02
Hoboken, NJ 07030-5774
Fax: (201) 748-6008

To whom it may concern,

I am in the process of preparing my dissertation in the Department of Watershed Sciences and Ecology Center at Utah State University and I plan to complete my degree by the Dec. 2010. I would like to ask permission to reprint the manuscript published as: Gebreab, S. Y., Gillies, R. R., Munger, R. G., Symanzik, J. (2008) Visualization and interpretation of birth defects data using linked micromap plots. Birth Defects Research Part A Clinical and Molecular Teratology, 82 (2):110 – 119, to be used as a chapter in my dissertation. Please note that reprinting it as a chapter may require making some revision. If you have any questions, please do not hesitate to contact me via e-mail.

Thank you for your cooperation,

Sincerely;

Samson Gebreab

APPENDIX B

Deviance Information Criterion

Deviance information criterion (DIC) is a Bayesian model comparison criterion proposed by Spiegelhalter et al. (2002). DIC is based on trade-off between model "goodness of fit" and "complexity" that is based on the posterior distribution of the deviance statistic:

$$D(\boldsymbol{\theta}) = -2\log f(\mathbf{y} \mid \boldsymbol{\theta}) + 2\log h(\mathbf{y})$$

where, $f(\mathbf{y} \mid \boldsymbol{\theta})$ is the likelihood function for the observed data vector $\mathbf{y}$ given the parameter $\boldsymbol{\theta}$ and $h(\mathbf{y})$ is some standardizing function of the data alone and has no impact on model selection. In this approach the model goodness of fit of the data is summarized by the posterior expectation of the deviance $\overline{D} = E_{\theta|y}[D]$, while the model complexity is captured by the number of effective parameters p_D, which is defined as the posterior mean deviance minus deviance evaluated at the posterior mean of the parameters:

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \overline{D} - D(\overline{\boldsymbol{\theta}})$$

The DIC is then defined analogously to AIC, i.e., the sum of model goodness of fit and the effective number of parameters:

$$DIC = \overline{D} + p_D = 2\overline{D} - D(\overline{\boldsymbol{\theta}})$$

Since a small value of ($\overline{D}$) indicates good fit while a small value of p_D indicates a parsimonious (simpler) model, therefore, a small value of DIC indicates that the model

is better supported by the data. DIC can be monitored in WinBUGS from Inference/DIC menu.

CURRICULUM VITAE

Samson Y. Gebreab

RESEARCH INTERESTS

Birth Defects Epidemiology, Infectious Diseases, Cardiovascular Diseases Epidemiology, Generalized Linear Models, Multilevel Analysis, Spatial Epidemiology, Spatial Statistics, and Geographic Information Systems (GIS).

EDUCATION

- **Ph.D. Emphasis in Spatial Epidemiology** **May 2010**
  Utah State University, Logan, UT.

  Dissertation Topic: *Spatial Epidemiology of Birth Defects in the United States and the State of Utah Using Geographic Information Systems and Spatial Statistics*
  Advisor: Dr. Robert R. Gillies.

- **M.Sc. in Statistics** **Dec 2007**
  Utah State University, Logan, UT.

- **M. Sc. in Geo-information Science** **Jan 2003**
  Wageningen University, Wageningen, the Netherlands.

- **B.Sc. in Soil and Water Conservation (***With Distinction*) **Sep 1999**
  University of Asmara, Asmara, Eritrea

RESEARCH EXPERIENCE

- **Research Assistant** **Aug 2003 – May 2008**
  Utah State University, Logan, UT

  Conducted research on the spatial distribution of birth defects, anthrax, West Nile virus, and cancer diseases, including spatial clustering and clusters analyses, hierarchical Bayesian modeling, and ecological analyses. Statistical software R, SaTScan, GeoDa, and WinBUGS were used to implement the analyses.

  Performed statistical visualization analysis of birth defects, West Nile virus, and chronic disease wasting data using linked micromaps plots, choropleth maps, and

various statistical plots. R/Splus and ArcGIS were used to implement the statistical visualization techniques.

Designed and implemented relational database for health and nutrition survey data using MS ACCESS/MS SQL to be utilized by the Center for Epidemiologic Studies, Utah State University, Logan, UT.

- **Visiting Scholar** **Aug 2002 – Jan 2003**
  Utah State University, Logan, UT.

  Conducted research on spatial distribution of impervious surface area (ISA) for the City of Atlanta, Georgia using Geostatistics and Satellite Remote Sensing Images (Landsat MSS, TM, and ETM+) for the period 1972-2002 as part of my MSc. thesis.

- **Graduate Research Assistant** **Aug 2000 – Aug 2001**
  University of Asmara, Asmara, Eritrea.

  Worked as research assistant and program coordinator at the College of Agriculture and Aquatic Sciences. Duties include scheduling class rooms, proctoring exams, tutoring, and supervising labs.

TEACHING EXPERIENCE

- **Instructor** **Spring 2007**
  Stat 2000 (Introduction to Statistical Methods)
  Dept. of Mathematics and Statistics, Utah State University, Logan, UT.

- **Instructor** **Fall 2006**
  Math 1010 (Intermediate Algebra)
  Dept. of Mathematics and Statistics, Utah State University Logan, UT.

- **Teaching Assistant** **Fall 2003, 2004, & 2005**
  AWER 6740 (Fundamentals of Remote Sensing Science)
  Tutored and supervised AWER 6470 Lab (hands-on Image processing with ERDAS Imagine software).
  Dept. of Watershed Sciences, Utah State University, Logan, UT.

- **Teaching Assistant** **Aug 2000 – Aug 2001**
  Supervised Soil Science Labs, graded homework, and proctored exams.
  Dept. of Soil and Water Conservation, University of Asmara, Asmara, Eritrea

PUBLICATIONS

- **Gebreab, SY**, Gillies, R., Munger, R., and Symanzik, J. (2010). Small Area Mapping and Ecological Analyses of Oral Clefts in the State of Utah Using Hierarchical Bayesian Models. *To be submitted to Geospatial Health.*

- **Gebreab, SY**, Gillies, R., Munger, R., and Symanzik, J. (2010). A Multi-scalar Approach to the Spatial Cluster Analysis of Oral Clefts in the State of Utah from 1995 to 2004: With Clues for Etiology. *To be submitted to Environmental Health.*

- **Gebreab, SY**, Gillies, R., Munger, R., and Symanzik, J. (2008). Visualization and interpretation of birth defects data using LM plot. *Birth Defects Research Part A: Clinical and Molecular Teratology, 82:110-119, 2008.*

- Goovaerts, P, **Gebreab, SY**. (2008). How does Poisson Kriging compare to the popular BYM model for mapping disease risks? *International Journal of Health Geographics 7:6.*

- Symanzik, J., **Gebreab, SY**, Gillies, R. and Wilson, J. (2003). Visualizing the spread of West Nile Virus. *Proceedings of American Statistical Association*, Alexandria, Virginia, CD.

CONFERENCES, PRESENTATIONS AND ABSTRACTS

- Bayesian approach to mapping the spatial distribution of Oral Clefts in Utah. Presented, URISA GIS in Public Health Conference, New Orleans, Louisiana, May 2007.

- Visualization of birth defects data using linked micromap plots. Presented, URISA GIS in Public Health Conference, New Orleans, Louisiana, May 2007.

- How to build R packages for Windows: Demo R package? Presented at the Statistical Computing Course, Department of Mathematics and Statistics, Utah State University, Logan, Utah, April 2007

- Relational database for dietary supplements using MS Access /MS SQL. Invited Presentation, Center for Epidemiologic Studies, Utah State University, Logan, Utah, Nov 2005.

- Classification and prediction of satellite imagery data for land use planning: Data Mining Applications. Poster, Department of Mathematics and Statistics, Utah State University, Logan, Utah, Apr 2005.

- Climate and Health. Workshop attended, National Center for Atmospheric Science, Boulder, Colorado, Jul 2004.

COMPUTER SKILLS

- Statistical software: Extensive experience with SAS, R/SPlus, WinBUGS, HLM, GeoDa, SaTScan, and SPSS.
- Programming Languages: Some experience with C/C++.
- GIS/Remote Sensing Software: ERDAS Imagine, ArcGIS, ArcView, ArcInfo, AML.
- Database Systems: MS Access, SQL.
- Text type settings: LaTeX, MS Office (World, PowerPoint, Excel, and Access).
- Computing platforms: UNIX, Windows 95/98/XP/VISTA.

AWARDS, HONORS, AND LEADERSHIPS

- Recipient of $5000 Dissertation Fellowships, Utah State University, Logan, UT, Aug 2007 – May, 2008.
- Recipient of $300 Graduate Student travel grants, Utah State University, May 2007.
- Recipient of fellowship to attend workshop on Climate and Health, National Center for Atmospheric Science in Boulder, Colorado, USA, July 21– 28 2004.
- Recipient of the Centre for Environment and Development (CDE) scholarship to study MSc in Geographic Information Science, Wageningen University, the Netherlands, Aug 2001 – Jan 2003.
- Gold medal award winner (ranked first) from the College of Agriculture and Aquatic Sciences, University of Asmara, Asmara, Eritrea, Sep 1999.
- Member of Student's Union Executive Committee, University of Asmara, Asmara, Eritrea, 1997-1999.

PROFESSIONAL SERVICE AND MEMBERSHIPS

- Reviewer for Journal of Computational and Graphical Statistics, 2008.
- Member American Statistical Association, 2007.