THE IMPACT OF DATA SOVEREIGNTY ON AMERICAN INDIAN SELF-DETERMINATION: A FRAMEWORK PROOF OF CONCEPT USING DATA SCIENCE

BY

JOSEPH CARVER ROBERTSON

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Computational Science and Statistics

South Dakota State University

2018

THE IMPACT OF DATA SOVEREIGNTY ON

AMERICAN INDIAN SELF-DETERMINATION:

A FRAMEWORK PROOF OF CONCEPT USING DATA SCIENCE

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Gary Hatfield 4-26-18

Gary Hatfield, Ph.D.
Dissertation Advisor
and Committee Chair Date

Kurt Cogswell 4-26-18

Kurt Cogswell, Ph.D.
Head, Mathematics and Statistics Department

Date

Kendel C Doerner 27 April 2018

Dean, Graduate School Date

I would like to dedicate this manuscript most certainly to my mother and my father, Ann and Creighton Robertson. There has never been a time that I can remember that you have lost faith in my dreams, aspirations, and struggles. I have always felt nurtured and respected despite the all of the grief I have caused you. Your patience has been the cornerstone of understanding that despite my immaturity at times, it was only a matter of time before I realized the wisdom of your acts.

When my father Creighton was unexpectedly taken from us during a crucial time in my PhD studies; it was unclear at times whether I could carry on. If it had not been for the deep connection to family during this time, then I would not have realized the sacrifice that was made for my accomplishments today. Dad, wherever you are, take comfort in all that you have done for me is a debt I will never be able to repay. And mom, I still haven't cleaned out your garage, I promise I will.

ACKNOWLEDGEMENTS

This manuscript has been a huge undertaking. This began with a search on the internet for the word data sovereignty. There wasn't much out there, so this is where I began. It was not long before this project began taking on a life of its own. When I think back to the Consider the Century talk I gave here at SDSU to introduce my initial idea, and now this manuscript in finished form....amazing. This journey has solidified my resolve to work on behalf of my communities to be a force for change, no matter how little that might be.

This would not have been possible without the love and care of many that dared me to dream big. First of all, to my advisor and committee chair, Dr. Gary Hatfield, we have been together through my master's studies and throughout this journey you have been a silent, but wise advocate for the radical ideas I presented intended to disrupt the norms of higher education. You have kept me in check and never doubted my higher aspirations. In addition, without your love for statistics, candor, and understanding, I am most certain this could never have been possible. Thank you.

I would also like to thank Dr. Kurt Cogswell, committee member and department head. Your wisdom and forward thinking in some very difficult times forced me to face and ultimately decide to take ownership of my studies using a nation building approach. This manuscript is a representation of this both metaphorically and literally. This is what it means to build a community around trust, understanding, and compromise. Thank you for your guidance and faith; I am honored by your commitment to this project.

To the rest of my committee, Dr. Weiwei Zhang, Dr. Craig Howe, and Dr. Larry Browning, what can I say? I could not have asked for a greater group of people ready and willing see my vision through. Dr. Zhang, your input helped me to look at the bigger picture and to refine my raw passion into a concise and honest look at leveraging science in a whole new context. Dr. Howe, it was no surprise your expertise was tremendously helpful in contextualizing my ideas of nation building and to honor cultural capital by being mindful of what is at stake. Dr. Browning, our partnerships over the years serving American Indians students in all things science has been an inspiration and guiding force for me here at SDSU. Your commitment to our students is unparalleled and I thank you for that.

Next, I would like to thank Dr. Thomas Horan and Dr. Brian Hilton at Claremont Graduate University. Dr. Horan, if it was not for your offer to allow me to contribute to the tribal traffic safety project; this manuscript most certainly would not be what it is today. You and Dr. Hilton have been an integral part of helping me develop my data sovereignty framework as a proof of concept and your patience and understanding has been very impactful.

Finally, I would like to thank my family and friends. To my mom, you have been a shining light and your wisdom goes beyond anything I can even begin to describe. You have helped me face adversity, held me accountable every step of the way and most importantly respected every part of who I am. To my late father, Creighton, dad wherever you are I now understand the price of historical sacrifice. And to my siblings and friends, I am forever in your debt in teaching me the value of friendship and unconditional love.

TABLE OF CONTENTS

ABBREVIATIONS

AI	Artificial Intelligence
CGU	Claremont Graduate University
CSR	Complete Spatial Randomness
ECSA	European Citizen Science Alliance
EDA	Exploratory Data Analysis
ESRI	Environmental Systems Research Institute is an International Supplier Of Geographic Information System Software, Web GIS and Geodatabase Management Applications
EV	Enhanced View Program
GIS	Geographic Information Systems
IRP	Independent Random Process
KDE	Kernel Density Estimation
MAF	Master Address File
MnCMAT	Minnesota Crash Mapping Analysis Tool
NGA	National Geospatial-Intelligence Agency
NNI	Native Nations Institute
R	R is a Free Software Environment for Statistical Computing and Graphics

RBF	Radial Basis Function
RKHS	Reproducing Kernel Hilbert Space
RSI	Road Safety Institute
SD	Standard Distance
SMART	An Acronym used to describe a Data Sovereignty Framework Metric
SVM	Support Vector Machine
SWO	Sisseton Wahpeton Oyate of The Lake Traverse Reservation is a Tribal Nation in South Dakota
UC	ESRI User Conference
UTC	Roadway Safety Institute is the Region 5 University Transportation Center
USIDSN	U.S. Indigenous Data Sovereignty Network

LIST OF FIGURES

Chapter 2

Chapter 3

[Tribal Traffic Safety Manuscript Brief]

Chapter 1

Chapter 2

Chapter 3

Chapter 4

END [Tribal Traffic Safety MANUSCRIPT Brief]

Chapter 4

LIST OF TABLES

Chapter 2

Chapter 3

[Tribal Traffic Safety MANUSCRIPT Brief]

Chapter 4

ABSTRACT

THE IMPACT OF DATA SOVEREIGNTY ON

AMERICAN INDIAN SELF-DETERMINATION:

A FRAMEWORK PROOF OF CONCEPT USING DATA SCIENCE

JOSEPH ROBERTSON

2018

The Data Sovereignty Initiative is a collection of ideas that was designed to create SMART solutions for tribal communities. This concept was to develop a horizontal governance framework to create a strategic act of sovereignty using data science. The core concept of this idea was to present data sovereignty as a way for tribal communities to take ownership of data in order to affect policy and strategic decisions that are data driven in nature.

The case studies in this manuscript were developed around statistical theories of spatial statistics, exploratory data analysis, and machine learning. And although these case studies are first, scientific in nature, the data sovereignty framework was designed around these concepts to leverage nation building, cultural capital, and citizen science for economic development and planning.

The data sovereignty framework is a flexible way to create *data domains*, around developed key indicators to integrate appropriate cultural capital when working with Native nations. This design is intended to put scientific theory into practice to affect everyday outcomes using data driven decision making. This framework is a proof concept and represents both applied and theoretical metrics in design strength.

Chapter 1

Introduction

The Foundations of the Data Sovereignty Initiative

My name is Joseph Robertson. I am an enrolled member of the Sisseton Wahpeton Oyate of the Lake Traverse Reservation located in northeast South Dakota. I have developed a concept called the *Data Sovereignty Initiative, Creating SMART Solutions for Tribal Communities.*

It is a native-centric horizontal governance framework designed to create SMART solutions for tribal communities. The fundamental reason that I have pursued this area of study is to provide an ethical, cultural, and community based consultancy that is designed by an American Indian, for nation building to assist tribal communities with economic development, strategic planning, and data driven decision-making.

The design of this manuscript is to provide an overview of the nature of higher education pedagogy, nation building, citizen science, and data science that establishes a foundation for new praxis in understanding American Indian tribal sovereignty. It is my hope to create data infrastructure from the *Data Sovereignty Initiative* to serve and advocate on behalf of American Indian people through the lens of data science.

Concrete examples of statistical design theory as it pertains to data driven projects can be accomplished through nation building only if well-designed data domains utilize the strength of scientific methods and inquiry first. This holistic approach to applying science and technology to the real problems communities face can be thought of

as a disruptive, but a necessary part of redefining how we use science in intelligent ways to affect outcomes in our communities.

The scope of the case studies contained in this manuscript is to develop a tribal governance framework using data sovereignty as a mechanism for promoting appropriate data collection and practice in the context of data science. The framework design is a multidimensional approach that is anchored in the concept of *tribal sovereignty*, which is extended to the concept of *data sovereignty*.

The concept of SMART solutions pays homage to the current technology sector lexicon in which smart devices, smart phones, and smart homes have revolutionized the way we do things in *intelligent* ways.

Tawansi (2012) commented on the role smart devices play in our everyday lives:

> Smart devices offer a wealth of applications that can address almost every need at the touch of a button. This easy accessibility for consumers and business users alike has played a major role in the popularity of the devices. Furthermore, the functionality that currently exists at the fingertips of hundreds of millions of people around the world is constantly opening up new means of communicating, collaborating, transacting, processing and even analyzing. (para. 3)

(The acronym SMART will be described in more detail in the context of this framework in Chapter 2.)

As the world has moved into this new era, the need to create systems of SMART solutions that could potentially benefit American Indian communities through purely

scientific endeavors seemed reasonable but not exactly practical. And although the scope of this manuscript is undoubtedly scientific, many other subjects of inquiry were designed to intersect the science in order to put theory into practice.

The disruptive nature of how fast technology has been changing the societal and cultural norms in the world has had many unintended consequences. By "changing the rules" of everyday traditions, a new era has unfolded, the world of data science. As we will see, there are a number of collaborative scientific fields of discipline that make up data science such as statistics and machine learning, database management, and distributed and parallel systems that provide computational infrastructure.

This manuscript utilizes data science as a flexible way to create the SMART solution concept, but also attempts to bridge a number of topics that are found throughout this document. It is important to note that the scientific inquiry contained in these case studies is mutually exclusive to data sovereignty framework I designed to accompany those inquiries. It is by the creation of a *data domain* that it is possible to integrate scientific methods and inquiry to serve a multidimensional purpose of nation building to create economic development outcomes that are data driven.

This approach was no easy task. There are definitely some opposing forces that had to be considered when integrating cultural capital and nation building with scientific inquiry. By creating a disruptive shift in cultural norms or traditions, I was forced to critically examine a number of topics that could have unintended consequences. The most sensible way to create balance between all of the concepts that are weaved together in this manuscript is to divide the manuscript into three types of dependent and independent

schools of thought: the scientific implications, the data sovereignty framework implications, and the cultural implications.

In the purest terms, the science stands on its own, but can be applied to cultural and framework objectives through the data domain concept. The framework implications are tied to methodological work in strategic management, business intelligence, and are connected to the cultural implications through the designed key indicators, *Tribal Governance* and *Tribal Communities and Culture*.

The cultural implications are a very tricky subject to approach. There has been a great deal of thought put into defining culture as it relates to the concepts of what the words Native, indigenous, American Indian, and Native nations mean in the context of nation building. Chapter 2 attempts to provide a context into the complexities of how and when we speak about unique indigenous or tribal nations.

Although this manuscript has a fundamental air of culture, the framework design nonetheless utilizes statistical and mathematical principles of scientific method at its core. It is no easy task to bridge scientific implications with any given set of cultural implications without a well thought out framework design. The data sovereignty framework was created to advance a broader dialogue in how nation building and a critical examination of community and culture help to formulate new ideas in economic development and planning. Thus, the task of integrating these implications into a framework using scientific methods of inquiry begins in next chapter where we will examine how this process unfolds.

Chapter 2

The Data Sovereignty Framework

Synopsis

The *Data Sovereignty Initiative* is a collection of ideas that forms the data sovereignty framework. The data sovereignty framework conceptually is an intriguing way to incorporate a nation building approach, coupled with higher education as the foundation for integration of data science throughout Indian country. The term Indian country is a term that has been defined as "country within which Indian laws and customs and federal laws (United States) relating to Indians are generally acceptable" (Cohen, 2012, p. 183).

Initially, the goal was to critically examine Federal Indian Law and policy to better understand how the use of data could be used in American Indian tribes' economic planning and development. The foundations of this framework have evolved into a proof of concept system of data collection and practice that involve the use of four key indicators addressing not only analytical data driven metrics, but cultural and tribal governance metrics. In a nation building approach, governing institutions match indigenous political culture.

Native nations' dependence on institutional practice that does not reflect an indigenous design strategy is crucial in understanding and organizing appropriate political capital that is community driven. Building legitimate institutions asserts that when cultural match is high, economic development tends to be more successful (Cornell & Kalt, 2007). Developing governing bodies on both a global and local level can resonate

with deeply held principles and beliefs about authority, and can meet more contemporary needs through strategic decision-making.

Leadership is not limited to elected tribal officials and often individual tribal citizens doing the important work bear responsibility for what their citizens make for the future of their nation. Replacing the outsider-generated, top-down standard approach with indigenously generated responses to each nation's challenges is deemed to be cultural match.

The data sovereignty framework maintains this crucial balance of understanding sovereignty in terms of governance, but also attempts to legitimize the role of local stakeholders. This helps tribal citizens understand the tasks in rebuilding nations as a matter of sovereignty. "The distinctive features of such leadership are public spiritedness and the conviction that empowering a nation as a whole is more important than empowering factions or individuals" (Cornell & Kalt, 2007, p. 27).

Although it has been conceived in the past that only administrative top-down models produce results that are measurable, this is simply not the case. In examining the collective nature of the balance between top-down and horizontal design strategies, one thing is clear: the Native-centric path of empowerment through tribal citizenry makes a more direct impact in the perceptions of what is possible, rather than depending exclusively on an administrative approach to governance.

The horizontal design of the data sovereignty framework allows for multiple strategies and positions for different groups in a tribal organization to work together or independently depending on the level of sophistication with respect to a specific

Leveraging Tribal Governance and Citizenry Information is Paramount in a Native-Centric Governance Strategy

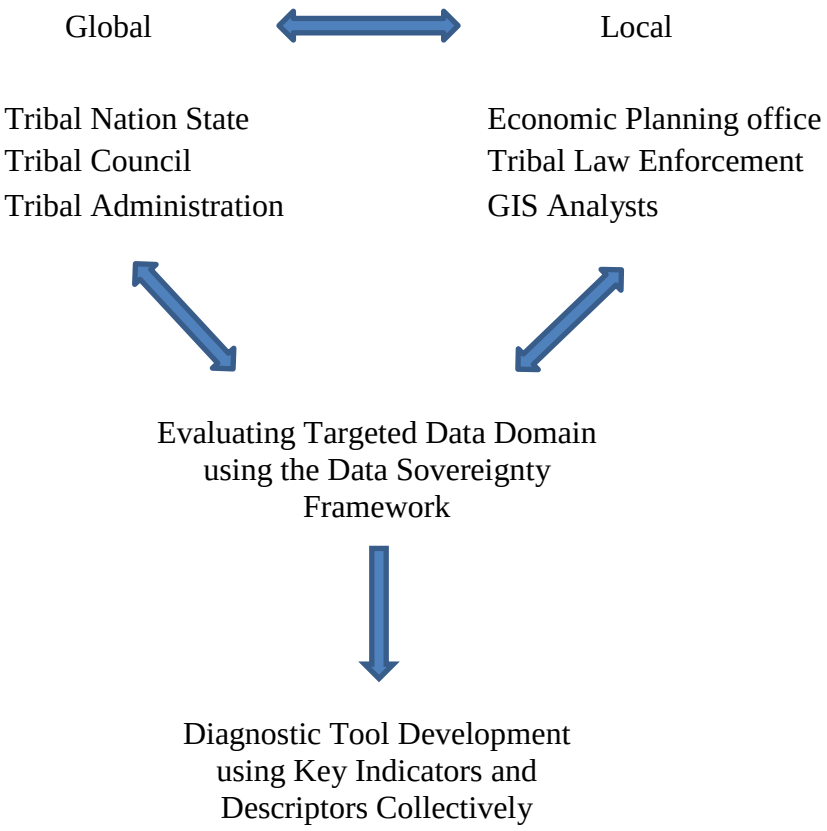


Figure 2.1 - An Example of Horizontal Shifting of Governance Strategies for Developing a Data Domain

community project. Figure 2.1 is an example of how a horizontal shift in governance provides a necessary departure from top-down strategies to assess the operational capacity of all levels in a tribal organization.

Kessler-Mata (2014) explores the concept of tribal sovereignty and "recognizes and promotes interaction between tribes and non-tribal governments as a central element of self-governance. This is in stark contrast to the widely and strongly held notion of

tribal sovereignty that views such interactions and negotiations as threats to tribes' sovereign status. While the ability for tribes to engage in intergovernmental relations is not the only element of self-governance or self-determination, my conception of tribal sovereignty positions this as, nonetheless, an essential one" (p. 36).

Equitable interaction and political coordination is the foundation for establishing data analysis as not only a policy stance, but for ultimately defining data sovereignty in strategic decision making. This forms a path moving forward using this as the foundation of nation to nation communique.

The task then is working with intergovernmental agencies (i.e. state and local governments) to adhere to principles of equitable interaction and respect for tribal sovereignty. The foundations of data sovereignty rely on a horizontal nation building approach which is flexible enough at its core to provide a voice to local tribal stakeholders that are almost always overlooked.

Evaluating each tribal group in terms of their governance as it relates to their strategic data collection and practice provides a method of exploratory analysis essential to establishing a baseline of four key indicators. Key descriptors are an extension of the four key indicators which provide additional dimensionality to the design which will be presented in more detail in the key indicator section.

The proposed concept is designed around the flexible key indicator called a data domain. A data domain is an extremely useful concept as it can be designed to encompass any task a stakeholder is interested in quantifying through data science. The framework is

in a highly developed form and is currently a proof of concept in the two case studies contained in this dissertation.

This framework provides a set of equitable solutions that address common shortcomings as it pertains to nation building throughout Indian Country, and particularly in advocating the use of data science as a matter of sovereignty.

So What is Data Sovereignty?

During the initial conception of the manuscript, the first question to ask was, How should data sovereignty be conceptualized? A nation building approach coupled with higher education deemed to be the most effective way to begin quantifying previously only qualitative assumptions about how American Indians communities operate as a sovereign polity. Data sovereignty in this context is the use of data science to leverage a data-driven outcome as a *strategic* act of sovereignty. What this means, is when data infrastructure is created using this framework, the call for strategic planning, business intelligence, data collection and practice are the drivers for not only ownership of the data, but strategic management of that data to act in a highly probabilistic way of affecting a policy outcome.

In developing this dissertation, some existing work on data sovereignty by another organization helped to expand the scope of this framework. The U.S. Indigenous Data Sovereignty Network (USIDSN) which "unites and advocates for Indigenous Data Sovereignty at the tribal, state, national, and international levels" (n.d., para. 4). The organization has published a number of recent policy briefs called *Data Governance for Native Nation Rebuilding*. These briefs outline a process of 'decolonizing' data. They

assert that exercising a right to data sovereignty occurs at the data system level where other non-tribal entities control a tribe's data.

Rainie, Rodriguez-LoneBear, & Martinez (2016) are correct when they assert:

> In the United States, the processes of colonization have led to a state of data dependency in Indian Country. Federal policies of assimilation, forced removal, relocation, residential schooling and other cultural ruptures led many tribes to rely on external sources of information about their communities' economic, environmental, and health status. This data dependency produces a paradox of scarcity and abundance: extensive data are collected about tribes, but rarely by tribes or for tribal uses.

As a result:

- Existing Indigenous data are inconsistent, inaccurate, or irrelevant to tribal goals;
- The collection, ownership, and application of Indigenous data are controlled by external entities;
- An extensive history of exploitative research and policies has left a legacy of mistrust of data; and
- A lack of data infrastructure and capability cripples tribal efforts to overcome these obstacles

> Indigenous data sovereignty is an aspiration. (p. 2)

This framework has used many of the concepts created by USIDSN as inspiration for this manuscript. After an exhaustive examination of the articles, policy stances, and academic papers on the USIDSN website; there was one thing that was absent from this

collection of information: There was no representation of how to transition data sovereignty from theory into practice. This manuscript represents that shift.

What is Nation Building?

The Native Nations Institute (NNI), a unit of the University of Arizona Udall Center for Studies in Public Policy, hosts the Network and defines nation building:

> Nation building refers to efforts Native nations make to increase their capacities for self-rule and for self-determined, sustainable community and economic development. Nation building involves building institutions of self-government that are culturally appropriate to the nation and that are effective in addressing the nation's challenges.
>
> It involves developing the nation's capacity to make timely, strategically informed decisions about its affairs and to implement those decisions. It involves a comprehensive effort to rebuild societies that work. In other words, a nation-building approach understands that tribes are not merely interest groups, but governing nations confronting classic problems of human societies. (NNI, n.d., para. 1-3)

As much of the world knows, many of the over five hundred American Indian nations live in poverty and this has put a huge strain on not only economic viability, but social and cultural systems. Over the last quarter century, many tribes have faced desperate economic conditions and throughout the twentieth century federal policy and

tribal efforts have been governed by top down models of strategies that "view indigenous culture as a primary obstacle to development" (Cornell & Kalt, 2007, p. 8).

This has been defined as the *standard approach.* Cornell and Kalt (2007) have outlined developing new nation building approaches with tribes that have overcome these challenges through their research. In a rapidly changing world, economic development should develop as an organic process rather than short term, nonstrategic forms of planning and management.

Many Native nations have invented very different approaches and nation building in this context serves two purposes: asserting Indigenous rights to govern themselves and building foundational, institutional capacity to exercise those rights effectively, thereby providing an environment for sustained economic development (p.18). This manuscript seeks to use this nation building approach to build capacity through the data sovereignty framework.

In addition, data domains that can be designed with this framework have been conceived through scientific principles to serve a nation building purpose. The idea is to create the digital infrastructure that can create outcomes, and as a matter of nation building. It will also be provided to any tribe who wish to use the framework. This act constitutes the basis for a strategic plan to unify data in Indian Country. Without nation building, this outcome cannot exist.

The Concept of Tribal Sovereignty

Deloria (1976) as cited in (McKinley Jones Brayboy, Fann, Castagno, & Solyom, 2010) asserts in moving past legal/political conceptions concerning sovereignty: "Sovereignty is a useful word to describe the process of growth and awareness that characterizes a group of people working toward and achieving maturity. If it is restricted to a legal-political context, then it becomes a limiting concept, which serves to prevent solutions. The legal-political context is structured in an adversary situation which precludes both understanding and satisfactory resolution of difficulties and should be considered as a last resort, not as a first instance in which human problems and relationships are to be seen" (p.28).

The concept of sovereignty has been well documented and established through over a century of United States Federal Indian Law and Policy. The extensive body of literature contained in Felix Cohen's *Handbook of Federal Indian Law* is a monumental task to not only understand the complexity the role the federal government plays in developing nation to nation communique with tribes, notwithstanding state and local jurisdictions that must also contend with precedence in interpreting American Indian Law (Kessler-Mata, 2014).

As mentioned in the introduction, the scope of the case studies contained in this manuscript is to develop a tribal governance framework using data sovereignty as a mechanism for promoting appropriate data collection and practice in the context of data science. The framework design is a multidimensional approach that is anchored in the concept of *Tribal Sovereignty*, which is extended to the concept of *Data Sovereignty*.

Recently, Kessler-Mata's (2014) *A Constitutive Theory of Tribal Sovereignty: The Possibilities of Federalism* explained:

> Claims by tribes in the United States for the rights to exercise self-determination and self-governance are most often made through an appeal to the concept of tribal sovereignty. Tribal sovereignty is supposed to serve as both a justification for these rights (i.e. 'as tribes, we are sovereign entities and, therefore, ought to be able to exercise these rights'), as well as a guiding principle that enables tribes to delineate boundaries and authorities between themselves and other polities (i.e. 'as sovereigns, we are empowered with these unencumbered rights of governance').
>
> To claim that tribal sovereignty embodies a right to self-determination or a right to self-governance is to put forward a concept that does much more in theory that it does in practice. The concept of tribal sovereignty is one that promotes intergovernmental relations with non-tribal governments and which takes the principles of equitable interaction and political coordination as central to its operation. (p. 35)

Discussion

After an extensive review of the impact of Federal Indian Law and Policy, the impact of the Harvard Project on American Indian Economic Development, Corntassel's ideology concerning forced federalism, and Kessler-Mata's forward thinking on the *Constitutive Theory of Tribal Sovereignty*, a solution must be constructed that constitutes

all the ideas of sovereignty in terms of policy so data science can affect appropriate policy objectives.

I agree with Kessler-Mata in exploring additional praxes relative to equitable interaction, however one thing remains to forward this thinking: data collection and practice as it relates to sovereignty is an issue that cannot be ignored. Self-determination has always been a subjective idea whether it is through ideological constructs through Federal Indian Law or federalism. Arguments come down to one issue: to achieve parity with an ever growing set of ideas governing policy requires data to govern the process going forward. And although this manuscript does not focus explicitly on theories of self-determination or sovereignty directly, the framework design relies on the fundamental understanding of these concepts in developing key indicators for not only tribal governance, but additional indicators crucial to advancing sovereignty as a matter of data analysis.

Presently, there is no precedent that effectively addresses the current state of how tribes aggregate, maintain, or share data as an act of sovereignty. Since every tribal, federal, state, and local government is trying to contend with the exponential growth of data as it relates to strategic decision-making; tribes are in a unique position to provide their own data and analysis to strengthen their position in data driven decision-making using this framework.

The foundations of this idea will be addressed further in the key indicator of Tribal Governance as it plays a crucial role in advocating for more progressive ideas in addressing tribes as not only as political polities, but Native nations with unique citizens.

Separating the Science

This last section is a critical examination of the implications of the data sovereignty framework design. It is important to make the distinction that culture, tribal identity, and nation building are mutually exclusive objects that do not affect the outcome of the scientific methods presented here; rather it serves to create framework implications that provide appropriate context for discussion. Now that the primers for the framework have been established, the next step is to examine the scientific concepts that drive the actual process.

So What is Data Science?

Wikipedia defines data science as:

> Data science, also known as data-driven science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining.
>
> Data science is a 'concept to unify statistics, data analysis, machine learning and their related methods' in order to 'understand and analyze actual phenomena' with data. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, classification, cluster analysis, uncertainty quantification, computational science, data mining, databases, and visualization.

> In 1996, members of the International Federation of Classification Societies (IFCS) met in Kobe for their biennial conference. Here, for the first time, the term data science is included in the title of the conference *Data Science, classification, and related methods*, after the term was introduced in a roundtable discussion by Chikio Hayashi. ("Data Science", n.d., para. 1)

This paper Hayashi (1998) wrote called *What is Data Science? Fundamental Concepts and a Heuristic Example,* explains the basic idea: "Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data" (p. 40).

Since then, there has been much debate as to what data science actually entails. The exponential growth in the technology sector has expanded the use of the term data science. In the last five years, data science has mostly referred to engineers and analysts in computer science fields. The term continues to evolve and there have been many organizations taking ownership of the term.

The American Statistical Association (ASA) in 2015, sought to clarify the role of statistics in data science, they write:

> The rise of data science, including big data and data analytics, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means

that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science:

- Database Management enables transformation, conglomeration, and organization of data resources;
- Statistics and Machine Learning convert data into knowledge; and
- Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

Certainly, data science intersects with numerous other disciplines and areas of research. Indeed it is difficult to think of an area of science, industry, commerce, or government that is not in some way involved in the data revolution. But it is databases, statistics, and distributed systems that provide the core pipeline. At its most fundamental level, we view data science as a mutually beneficial collaboration among these three professional communities, complemented with significant interactions with numerous related disciplines. For data science to fully realize its potential requires maximum and multifaceted collaboration among these groups.

Statistics and machine learning play a central role in data science. Framing questions statistically allows us to leverage data resources to extract knowledge and obtain better answers. The central dogma of statistical inference, that there is

> a component of randomness in data, enables researchers to formulate questions in terms of underlying processes and to quantify uncertainty in their answers.
>
> A statistical framework allows researchers to distinguish between causation and correlation and thus to identify interventions that will cause changes in outcomes. It also allows them to establish methods for prediction and estimation, to quantify their degree of certainty, and to do all of this using algorithms that exhibit predictable and reproducible behavior. In this way, statistical methods aim to focus attention on findings that can be reproduced by other researchers with different data resources. Simply put, statistical methods allow researchers to accumulate knowledge. (pp. 1-2)

This manuscript is in line with this assertion. The constructs of these case studies utilize many of these collaborative definitions. For instance, in creating a smart solution in *Case Study 2*, these fundamental concepts allow machine learning to extract pixel data from raster images of high resolution satellite and drone imagery. The entire underlying process is mathematical and statistical in nature.

The Support Vector Machine (SVM) used to create the image classifications are algorithms run in software that performs the necessary calculations and then produces a visual representation of the classification. According to the definitions presented, this collection of tools used to exact this outcome is almost certainly an act of data science.

When I set out to create this framework using the term data science, this was not a coincidence. The proper use of modern terms and techniques are crucial when developing new educational praxis. Because data science has a very wide reach in data driven

decision making, creating a data domain through the data sovereignty framework is to actualize this concept in a wide variety of measures.

And although the case studies presented here involve spatial and Geographic Information Systems (GIS) analysis, the framework itself is not limited to only one subject of study. I would like to further comment on this.

As to what is statistical design theory and how does it specifically relate to the data sovereignty framework:

Statistical design theory was a generic way to emphasize the importance of data driven decision making. Too often, data is collected with no regard as to how the data should be analyzed, rather than could be analyzed. The framework key indicator Data Management encompasses the idea of design of experiments, data collection and practice, as well as establishing ways of continually refining and testing collected data to optimize explanatory power and minimizing error variance.

This was what is meant by statistical design theory: using current and established statistical principles to guide all types of data analysis. Data science is the act of incorporating these principles to exact an outcome. The benefits of this approach will undoubtedly provide more meaningful insights to data driven decision making. Since there are many stages of data collection or practice that are always ongoing in tribal organizations; I think it is important to consider strategic management and planning when considering how to leverage citizen science in a way that benefits the citizens doing the work. The brief workflow below is an example how I would approach an initial consultation with any client needing the expertise of a doctoral level data scientist.

To evaluate the strengths and weaknesses of any statistical process, it is imperative to develop a statistical methodology framework which usually follows a general hierarchy of:

1. Examining any descriptive statistics pertinent to a particular study
2. Develop and test a number of Exploratory Data Analysis (EDA) techniques that provide a basis for more complex methodology and inference
3. Using the information from the EDA, to create a formal inferential hypothesis for examining more complex processes that may exist beyond the first stage of the current project.
4. Repeat the process until a strategic assessment has been developed
5. Design a strategic plan to implement the data domain

In conclusion, the goal is to move beyond simple descriptive measures and begin a more robust process to help tribes attain sovereignty and policy outcomes that favor data driven decision making using citizen science. It is imperative for tribes to have advocates that not only promote their cultural capital, but design systems of data infrastructure that allow for realistic and achievable economic development and planning.

What is Citizen Science?

Citizen Science refers to the general public engagement in scientific research activities when citizens actively contribute to science either with their intellectual effort or surrounding knowledge or with their tools and resources.

The ongoing work in citizen science is another new constantly evolving concept. The European Citizen Science Association (ECSA) is one of the active organizations in

developing ethics and principles that guide best practices. The ECSA asserts “Sharing best practices and building capacity” in citizen science is a flexible concept that can be adapted and applied within diverse situations and disciplines.

Their key principles *Ten Principles of Citizen Science* (2015) have been developed by a number of their Association members to provide a set of guidelines when undertaking a citizen science task.

1. Citizen science projects actively involve citizens in scientific endeavor that generates new knowledge or understanding.
2. Citizen science projects have a genuine science outcome.
3. Both the professional scientists and the citizen scientists benefit from taking part.
4. Citizen science may, if they wish, participate in multiple stages of the scientific process.
5. Citizen scientists received feedback from the project.
6. Citizen science is considered a research approach like any other, with limitations and biases that should be considered controlled for.
7. Citizen science project data and metadata are made publicly available and where possible, results are published in open access format.
8. Citizen scientists are acknowledged and project results and publications
9. Citizen science programmers are evaluated for their scientific output, data quality, participant experience and wider societal or policy impact.
10. The leaders of citizen science projects take into consideration legal and ethical issues surrounding copyright, intellectual property, data sharing agreements, confidentiality, attribution, and the environmental impact of any activities. (p.1)

Using this set of guiding principles, the data sovereignty framework will attempt to leverage this concept in the context of empowering tribal citizens as members of their

nation to undertake the tasks defined by a given data domain. In the data domain design process, the digital infrastructure designed will be open source in nature and the detailed instructions can be constructed to allow for a stakeholder to undertake the task, while leaving my expertise to interpret and provide guidance.

Defining Citizenry in a Tribal Context

It is no easy task to define a term like Native-centric without acknowledging that the concept could be interpreted to promote exclusion. Rather the term reflects an inherent set of ideas that help create appropriate dialogue that defines examination in the context of governing philosophies brought on by culture.

Citizen science is an excellent way to define the idea of citizens doing work on behalf of their nation, and is very specific to that nation's set of expectations. It is a common misconception when referring to Native America or American Indian tribes or culture that exchange a set of global ideas without knowingly first understanding that by doing this, we have marginalized the unique aspect of tribes as independent nations.

For instance, how do we make the distinction between a framework that is indigenously inspired; but is not nation specific? How do we bridge the divide between being Native but also aware of a specific tribal nation's history and culture with respect to their citizens? We must be careful when speaking broadly about 'American Indian culture' in the context that there are hundreds of unique nations federally recognized by the United States government. We must also be mindful that in referencing the term indigenous as it relates to a broader context of not just American Indians, but first nations that located all over the world.

Furthermore, I am a citizen of my specific tribal nation. The question is; can I also belong to American Indian culture which in context marginalizes the uniqueness of each tribal nation? Thus, creating smart solutions for tribal communities means first creating scientific inquiries rooted in data science that is independent of any community or culture. The citizen science concept can capitalize on honoring tribal citizens doing the work on behalf of their nations rather than doing this on behalf of 'American Indian culture'.

A Simple Citizen Science Example

For instance, in the simplest case, say a tribal citizen working to understand how to create a cross-tabulation table for a report. Excel is not an efficient way to query that type data thus a few lines of code in R and instructions on how to run the code can be constructed to teach a stakeholder how to conduct the analysis on their own.

As tribes need specific data domains for solving data related problems, I will begin creating the SMART solution they require. It is the hopes that as a compendium of these solutions begin to grow; more tribes can utilize the solution through nation building.

The project outline in the next section outlines the path for further framework development.

Project Outline

These conceptual ideas in the introduction are the cornerstone of The Data Sovereignty Initiative development framework. To summarize, the key issues in building smart solutions using the framework and cultural implications are:

- American Indian History & Federal Indian Law and Policy
- Nation Building
- Higher Education
- Citizen Science

Assessing these topic creates will create a SMART solution. Quantifying solutions using computational statistics through data science is to adhere to the scientific implications outlined in the introduction.

Among some of the questions I sought to answer were:

- So how can we translate data science into the conceptualization of self-determination?
- How do we empower our Tribal communities to undertake data science tasks on their own?
- How do we unify a data platform across all of Indian Country?
- How is this accomplished using nation building?
- How important is higher education in using credible research in not only data analysis, but policy decisions?

The next section will begin the development of the framework using these design metrics to create multidimensionality through the use of key indicators, key descriptors, and research objectives.

Data Sovereignty Framework Development

Data sovereignty can be thought of as an initiative to provide a set of tools or smart solutions that when constructed as a collective framework, can empower Tribal governments to use data as matter of self-determination. Collectively, this idea can encompass any number of data driven decision making tools such as designing and implementing a Tribal census, managing natural resources and sacred sites, or developing data collection through statistical analysis to further a tribe's ability to make effective planning decisions. The outline of this exploratory process is organized around four key indicators. In addition, key descriptors are an extension of the four key indicators provided by the data sovereignty framework. First, we begin with the research objectives.

Research Objectives

The proposed design is a set of diagnostic tools used in the data sovereignty initiative framework to better understand issues not always addressed to the specific needs of each individual tribal group. Thus, a critical examination from a global and local perspective is necessary. The conceptual framework will align to the following objectives to test the validity of the proposed design.

Purpose - The framework design is to implement a model allowing more flexibility than previous and well-documented top-down models. Corntassel et al. (2008) maintains that in order to construct successful social and economic development systems; we must not exclude individual entrepreneurship and informal systems of government (i.e. clans, societies, etc.) that drive appropriate cultural practice on the local level. The framework includes global ideas of self-governance but also examines methods practical

implementation of diagnostic tools to measure success for each unique tribe through key indicators.

Dimension - The key to effective framework design is to allow for a multidimensional approach that provides depth in evaluating such a complex set of indicators aimed at not only understanding data management, policy analysis, and any number of associated data domains; but how this applies to the diversity of social, economic, and political structures on American Indian reservations. There is no feasible way to implement a top-down catch all model that maintains all tribal entities are the same.

However, the power of statistical data collection and practice provides a more practical approach to understanding the nature of dimension: some data provides descriptive measure, some data allows for more complex inference and prediction, and some data allows for machine learning. The key to applying multidimensionality is to understand the depth of knowledge of the stakeholders at each phase of any analysis to produce an effective design strategy to unify this information strategically.

Expected Use - In nation building, we can use a set of established guidelines to delineate administrative (tribal governance) versus local stakeholder interaction using a horizontal approach to design. It is clear designing strategies that do not take into account the dimensionality and complexity of each tribal group is what is at stake here. The goal is to orient the values and opinions of tribal stakeholders that represent the contingency of individuals that are doing the work, and to work for more effective design strategies that provide an administrative way to incorporate all global and local stakeholders in all levels of governance. It is important to account for all levels of interaction and to build sensible

policy from information gathered at all levels to be analyzed for the benefit of the community, not necessarily the individual.

Expected Benefit

The benefit to tribes in capitalizing on a framework that has been designed from the bottom up with a Native-centric view of policy, experience, and data analysis is paramount. At the heart of this analysis are tribes' ability to assert self-determination and sovereignty in a way that has not been tested: data driven policy initiatives challenge the anecdotal nature of what has historically never been a zero-sum game for tribes relative to federal, state, and local authority challenges to that sovereignty. It is the hope that this framework provides insight as to how to use data driven decision-making to enact meaningful and effective approaches to data sovereignty as it applies to any number of data domain relative to policy.

Framework Objectives

The goal of this analysis is to aid further development of the framework with a set of diagnostic tools to better understand how a data domain along with the three key indicators can be managed consistent with tribal sovereignty principles. The framework has been designed with ease of implementation in mind to provide more flexibility than previous and well-documented top-down models of governance and management strategies. The basis for this proof of concept framework will align to objectives aimed at testing the validity of the diagnostic tools by acquiring appropriate tribal and non-tribal stakeholder feedback concerning the design (as described in Case Study 1).

Analysis Steps

Step 1: Developing a Preliminary Framework

First, based on a review of the literature, four key indicators have been devised to capture critical elements of the Data Sovereignty Framework: Tribal Community and Culture, Tribal Governance, Data Management, and Data Domain Structures. Within each of these dimensions, key indicators have been identified and a series of questions devised to guide development of tribal specific framework metrics. Thus, the framework is intended to inductively generate a number of Tribal specific issues for data sovereignty and how this relates to the data domain of tribal transportation safety.

Step 2: Diagnostic Evaluation of Key Descriptors

Once identification of key indicators has been established, an examination of the key descriptors can begin. Every key indicator's descriptors remain constant to provide a way to uniformly assess the interaction of governance, community and data practices. The next step is to evaluate the level of fulfillment each descriptor plays in contributing to the selected data domain. For instance, if the tribal government has an existing agreement with a stakeholder in analyzing traffic data, then not only is nation to nation communique fulfilled, but data ownership, security and privacy may also have been established.

Step 3: Analysis of the Targeted Data Domain

Once the key descriptors have been examined, a more in depth analysis of the tribal traffic safety domain can begin. In design theory, an exploratory analysis will

provide an assessment of information already collected or need to be collected, design options, looking at operational capacity, or how to unify existing structures with the purpose of creating a unified set of practices to produce specific data driven outcomes. After examining the specific metrics of interest, a comprehensive plan can be produced to specifically address data sovereignty infrastructure as it pertains to a data domain.

The next section examines the preliminary framework design.

How to Develop a SMART Solution for Tribal Communities

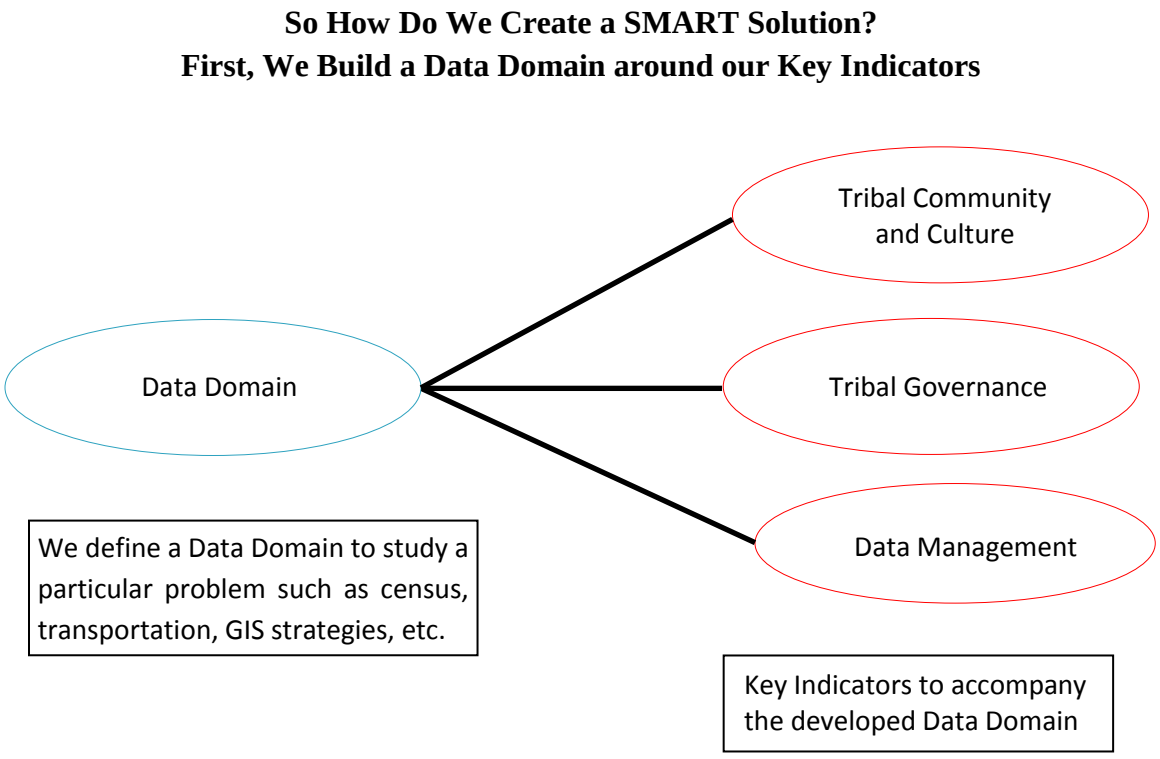


Figure 2.2 - An Example of How to Develop a SMART Solution for Tribal Communities

The Definition of a SMART Solution

The introduction mentioned when I first created the idea of the data sovereignty framework, the idea of a smart solution was taken from the current lexicon of the technology sector, which was inferred from the current generation *smart* devices being created and named as such.

As the framework was developing with the help of Case Study 1, the smart idea became SMART, an acronym for developing an administrative approach to inductively

assessing the operational capacity of any data domain and key indicators involved in an act of sovereignty.

This SMART technique is specifically used in constructing a data domain with achieving practical goals using statistical design theory at its heart. Once the data domain has been developed using this method, pairing a preliminary evaluation with the key indicators in Figure 2.2 is what creates a SMART solution.

Further development of what the SMART acronym represents comes from an administrative technique Doran (1981) developed in strategic management:

Specific	Target a specific area for improvement.
Measurable	Quantify or at least suggest an indicator of progress.
Achievable	State what results can realistically be achieved, given available resources.
Responsible	Specify who will do it.
Time-related	Specify when the result(s) can be achieved.

Realistic goal setting is what defines the feasibility of a project and this inductive approach that lays the groundwork for an initial assessment. The proposed case studies would follow a set of instructions to provide a more methodological review of how this process works in practice. Since strategic planning and business intelligence plays a major role in management decisions; this technique can provide an introductory approach to goal setting. All these ideas are again part of collective ideas that are at the Data Sovereignty Initiative's core set of procedures and philosophy. Although it may seem these concepts are disjointed, it is important to see these steps as a collective and holistic approach to nation building.

Preliminary Data Sovereignty Framework

Defining Key Indicators using the Data Sovereignty Framework

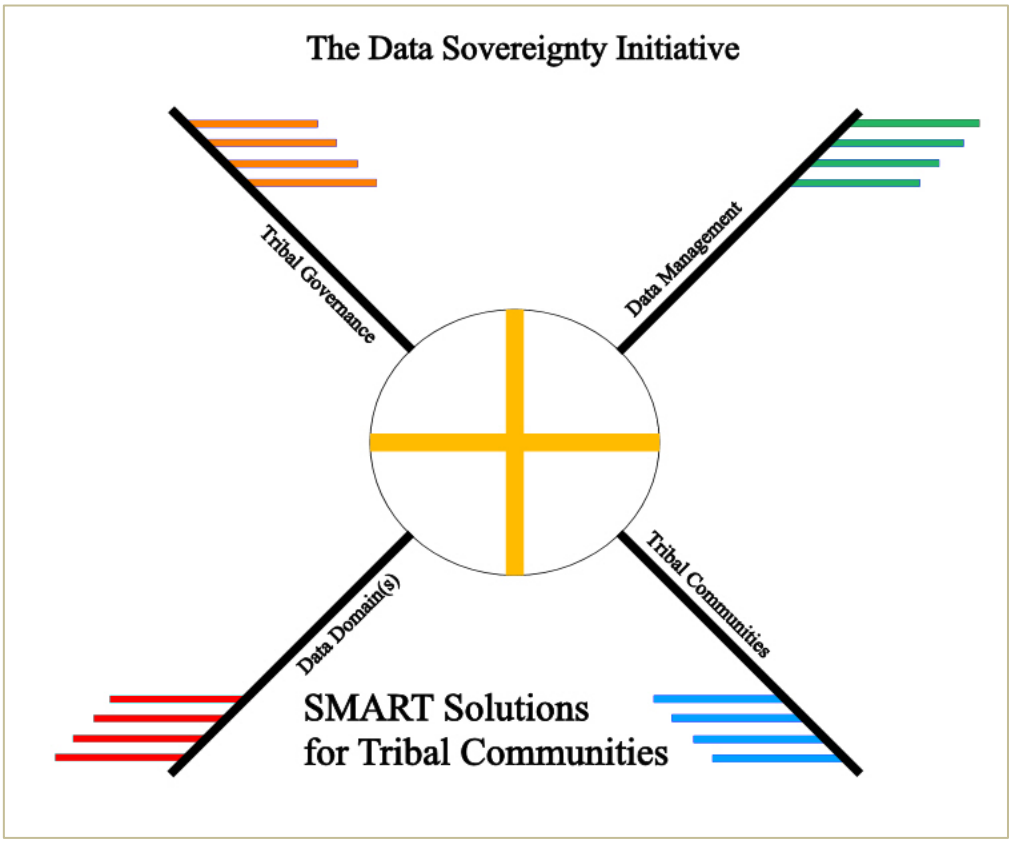


Figure 2.3 - Data Sovereignty Initiative Design Metrics

The framework design proposed is differentiating two sets of metrics that contain both data and cultural design aspects crucial to maintaining both dimension and scope when developing policy based decision-making. Each respective design metric has been developed to maintain unity in cultural practice that has a direct effect on Indigenous framework design theory (Cornell & Kalt, 2007).

Baseline metrics play the role of engaging stakeholders in an introductory exploration and requires a set of diagnostic tools to assess the strength of engagement across a tribal organization.

The Data Sovereignty Framework has four key Indicators:

1. Tribal Community and Culture
2. Tribal Governance
3. Data Management
4. Data Domains

Each dimension drives a decision-making process to allow for specific framework diagnostic questions to be developed using the paired key descriptors in Table 2.1.

Data Sovereignty Key Descriptors with Each Respective Key Indicator

Tribal Community and Culture	Tribal Governance	Data Management	Data Domain: Tribal Transportation Safety
History	Federal Indian Law and Policy	Data Collection and Practice	Quantitative Transportation Data
Culture	Nation to Nation Communique	Data Analysis	Qualitative Transportation Data
Cultural Values	Sovereignty	Data Ownership	GIS and Analytical Infrastructure
Citizenry	Self-Determination	Security and Privacy	Tribal Plans and Priorities

Table 2.1 - Data Sovereignty Framework Indicators with Key Descriptors.
the Data Domain was developed for Case Study 1

The design templates developed for the diagnostic tools in the next section rely on the key indicators and descriptors to remain fixed with the exception of data domain;

which can be regarded as the indicator that can encompass any task a tribe needs to analyze. This maintains the integrity of the framework design.

In addition, it is important to prioritize the cultural design metrics first. Allowing tribal citizens to contribute in an overall governance strategy is an important aspect in indigenous nation building: to be stable and effective in self-governing, governmental systems have to fit with ways tribal culture answers questions of who, what, where, and how. This is called cultural match: a fit with the shared norms of the community. In addition, cultural grounding is a critical element in legitimacy that makes wielding governmental authority a sacred trust, a sacred responsibility to serve the people and their interest in an appropriate way.

Although most organizations stress the importance of data driven metrics; too often analyses are compiled by organizations outside of the tribal community and the disconnect from what is perceived to be important in tribal communities may be much different than what the actual needs of tribal citizenry are. The next section defines this cultural and data relationship in broader terms.

The Four Key Indicator Definitions

These indicators are aligned to adhere to the research objectives and provide theoretical background information that is the basis of the diagnostic tool constructs in the following section. These definitions are a set of guidelines established by the current literature review, whether it is Federal Indian Law, governance strategies, or understanding research design principles.

Key Indicator 1: Tribal Community and Culture - Culture Does Matter

Conceptually, the reason many projects fail to deliver in many Tribal communities is that the one-size-fits-all model of governance is inappropriate given the diverse and complex nature of each tribal groups' history, culture, and identity. Ignoring this critical aspect in Indian country has had devastating effects and decentralizes important power structures defined by each tribe in their everyday affairs.

More broadly, the importance of *cultural match* maintains a necessity with consonance (match) between the structure of a society's formal institutions of governance and economic development and its underlying norms of political power and authority (culture) for those institutions to function and serve effectively. For this concept to "work", institutions must meet two tests: legitimacy in eyes of the citizens and practical efficacy (Kalt, et al., 2008)

The citizens of tribal nations often have the daunting task of compiling information for any number of things happening in day to day operations. The roles these individuals undertake can be complex and sometimes one individual is solely responsible for multiple jobs involving GIS analysis, IT, data analysis; thus human capital often gets stretched thin. The proposed framework was designed with tribal citizenry at its center for a number of purposes:

- To honor tribal cultural capital
- To obtain information from citizens at the ground level
- Through citizen science, information can be provided to aid in reducing workload

Thus, the unification of tribal voices may seem cliché; however, history has shown that the exclusion of citizens at any level creates fragmentation and the aim here is to study the quantitative effect of inclusion through these key indicators.

Key Indicator 2: Tribal Governance - Sovereignty Matters

Undoubtedly, matters of governance and culture in tribal communities relies on a careful assessment of "the process by which a community or nation improves its economic ability to sustain its citizens, achieve its sociocultural goals, and support its sovereignty and governance processes" (Begay, Cornell, Jorgensen, & Kalt, 2007). One strategy is to allow self-determination to be the vehicle that drives economic development that reflects a unique tribe's agenda in achieving those goals.

The scope of governance is functionally diverse and developing priorities that support development are not always achievable, so it is important to understand what strategies exist in each tribal government that maximizes the relationship between tribal governance policy and the community driving the work that reflects this dynamic.

Thus, this indicator is designed to understand the scope of each individual tribal system of governance. The delicate balance between the theory and practice in asserting sovereignty through self-determination is no easy task. But what is most important is this notion of equitable interaction, which is conceptually asserting sovereignty through agreements made with non-Indian entities and maintaining a more powerful position of negotiation through data collection and practice.

Key Indicator 3: Data Management – Data Ownership and Management Matters

The descriptors in the Data Management key indicator are a direct result of governance strategies that regard data and *Data Sovereignty* as the next step in providing safeguards for tribes in pursuing data ownership, security, and privacy. These framework inter-dependencies are the result of Native-centric design principles geared towards assisting Tribes with understanding the power of data analysis to bolster government to government communication, written agreements, and sustainable economic development.

Data sovereignty represents the highest quantifiable standard to which governance reflects native nation building. Data management is a hierarchy of processes that utilize the foundations of statistical design theory in managing the overall framework stability. Since data collection often begins with the need for situated, qualitative, and collaboratively produced data, the natural order of pursuing more advanced data techniques such as survey design, statistical modeling is an example of how key descriptors in this key indicator add dimensional structure as the framework continues to advance beyond simply an exploratory process.

Thus, data management is defined to be more than just information collection; it is the cornerstone of providing crucial data driven metrics important in asserting sovereignty through governance and economic development. Although this dissertation predominantly examines spatial statistical theories, the framework is designed to encompass all data science related theory and practice.

Key Indicator 4: Data Domains: An Open Way to Examine Data-driven Decision-Making

Finally, data domains are defined as any data collection process that occurs in Tribal communities that can be understood to have specific meaning in the context of governance, economic development, or operational capacity. Domains such as health care, transportation, education, enrollment, historic preservation, and census are all considered data domains. Since data domains can represent any project a tribe is interested in, it is implied that the four key descriptors will change depending on the choice of data domain. The dimensionalities of these domains are vast and can encompass any level of analysis from descriptive to analytical. The choice of descriptors is then a matter of assessing each indicator as it pertains to the domain.

Discussion

As to the feasibility and potential impact of creating and utilizing data domains that encompass data originating with multiple tribal groups, possibly even groups that are national in scope:

When I first began exploring the possibilities of using data as a matter of sovereignty, I had not realized how important this concept was in strategic planning. Data domains in this framework have key indicators designed specifically to address feasibility from two key points of view.

1. Any data domain developed is mutually exclusive from the data collected for each unique tribe using that particular data domain. However, once a data domain is created the nation building component from the framework assures the use of that

data domain as a model for other tribal groups to use to collect similar information. For example, if a data domain is created in order to provide a data collection process for census, then this digital infrastructure becomes available to other groups to utilize should they want to pursue their own census. The goal is to eventually host data domains as an open source digital infrastructure on Github for anyone to access and utilize in their own capacity should they desire that. Thus, any national organization can access this information, not just tribes.

2. Cultural key indicators represent one way to address data originating from an individual tribe; however it is up to the tribal group to decide what information is shared or kept private. The data management key indicators have two specific key descriptors that address data ownership, security, and privacy. The hope is through equitable interaction, these key indicators pave the way for meaningful policy on how tribes share data in nation building but have safeguards in place to share only information they choose.

Thus, the potential impact is a unified digital infrastructure that can be used by anyone to advance data collection and practice with a particular statistical design structure so the data can be compared across any organization that utilizes a given data domain.

As to the political, technical, statistical, and other barriers that might stand in the way of this:

Like with any proof of concept, there are potentially many barriers that could have unintended consequences. Tribal government structures have unique and often

volatile political substructures that cannot be simply tackled from a conventional standpoint. Another potential barrier is data collection and practice from a data sovereignty position is very new and adoption of high level statistical thinking has a high potential for resistance amongst communities with low levels of educational attainment. Lastly, the technology needs of each tribal group is potentially diverse and developing open source code base in R for deployment of more complex data analyses could present too high a learning curve for individuals conducting the research on behalf of their tribe.

The data sovereignty framework design has considered many of these barriers. The key indicators and concepts provide more rigorous systems of evaluation such as examining tribal governance, involving tribal citizenry in the data domain design as well as leveraging citizen science to provide advanced statistical techniques in a workable form for individuals to implement the data domain, rather than design it from the bottom up.

As to specific situations or classes of situations for which the potential benefits would justify expending the resources needed to overcome those barriers:

Data is so crucial to economic development and planning since federal funding formula calculations are now highly dependent on measurable outcomes. The data sovereignty framework is designed as cost effective set of solutions. The nation building component was designed to give the intellectual property away at no cost as part of a creative common license. A good example of this concept is case study two. The potential benefit for tribes obtaining a master address file is not simply for census. The points in space not only provide locations of critical infrastructure, but can also be used to

map other processes at the same time such as tribal housing utilities, 911 locations, propane locations, etc. In addition, more advanced techniques of spatial modeling can be applied for examining clustering trends, simulations, and predictions. Thus, expending resources on a high resolution satellite imagery or drone imagery to help in the analysis would be a great financial return on investment.

Diagnostic Tool Development Templates for Evaluation

The construction of diagnostic tools to accompany the data sovereignty framework is to put theory into practice. The set of tools were designed to understand the collective nature of the connectedness of tribal cultural values, the importance of governance, and the impact of quantifying information as a matter of sovereignty: making data driven decisions a strategic measure in how a tribe asserts its authority is no different than any organization wishing to use analytics as the foundation of providing irrefutable evidence governing their decision making for sound policy decisions.

The tools created below are general guidelines to provide a position to begin evaluating the strengths and weakness of a tribe's information technology and how this plays a role in quantifying their current techniques with what is currently available. The overarching goal has always been to unify a system of information designed for nation building.

Diagnostic Tools for Data Sovereignty Framework

The following indicators contain a narrative that aligns to the key indicator definitions presented in the last section.

Each indicator attempts to address multidimensional diagnostic topics for evaluation namely:

- The importance of the key indicators collectively
- The current condition relative to the data domain
- The implications of current policy decision making
- Possible actions to accompany the future decision making
- Question clouds are utilized for developing a sound hypothesis or research question.

The data domain presented in this dissertation is specific to tribal traffic safety, so it is important to clarify the data domain is in general the subject of any tribal related project; and the name and key descriptors can reflect any number of data related topics. Using the evaluation techniques presented should provide enough information to devise a preliminary synopsis of the state of the data domain and how it can be used to assert sovereignty in data related activities. The next section illustrates how each key indicator functions in practice.

[Key Indicators Section]

Key Indicator 1: *Tribal Community and Culture*

Key Descriptors:

History	Culture	Cultural Values	Citizenry

Why is it important? Culture Does Matter

Cultural values play an important in decision making. The importance of consulting elders, stakeholders, and citizens is to be Native. The history of excluding voices that speak for local clans, districts, or inter-tribal groups is a product of assimilation and acculturation. To understand the voices of the community requires a commitment to knowing each tribe as unique entities with a set of cultural values and citizenry that when given a proper voice would contribute to a betterment of their community if given the opportunity. Each unique history provides context to decision making.

Communities still face issues of historical trauma, lack of stable governments, poverty, and lack of representation. This framework provides a well-developed process of data collection and practice aimed at collecting information from tribal citizens that do the work of the community in addition to providing input to tribal officials.

Descriptors:

History

What particular tribal history is important to understand when developing a data domain?

How do non-Indian stakeholders hold themselves accountable to understanding cultural history and context in working with tribes?

How do tribes maintain the context of relevant historical record when making planning decisions?

Key Indicator 1: *Tribal Community and Culture*

Key Descriptors:

History	Culture	Cultural Values	Citizenry

Descriptors:

Current Culture

What cultural beliefs should be understood when making decisions?

What is current state of community culture and perceptions?

How can we support these stakeholders through data-driven initiatives using cultural capital?

Cultural Values

Who are the elders you think can contribute most effectively injecting additional cultural values in decision making?

Citizenry

How is the citizenry divided (i.e. groups, clans, or districts)?

What is the relative size of the citizenry?

When it comes to local stakeholders, who are these stakeholders (i.e. GIS analyst, tribal law enforcement, or local citizens)?

What are potential challenges local stakeholders have in tribal administrative involvement?

Key Indicator 2: *Tribal Governance*

Key Descriptors:

Federal Indian Law & Policy	Nation to Nation Communique	Tribal Sovereignty	Self-Determination

Why is it Important? Sovereignty Matters.

Tribal governments who assert sovereignty through self-determination face innumerable challenges that can strengthen or weaken their position in negotiations. Tribes do have inherent rights to manage their affairs as they see fit, but in the era of forced federalism, the need for equitable interaction and political capital is an absolute necessity when making agreements with non-Indian stakeholders.

In the modern era, sovereignty and self-determination are tools tribes utilize to maintain their unique nationhood directly from the body of Federal Indian Law established from two centuries of negotiation, sacrifice, and cultural identity. This historical precedence is why tribal governments fight so hard for their nationhood.

Tribes who assert sovereignty in nation to nation communications by forming compacts, memoranda of understanding, or informal agreements must fully understand there are policy tools non-Indian policymakers use that have a direct effect on not only policy, but political capital as well. The two types of policy tools that have a direct effect on negotiations are regulatory and capacity-building.

Regulatory policies are used when policymakers view emerging contenders as a threat to economic or political well-being; while in contrast capacity-building tools are intended to strengthen communities by enhancing tribal powers of self-determination. Capacity-building tools are the cornerstone of the data sovereignty framework because it comes the closest to describing Indigenous nation building strategies.

Key Indicator 2: *Tribal Governance*

Key Descriptors:

Federal Indian Law & Policy	Nation to Nation Communique	Tribal Sovereignty	Self-Determination

Descriptors:

Federal Indian Law and Policy

Is there appropriate Federal Indian Law precedence in exploring the current data domain?

What is the position the tribal government takes in data-driven policy initiatives?

Nation to Nation Communique

Are their existing MOA/MOU/Compact agreements that the tribe has used in the past?

Sovereignty

What is the governance structure of the Tribal government?

How can data sovereignty assist tribes with a stronger position in data driven decision making?

Self-Determination

What types of sovereignty are easier to assert than others?

Are there cases where a policy victory can be guaranteed using self-determination?

Are there specific concerns tribes can be assisted with in asserting self-determination from positions of Federal Indian Law and Policy, established compacts, or Memoranda of Understanding (MOU)?

Key Indicator 3: *Data Management*

Key Descriptors:

Data Collection Practices	Data Analysis	Data Ownership	Security and Privacy

Why is it Important? Unifying Data in Indian Country is Paramount

Data management is crucial in organizing information so meaningful outcomes can be achieved. The framework was designed with the intent of creating a structured look at data in developmental stages of capacity. The key descriptors can be divided into two groups: quality of data practices as it relates to the data domain and management of the data as for security purposes. In addition, this indicator also attempts to identify the types of technologies that are used in a tribe's data collection process which could be data storage systems or software related to analysis such as ArcGIS, SAS, or R. The importance of these baseline metrics provides informational capacity so when an exploratory process is designed, it aligns to the best possible outcome for the not only the project, but governance strategies as well.

At every level tribes are collecting data. Some the data is structured some it is unstructured. This framework seeks to understand a tribe's operational capacity of data collection, practice, as well as security and privacy. In addition, all tribes have some sort of technological infrastructures in place, and so the purpose of this indicator is to understand the strengths and weaknesses as it relates to the data domain. Thus, much of data management revolves around structuring data based on a tribe's current infrastructure to investigate which stage of development is appropriate from the analysis processes described in the last section (i.e. descriptive, exploratory or advanced analysis).Learning how to create strategic solutions from either descriptive or inferential topics in strategic planning will help tribes implement data solutions to minimize errors in measurement, or to invest software platforms capable of making data collection or

Key Indicator 3: *Data Management*

Key Descriptors:

Data Collection Practices	Data Analysis	Data Ownership	Security and Privacy

analysis easier. Data sovereignty conceptually is creating platforms to utilize data as an Indigenous nation building tool for equitable interaction.

Descriptors:

Data Collection Practices

What are the current data collection practices (i.e. manual or digital data collection)?

Who collects the data? Where is the data stored?

Data Analysis

What is the quality of tribal data analyses?

Do tribes rely on consultants or in-house personnel for analyzing data?

Does the tribe have any software platforms for data collection, analysis, or reporting?

Data Ownership

Do tribes have a policies governing data ownership, whether it is with existing or future data collection?

Is there Federal Indian Law precedence that prevents tribes from taking full ownership of data when engaged in federal, state or local jurisdictional agreements?

Who uses and reports specific data to the tribe?

Security and Privacy

Are there privacy policies in place? Are their data sharing policies in place?

Are there data encryption protocols in place?

Are there protocols in place in regard to theft of data through cyber-attacks?

Key Indicator 4: *Specified Data Domain*

Key Descriptors: Tribal Transportation Safety

Quantitative Transportation Data	Qualitative Transportation Data	GIS and Analytical Infrastructure	Tribal Plans and Priorities

Why is it Important? Data Ownership and Management Matters

An accompanying report, *Using GIS to Improve Tribal Traffic Safety: A Statistical Evaluation of Hot Spots on Minnesota Tribal Reservation Areas* was the basis for developing these descriptors. Hot spot analysis was a prototype application developed as a matter of assessing the Tribes interest in implementing GIS applications for planning, analysis, and programming in transportation safety. The report also provided additional inferential topics to strengthen further development of not just tribal traffic analysis, but point process model development and network analysis in current transportation safety literature.

The American Indian reservation system is not entirely disjoint from the regular business that occurs in areas within a reasonable distance to Tribal affairs, and traffic related accidents are very relevant to the location of services within the immediate vicinity of townships that border the reservation. Development of this data domain into a proof of concept was the purpose of the report and to aggregate any number of considerations that create a unified framework in order to work with Tribal governments in developing GIS analysis as a practical and efficient way to improve transportation safety through equitable interaction with state and local officials.

Descriptors:

Quantitative Transportation Data

What is the source of data used for Tribal safety planning and analysis?

Do Tribes have ready access and capacity to analyze this data?

Key Indicator 4: *Specified Data Domain*

Key Descriptors: Tribal Transportation Safety

Quantitative Transportation Data	Qualitative Transportation Data	GIS and Analytical Infrastructure	Tribal Plans and Priorities

Descriptors:

Qualitative Transportation Data

What qualitative data is available for understanding transportation safety issues?

Do Tribes have read access and capacity to analyze this data?

GIS and Analytical Infrastructure

In developing traffic safety, what tools are necessary?

Does the tribe have a GIS department?

Do tribal stakeholders have access to MnCMAT should data sharing?

Has there been on-going data collection on traffic issues within the reservation boundaries?

How can tribes use hot spot analysis as a potential prototype for tribes to provide input?

Tribal Plans and Priorities

What priorities do tribes have for traffic safety?

How well has the data been used to inform these priorities?

Have traffic safety plans and programs been developed using this data?

Final Discussion

The *Data Sovereignty Initiative* framework has been a continued work in progress. As I began the design of this framework, I had presented an initial talk here at SDSU called Consider the Century. It was an annual set of talks based Native American perspectives. I outlined the ideas of honoring American Indian history and cultural through developing a statistical framework. As this idea continued to develop, I started attending different conferences to maybe get some more ideas how to proceed.

As I will describe in the next chapter; there was a turning point in my research when I attended the 2016 Environmental Systems Research Institute (ESRI) User Conference. During the conference, I had made contact with a professor who became interested in my work. As they say, the rest is history. The current partnership I have with Claremont Graduate University and the Road Safety Institute has played an integral part in developing this framework as a proof of concept.

And as such, two of the components that I have designed contributed to the project: *Using GIS to Improve Tribal Traffic Safety* and were also two key components of this manuscript. The first component was designing the framework around the key indicators I had been working on and then developing a data domain around traffic safety which is presented as a prototype in the key indicators section.

The second component is the work I present in the next chapter. It is a *Tribal Traffic Safety Manuscript Brief*. As I first began work on the traffic safety project, an initial hot spot analysis had already been completed. The brief I wrote was to perform additional statistical analyses to evaluate the original analysis. I essentially created a

literature review and performed a number of exploratory data analyses (EDA) which was designed to explore spatial point patterns and processes further. The next section, presents this analysis with a brief literature review to introduce some concepts contained in the brief.

CHAPTER 3

CASE STUDY 1: Using GIS to Improve Tribal Traffic Safety

A research project funded by the Road Safety Institute (RSI) through the Center for Information Systems and Technology, Claremont Graduate University.

Dr. Thomas Horan, Principal Investigator

Dr. Brian Hilton, Clinical Associate Professor

Background

As I was in the early stages of developing the Data Sovereignty Initiative, I had elected to attend the 2016 ERSI User Conference (UC), which is the company's yearly gathering of GIS professionals from all over the world. Since my research had been in spatial point pattern, processes, and modeling; I felt this conference would provide more ideas as to where to take my research. During the registration process, I was asked if I would like to comment on anything ESRI would like to know about me.

I candidly described my role as an American Indian data scientist working towards creating SMART solutions for tribal communities and why I was attending the conference. ESRI president Jack Dangermond expressed his interest in my work, and I was encouraged to submit a story map for the ESRI UC's Tribal Story Map Challenge. The map I created, *The Impact of Data Sovereignty on American Indian Self-Determination* was a fundamental primer in constructing this manuscript and is mentioned in the appendix.

The map itself was an introduction to my development framework mainly looking at American Indian history and understanding context in moving towards a unified idea of quantifying this information using computational statistics.

The efforts in the story map challenge resulted in a third place finish for my map. This recognition was what catapulted my original idea for this manuscript into reality. As I described before, my department head, Dr. Kurt Cogswell has been extremely supportive of designing innovative projects, particularly with the goals of creating a support system to serve local American Indian students in succeeding within the mathematics and statistics department.

I have previously described the importance of context when our higher education institutions seek to develop new strategies in supporting underrepresented minorities. The results of my attendance of the 2016 ESRI UC also resulted in a partnership with Claremont Graduate University (CGU) in Claremont, California. Dr. Thomas Horan, the Dean of the Drucker School of Management and Director of the Center for Information Systems and Technology at CGU had presented the results of a partial analysis of *Using GIS to Improve Tribal Traffic Safety* project in the Tribal Tract section of the 2016 ESRI UC.

Dr. Horan is the principal investigator of this project in conjunction with the Road Safety Institute (RSI). The Roadway Safety Institute is the Region 5 University Transportation Center (UTC) led by the University of Minnesota. The institute conducts research, education, and technology driven by goal of preventing crashes to prevent fatalities and life-changing injuries (RSI Overview, n.d.).

During the 2016 ESRI UC, I had an opportunity to speak with Dr. Horan about his methods during the conference. Subsequently, the results of our interaction resulted in an offer for me to contribute my expertise to the project. Dr. Horan expressed interest in incorporating my data sovereignty framework ideas into the on-going design of the *Using GIS to Improve Tribal Traffic Safety* project. This subsequent partnership has been quintessential in bringing this manuscript from theory into proof of concept.

The basis for the analysis I present in this chapter is a statistical evaluation of a hot spots that Horan and Hilton (2016) had conducted as one metric in the project.

The two goals of working with this grant, was to,

- Provide additional statistical methods to support the existing Getis-Ord G_i^* analysis
- Provide recommendations for a more robust way of analyzing tribal traffic safety frameworks for tribes
- Design a proof of concept from the data sovereignty framework to obtain feedback from tribal interviews as to the strengths and weaknesses of the framework.

This dissertation has presented a number of interesting challenges. First, my intent with this dissertation was to create something that could be used in real life. Second, since I had not anticipated that my contribution in this project would drive the main design of this manuscript; careful thought was needed as to what to present.

Since I have already presented the influence of what this project has had on the framework I presented in the last chapter; I have elected to present the findings of the

Tribal Traffic Safety Manuscript Brief I initially submitted to Dr. Horan in order to further examine the statistical properties of the point pattern data.

The manuscript brief was designed as an *Exploratory Data Analysis* to accompany the original analysis presented at the 2016 ESRI UC, and to fulfill the project objectives of the Road Safety Institute requirements.

I had debated whether to provide in-line additional literature review within the original document but since this was part of another project, I have elected to keep the manuscript brief intact to stand on its own. The additional literature review I have provided in the next section outlines the topics covered in the manuscript brief, so there is a more complete theoretical examination of each exploratory technique used in further examining the results of Getis-Ord G_i^*.

Statistical Literature Review of the *Tribal Traffic Safety Manuscript Brief*

This review sets the general order of how a spatial process is analyzed. One question that arose from the original analysis was to understand if the crash data obtained from the Minnesota Crash Mapping Analysis Tool could be used in developing traffic safety on or near Minnesota tribal reservation areas.

Any test for spatial association should carefully examine the dependence structure in spatial patterns (Getis & Ord, 1992). The manuscript brief is comprehensive examination of many topics that were used to assess the Getis-Ord G_i^* hot spot results. GIS based methodology for fatal, injury, and pedestrian crash cluster mapping and analysis of crash data begins with understanding the underlying spatial dependence of points in space.

Methods for detecting *hot spots* of point processes are divided into two classes, first order and second order effects. The first set of effects focus on the underlying properties of point events and measures the variation in the mean value of the process. The second order effects focuses mainly examining the spatial interaction or dependency structure of point events for point patterns on the local level (Xie & Yan, 2008).

Understanding Point Patterns

Baddeley et al. (2016):

> A 'spatial point pattern' is a dataset giving the observed spatial locations of things or events. Examples include the locations of trees in a forest, gold deposits mapped in a geological survey, stars in a star cluster, road accidents, earthquake epicenters, mobile phone calls, animal sightings, or cases of a rare disease. The spatial arrangement of points is the main focus of investigation.
>
> Interest in methods for analyzing such data is rapidly expanding across many fields of science, notably in ecology, epidemiology, geoscience, astronomy, econometrics, and crime research.
>
> Statistical analysis of the spatial arrangement of points can reveal important features, such as a tendency for gold deposits to be found close to a major geological fault, or for cases of a disease to be more prevalent near a pollution source, or for bird nests to maintain a certain minimum separation from each other. Analysis of point pattern data has provided pivotal evidence for important research on everything from the transmission of cholera to the behavior of serial killers to the large-scale structure of the universe.

Marked Point Patterns

> The points in a point pattern may carry all kinds of attributes. A forest survey might record each tree's location, species, and diameter; a catalogue of stars may give their sky positions, masses, magnitudes, shapes, and colors; disease case locations may be linked to detailed clinical records. Auxiliary information attached to each point in the point pattern is called a mark and we speak of a marked point pattern (pp. 3-7).

The simplest example of a marked point pattern arises when the mark attached to each point is a single categorical value. In a hotspot analysis, this would value would be the type crash or the number killed in the accident. Multiple marked point categories are called a multi-type point pattern.

Understanding Spatial Dependence

Conventional statistical analysis frequently imposes a number of conditions or assumptions on the data it uses. Foremost among these is the requirement that samples be random. The most fundamental reason that spatial data are special is that they almost always violate this requirement. The nonrandom distribution of phenomena in space has various consequences for conventional statistical analysis.

For example, the usual parameter estimates based on samples that are not randomly distributed in space will be biased toward values prevalent in the regions favored in the sampling scheme. As a result, many of the assumptions we are required to make about data before applying statistical tests become invalid (O'Sullivan and Unwin, 2010).

Specifically, spatial dependence is a property of a spatial stochastic process in which the outcomes at different locations may be dependent. The technical term describing this problem is spatial autocorrelation. Spatial autocorrelation is probably one of the most well developed concepts in geographic information analysis and is particularly important when taking into account effects of temporal and spatial proximity in spatial analysis.

Anselin and Bera as cited in (Plant, 2012) provide a concise verbal definition: "spatial autocorrelation can be loosely defined as the coincidence of value similarity with locational similarity" (p. 59). Plant (2012) provides a more formal definition as well: "A nonzero spatial autocorrelation exists between attributes of a feature defined at locations *i* and *j* if the covariance between feature attribute values at those points is nonzero...

If this covariance is positive (i.e., if data with attribute values above the mean tend to be near other data with values above the mean), then we say there is positive spatial autocorrelation; if the converse is true, then we say there is negative spatial autocorrelation. Positive autocorrelation is much more common in nature, but negative autocorrelation does also exist" (p. 59).

The theory behind calculating spatial autocorrelation relies on a measure of neighbor distance so that when the spatial lags of distance *d* reaches a certain distance the autocorrelation goes to zero. It is entirely possible to surmise that if we have a cluster of points in say, a township and we begin looking for neighbors of say a radius *d*, that it will not take long before each defined spatial lag goes farther and farther out into an area where there are no related set of points that are similar to the original.

The conclusion that can be drawn from this observation is that the point process is most likely not homogeneous, and that there exists some sort of spatial dependency between the choice of boundaries and proximity of related events.

Spatial autocorrelation is the foundation that allows for examination of many other exploratory techniques that form the basis of spatial analysis. Techniques such as tests for complete spatial randomness (CSR) using quadrat analysis, kernel density estimation including hot and cold spot techniques using the Getis-Ord Gi^* are all examples.

Getis-Ord G_i^* Statistics

The original analysis of focused solely on the spatial analysis of hot spots. The hot spot tool in ArcMap uses the Getis-Ord G_i^* statistic for each feature in dataset. It is a technique that measures the local spatial association of a concentration of weighted points and all other points within a radius of distance d .

The Getis and Ord (1992) defines the statistic as:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij} x_j - \bar{X} \sum_{j=1}^{n} w_{i,j}}{S \frac{\sqrt{n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} X_{i,j} \right)^2}}{n-1}} \qquad 3.1$$

where x_j is the attribute for feature j, $w_{i,j}$ is the binary spatial weight between i and j , n is equal to the total number of features with:

$$\bar{X} = \frac{\sum_{j=1}^{n} X_j}{n} \qquad 3.2$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} x_j^2}{n} - (\bar{X})^2} \qquad 3.3$$

The resulting G_i^* statistic is a z-score. The score tests significance against the following hypothesis:

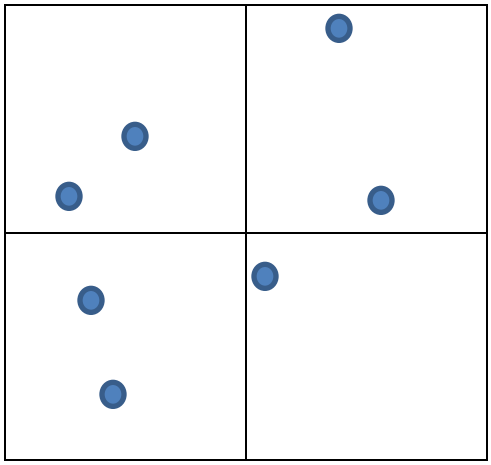
H_0 : No association at site *i*, and its neighbors of distance *d*
H_1 : An association does exist

The resultant z-scores and p-values indicate where features with either high or low values cluster spatially. The G_i^* statistic returned for each feature in the dataset is a z-score. For statistically significant positive z-scores, the larger the z-score is, the more intense the clustering of high values (hot spot). For statistically significant negative z-scores, the smaller the z-score is, the more intense the clustering of low values (cold spot).

In original analysis, the hotspot analysis calculated the number of people killed in all accidents in the roadways. Thus, a cold spot represents when an accident occurs and no one was killed. A hot spot indicates clustering of fatal accidents of one or more fatal accidents. The next technique performed was a test for complete spatial randomness called a quadrat analysis.

Quadrat Analysis

When a point pattern has an underlying stochastic process, we call this a process a description of how a spatial pattern might be generated. If we sketch a region of space A (Figure 3.1) which has been subdivided into tiny squares, the number $n(\mathbf{X} \cap B)$ of random points falling in A is equal to the sum of the numbers falling in the squares inside B.



Region B

Figure 3.1 - A Point Pattern of Quadrats Used in Determining CSR

Assuming, as above, that the outcomes in different squares are independent, and that there is negligible chance that some squares have more than one point, $n(\mathbf{X} \cap B)$ is the number of successes in a large number of independent trials, each trial having a small probability of success. By a famous theorem in probability theory, this means that $n(\mathbf{X} \cap B)$ has a Poisson distribution.

This distribution is used in finding rare events

$$P\{X = x\} = \frac{\lambda^x}{x!}\exp(-\lambda) \text{ where } i = 0,1,2... \qquad 3.4$$

where λ is the intensity parameter of a random variable X

A homogeneous Poisson point process is referred to *complete spatial randomness* (CSR) with intensity $\lambda > 0$ is governed by three properties:

- Homogeneity, the number of $n(\mathbf{X} \cap B)$ of random points of falling into a test region B has mean value $E(\mathbf{X} \cap B) = \lambda |B|$
- Independence, the test regions $B_1, B_2, ...B_m$ which do not overlap, the counts $n(\mathbf{X} \cap B_1)...n(\mathbf{X} \cap B_m)$ are independent random variables
- Poisson distribution, the number of $n(\mathbf{X} \cap B)$ of random point falling in a test region B has Poisson distribution (3.4)

A simple way to check for inhomogeneity is to check whether regions of equal area contain roughly equal numbers of points (as they must do if the point process is homogeneous). In quadrat counting, the observation window A is divided into sub-regions $B_1, ..., B_n$ called quadrats. For simplicity, suppose the regions have equal area.

We count the numbers of points falling in each quadrat, $n_j = n(\mathbf{x} \cap B_j)$ for $j = 1, ..., n$ Since these counts are unbiased estimators of the corresponding expected values $E\left[n(\mathbf{X} \cap B_j\right]$, they should be equal 'on average' if the intensity is homogeneous. Thus, the null hypothesis is that the intensity is homogeneous,

and the alternative hypothesis is that the intensity is inhomogeneous in some unspecified fashion (Baddeley et al., 2016, p. 134-135).

In the manuscript brief, this technique was used to visually assess the homogeneity of the traffic crashes, but did not do a formal test for CSR, although there is an easy way to detect this.

A measure for how well an observed distribution of quadrat counts fit a Poisson prediction is based on the property that the mean and variance are equal ($\mu = \sigma^2$). The *variance-mean ratio* (VMR) is defined a

$$VMR = \frac{Var(X)}{mean(X)} \qquad 3.5$$

and the expected value of a VMR values is 1.0 if the distribution is Poisson (O'Sullivan & Unwin, 2010). The hypothesis test is as follows:

H_0 : The underlying process is CSR, or homogeneous
H_1 : The underlying process is inhomogeneous in some specified pattern

The product $(n-1)\text{VMR}$ where n is the number of quadrats is a chi-square test statistic and follows a chi-squared distribution. The result of this test is the probability that a value as large or larger than the observed VMR could occur under the null hypothesis.

Kernel Density Estimation

Kernel Density Estimation (KDE) is a popular method for analyzing first order properties of a point event distribution. This technique is common for analyzing traffic accidents. Conceptually,

The general form of a kernel density estimator in a 2-D space is:

$$\lambda(s) = \sum_{i=1}^{n} \frac{1}{\pi r^2} k\left(\frac{d_{is}}{r}\right) \qquad 3.6$$

where $\lambda(s)$ is the density of at location s, r is the search radius or bandwidth of KDE. Only the points within r are used to estimate $\lambda(s)$ and k is usually modeled as a function called a kernel of the ratio between d_{is} and r. Instead of choosing a uniform function that gives equal weight to all points with bandwidth r, the idea is to model a "distance decay effect". The farther away a point is from location s, the less a point is weighted for calculating the density (Xie & Yan, 2008).

In this case study the spatial analyst toolbox in ArcMap, *How Kernel Density Works* (n.d.) uses the following algorithm:

> The algorithm used to determine the default search radius, also known as the bandwidth, is as follows:
>
> 1. Calculate the mean center of the input points. If a Population field other than NONE was selected, this and all the following calculations, will be weighted by the values in that field.
> 2. Calculate the distance from the (weighted) mean center for all points.
> 3. Calculate the (weighted) median of these distances, D_m.
> 4. Calculate the (weighted) Standard Distance, *SD*.

5. Apply the following formula to calculate the bandwidth:

$$\text{SearchRadius} = 0.9 \cdot \min\left(SD, \sqrt{\frac{1}{\ln(2)}} \cdot D_m \right) n^{-\frac{1}{5}} \qquad 3.7$$

where SD is the standard distance and D_m is the median distance, and n is the number of points if no population field is used, or if a population field is supplied, n is the sum of the population field values. Kernel Density calculates the density of point features around each output raster cell.

Conceptually, a smoothly curved surface is fitted over each point. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the Search radius distance from the point. Only a circular neighborhood is possible. The volume under the surface equals the Population field value for the point, or 1 if NONE is specified. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell center.

The kernel is based on the quartic bi-weight kernel function described in Silverman (1986):

The multivariate kernel density estimator with kernel K and window h is defined by:

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left\{ \frac{1}{h}(\mathbf{x} - \mathbf{X}_i \right\} \qquad 3.8$$

The kernel function $K(\mathbf{x})$ is now a function, defined for a *d*-dimensional $\mathbf{x}$ satisfying

$$\int_{R^d} K(\mathbf{x})d\mathbf{x} = 1 \qquad 3.9$$

with a kernel defined for $d = 2$, the quartic bi-weight kernel is

$$K_2(\mathbf{x}) = \begin{cases} \frac{3}{\pi}(1-\mathbf{x}^\tau\mathbf{x})^2 & \text{if } \mathbf{x}^\tau\mathbf{x} < 1 \\ \mathbf{0} & \text{otherwise} \end{cases} \qquad 3.10$$

The results in the manuscript brief were calculated to examine all of the crashes in the data set. This visualization was used as an additional clustering test, to further study the relationship between the hotspots.

The next section contains the *Tribal Traffic Safety Manuscript Brief*, which I described at the beginning of this section as only a part of a greater study in *Using GIS to Improve Tribal Traffic Safety*. It was decided to let this document stand on its own and I will outline the state of the project at the end of this chapter following the manuscript brief.

BEGIN
TRIBAL TRAFFIC SAFETY
MANUSCRIPT BRIEF

Using GIS to Improve Tribal Traffic Safety

A Statistical Evaluation of Hot Spots On Minnesota Tribal Reservation Areas

Joseph C. Robertson
Statistical Consulting Services

Submitted on Behalf of
Claremont Graduate University
Center for Information Systems and Technology

Thomas A. Horan, PhD
Brian Hilton, PhD

February 3, 2017

Scope of the Project

The purpose of this analysis is to provide additional statistical insight and methodology to the existing preliminary assessment of geo-related Tribal traffic safety information. The objective is to design traffic safety framework of prototypes for potential use by Tribal governments through this initiative. In my initial consultation with Claremont Graduate University, Dr. Tom Horan, Dr. Brian Hilton and I discussed the various metrics required for completion of the project.

Scope of Work

Task 1: Tribal Data Analysis Spatial analysis and testing of the traffic data for additional metrics to strengthen the current results.

- **Task 1.1.** Tests for complete spatial randomness (CSR) such as quadrat analysis, Ripley's K simulation envelopes, and point pattern analysis.
- **Task 1.2.** Further investigation of possible Poisson related spatial point processes and distribution assessment.
- **Task 1.3.** EDA of possible links to provided covariates whenever appropriate, with additional spatial autocorrelation techniques

Whenever possible, I have designed this report with the tasks required to address the scope of work agreed upon.

Preliminary Analysis Synopsis: *State of Minnesota Tribal Crash Analysis*

The initial results of the hot spot analysis used a variety of techniques to spatially assess the severity of injuries sustained in traffic crashes that were recorded from 2005-2014 using the Minnesota Crash Mapping Analysis Tool (MnCMAT). Descriptive statistics obtained from the crash point data provided four types of injuries sustained as well a number of additional covariates detailing metrics such as month and year of crashes, weather type, and road conditions.

A simple comparison of the percentage of fatal and incapacitating injuries, as well as sufficient sample size was the criteria used in selecting four of the fifteen federally recognized tribes in Minnesota for the initial spatial analysis: Leech Lake, White Earth, Red Lake, and Mille Lacs. Culturally, there are many aspects of reservation life that are unique to each American Indian group living on their respective sovereign territory; and when developing a framework to assess how strong the correlation is with respect to those processes, it was important to begin with an overall global view of the initial point process.

Horan and Hilton (2016), in their initial analysis determined the most sensible way to accurately assess point processes on the selected reservations was to integrate existing road segment data to include an additional drive time polygons adjacent to the reservation boundaries to "evaluate the accessibility of a point with respect to some other features." In creating intersected tribal boundaries with a 15-minute drive time in the surrounding areas more accurately models a preliminary process, independent of the culture and the reservation boundaries.

This design was intended to make a number of comparisons to better understand the nature of clustering of fatal injuries within reservation boundaries combined with the outlying drive-time area. The subsequent Getis-Ord Gi* Z statistics were generated for each area to identify hot and cold spots. The resultant Z score identifies where features with either high or low values cluster spatially. This tool works by looking at each feature within the context of neighboring features. A feature with a high value is interesting, but may not be a statistically significant hot spot.

To be a statistically significant hot spot, a feature will have a high value and be surrounded by other features with high values as well. The local sum for a feature and its neighbors is compared proportionally to the sum of all features; when the local sum is much different than the expected local sum, and that difference is too large to be the result of random chance, a statistically significant Z score results.

Figure 1 below is an example of the calculated Z scores from the Leech Lake reservation. The points allowed for a visible confirmation of possible hot spots with the intent of providing an additional normalization factor (i.e. road miles, annual average of daily traffic) to make an ad hoc comparison of the point process across the four reservation areas. The reasoning for normalization relies on the fact that each reservation and surrounding area has differing sizes in both geographic area, as well as road network miles; so in order to make a direct comparison between each area we must account for the size and length of the network. The ratio of the average Gi* Z-score / total road miles for each region provides a rank for each area, which was then compared.

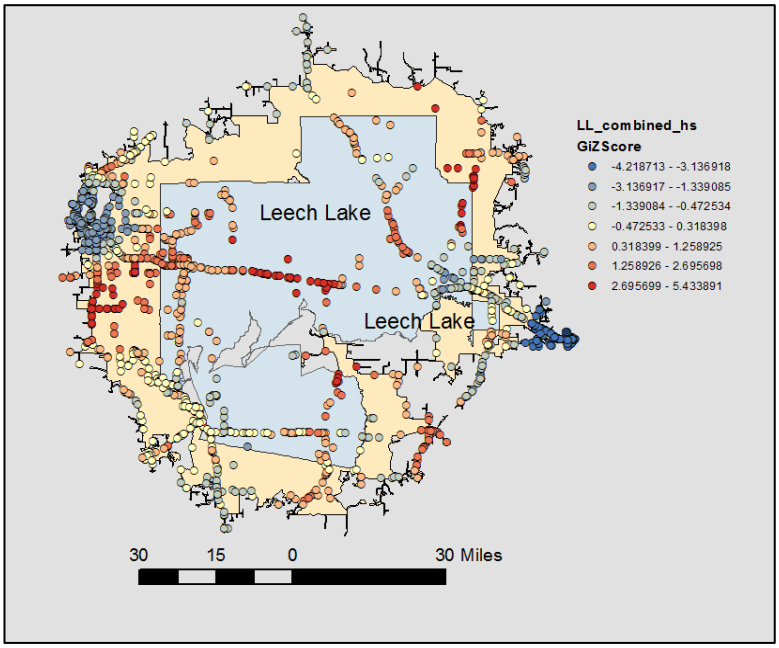


Figure 1

To further clarify this choice of methodology, let us assume we have two companies who have a number of insurance claims they been have processed for payment, say Company 1 has 50 claims and Company 2 has 100 claims. From this information, it appears that Company 2 processes twice as many claims in favor of the customer. However, making a direct comparison of this information is problematic due to the fact that unless we understand that the actual number of claims filed, say Company 1 is 100 and Company 2 is 1000, the information is misleading. Normalizing by this factor produces a more meaningful comparison, that in fact Company 1 has a 50% claim rate versus Company 2 at 10%.

In accounting for the size of the road network and the average daily traffic, Table 1 is the results Horan and Hilton (2016) obtained and ranked according to this

methodology. Although the average Z-score may not be a strong measure in assessing such a complex spatial process from an inferential statistics point of view; nonetheless, it provides a starting point in developing methodology to assist American Indian Tribes with insight into their lives on and off the reservation.

Hot spot analysis is one of six current prototype applications Horan and Hilton have developed as a matter of "assessing the Tribes interest in implementing GIS applications for planning, analysis, and programming in Tribal safety planning." This analysis provides additional inferential topics in such a way, as to strengthen further development of not just tribal traffic analysis, but the growing body of literature in point process model development and network analysis.

Comparison Rank of Each Reservation Area using a Normalization Factor

Z Score Only	Tribal Area Avg Z Score	Adjacent Area Avg Z Score	Absolute Value Z Score	Rank in Terms of Severity
Leech Lake	-0.217363	-0.872345	0.654982	4
Mille Lacs	0.338127	-0.394172	0.732299	3
Red Lake	0.072112	-1.606956	1.679068	1
White Earth	0.010559	-0.761449	0.772008	2

Z Score / RM	Tribal Area Avg Z Score	Adjacent Area Avg Z Score	Absolute Value Z Score	Rank in Terms of Severity
Leech Lake	-0.000121	-0.000342	0.000221	4
Mille Lacs	0.001556	-0.000151	0.001707	1
Red Lake	0.000139	-0.000474	0.000613	2
White Earth	0.000007	-0.000276	0.000283	3

Z Score / AADT	Tribal Area Avg Z Score	Adjacent Area Avg Z Score	Absolute Value Z Score	Rank in Terms of Severity
Leech Lake	-0.000001	-0.000001	0	4
Mille Lacs	0.000003	-0.000001	0.000004	2
Red Lake	0.000002	-0.000004	0.000006	1
White Earth	0	-0.000002	0.000002	3

RM= Total Road Miles AADT= Annual Average Daily Traffic

Table 1

Discussion

There are any number of ways to begin the evaluating the strengths and weaknesses of any statistical process. As I have outlined in the previous flowchart above, developing a statistical methodology framework usually follows a general hierarchy of:

1. Examining any descriptive statistics pertinent to the study
2. Develop and test and number of *Exploratory Data Analysis* (*EDA*) techniques that provide a basis for more complex methodology and inference
3. Finally, using the above information to create a formal inferential hypothesis for examining more complex processes that may exist beyond the first stage of the current project.
4. Repeat the process whenever necessary

To better understand the nature of the point process and how we develop additional metrics to strengthen the original Getis-Ord analysis, we must first begin by examining some fundamental aspects of the how the analysis was conducted, the assumptions, and how the outcomes provide additional EDA techniques to provide a final set of recommendations to help strengthen the framework when consulting with Tribes.

Literature Review of Point Processes and GIS

The nature of the traffic data and its associated coordinates in 2-space undoubtedly constitute some sort of spatial point process. In geographic information analysis, there many techniques that can be employed to assess the nature of the point process; however one must take care in making the correct assumptions when designing an inferential test.

Baddeley, et al. (2016) write:

Advantages of Statistical Modeling

Statistical modeling is a much more powerful way to analyze data than simply computing summary statistics. Formulating a statistical model, and fitting it to the data, allows us to adjust for effects that might otherwise distort the analysis (such as uneven distribution of survey effort, and spatially varying population density) by including terms that represent these effects. By fitting different models that include or omit a particular term, and comparing the models, we can decide which variables have a statistically significant influence on the intensity.

A great advantage of statistical modeling is that the assumptions of the analysis are openly stated, rather than implicitly imposed. All model assumptions are 'on the table' and are amenable to criticism (Baddeley, et. al., 300).

O'Sullivan and Unwin (2010) provide additional insight to better understand some fundamental aspects of how to understand the nature of spatial data:

Objects Are Not Always What They Appear to Be

Students often confuse the various cartographic conventional representations with the fundamental nature of objects and fields. For example, on a map, a cartographic line may be used to mark the edge of an area, but the entity is still an area object. Real line objects represent linear entities such as railways, roads, and rivers. On topographic maps, it is common to represent the continuous field variable of height above sea level using the lines we call contours; yet, as we have discussed, fields can be represented on maps in many different ways.

This is consistent with understanding that although there are unique areas in Indian country with respect to Tribal boundaries and the culture contained within, notwithstanding the nature of processes we are investigating are still an area object and investigating the nature of the covariates that make this space unique requires additional planning and careful analysis as to how this might affect the point process; or whether such a relationship even exists.

Objects Are Usually Multidimensional

Very frequently, spatial objects have more than the single dimension of variability that defines them. We might, for example, locate a point object by its (x, y) coordinates in two spatial dimensions, but in many applications it would be much better to record it in three spatial dimensions (x, y, z), with depth or height as a third dimension.

Not only is the nature of data multivariate with respect to geographic properties that remain constant over the area of interest, but we must also account for the multidimensional covariates that assumedly make each point in the point process unique; although this may not always be the case.

Objects Don't Have Simple Geometries

Some aspects of geographic reality that we might want to capture are not well represented in either the raster/vector or object/field views. The obvious example here is a transport or river network.

As we will see in Chapter 4, the nature of the point process assessed using Getis-Ord, does not take into account that road accidents and subsequent road injury severity is completely dependent on the road network and not necessarily the entire area we are assuming to be part of a homogeneous area of study.

Objects Depend on the Scale of Analysis

Different object types may represent the same real-world phenomenon at different scales. For example, on his daily journey to work, one of us used to arrive in London by rail at an entity called Euston Station. At one scale this is best represented by a dot on a map, which in turn is an instance of a point object that can be represented digitally by its (x, y) coordinates. Zoom in a little, and Euston Station becomes an area object, best represented digitally as a closed string of (x, y) coordinates defining a polygon. Zooming in closer still, we see a network of railway lines (a set of line objects) together with some buildings (area objects), all of which would be represented by an even more complex data description. Clearly, the same entity may be represented in several ways. This is an example of the multiple representation problem in geographic information analysis. Its main consequence is to make it imperative that in designing a geographic information database and populating it with objects of interest, it is vital that the type of representation chosen will allow the intended analyses to be carried out.

Scale is particularly important due the nature of point processes: “The nonrandom distribution of phenomena in space has various consequences for conventional statistical analysis. For example, the usual parameter estimates based on samples that are not

randomly distributed in space will be biased toward values prevalent in the regions favored in the sampling scheme. As a result, many of the assumptions we are required to make about data before applying statistical tests become invalid" (O'Sullivan and Unwin, 2010).

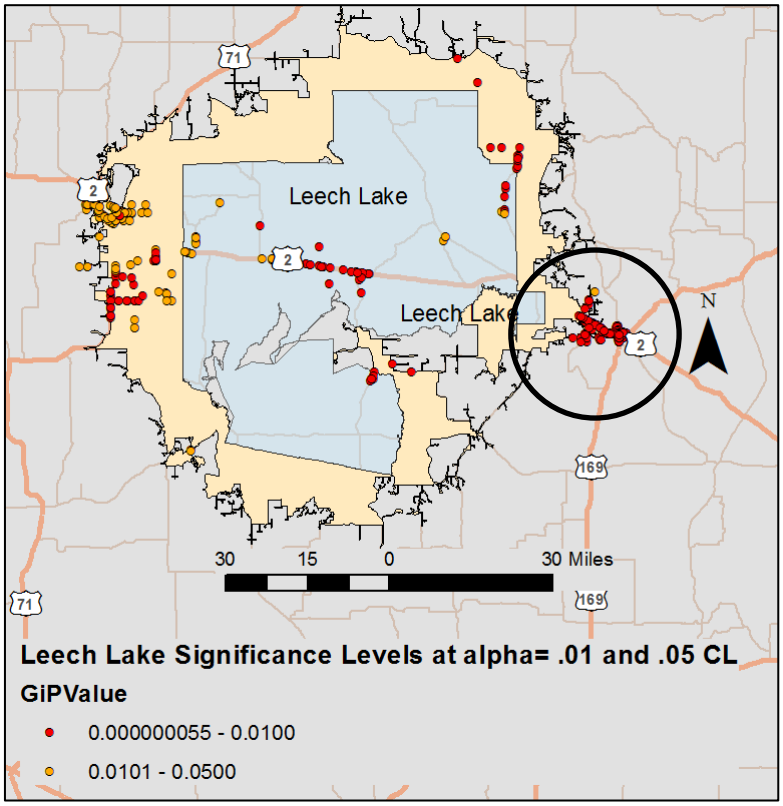


Figure 2

In examining Figure 1 on page 75, the cold spots (blue) in the far central-east of the network area contains a cluster of points. In the above figure, we can see the proportion of points that are significant at the $\alpha = .01$ and $.05$ level. There are a number interesting things happening on a number of scales in this example. First, the Getis-Ord Gi* statistic (Figure 1) provided a point process of the calculated Z scores, which in turn allowed for aggregating the associated p-values that were significant.

As you can see from Figure 2, the area enclosed by the circle on the easternmost boundary of the study area has what appears to be a number of statistically significant points clustered together. One of the objectives of the original analysis was to gain insight of potential severe and fatal accidents. A person with a limited background might construe this area within the circle to be a hot or cold spot of severe and/or fatal accidents. Figure 3 shows the result of zooming in closer to the area of interest.

Understanding that spatial phenomena are aggregated in certain contexts *can* produce statistically biased results and is why EDA is so important for inference decision-making

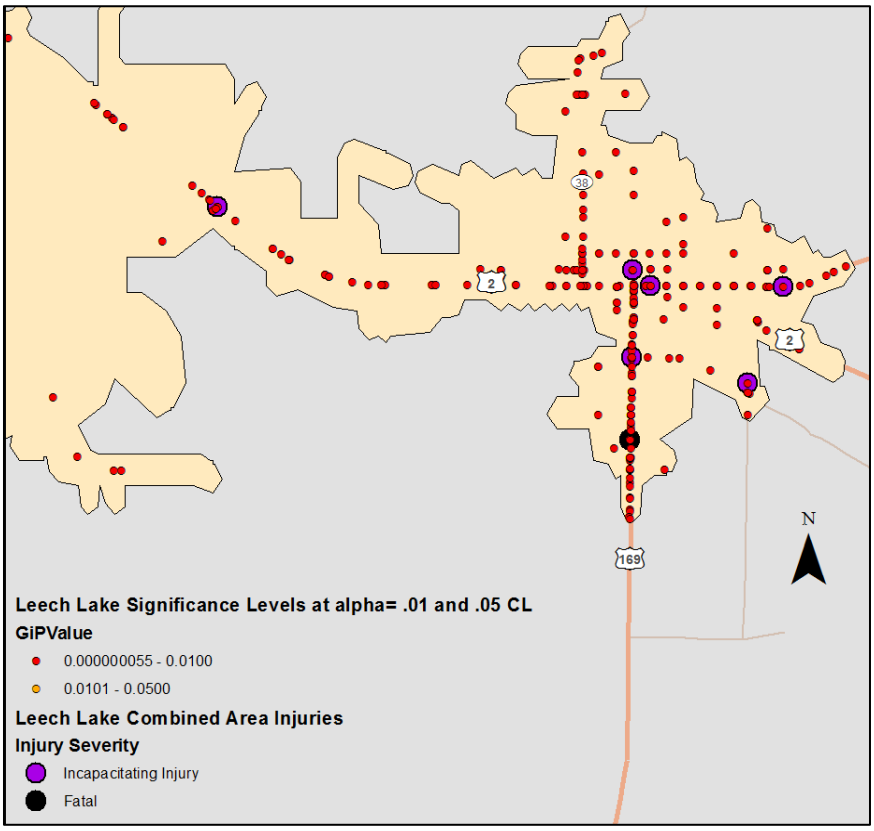


Figure 3

processes. This is what is referred to as the modifiable areal unit problem (MAUP). The choice of drive-time boundaries allowed the inclusion of additional townships in the rural space which naturally contain more accidents, and thus the influence of these townships, must be accounted for.

The spatial points with statistical significance were placed with the corresponding severe and fatal crashes and the study area revealed in Figure 3:

1. The significant cold spots are a result of an over-inflation of zeros representing non severe / fatal accidents.
2. The area of interest actually represents a collection of accidents, non-homogeneous in nature to rest of the study area, as this area is a township vs. a rural area.
3. The points we are interested in i.e. severe and fatal crashes are small in size and the overinflated zeros representing the non-fatal accidents overtake all other points and produce a cold spot.
4. This area is contained in the additional drive-time areas outside the reservation and thus further work must be done to assess if this information would be useful in Tribal traffic safety.

The context in which information is presented can often help inform additional stages of planning and development. It has been shown a *simple* test for clustering can result in additional information important for designing additional EDA techniques and future modeling. Engaging tribal stakeholders in this process allows for more refined area

of study that tailors the analysis to results that are meaningful to Horan and Hilton's initial analysis:

"Where are there unexpectedly high spatial clusters of injuries, in particular, fatal injuries, given all injuries?"

To begin answering this question, let us consider the following:

Is the Point Pattern a Realization of a Point Process?

The general assumption of a point pattern is that it is generated by some deterministic or stochastic process that can be modeled in some way. Secondly, given that process, is it possible to model the phenomena is some meaningful way? "Geographic data are rarely deterministic in this way. More often, they appear to be the result of a chance process, whose outcome is subject to variation that cannot be given precisely by a mathematical function. This apparently chance element seems inherent in processes involving the individual or collective results of human decisions…Whatever the reason for this chance variation, the result is that the same process may generate many different results" (O'Sullivan and Unwin, 2010).

This realization is defined as an independent random process (IRP), or commonly known as complete spatial randomness (CSR). "The IRP is mathematically elegant and forms a useful starting point for spatial analysis, but its use is often exceedingly naive and unrealistic. Many applications of the model are made in the expectation of being forced to reject the null hypothesis of independence and randomness in favor of some alternative hypothesis that postulates a spatially dependent process. If real-world spatial patterns were indeed generated by unconstrained randomness, then geography as we understand it

would have little meaning or interest and most GIS operations would be pointless ” (O’Sullivan and Unwin, 2010).

As with any analysis, a careful construction of assumptions is paramount to understanding the nature of the process generating the point pattern. Tests designed to test some departure from CSR is a seemingly whimsical assumption about a process that is clearly perceived to be a non-random process. Nonetheless, merely assuming a point process is not random with any evidence to the contrary is problematic. In the case of this traffic study, is there evidence to suggest this point pattern is random?

As we have seen from the figures earlier in this chapter, there is evidence to suggest the point pattern exhibits some sort of clustering process, but more formally, is there a way to test this with complete confidence? The answer is yes and no. As this analysis moves past simple descriptive statistics, such a simply viewing the point pattern on a map, we must carefully assess what are the underlying assumptions before performing a formal statistical test.

A common way to assess ISR/CSR is what is known as a quadrat count test. Again, we begin this process by making the following assumptions about the point process.

1. The condition of equal probability. We assume the events in question can occur anywhere in the study area, or any subarea has an equal chance of receiving an event.
2. Independence. Each point in the pattern is independent of any other events in point pattern.

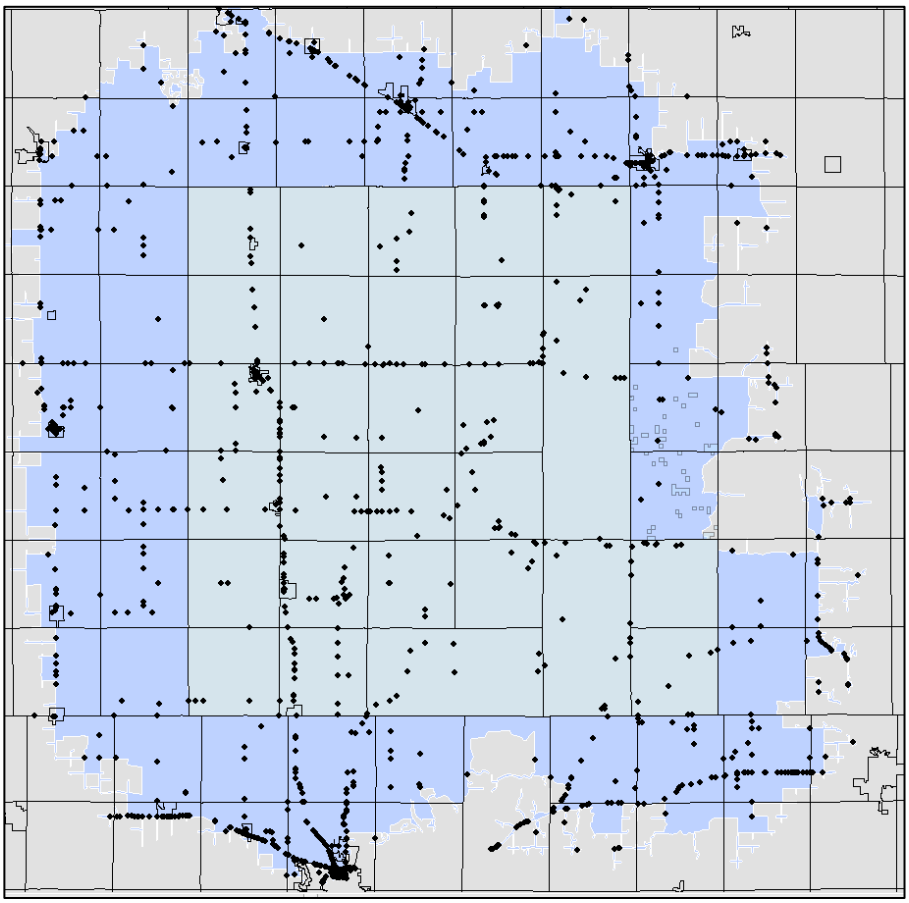


Figure 4

Let us consider the point pattern from the White Earth reservation and surrounding area. To perform a quadrat count analysis, we are interested in determining if the counts contained in the grid are the result of complete spatial randomness. Ignoring the MAUP for the moment, the grid was generated conveniently from the established townships geometry for each state from the U.S. Census Bureau. Although crude when compared to a more uniform grid of each length and width, we are interested in looking visually if there are discernible areas of clustering and if this test appropriate given the assumptions.

What is interesting in this figure is although points fall conveniently into the township grid, it is hard to ignore there are collection of points which exhibit a north-south and non-linear consecutive patterns of alignment despite the grid.

If we look more closely at Figure 5 below after overlaying the road network, we can see this point process violates at least one assumption: The counts of points occurring in each quadrat may be independent with regard to the event occurrence, but clearly the events are contingent on the location of the network of roads. This means that if we assume that the intensity has equal probability to occur anywhere in the grid space, this is clearly not the case.

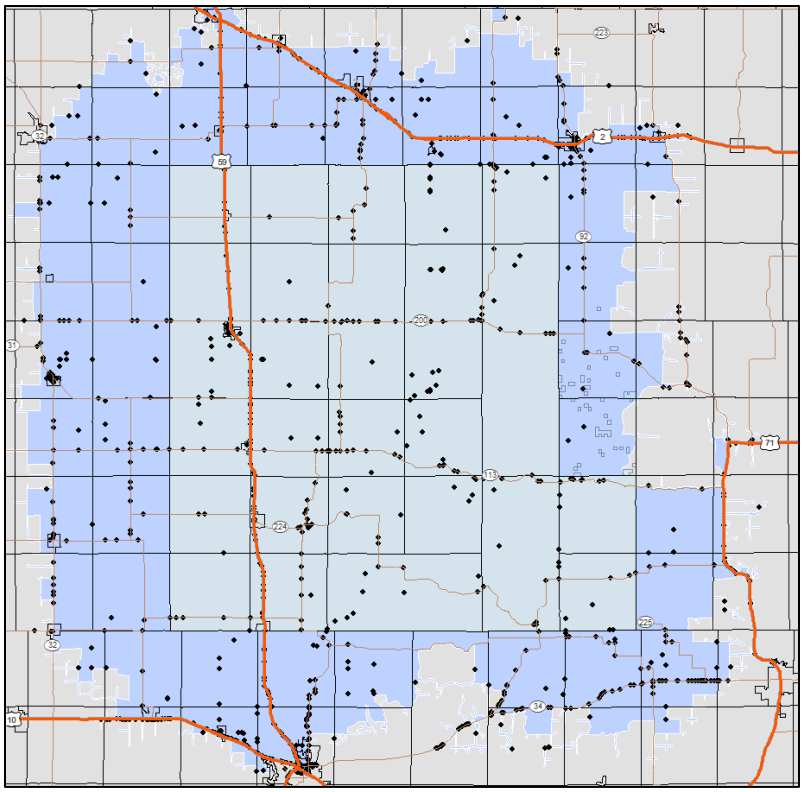


Figure 5

This is an example where if we misclassify the point process, our conclusions become problematic. In fact, because of the point process' dependence on the linear network of roads, using an exploratory quadrat test is inappropriate, due to the multidimensionality of the point process. This will be explained in more detail in Chapter 4.

Spatial Autocorrelation

As we have seen from the previous example, designing and testing point patterns using the most basic methods grows increasingly complex the more assumptions we make with regard to the point pattern, point process, and choice of boundaries. Spatial autocorrelation is probably one of the most well developed concepts in geographic information analysis and is particularly important when taking into account effects of temporal and spatial proximity in these statistical analyses.

Anselin and Bera (1998) provide a concise verbal definition: "spatial autocorrelation can be loosely defined as the coincidence of value similarity with locational similarity." They also provide a more formal definition as well: "A nonzero spatial autocorrelation exists between attributes of a feature defined at locations i and j if the covariance between feature attribute values at those points is nonzero.

If this covariance is positive (i.e., if data with attribute values above the mean tend to be near other data with values above the mean), then we say there is positive spatial autocorrelation; if the converse is true, then we say there is negative spatial autocorrelation. Positive autocorrelation is much more common in nature, but negative autocorrelation does also exist" (Plant, 2012).

In the previous quadrat test, we were interested in identification of possible spatial clustering through a simple non parametric test based on two assumptions. One fundamental issue that spatial autocorrelation attempts to address is: Does the point process exhibit some overall global pattern of autocorrelation, or is the process more complex when considering local area of neighborhoods near a particular point of interest?

The choice to overlay a township grid in Figure 5 was a simple way to assess the point pattern where a grid containing counts of traffic accidents could be assessed with some nearest neighbor attributes. Although it was established this test is inappropriate for the study at hand, nonetheless, this technique allows for a simple geometric assessment of neighbors in such a way, that allow for a more simple calculation using a spatial weights matrix of neighbors defined by shared common borders and/or intersection of vertices within the grid. This is known as Moran's I, which is the most common way to test both global and local autocorrelation.

Horan and Hilton's analysis did not contain this metric and from a practical point of view it most likely would not be effective unless a grid could provide a well-defined area of calculating distances to neighbors in a more structured manner. For instance, the Leech Lake reservation and surrounding area example showed clustering due to the location of a township and natural clustering of accidents will have a higher likelihood to occur when compared to rural areas where there are far fewer human inhabitants.

The theory behind calculating spatial autocorrelation relies on some neighbor distance that when the spatial lags of distance d reaches a certain distance the autocorrelation goes to zero. It is entirely possible to surmise that if we have a cluster of

points in a township and we begin looking for neighbors of say a radius *d* that it will not take long before each defined spatial lag goes farther and farther out into an area where there are no accidents whatsoever. The conclusion that can be drawn from this observation is that the point process is most likely not homogeneous, and that there exists some sort of spatial dependency between the choice of boundaries and the network of roads the accidents occur on.

Yamada and Thill (2004) assert: "Traffic accidents are commonly anticipated to form clusters in the geographic space and over time for the reason that their occurrence is tied to traffic volumes, which themselves exhibit distinct spatial and temporal patterns, as well as because of their link to natural environmental factors such as snow and fog, configuration of highway networks such as locations of access and egress points, and deficient design and maintenance of highways."

Levine et al. (1995a) divide research on spatial dependence of traffic accidents into four main categories. The first category of studies compares different types of spatial environments, such as urban and rural settings on the basis of accident prevalence, and usually involves highly aggregated data and large geographical units. The second looks for causal relationships between traffic accidents and attributes of the roadway system, for example, traffic volumes and roadway types.

This category would include the identification and analysis of locations producing more accidents than other locations in a given network, also known as ''hot spots'' or ''blackspots'' (McGuigan, 1981). Studies of the third type examine accidents in particular areas or corridors while emphasizing socially and ecologically integrated

analysis units. The last line of research focuses on system-wide variations in traffic accidents, in other words, how local patterns of accidents compose a global-scale pattern (Yamada and Thill ,2004).

As we have seen from even the limited number of considerations in this report, there are a number of methods that can help determine whether an observed pattern of events results from a random pattern, or it follows from some systematic process so as to form a clustered or regular pattern (Yamada and Thill, 2004). This global test for overall variation in the mean value of the process is called a *first order effect*; which gives us some sort of information about what is happening in the overall area of study. *Second order effects* refer to local deviations of the process from its mean value caused by its spatial dependence structure.

It is typical to make exploratory comparisons of metrics such as Moran's I and Getis-Ord Gi*for similar and hypothetical comparisons. The original analysis can be strengthened with this design theory. "When used in conjunction with a statistic such as Moran's I, they deepen the knowledge of the processes that give rise to spatial association, in that they enable us to detect local 'pockets' of dependence that may not show up when using global statistics" (Getis and Ord, 1992).

Horan and Hilton's original analysis using the Getis-Ord Gi* is fundamentally tied to second order effects in that to determine a measure of significance to produce hot and cold spots, some local measure of distance must be established. The original analysis revealed that severe and fatal crashes were a somewhat rare event, and as such, the inflation of zeros used in the calculation of a global autocorrelation measure could be

misleading. However, the purpose of additional measures provides a stakeholder with additional information that should be designed to incrementally strengthen the value of any results with as many credible measures as possible.

In regard to second order effects, the literature review presented here such as the quadrat analysis are only one of a number of well-established techniques to assess local spatial dependency. In the course of this analysis, I performed a number of diagnostic tests in addition to quadrat analysis. Given the complex set of assumptions every researcher faces when determining the appropriate course, Ripley's K-function was another technique considered. "Among methods that deal with the second order intensity of the point process, Ripley's K-function is regarded as one of the most effective and comprehensive methods because it tests point patterns at various spatial scales, handles all event-event distances, and does not aggregate points into areas" (Yamada and Thill, 2004). Assumedly, claimed to be "free" of the modifiable area unit problem.

However, upon further examination of this metric, one critical consideration that cannot be sidestepped is that the point process is undoubtedly constrained to a network in geographic space. The K-function assumes an infinitely continuous planar space where distances are measured as a straight line.

Yamada and Thill (2004) write:

> The planar space assumption is problematic first because many events are bound to happen only in a subset of this space. For instance, traffic accidents occur only on streets and highways; retail businesses such as fast-food restaurants or gas stations are usually located along major streets. As a matter of fact, any data that

> are geocoded on the basis of a street address or a milepost are inherently constrained by the street network. With events constrained by a one-dimensional subset of the planar space (such as a transportation network), the proper research question should bear on the existence of non-random patterns of events embedded within this one-dimensional subset instead of the planar space itself. Not only are the locations of events constrained by the transportation network, but so is also the movement between them. Consequently, the distance between any two network constrained events is more adequately represented by a shortest-path (network) distance than by a Euclidean distance measure. (p.149)

In addition, Lu and Chen (2007) contribute additional insight:

> Most applications of K-function, including those mentioned above, use Euclidean distance to represent spatial separation between points. This is not a problem when the point pattern is a continuous and unrestrained distribution over a Euclidean plane. However, Euclidean distance is no longer an accurate measurement when the point distribution is subject to certain restrictions. A set of points distributed along urban streets is but an example. From the definition of K-function, we can see that the measure of distance plays a critical role in evaluating the clustering situations in a point set. Using different measures of distance would result in the K-function behaving differently, leading to different conclusions regarding the patterns of the same point set. (pp. 614-615)

The purpose of this literature review was to present a number of considerations to accompany Horan and Hilton's original analysis. In statistical design theory, the most

effective way of determining the strengths and weaknesses of any analysis is examining the body of evidence that balances cost effective strategies in descriptive and EDA measures versus more complex and computationally intense techniques, while appropriate theoretically but perhaps practically inefficient.

Considerations for Future Modeling Crashes within a Framework

The purpose of this report was to aggregate any number of considerations that create a unified framework for work with Tribal governments in developing GIS analysis as a practical and efficient way of engaging stakeholders in meaningful ways to improve community economic planning and development. The American Indian reservation system is not entirely disjoint from the regular business that occurs in areas within a reasonable distance to Tribal affairs, and traffic related accidents are very relevant to the location of services within the immediate vicinity of townships that border the reservation.

When constructing a framework to address possible ways to address these issues in traffic safety, I have provided a list of considerations statistically that can possibly contribute to a more refined method of engaging Tribal stakeholders.

1. Is this point pattern a result of a point process?
2. Can this process be modeled as a Poisson point process?
3. Is the point pattern data deemed to be homogeneous or inhomogeneous?
4. Are there temporal (consideration of time) implications embedded in the raw data?

5. Additional network considerations must address additional dimensionality. Can we reliably test statistical significance using a Euclidean 2-D system of assumptions?
6. Getis-Ord Gi*: Are the realized values an effective measure in assessing Tribal traffic safety without first considering clustering of accidents that fall out of the immediate reservation area that may or may not explain the variation captured by the point process?
7. Can we use the original report conclusions to better understand if in fact the reservation and surrounding areas are correlated in some way?
8. To assess this, we need to understand how Tribal infrastructure plays a role in this preliminary assessment. Dependent location such a Tribal casino jobs, transportation, Tribal housing, health care, and everyday needs are dependent on the routes required to acquire these services are a necessary consideration.

We have presented here a rather exhaustive set of techniques, assumptions, and case study to better understand that a geographic information analysis using well established design theory such as Levine et al. can provide effective framework design.

Chapter 2

Additional Descriptive Measures of the Getis-Ord Analysis

Xie and Yan (2008) write: “To reduce traffic accidents and improve road safety, it is crucial to understand how, where and when traffic accidents occurred. An improved understanding of spatial patterns of traffic accidents can make accident reduction efforts more effective. For instance, by knowing where and when traffic accidents usually occur, law enforcement can conduct more efficient patrols and highway departments can disseminate more effectively to drivers the critical information about roadway conditions. In reality, the occurrences of traffic accidents are seldom random in space and time. In most cases, traffic accidents form clusters (known as ‘‘hot spots”) in geographic space.”

Maps convey powerful messages to their readers, and most are not knowledgeable of the technicalities of complex spatial processes. The points in the Getis-Ord analysis convey a message about potential hot spots of a process. Descriptive statistics are equally powerful by providing visual representations.

In developing GIS based traffic safety prototypes for Tribal stakeholders, it is important to convey as much information that is easy to understand for decision making. The structure of the traffic data contained additional covariates that can provide additional trends when investigating severe and fatal accidents. The Getis-Ord design metrics provided a normalized rankings system to investigate which reservation and surrounding areas contained the highest proportion of hot spots.

This chapter focuses on providing additional measures to consider metrics of temporal variability which when compared to the proportion of hot spots, may show

central tendencies or trends in each respective area. Cressie and Wikle (2011) write, "Traditionally, much of time series literature has focused on what is described as the 'process model'. This is a statistical model for a (hidden) process, indexed by time that either implicitly or explicitly conditioned on underlying parameters that describe the process evolution and/or dependence structure. A data set can be thought of as a window of through which knowledge can be obtained, enough to infer answers to the 'why' question. What stops a scientist from truly deducing answers, rather than inferring them, is the ubiquitous presence of uncertainty."

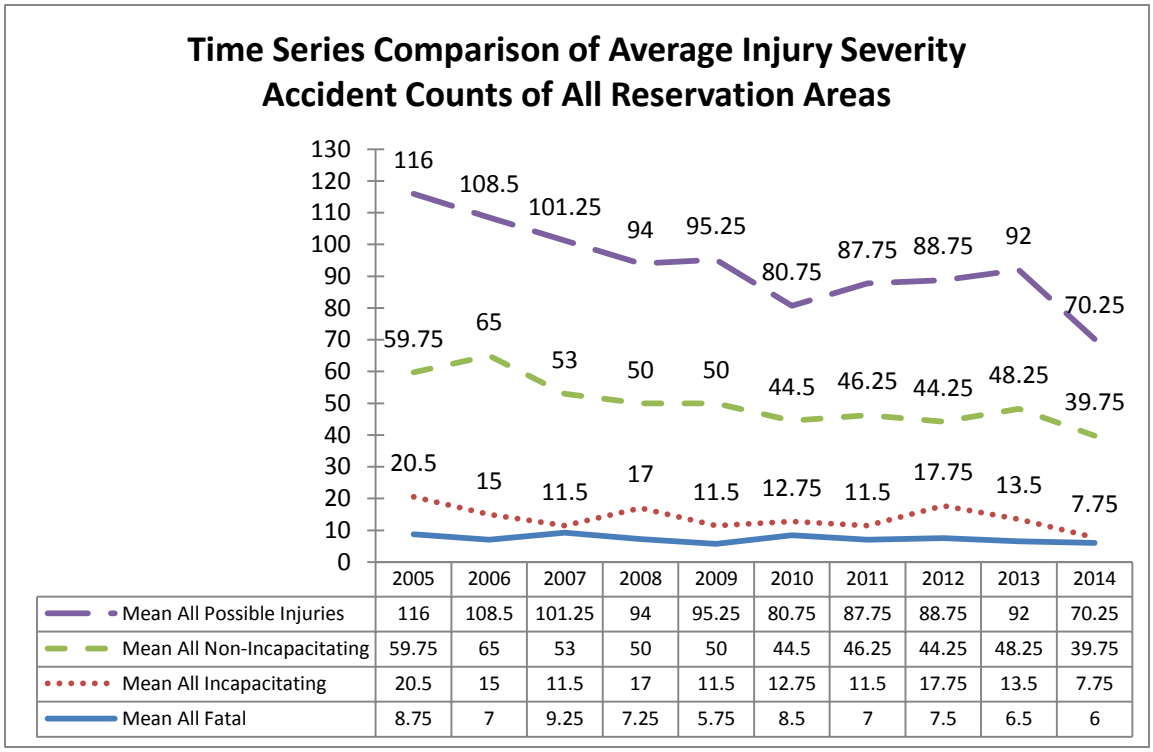
Statistics can account for this uncertainty through development frameworks that are aimed at looking for optimal procedures to explain this uncertainty, and then developing optimal procedures to quantify uncertainty in a meaningful way. "In addition to assumptions of stationarity, probabilistic properties of time series are typically specified in order to simplify their characterization. The fact that the time series literature emphasizes such process-model-based descriptions is motivated by science. This is in contrast to much of the literature in geostatistics, where, for historical reasons, more descriptive and less explanatory models are used to represent the data" (Cressie and Wikle, 2011).

Discussion

This chapter has been designed with simple but powerful visualizations concerning the temporal aspect of the data. It is not uncommon to aggregate many years of data which in turn provides any number of summary statistics. The focus of these descriptors is to assess and evaluate any trends that might be present to help explain

Horan and Hilton's ranking of hot spots in their original analysis. One interesting result was the Mille Lacs reservation and its surrounding areas ranked the highest in proportions of hot spots in all normalization tests for road miles and AADT.

When looking at a time series, it was important to put all of the areas on the same scale. The reason for this allows any stakeholder to observe spatial and temporal variability present over time. As we will see, the resulting time series of accident type and severity reveal some interesting results with respect to variability.



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Mean All Possible Injuries	116	108.5	101.25	94	95.25	80.75	87.75	88.75	92	70.25
Mean All Non-Incapacitating	59.75	65	53	50	50	44.5	46.25	44.25	48.25	39.75
Mean All Incapacitating	20.5	15	11.5	17	11.5	12.75	11.5	17.75	13.5	7.75
Mean All Fatal	8.75	7	9.25	7.25	5.75	8.5	7	7.5	6.5	6

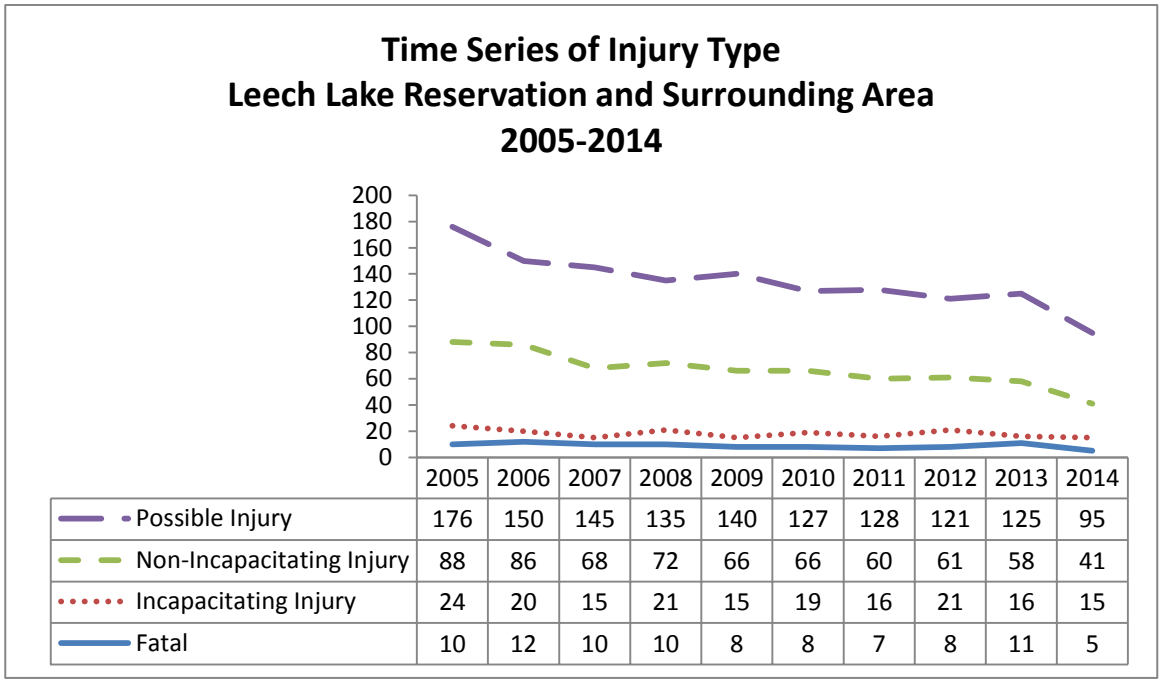
Figure 6

Each time series has been divided into its respective year so we may look at the trends over an entire decade. We are most interested in looking at the how big the spread is over time of each area, particularly with respect to severe and fatal crashes. To make

interpretation as simple as possible, the time series always shows from least to most severe going from the top to the bottom. For example, in the case we are investigating severe and fatal crashes, the red and blue lines can be compared directly to any other red and blue line in this group to investigate variability trends.

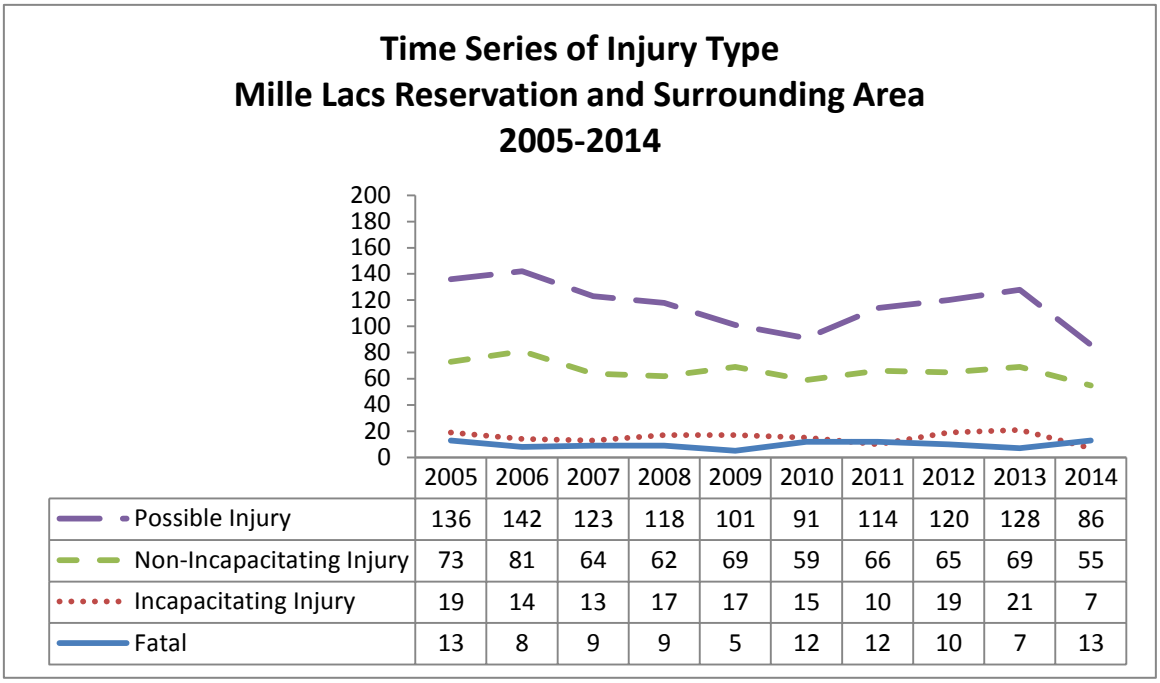
Global Versus Local

Figure 6 provides an ad hoc way of looking at overall trends by aggregating all regions and calculating their respect means over time. This visualization gives a general spread across all injury severities but actual local trend with the associated spatial data points goes beyond this simple descriptive method. Additional design metrics are required to make any meaningful inferences beyond what is shown. These sections will simply present these metrics for study and summarize the general trend at the end of each time series.



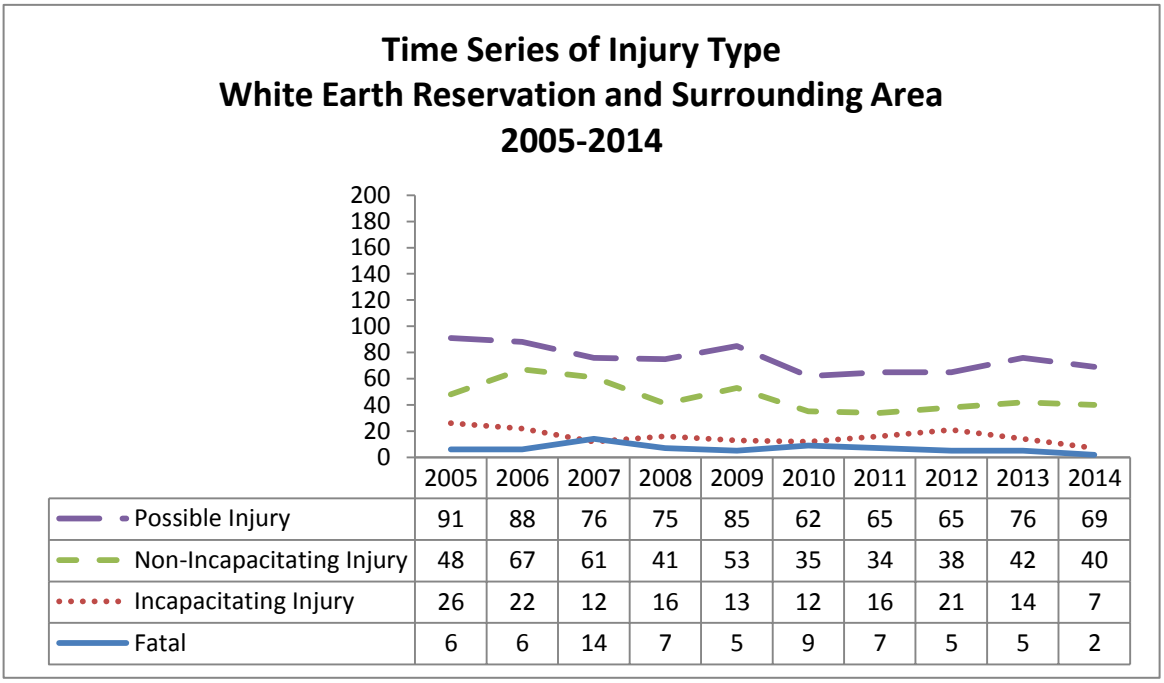
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Possible Injury	176	150	145	135	140	127	128	121	125	95
Non-Incapacitating Injury	88	86	68	72	66	66	60	61	58	41
Incapacitating Injury	24	20	15	21	15	19	16	21	16	15
Fatal	10	12	10	10	8	8	7	8	11	5

Figure 7



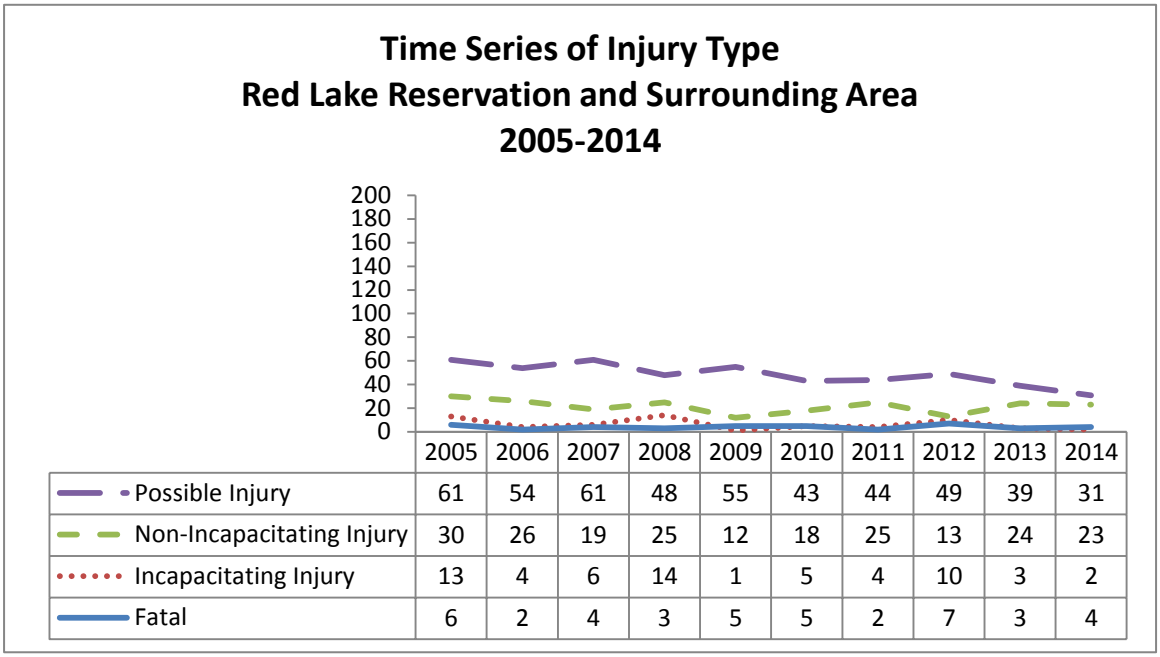
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Possible Injury	136	142	123	118	101	91	114	120	128	86
Non-Incapacitating Injury	73	81	64	62	69	59	66	65	69	55
Incapacitating Injury	19	14	13	17	17	15	10	19	21	7
Fatal	13	8	9	9	5	12	12	10	7	13

Figure 8



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Possible Injury	91	88	76	75	85	62	65	65	76	69
Non-Incapacitating Injury	48	67	61	41	53	35	34	38	42	40
Incapacitating Injury	26	22	12	16	13	12	16	21	14	7
Fatal	6	6	14	7	5	9	7	5	5	2

Figure 9



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Possible Injury	61	54	61	48	55	43	44	49	39	31
Non-Incapacitating Injury	30	26	19	25	12	18	25	13	24	23
Incapacitating Injury	13	4	6	14	1	5	4	10	3	2
Fatal	6	2	4	3	5	5	2	7	3	4

Figure 10

Summary

This set of time series looks at the spread of all types of injuries by each reservation and surrounding area.

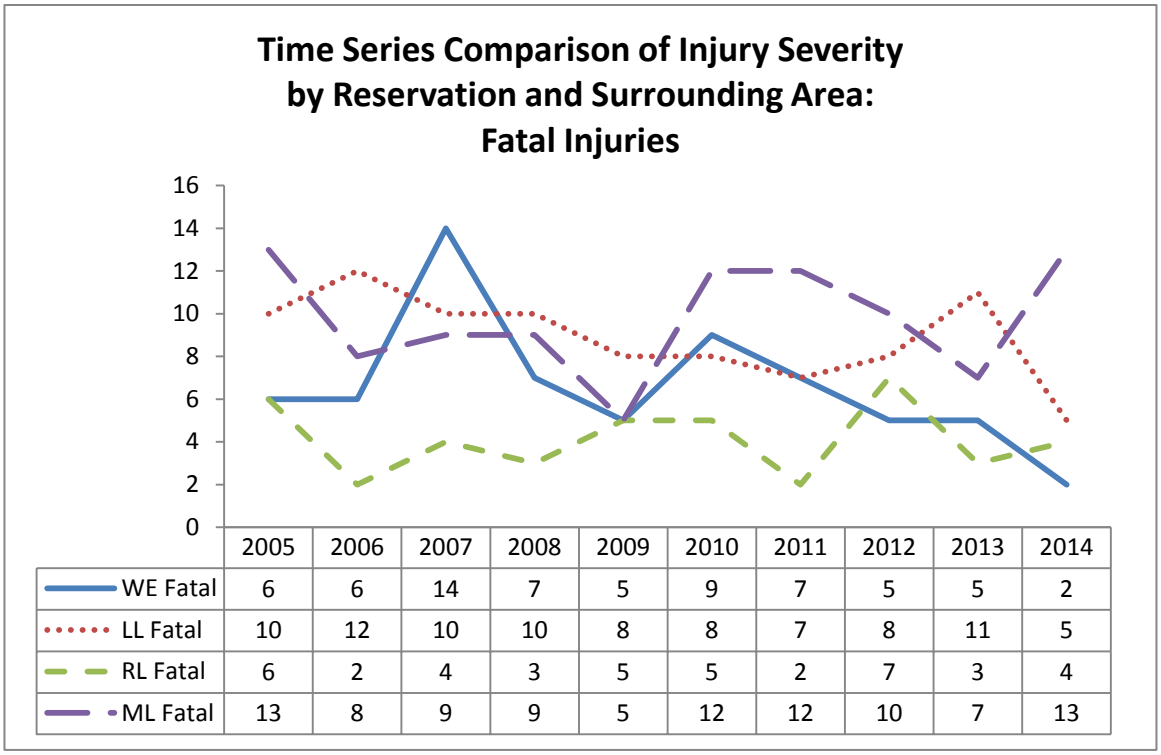
The general trends are as follows:

- Using the same scale to understand where each region falls with respect to the number accidents shows Mille Lacs and Leech Lake have similar spreads in variability possibly due the higher number of accidents
- White Earth and Red lake also share a similar and tighter spread, again due to smaller number of accidents over time
- The spread is most due to the variability of the number of non-severe accidents recorded in each region, while the severe accidents remain relative close to together

Does this reveal any information as to why Mille Lacs had the highest proportion of normalized hot spots?

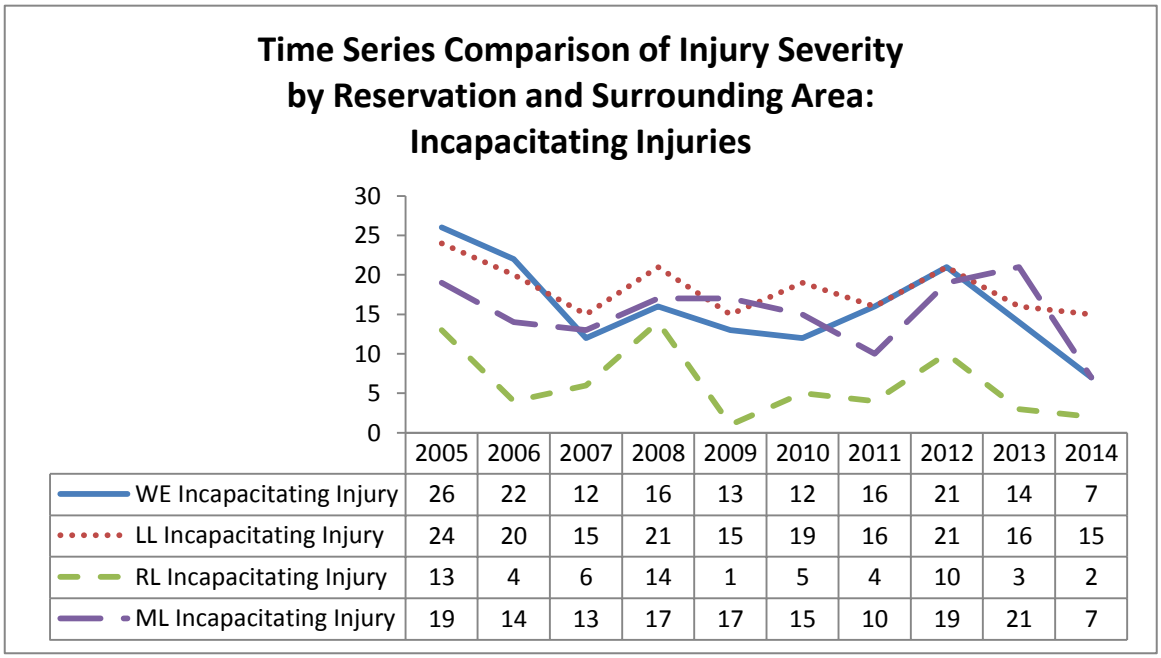
The results are somewhat inconclusive, however taking into account the drive-time area contained two townships outside the reservation area possibly could have contributed to a natural increase in accident frequency.

The next set of visualizations aggregates each injury type by reservation and surrounding areas to observe any distinctive trends with one specific injury type only. This allows for a more refine way of looking at the trends from time series 2.



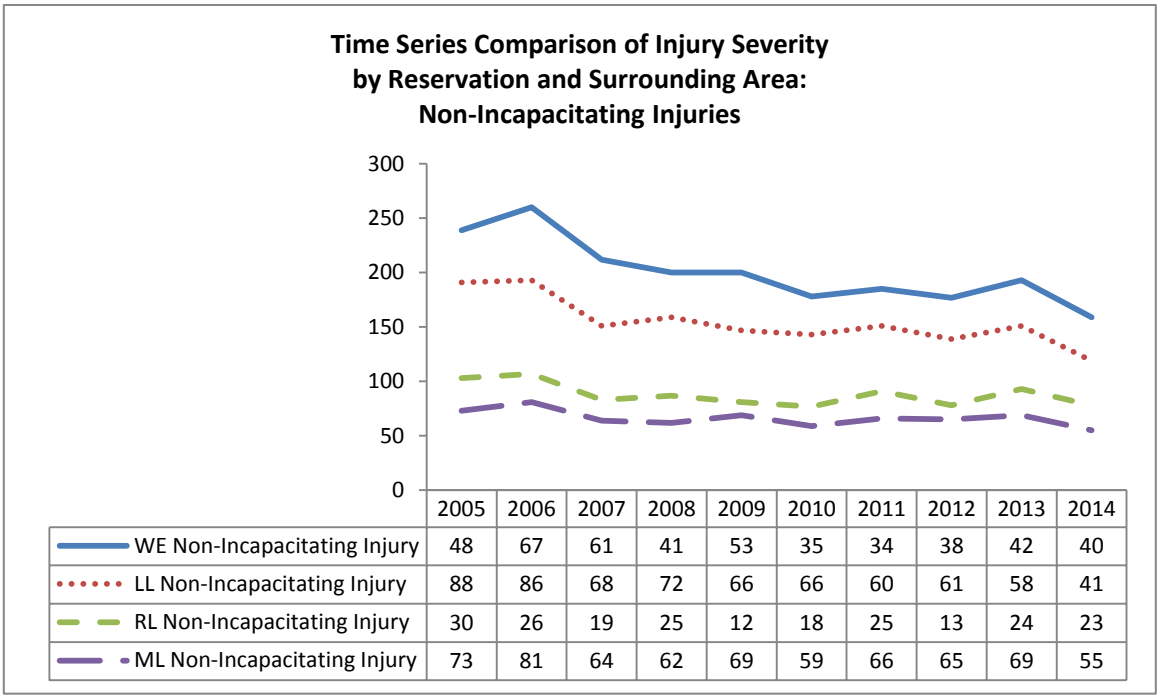
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
WE Fatal	6	6	14	7	5	9	7	5	5	2
LL Fatal	10	12	10	10	8	8	7	8	11	5
RL Fatal	6	2	4	3	5	5	2	7	3	4
ML Fatal	13	8	9	9	5	12	12	10	7	13

Figure 11



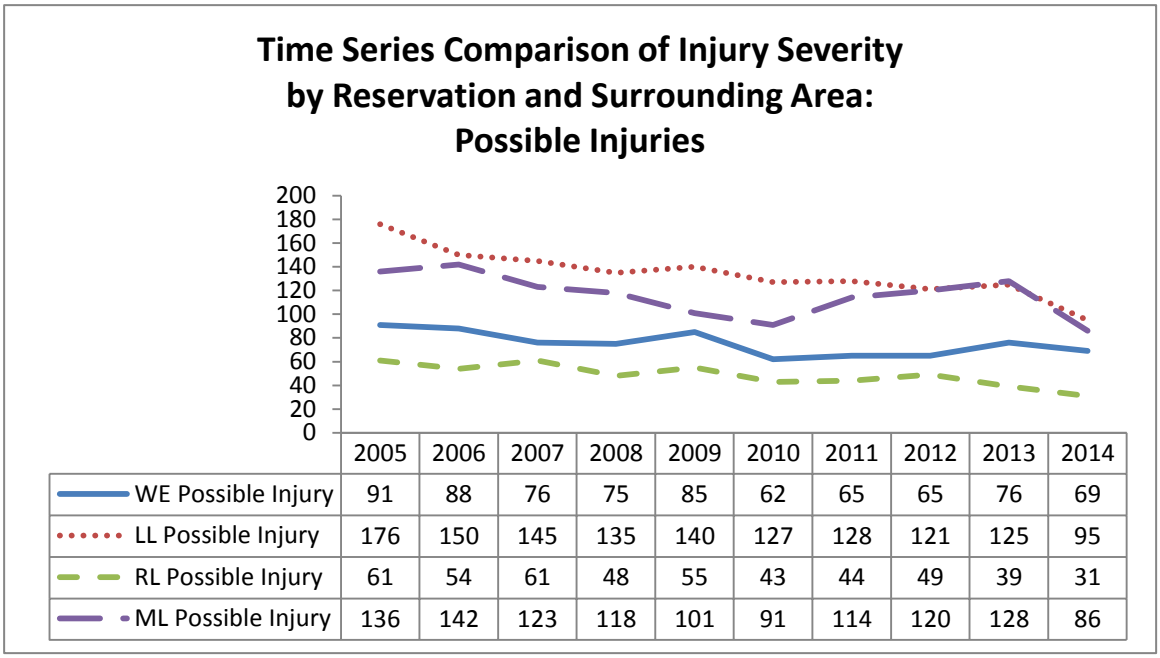
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
WE Incapacitating Injury	26	22	12	16	13	12	16	21	14	7
LL Incapacitating Injury	24	20	15	21	15	19	16	21	16	15
RL Incapacitating Injury	13	4	6	14	1	5	4	10	3	2
ML Incapacitating Injury	19	14	13	17	17	15	10	19	21	7

Figure 12



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
WE Non-Incapacitating Injury	48	67	61	41	53	35	34	38	42	40
LL Non-Incapacitating Injury	88	86	68	72	66	66	60	61	58	41
RL Non-Incapacitating Injury	30	26	19	25	12	18	25	13	24	23
ML Non-Incapacitating Injury	73	81	64	62	69	59	66	65	69	55

Figure 13



	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
WE Possible Injury	91	88	76	75	85	62	65	65	76	69
LL Possible Injury	176	150	145	135	140	127	128	121	125	95
RL Possible Injury	61	54	61	48	55	43	44	49	39	31
ML Possible Injury	136	142	123	118	101	91	114	120	128	86

Figure 14

Summary

This set of time series looks at the spread of specific types of injuries by each reservation and surrounding area.

The general trends are as follows:

- As you can see both types of severe injuries behave rather erratically. This could suggest that the nature of fatal and incapacitating injuries are far more random than the non-severe accidents
- The interesting trend in non-severe accidents in the last two figures shows these accidents run nearly parallel with each study area. The accidents also are similar in shape which may reveal a common point process all of these regions share despite their location.

Overall, as we can see the numbers of accidents are in steady decline. This is somewhat counter-intuitive to Horan and Hilton's original assessment that among Native populations motor vehicle-related accidents are on the increase, while nationally accidents are in decline. A careful assessment of these metrics must be noted because descriptive measures can be misleading.

For instance, due to the included drive-time areas, the conclusion that Native vehicle accident are not increasing could be due to additional non-Native accidents included the study. Horan and Hilton also maintains that due to lack of timely data, and available tools, Tribal stakeholders simply do not have an accurate data set to show to the contrary. This is why this report is so important in identifying problem areas in analysis, so that when a future proposed framework is provided to Tribes for their input, their own Tribal data could help in alleviating the variability in the data aggregates. This will play a future

role in providing prototypes of possible data collection tools that make it possible for Tribes to collect traffic safety data with the same structure as MnCMAT for meaningful comparisons.

Road Safety: Rural Versus Urban

Finally, road safety research is argued to be still in its infancy. It is suggested we segregate various trends of approach to road safety and the analysis of collisions. To further hypothesize about the accident trends in Indian country, it is important to understand that despite the metrics reported in this chapter paint a picture of accident decline, one fundamental outcome that can help explain the nature of Tribal fatalities is the rural versus urban divide: By the percentages, accidents and fatalities in rural areas are higher than urban areas.

Loo and Anderson (2016) explain this in more detail:

> The rural–urban divide in road safety has been recognized worldwide. In a road safety report on European countries, it was found that 50%–75% of the traffic collisions causing injuries happened in urban built-up areas (OECD 2002). Nonetheless, more than 60% of the fatalities in traffic collisions happened in the rural areas. The risk of fatality in collisions was much higher on roads in rural areas than in urban areas. Furthermore, there seems to be a distinctive risk-taking driving culture in the rural areas (Eiksund 2009).
>
> In North America, Mueller et al. (1988) found that the rate of motor vehicle–pedestrian collisions was higher in urban areas, but the death rate in collisions was generally higher in the rural areas. The rural–urban divide was related to vehicle

> speed, availability of emergency care, age and sex distribution of the population, and proximity to definitive medical care. Moreover, the National Highway Traffic Safety Administration of the United States (2008) showed that road fatality rates were higher in the rural than urban areas from 1997 to 2006. In 2006, 56% of all fatal collisions in the country happened in rural areas, but only 23% of the population lived there. In addition, more rural drivers were found to have been drunk-driving, speeding, and driving unrestrained than urban drivers (p. 301).

As we move forward it is important to take into account the body of literature that exists in designing a framework for Tribal traffic safety so that stakeholders fundamentally understand the nature of the problem.

Chapter 3

Additional Exploratory Data Analysis: Kernel Density Estimation

We have established first (global) and second order (local) effects when paired together complement an EDA by allowing a researcher to identify processes that may not be entirely visible from a global view. The results of the Getis-Ord analysis were a good starting point in understanding the fundamental point process.

Typically, a researcher may use a number of techniques to better understand the nature of a preliminary point process. Chapter 4 outlines the need for refined linear network analysis in order to make comparative judgments about planar methods that do not take into account the dependent structure of point processes constrained to a network. This chapter focuses on another common exploratory technique to begin formulating a more refined set of statistical hypotheses to recommend for future analysis.

Yamada and Thill (2007) write:

> Kernel Density Estimation (KDE) is one of the most popular methods for analyzing the first order properties of a point event distribution partially because it is *easy to understand and implement*. Some KDE tools are already made available in some leading commercial GIS software, e.g., the Spatial Analyst Extension of ESRI's ArcGIS, as well as some popular spatial statistical analysis software, such as CrimeStat. The planar KDE has been used widely for traffic accidents ''hotspots" analysis and detection. The recent examples include study of urban cyclists traffic hazard intensity, pedestrian crash zones detection, wildlife-vehicle accident analysis, highway accident ''hot spot" analysis, etc.

> The purpose of KDE is to produce a smooth density surface of point events over space by computing event intensity as density estimation. In planar KDE, the space is characterized as a 2-D homogeneous Euclidian space and density is usually estimated at a large number of locations that are regularly spaced (a grid). However, in analyzing the spatial pattern of traffic accidents, which usually occur on roadways and inside a network, the assumption of homogeneity of 2-D space does not hold and the relevant KDE methods are not readily applicable. Special considerations are thus needed for measuring such point events occurring in network spaces.

My initial thoughts after I had read Horan and Hilton's results from the Getis-Ord hot spot analysis was similar to Yamada and Thill's supposition: because I was seeing hot and cold spots manifesting in and around townships, I wanted to better understand if fatal accidents exhibited a similar pattern as the hot spots analysis.

Using ArcMap, I was able to produce a kernel density estimate for each reservation and surrounding area to get an initial look what was happening spatially. Since the clusters of the hot spots were an aggregation of the number killed given all accidents, I surmised the if we could combined all fatal accidents as one point pattern, we might be able to gain additional insight as to the effect of a zero-inflated situation of non-fatal accidents had played in the hot spot analysis.

The following figures are the result plotting all fatal accidents, regardless of the number killed. The data set indicated fatality with how many persons had been killed on a scale from one to four; I simply aggregated all of these points for this analysis.

Some key points of interest were to investigate the following:

- Examine the locations of fatal accidents with respect to the reservations boundaries and the outer drive-time areas for comparison
- To better understand if more accidents occurred on the reservation
- To surmise that if the majority of fatal accidents did fall outside the reservation boundaries; do the surrounding townships, possible work sites, or any other factors influence why this so?
- Obtain a better understanding of the maximum neighbor distances in square miles to assess how far does the planar KDE measure the density of fatal accidents

After obtaining the KDE estimates, I have indicated on the maps points in interest that are of increasing intensity. In addition, the black star indicates the location of each respective Tribe's casino(s). Like in Chapter 2, these figures have been provided to reader for study. There is a brief synopsis at the end to summarize each KDE

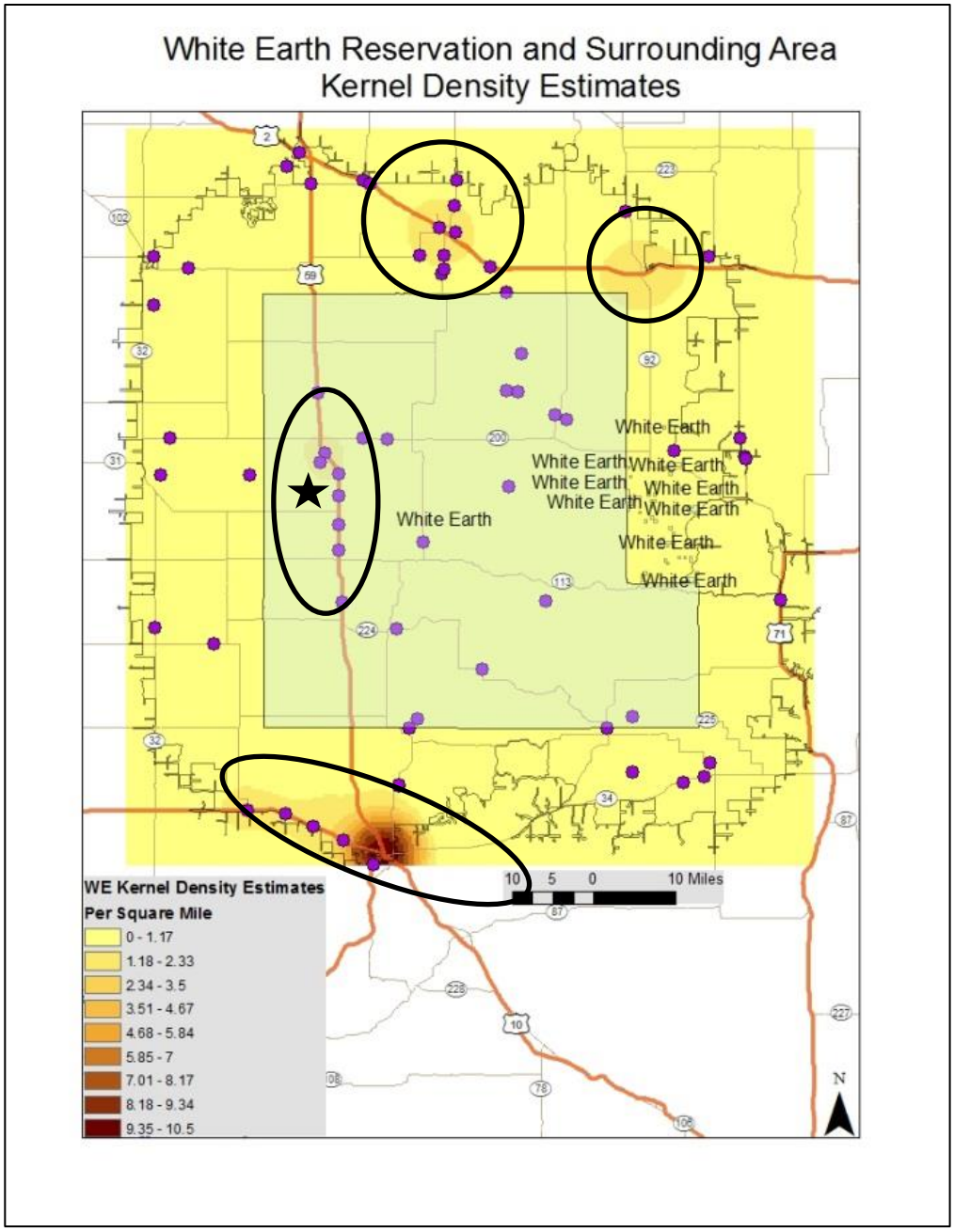

White Earth Reservation and Surrounding Area
Kernel Density Estimates
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
White Earth
10 5 0 10 Miles
WE Kernel Density Estimates
Per Square Mile
0 - 1.17
1.18 - 2.33
2.34 - 3.5
3.51 - 4.67
4.68 - 5.84
5.85 - 7
7.01 - 8.17
8.18 - 9.34
9.35 - 10.5
N

Figure 15

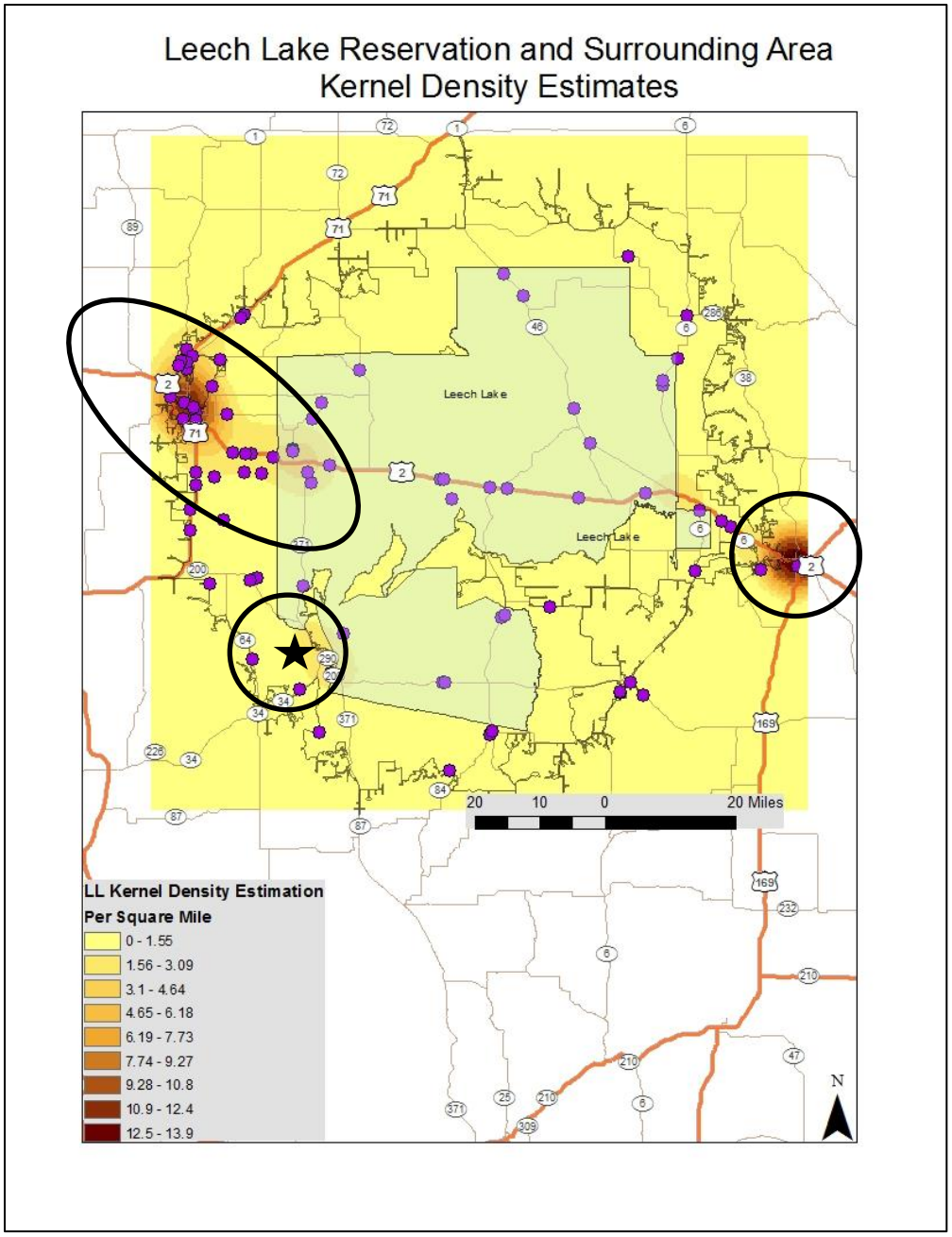

Leech Lake Reservation and Surrounding Area
Kernel Density Estimates
Leech Lake
Leech Lake
20 10 0 20 Miles
LL Kernel Density Estimation
Per Square Mile
0 - 1.55
1.56 - 3.09
3.1 - 4.64
4.65 - 6.18
6.19 - 7.73
7.74 - 9.27
9.28 - 10.8
10.9 - 12.4
12.5 - 13.9
N

Figure 16

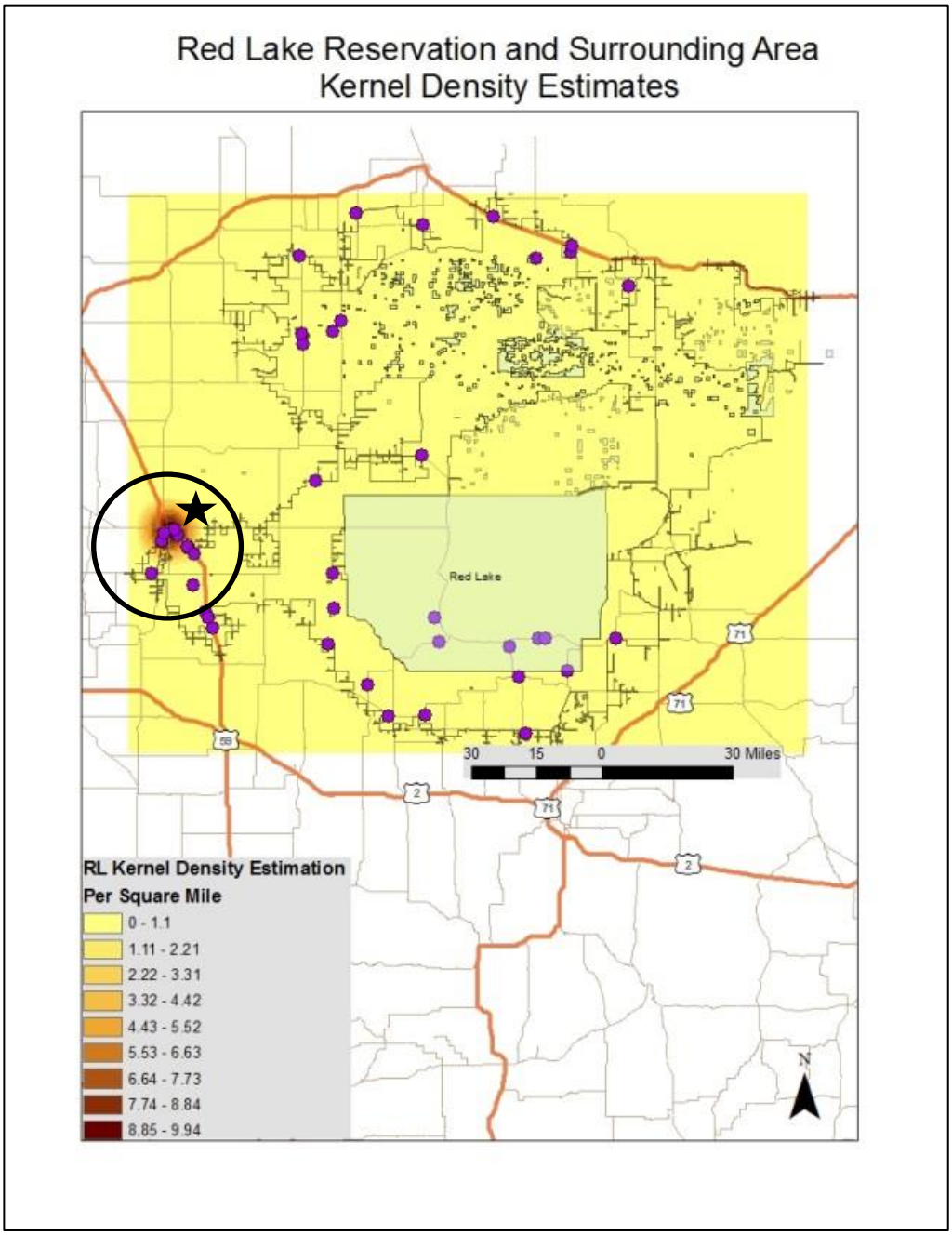

Red Lake Reservation and Surrounding Area
Kernel Density Estimates
Red Lake
30 15 0 30 Miles
RL Kernel Density Estimation
Per Square Mile
0 - 1.1
1.11 - 2.21
2.22 - 3.31
3.32 - 4.42
4.43 - 5.52
5.53 - 6.63
6.64 - 7.73
7.74 - 8.84
8.85 - 9.94
N

Figure 17

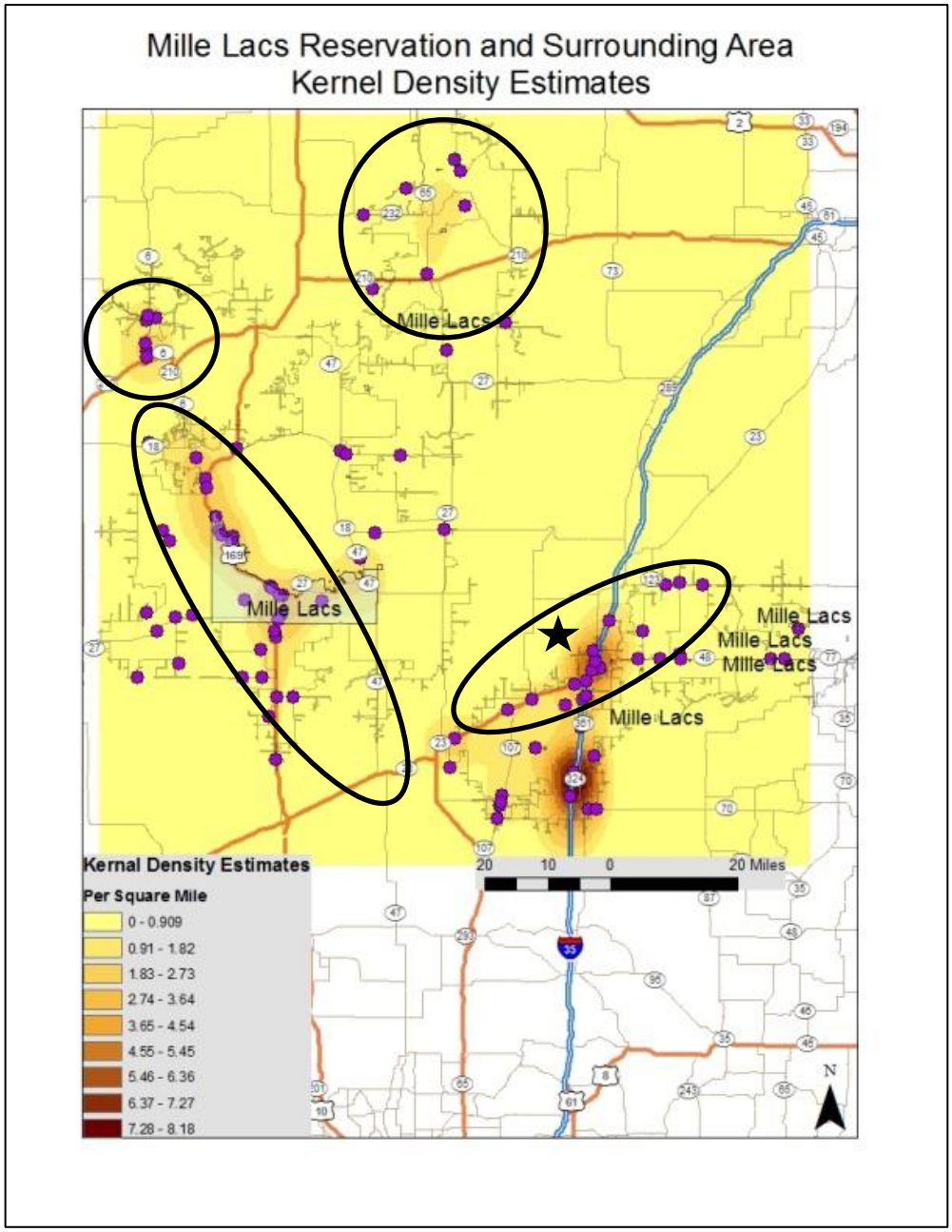

Mille Lacs Reservation and Surrounding Area
Kernel Density Estimates
Mille Lacs
Mille Lacs
Mille Lacs
Mille Lacs
Mille Lacs
Mille Lacs
Kernal Density Estimates
Per Square Mile
0 - 0.909
0.91 - 1.82
1.83 - 2.73
2.74 - 3.64
3.65 - 4.54
4.55 - 5.45
5.46 - 6.36
6.37 - 7.27
7.28 - 8.18
20 10 0 20 Miles
N

Figure 18

Summary

As we can see, each reservation and drive-time areas are unique not only in point distribution of fatal accidents but also on geometry. One of the most interesting things captured by these images is the distinct fatal points in the rural versus urban areas. It is not hard to see the deepest intensities lie within townships with radii of less than fourteen square miles. Because of the unique geometry of Mille Lacs, you can see the drive-time network was much more spread out and thus included many more townships. This could explain some of the explicit variation found in the analysis: the aggregations of points from numerous neighboring townships versus other reservation areas have one neighboring township only.

Since we added an additional covariate such as the location of the casino it may be easier to understand why Horan and Hilton chose an included drive-time area. Further analyses could look at Tribal housing and a theoretical transportation network could be established to further study how fatal accidents might be tied to commuting routes for Tribal members going to and from work. In addition, the fatal accident points in the rural on reservation networks can be further studied with the understanding from Chapter 2 regarding rural versus urban divide methods for road safety. Smaller on network analyses can be performed to investigate clustering as well as modeling.

Yamada and Thill (2004) maintain that planar KDE can create false detection associated with on network phenomenon and use of a planar K-function "entails a significant chance of over-detecting clustered patterns." Chapter 4 discusses ways of designing an on network K-function for comparison and since current methodology

already exists in the literature, this could prove useful should Tribes want specific results similar to what was suggested in the previous paragraph.

Chapter 4

Future Modeling Recommendations: Spatial Analysis along Networks

Throughout the previous chapters, we were focused on establishing a hierarchy of data analysis to demonstrate a number of statistical techniques that naturally move towards forming a more complex set of recommendations to present to Tribal stakeholders interested in tribal traffic safety. The emphasis has been on building more meaningful set of assumptions using an extensive literature review, descriptive measures, and exploratory data analysis to recommend future design techniques that allow for robust set of spatial modeling techniques for future analyses. Horan and Hilton's analysis was the cornerstone in expanding the existing knowledgebase.

The proposed EDA was designed to examine CSR, explore possible point processes using appropriate statistical methodologies, introduce additional EDA techniques to strengthen the original results, and provide links to additional covariates as it relates to GIS analysis for model building. This final chapter provides an overall assessment of the metrics to demonstrate careful assessment of assumptions is paramount to quality inferential model building.

Let us review a few key points of the project tasks:

Task 1: Tribal Data Analysis: Spatial analysis and testing of the traffic data for additional metrics to strengthen the current results.

Task 1.1 Tests for complete spatial randomness (CSR) such as quadrat analysis, Ripley's K simulation envelopes, and point pattern analysis.

As we explored in chapter one, non-parametric diagnostics tests such as quadrat rely on key assumptions such as independence and homogeneity. The results of some initial exploration revealed that assuming uniformity of accident crashes across the entire reservation and surrounding areas is problematic because:

- The unit area of each space contains inhomogeneous pocket clusters of accidents due to included townships in the surrounding drive-time areas
- A simple exploratory quadrat used townships geometry to loosely assess the counts using first and second order neighbors revealed clear non-random patterns such as linear and non-linear pattern within a given quadrat. When roadways were overlaid onto the quadrat, clearly a 1-D network of roads was influencing the actual locations of the point patterns.
- The point pattern exhibits some sort of point process; this was done through visual assessment.

The results indicate that care must be taken to properly investigate results obtained from this analysis. The literature review in Chapter 2 referenced that historically simple descriptive measures are favored over more complex modeling techniques, so if

presenting these methods to Tribal stakeholders, it is important to understand any assumptions we might violate when setting up a test for spatial clustering.

The body of literature on Ripley's K-function is extensive and for lack of computational resources for network analysis, this could be the most suitable test since we are simulating envelopes to view an estimate of the expected $K(d)$, but care must also be taken in this situation as well.

Lu and Chen (2007) quoted previously:

> Most applications of K-function, including those mentioned above, use Euclidean distance to represent spatial separation between points. This is not a problem when the point pattern is a continuous and unrestrained distribution over a Euclidean plane. However, Euclidean distance is no longer an accurate measurement when the point distribution is subject to certain restrictions. A set of points distributed along urban streets is but one example. From the definition of K-function, we can see that the measure of distance plays a critical role in evaluating the clustering situations in a point set. Using different measures of distance would result in the K-function behaving differently, leading to different conclusions regarding the patterns of the same point set. (p. 614)

The results of our investigation show the point pattern is undoubtedly restricted to a network of lines. This prevents the intensity (accidents) from having equal probability to occur over the entire space; a small cross section of the White Earth reservation validates this observation:

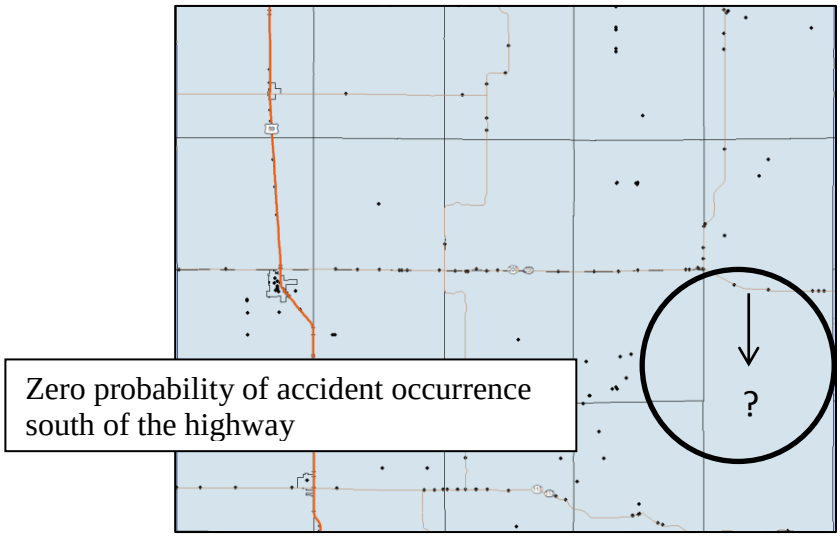


Figure 19

As we can see, the assumption of a point pattern being continuous planar over a Euclidean plane is violated because any departure from the road network has zero likelihood to occur. The circle in Figure 19 is to draw attention to distance departure from the roads will have no accidents because the likelihood an accident occurring say, in the middle of an agricultural field in the most practical sense is essentially zero. Although it is not uncommon to use these tests for exploratory purposes and comparisons; understanding what assumptions an investigator is choosing to violate for the results is important to document.

Task 1.2. Further investigation of possible Poisson related spatial point processes and distribution assessment.

In assessing whether the point process can modeled from a Poisson point process, again care must be taken to properly account for homogeneity, uniform intensity, and

independence. In addition, the injury severity contained in the data is what is referred to as s multi-type point pattern, where the pattern of point points is of several different types. The point pattern might be better represented when the marks are defined as categorical, rather than the number killed.

Fitting a Poisson process model to a point pattern is an effective technique due to the flexibility of allowing adjustments of model assumptions to accommodate inhomogeneity, additional covariates, interactions, etc. In addition to the injury severity, the traffic data contained additional covariates that could be used to further investigate other variability if the point process. Poisson modeling could be an effective tool for future traffic analysis if the data can be structured to validate most of the model assumptions.

Task 1.3. EDA of possible links to provided covariates whenever appropriate, with additional spatial autocorrelation techniques

The covariates such as injury type with respect to day, month, and road surface have been included in the appendix. As we talked extensively in Chapter 1, the nature of spatial autocorrelation is best explained by Waldo Tobler who made famous his "first law of geography," which states that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970).

There is undoubtedly more work to be done in this analysis in formulating additional metrics to discuss further design of spatial components that properly account for spatial autocorrelation. As we have examined this concept through the lens of GIS, it is important to formulate a hypothesis that is accurate in explaining not just the

relationship between two spatial points, but if this relationship is the result of an attainable point process that can be modeled for further statistical development.

Although, this underlying concept was not directly used in Horan and Hilton's original analysis, much work can be done to examine more closely local indicators of spatial association (LISA). It is with this collective of ideas; I would like to recommend the best possible methodology for future study and to present these findings to Tribal stakeholders.

Spatial Analysis along Networks

Traffic accidents are a point process of events that are strongly constrained to network. Okabe and Sugihara (2012) refer to these events as *network constrained events*. This is different than what are defined as *alongside-network events* such as business located alongside a roadside network. As we have seen, planar spatial analysis has its limitations due to the following assumptions:

1. Events occur on an unbounded continuous plane
2. Distances are measured by Euclidean distance.

These assumptions are the result of the most common and convenient way of computing a distance as well as the shortest path distance has been thought to be best approximated by Euclidean distance. Throughout this analysis, I have maintained that through the process of exploring point patterns, any number of techniques may be used, however each decision made is dependent on which assumptions a researcher is willing to violate to obtain their results.

As we analyzed the KDE in Chapter 3, we begin to see how planar density estimates tend to cluster in townships, rather than collectively assess the assumption of uniformity throughout the planar space. Each density plot revealed where townships occur, the per square mile radius cluster of fatal accidents remains in less than a 14 square mile radius; with not much occurring in rural sites capturing other fatal accidents. This information is indicative of the inhomogeneity of the process occurring in urban versus rural areas and again the points were constrained to network of local roadways located within city limits.

Given the set of limitations in realizing a network constrained point process test for spatial clustering, Okabe and Sugihara (2012) have outlined a methodology to address this issue:

> To overcome the above limitations of planar spatial methods, we now introduce a new type of spatial analysis that assumes:
>
> AN1: Events occur on and alongside a network.
>
> AN2: If a method for analyzing the events includes distance variables, the distance are shortest path distances.
>
> We make a few remarks on the above two assumptions, AN1 and AN2. The first assumption AN1 describes places where events occur. The on-network relation is obvious. Events occur exactly on a network, such as traffic accidents. The second assumption, AN2, specifies distance variables included in spatial methods. We notice from the above definition of network spatial analysis that it has salient features distinct from those of planar spatial analysis.

- First, by definition, network spatial analysis can properly analyze events occurring on and alongside a network.
- Second, network spatial analysis can easily take account of directions, such as directions of current in a river and traffic flow regulation on a street network.
- Third, network spatial analysis can treat detailed networks using a common data structure.
- Fourth, network spatial analysis can easily treat networks in three-dimensional space, such as underpaths and crossover bridges
- Fifth, as will be shown in Section 2.3, network spatial analysis can treat non-uniform activities on a network more easily than planar spatial analysis can.
- Sixth, network spatial analysis gains analytical tractability because a network consists of one-dimensional line segments. Mathematical derivations on a one-dimensional space are more tractable than those on a two-dimensional space. (pp. 6-7)

Baddeley, et al. (2016) provides additional insight:

> For such data, it is clearly not appropriate to use statistical techniques designed for point patterns in two-dimensional space, such as Ripley's K-function. The analysis needs to take into account the geometry of the network. In the last decade, substantial research effort has been addressed to this problem by A. Okabe and collaborators.
>
> The dependence between points in a point process is more difficult to study on a linear network than in the two-dimensional plane. A linear network is not a homogeneous space: different spatial locations on a network are surrounded by

different configurations of lines. Recently it was discovered how to ‘adjust’ for the geometry of the network when defining quantities like the K-function (p. 711).

Danger in using the two-dimensional K-function

To measure correlation between points on a linear network, it is clearly not appropriate to apply Ripley’s K-function to the two-dimensional spatial locations of the points. At least, it would be fallacious to take a point pattern on a linear network, forget the linear network and retain only the spatial (x, y) coordinates, compute the empirical Ripley K-function of these points, and compare this with the theoretical K-function for a completely random pattern in two dimensions.

The apparent discrepancy is an artifact, arising because we have chosen the wrong null hypothesis: the envelopes are computed from simulations of CSR in two dimensions, while the appropriate null hypothesis is a Poisson process on the linear network.

The fundamental aspect of the K-function relies on ***pair correlation*** defined for a two-dimensional point process. To define pair correlation on a linear network we need to define the pair correlation we first introduce the second moment intensity (or product density) $\lambda_2(u,v)$. Heuristically, given any two distinct locations u, v on the network, we consider a very short piece of the network around location u of length d_1u, and similarly a short piece of network around v of length d_1v, sketched in Figure 20. The probability $p(u,v)$ that both of these short segments contain at least one random point is assumed to be $p(u,v) = \lambda_2(u,v)d_1ud_1v$.

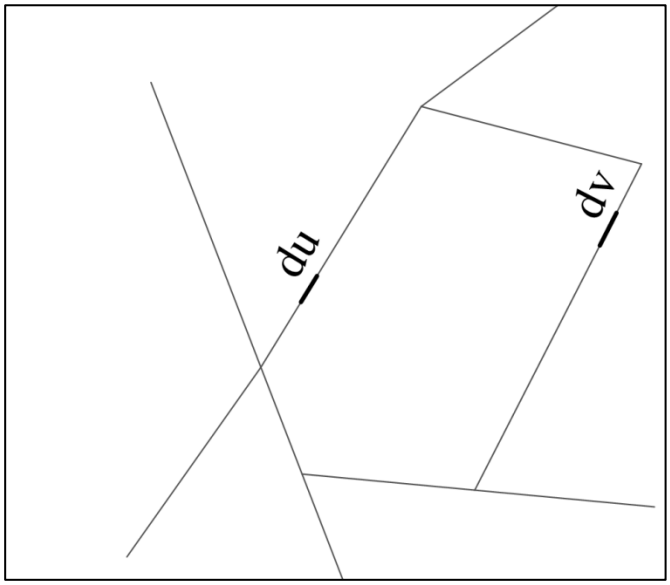


Figure 20

Then for any two line segments $A, B \subset L$ that's are disjoint $(A \cap B = 0)$ we have:

$$\mathrm{E}[n(\mathbf{X} \cap A)n(\mathbf{X} \cap B) = = \int_A \int_B \lambda_2(u,v) d_1 u d_1 v$$

Analogous to adjusting for inhomogeneity in the 2-D Ripley's K-function, the second moment intensity function can be thought of as instead of calculating the probability of a two random points falling at locations u and v.

The general pair correlation function g becomes:

$$g_2(u,v) = \frac{\lambda_2(u,v)}{\lambda(u)\lambda(v)}$$

Where λ is the intensity function. The general pair correlation function on a linear network has the same interpretation as it does in the 2-D case.

Finally, modifying the two-dimensional point pattern where $g_2(u,v)$ is defined to depend only on the Euclidean distance from any two locations:

$$g_2(u,v) = g(\|u-v\|)$$

We assume $g_2(u,v)$ depends only on the shortest-path distance $d_L(u,v)$, and thus:

$$g_2(u,v) = g(d_L(u,v))$$

With $g(r)$ is the linear network pair correlation function.

Let us look at an example: By definition, a Poisson process on a linear network has $g(r) = 1$. A value $g(10) = .5$ would mean that for a pair of locations u and v separated by a shortest-path distance of 10 units, the probability that random points fall in both locations is half as much as it would be for a Poisson process with the same intensity. The pair correlation $g(r)$ is defined for all distances $r \le D$, where D is the diameter of the network (maximum possible shortest-path distance between any two points in the network) (Baddeley et al. ,2016, pp. 732-736).

Figure 21 is an interesting location where a linear network in small scale applications might be of interest. When considering prototype applications of GIS analysis to present to Tribal stakeholders, the network between a Tribal place of business and its relationship to near Off Reservation Trust land could be of interest. The use of the linear network correlation function could be used to assess and model the point process in network commute times, or transportation. Although the realization of this concept is beyond the

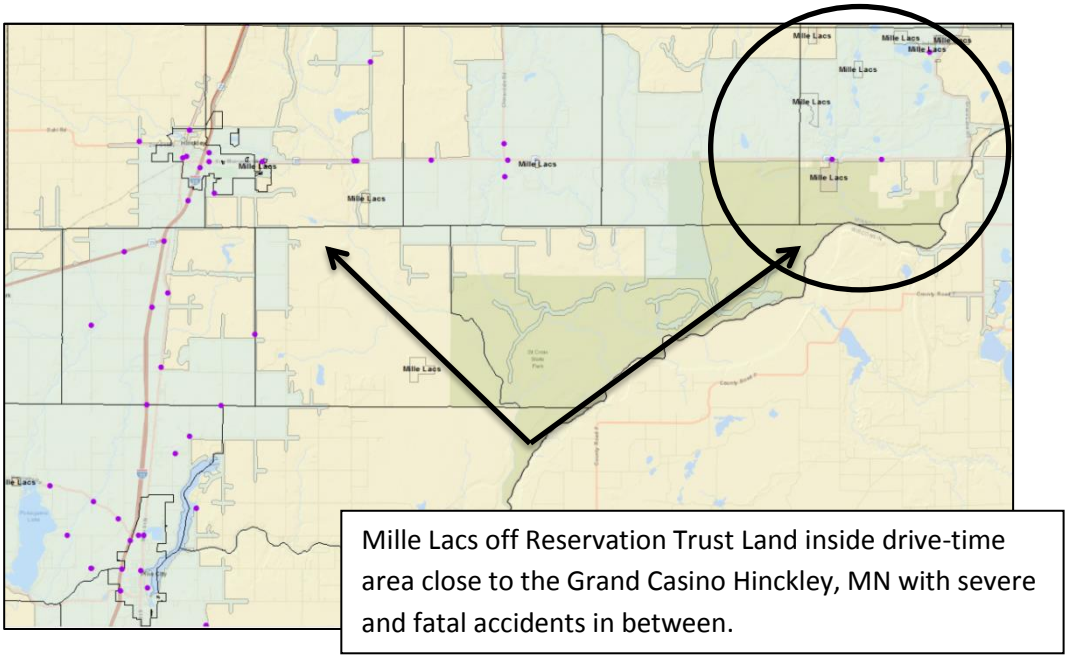


Figure 21

scope of this reporting, spatial analysis of this type is feasibly possible. One drawback is the computation network of lines and point to make the network is computational intensive, so small scale analyses are recommended unless access to high performance computing platforms is considered.

As we have seen, the number of tools available to create a robust GIS framework for tribal traffic safety begins with adhering to preliminary results, exploring descriptive measures, formulating exploratory data analyses, and finally developing all of this information into the most effective way to spatially model a point process. It my hope, this allows for an in-depth discussion between stakeholders as to what type of spatial and data analyses are available and how we are able to empower Tribal communities to use these tools for their own traffic safety studies.

Chapter 5

Conclusions and Recommendations

Roadmap to Effective Modeling of Traffic Safety

This document contains an exhaustive review of how point patterns, processes and modeling can be constructed to utilize GIS in traffic safety. At times, this may be construed to be overkill in terms of how an exploratory process unfolds, but too often Tribes are never consulted in these processes, and as an enrolled member of a federally recognized Tribe; I am honored to lend my expertise to this current initiative. The level of expertise I possess is precisely why I am able to advocate for an even higher expectation as to the quality data science Tribes are owed when we choose to work them on community development.

The design of this report was to provide a baseline of metrics as a way to critically examine the data quality needed when forming a partnership with Tribal communities for traffic safety: the data is just as important as the policy. As I mentioned briefly in executive summary, the results of this ongoing project is a realization of the concept of *data sovereignty* I have been developing through my doctoral dissertation in computational science and statistics. The collective framework is intended to empower Tribal stakeholders to use data as a matter of self-determination. Collectively, this idea can encompass any number of data driven decision making tools to assist all stakeholders in nation building through statistical design and analysis.

The next phase of the project is to use this document and other developed prototypes to begin a design strategy using data sovereignty as the model framework to

present traffic safety to Tribal stakeholders for their input and finally developing data solutions that allow for Tribal data collection and analysis to allow for further representation in traffic study initiatives.

As we continue with the next phase of this project, I wanted to conclude this analysis with the building blocks of any good design, so that we are reminded that roadmaps are not just realizations of point processes, but also frameworks for all that we do with data. I think it speaks for the level care needed in working with Tribal communities to the better outcomes through extensive thinking and praxis.

Baddeley et al. (2016) elegantly outline this as the *Scope of Inference:*

> There is a choice concerning the scope of statistical inference, that is, the 'population' to which we wish to generalize from the data.
>
> At the ***lowest level*** of generalization, we are interested only in the region that was actually surveyed. In applying precision agriculture to a particular farm, we might use the observed spatial point pattern of tree seedlings, which germinated in a field sown with a uniform density of seed, as a means of estimating the unobservable, spatially varying, fertility of the soil in the same field. Statistical inference here is a form of interpolation or prediction. The modeling approach is influenced by the prediction goals: to predict soil fertility it may be sufficient to model the point process intensity only, and ignore inter-point interaction.
>
> At the ***next level***, the observed point pattern is treated as a 'typical' sample from a larger pattern which is the target of inference. To draw conclusions about an

entire forest from observations in a small study region, we treat the forest as a spatial point process X effectively extending throughout the infinite two-dimensional plane. In order to draw inferences based only on a sample of X in a fixed bounded window W, we might assume that X is stationary and/or isotropic, meaning that statistical properties of the point process are unaffected by vector translations (shifts) and/or rotations, respectively. This implies that our dataset is a typical sample of the process, and supports nonparametric inference about distributional properties of X such as its intensity and K-function. It also supports parametric inference, for example about the interaction parameter γ of a Strauss process model for the spatial dependence between trees.

At a ***higher level***, we seek to extract general 'laws' or 'relationships' from the data. This involves generalizing from the observed point pattern to a hypothetical population of point patterns which are governed by the same 'laws' but which may be very different from the observed point pattern. One important example is modeling the dependence of the point pattern on a spatial covariate (such as terrain slope). This is a form of regression. We might assume that the intensity $\lambda(u)$ of the point process at a location u is a function $\lambda(u) = \rho(Z(u))$ of the spatial covariate $Z(u)$. The regression function ρ is the target of inference. The scope of inference is a population of experiments where the same variables are observed and the same regression relationship is assumed to hold. A model for ρ (parametric, non-, or semi-parametric) is formulated and fitted. More detailed inference requires either replication of the experiment, or an assumption such as

joint stationarity of the covariates and the response, under which a large sample can be treated as containing sufficient replication.

At the ***highest level***, we seek to capture all sources of variability that influence the spatial point pattern. Sources of variability may include 'fixed effects' such as regression on an observable spatial covariate, and also 'random effects' such as regression on an unobserved, random spatial covariate. For example, a Cox process is defined by starting with a random intensity function $\Lambda(u)$ and, conditional on the realization of Λ, letting the point process be Poisson with intensity Λ. In forestry applications, Λ could represent the unobserved, spatially inhomogeneous fertility of soil, modeled as a random process. Thus Λ is a 'random effect'. Whether soil fertility should be modeled as a fixed effect or random effect depends on whether the main interest is in inferring the value of soil fertility in the study region (fixed effect) or in characterizing the variability of soil fertility in general (random effect).

I look forward to assisting any stakeholder interested in helping American Indian Tribes with the complex issues they face, and I hope to continue to make an impact moving forward.

This ends the report.

END
TRIBAL TRAFFIC SAFETY
MANUSCRIPT BRIEF

Final Thoughts and Current State of the Project

As of this writing in April 2018, the Using GIS to Improve Tribal Traffic Safety is still ongoing. The purpose of the manuscript brief was to explore further topics that related to traffic related crashes in and around selected Minnesota reservations. My primary focus was to provide an exploratory data analysis of the original hot spot analysis and to contribute the data sovereignty framework as a way to promote tribal traffic safety.

During the design of the framework, one of the objectives was to present the framework to appropriate stakeholders for evaluation. This task ended up being slightly difficult, since the framework wasn't designed to ask tribal stakeholders what they thought; rather its strength was in designing a data domain and work *with* stakeholders to create the framework for a particular task.

However to clarify, the process that is ongoing has been very useful in interviewing some tribal officials as to nature of the data sovereignty as a framework. The prototype nature and feedback we have received has played a crucial role in how I have addressed the data sovereignty framework moving forward.

The data sovereignty framework served as a developmental prototype in this case study, and now that this proof of concept can be applied in a more practical context. The next chapter utilizes these findings to increase the scope of using this framework to create a smart solution. The machine learning technique I have chosen was designed specifically perform a task that would assist tribes with obtaining digital infrastructure with the intent of alleviating the potential economic cost of sending personnel into the field to collect data.

As we will see, the technique creates a simple, yet powerful way to obtain a *Master Address File* that can be used for several geospatial projects to increase capacity, alleviate budgets, and assert ownership of data that is not exclusively in the tribe's power to obtain. Further topics will be discussed as to how to get this SMART solution into the hands of tribal stakeholders who could make use of it.

Chapter 4

Case Study 2: Using Machine Learning in GIS to Create a SMART Solution in Tribal Census and Other Spatial Outcomes

So What is Machine Learning?

A comprehensive body of work by Izenman (2008) outlines the basics of machine learning and through my research on the topic provides some of the most precise explanations of the topic. The idea of machine learning:

> Machine learning evolved out of the subfield of computer science known as artificial intelligence (AI). Whereas the focus of AI is to make machines intelligent, able to think rationally like humans and solve problems, machine learning is concerned with creating computer systems and algorithms so that machines can "learn" from previous experience. Because intelligence cannot be attained without the ability to learn, machine learning now plays a dominant role in AI.
>
> The machine-learning community divides learning problems into various categories: the two most relevant to statistics are those of *supervised learning* and *unsupervised learning*.
>
> Supervised learning: Problems in which the learning algorithm receives a set of continuous or categorical input variables and a correct output variable (which is observed or provided by an explicit "teacher") and tries to find a function of the input variables to approximate the known output variable: a continuous output

variable yields a regression problem, whereas a categorical output variable yields a classification problem.

Unsupervised learning: Problems in which there is no information available (i.e., no explicit "teacher") to define an appropriate output variable; often referred to as "scientific discovery." The goal in unsupervised learning differs from that of supervised learning. In supervised learning, we study relationships between the input and output variables; in unsupervised learning, we explore particular characteristics of the input variables only, such as estimating the joint probability density, searching out clusters, drawing proximity maps, locating outliers, or imputing missing data.

In this analysis, the machine learning technique called a *Support Vector Machine* (SVM) is used. It is a supervised learning technique in which a set of training data is used to 'train' the algorithm to identify objects in a geospatial raster image for image classification. The reason this technique was chosen was due to the strength of its predictive power through *pre-learning* the nature of the problem at hand. Although, there are a number of classification techniques in both supervised and unsupervised learning such as k-nearest neighbor clustering, neural networks, or hierarchical clustering; a SVM was simply the most efficient way to create the polygons using hyperplanes to isolate the very specific infrastructure I was interested in.

Introduction

The data sovereignty framework allows for many data domains, in this case Tribal GIS outcomes. Using this framework, I have developed a machine learning technique to allow

for tribes to use GIS and a *Support Vector Machine* to train and predict housing domiciles through high resolution satellite and drone imagery. This will provide tribes with a method of obtaining point patterns and polygons to create a Master Address File (MAF) for use in many spatial dependent systems, census being one of them.

The potential for tribes obtaining accurate geographic information about crucial infrastructure through this technique is extremely powerful as it has other implications. Examples include cross-referencing existing GIS locations with no need to visit every location, and providing these point patterns for additional use such as transportation, emergency 911, or utility locations. To be cost effective, I am currently working in a small area of one of the nine reservations in South Dakota that contains census blocks of tribal housing that is defined by the U.S. Census Bureau shape files of tribal boundaries.

The companies that provide this imagery can be quite expensive, so to make imagery cost effective, a preliminary analysis of a small tribal housing development was considered. There are ways to obtain the maximized areas of consideration without sacrificing dwellings we are interested in such as drone mapping. Notwithstanding private satellite image companies; tribes can also obtain geospatial information through the U.S. Department of Interior, Bureau of Indian Affairs Branch of Geospatial Support using the Enhanced View program (EV) from Digital Globe as part of the National Geospatial-Intelligence Agency (NGA) at no cost.

As a matter of sovereignty, this SMART solution creates a template that will allow any tribe to scale their reservation boundaries to maximize the spatial areas of interest, while minimizing the cost of the satellite imagery. The findings of this analysis

will be presented formally after the literature review in this chapter. Again, this process is being designed as a template to give any tribal group a method to obtain spatial data outcomes without the need for ineffective costs associated with sending individuals into the field that a machine learning algorithm could provide if a high resolution image of an area could be obtained.

While I was initially conducting research on tribal sovereignty issues, I did not realize that under Title 13, the U.S. Census Bureau is not allowed to share Master Address File systems with any entities outside the bureau, including tribes. The previous consulting work I have done with tribes regarding tribal census indicated the need for a more streamlined way of unifying addresses of tribal citizens living on or near the reservation. The point pattern created from the addresses collected in the tribal census was highly inaccurate due to human error, and the data collection protocols had also produced a number of repeated household visits.

After consulting with tribal officials, it was explained that a very small team of workers were assigned to collect as many censuses as possible; but given the size and scope of the reservation, it would have taken many teams of individuals at great economic cost to make a complete census collection realistically possible.

At the time, the tribe's planning department had not anticipated that accurately collecting geographic data related to census was an utmost priority. There are tribes whose budgets are governed by funding formulas that do not always allocate adequate resources for MAF procurement since the census bureau already maintains a record of

addresses separately, thus allocating resources for a governmental task already in place does not make sense

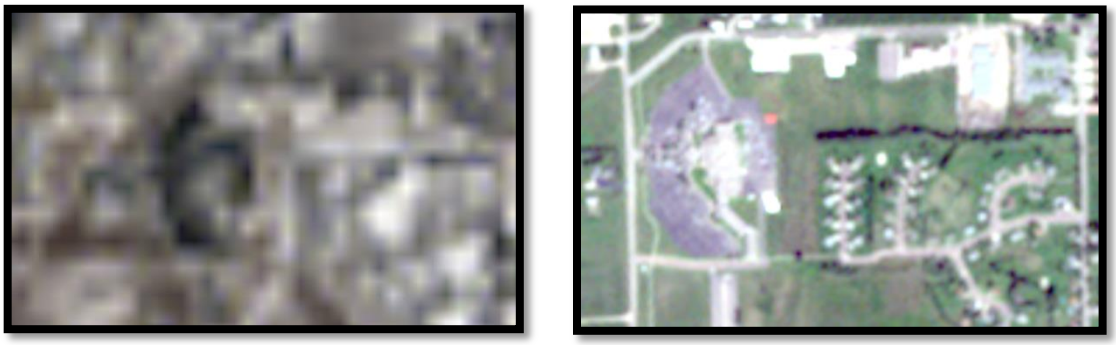
Tribes do however have the desire to act on their sovereignty to collect census data because of undercounts and lack of representative funding allocations are a result of those inaccuracies. The advent of advanced GIS database systems, machine learning, satellite and drone imaging have made tasks of constructing an MAF much easier. This is the reasoning behind this case study.

Spatial data has undergone a very fast transformation with increased computing power and software availability. Image classifying machine learning techniques are almost certainly being used in some capacity. Master Address File is simply a term that refers to spatial data related to the position of people or infrastructure. The images and corresponding point pattern, process, or models all come from how well the training data set of images best reflects the pattern of interest. This is not limited to tribes or people.

Image classification requires adequate resolution since pixels that are too large will render an image unrecognizable; but assuming a machine algorithm has an adequate training set, it makes no difference if it is tied to people or objects. A machine can be trained to classify anything such as trees, roads, shrubs, or types of crops in agricultural fields. Again, spatial scale is important to allow the training classifier to obtain enough results for accuracy; thus the more samples the classifier has, the better it can predict what it is looking for.

In personal communication with the lead developer of the Environmental Systems Research Institute (ESRI) machine learning division during the 2017 ESRI User

Conference (UC), I presented this issue and it was confirmed there was no possible way to use image classification on a scale of five meters or above because a support vector machine could not differentiate shapes due to the lack of pixel density, thus the accuracy of the classification was untenable. Figure 4.1 demonstrates visually why this is not possible. The images are the exact same area. Notice the image quality increases as the square meter resolution decreases.



Landsat7 30 meter

Rapid Eye 5 meter

Figure 4.1- Pixel Comparison of Dimensionality of a Tribal Census Tract
Each picture represents the same proposed area of interest

The benefit of obtaining a MAF has many useful applications. The point pattern created by the machine learning process will drastically reduce the cost and man power associated with sending personnel into the field to collect physical addresses. Success can be evaluated through a strategic plan. The accuracy of the point pattern can be assessed by the calculation of a confusion matrix to assess the accuracy of the pixel selections by choosing random points from the original sample and comparing them to the predicted classification images. In addition, by integrating other crucial projects associated with

tribal infrastructure such as developing 911 networks, transportation networks, and utilities these locations can be further vetted.

Smart phones can have the MAF integrated into any project so when personnel go out into the field for projects described above, the ESRI *Workforce* app can be integrated with all GPS coordinates and the accuracy can be verified a second time using a smart phone's real-time GPS coordinate systems for ongoing data collection that may be ongoing for another project. Thus, success can be measured through real projects a tribe needs to undertake rather than just going out into the field to map an area with no plan as to how to use this point pattern information.

Support Vector Machine Theory

Synopsis:

As described above, the purpose of this case study is to explore machine learning techniques that make SMART solutions accessible and cost effective to tribal members who are doing the work for their community. Image classification is a feasible way to integrate technology into developing digital infrastructure. Specifically, the data sovereignty framework regards this case study as a *specified Data Domain Key Indicator*. The other three key indicators are developed in counsel to address the specific needs of the tribal nation that chooses to use this technique.

An extensive literature review is required to understand how the geoprocessing tool in ESRI ArcMap's Support Vector Machine Training Classifier works. This literature review is quite extensive and the underlying mathematical theory is not typically documented in ERSI's general catalog of help topics in the software.

Nonetheless it is important to establish how the image classifier uses these principles to perform the actions many GIS professionals take for granted when a 'black box' algorithm is used to perform computationally intense calculations.

"Support Vector Machines (SVM) belongs to a class of kernel methods and are rooted in statistical learning theory. As all kernel-based learning algorithms they are composed of a general purpose learning machine (in the case of SVM a linear machine) and a problem specific kernel function. Since the linear machine can only classify the data in a linear separable feature space, the role of the kernel-function is to induce such a feature space by implicitly mapping the training data into a higher dimensional space where the data is linearly separable.

"SVMs have been successfully applied to classification problems as diverse as handwritten digit recognition, text categorization, cancer classification using microarray expression data, protein secondary-structure prediction, and cloud classification using satellite-radiance profiles" (Izenman, 2008, p. 369).

Since the general purpose learning machine and the kernel function can be used in a modular way, it is possible to construct different learning machines characterized by different nonlinear decision surfaces (Hofmann, 2006).

Outline of SVM Methodology Literature Review

There are cases that fundamentally form the mathematical foundation of support vector machine methodology:

- The simplest case is when two groups are completely separable. Support vectors are called a linearly separable case

- The case where two groups are linear, but non-separable
- Defining a kernel and the *kernel trick*
- Non-Linear Transformations
- The Radial Basis Function (RBF) kernel
- Grid search for parameter of the RBF
- Construction of a multi-class support vector machine

The results of the machine learning image classification were done in the Environmental Systems Research Institute (ESRI) platform, ArcGIS and ArcGIS Pro. This literature provides a strong background of the mathematical foundations of SVM theory; and then outlines the specific techniques used by the software itself.

Prior to looking at theory behind how a support vector machine functions in the hierarchical structure of creating an algorithm for classification; I will discuss briefly previous research I had undertaken in a precision agriculture initiative where one of the topics covered worked in analyzing topics of dimensionality as it related to computational limitations when addressing spatial dependence.

Understanding Dimensionality

During the academic 2016-2017 academic year, I conducted research in evaluating pixel image data and the effect of dimensionality. This was a one year position and was a partnership between South Dakota State University J. Lohr College of Engineering and the College of Agricultural and Biological Sciences to: “turn vast data-generation capabilities of precision agriculture into information that can drive decision-making in the field (SDSU, 2016, p. 11).

This inter-departmental collaboration was aimed at developing spatial-temporal risk models in examining white mold pathology research using historical agricultural and GIS related satellite data. My advisor Dr. Gary Hatfield was the lead spatial statistician, and he and I worked extensively on understanding how to integrate spatial data at different scales anywhere from 1 meter to 30 meters or larger, and the challenges 30-meter vegetation indices have on model prediction (SDSU, 2016, pp.10-13).

One of the objectives of this initiative was to examine the initial accuracy of making predictions by simply examining the spatial auto correlation between agricultural fields defined in space. "In a pixel image, the spatial domain is divided into a grid (of picture elements or 'pixels'), and a value is associated with each pixel. The pixel value could represent brightness (in a digital camera image or a remotely sensed image), terrain elevation (in a digital terrain model), soil pH or magnetic field strength (in a spatial survey), and other measurable quantities. Pixel values can be categorical values, representing a classification of space into different rock types, cell types, administrative regions, or land use types. Other types of spatial data can be converted into pixel images, so that the pixel value could represent (say) the distance from that pixel to the nearest geological fault. Many calculations in spatial statistics produce a pixel image as a result for example, a kernel estimate of point process intensity" (Baddeley et al., 2016).

One of the topics that came up in discussion was if we could identify white mold vector pathogens through the use of temporal spatial images throughout the growing season; and if so, could the vector be identified simply by the pixel value. The answer depended on the resolution. Figure 4.1 on page 141 gives a visual representation how ineffective low resolution images are in precision agriculture. Clearly it is not possible to

identify deterministically a small patch of white mold if the resolution is not sufficient; hence a more refined approach is necessary.

This also applies to using a support vector machine in this case study. First, we can examine the dimension of pixels as it relates to the scale as resolution increases, the dimension also increases. Figure 4.2 shows a pixel density scale to understand this. The image shows a standard 30-meter Landsat7 dimension relative to overlaying multiple pixel images of higher resolution to understand the 'curse of dimensionality'.

"The colorful phrase the 'curse of dimensionality' was apparently coined by Richard Bellman, in connection with the difficulty of optimization by exhaustive enumeration on product spaces. Bellman reminded us that, if we consider a Cartesian grid of spacing 1/10 on the unit cube in 10 dimensions, we have 10^{10} points; if the cube in 20 dimensions was considered, we would have of course 10^{20} points. His interpretation: if our goal is to optimize a function over a continuous product domain of a few dozen variables by exhaustively searching a discrete search space defined by a crude discretization, we could easily be faced with the problem of making tens of trillions of evaluations of the function. Bellman argued that this curse precluded, under almost any computational scheme then foreseeable, the use of exhaustive enumeration strategies, and argued in favor of his method of dynamic programming" (Donoho, 2000, p. 18).

As we can see definitively in Figure 4.2, a one pixel image from Landsat7 varies vastly from a 5-meter and 1-meter resolution in terms of the number of pixels as the resolution increases. When considering precision, the overlay of identical images at different resolutions creates a clear paradox in accurate prediction because multiple

images of interest may be completely contained within one pixel rather and multiple pixels.

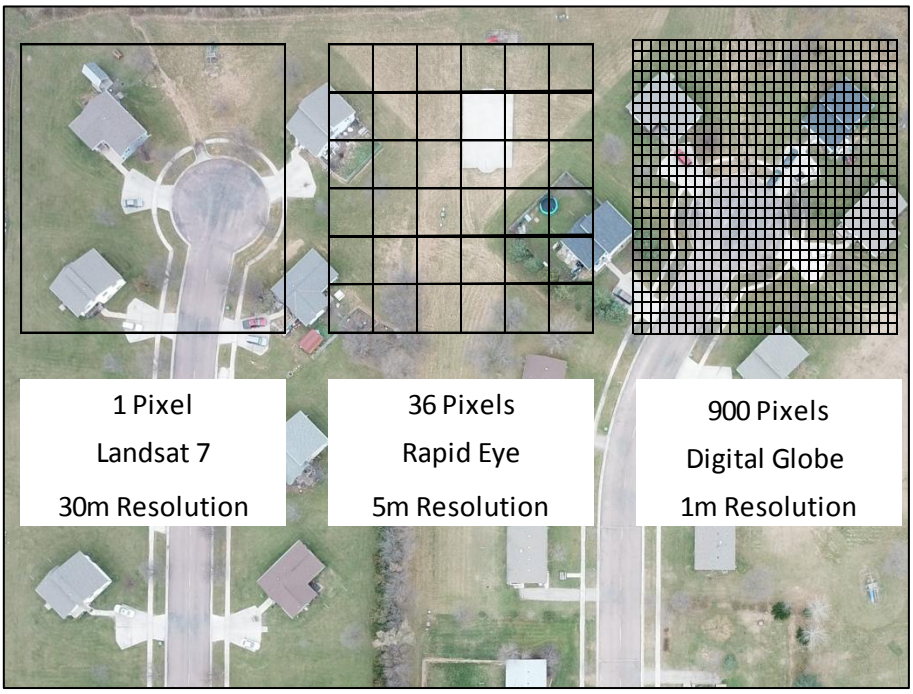


Figure 4.2 - Pixel Comparison of Dimensionality (Not to Scale)

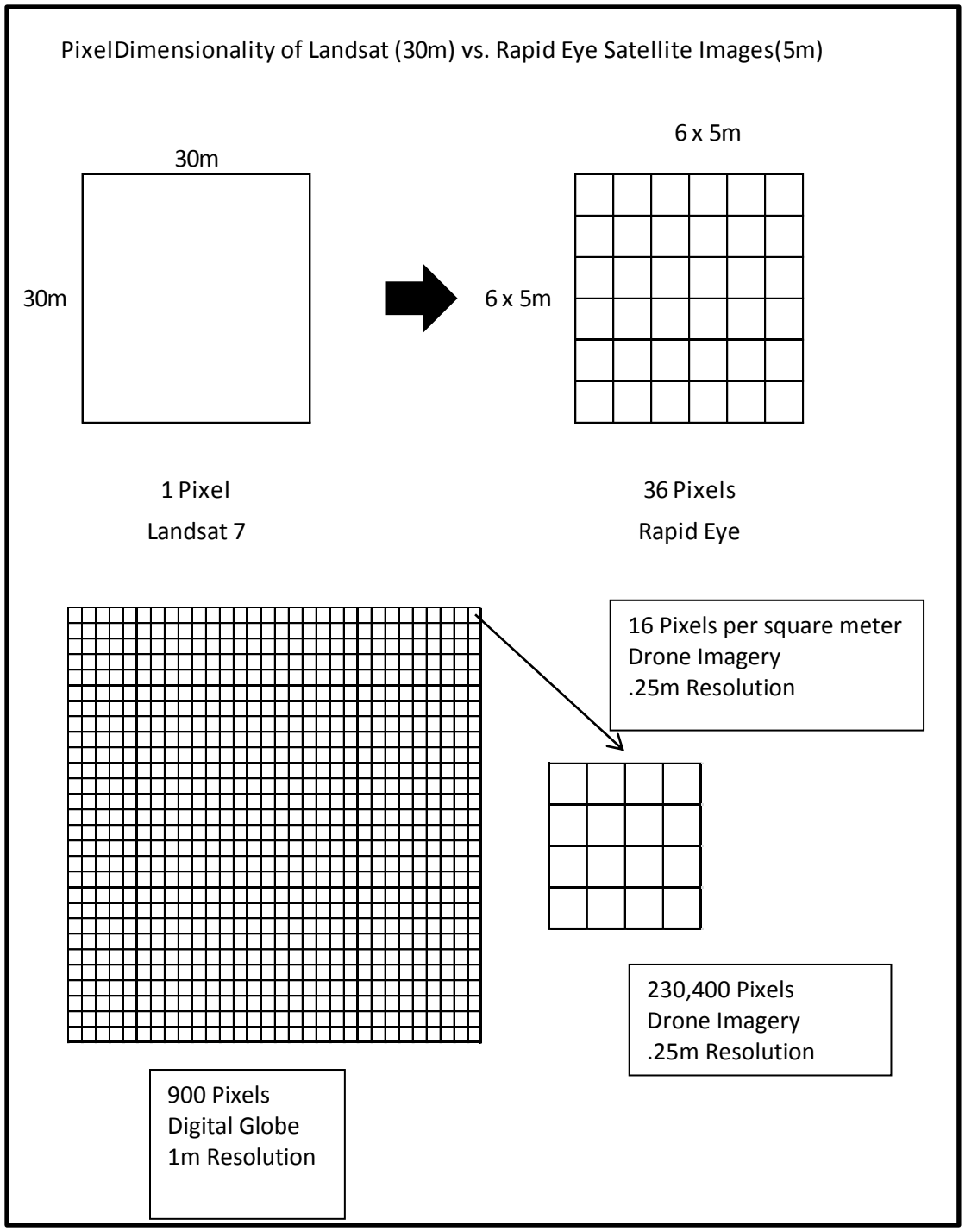


Figure 4.3 - Pixel Dimension Equivalence Comparisons to Landsat 7 (1 Pixel)

Discussion

Drone imagery in Figure 4.2 could not be included in the visualization due to the dimensionality discussed. Figure 4.3 on the previous page provides how quickly high resolution becomes untenable. One pixel in the one meter resolution Digital Globe image above would scale the image to an additional sixteen pixels per square meter which is simply not possible in this context. Figure 4.3 shows that visual scalability is not possible as the pixel size become infinitesimally small as the resolution increases; to show a 30-meter square pixel relative to a .25-meter in the same area would require 230,400 pixels. The higher dimension provides a better training set for machine learning at the price of computational resources in determining housing polygons.

Although Figure 4.2 is not exactly to scale, considering a 30m resolution pixel image from Landsat 7 is 900 square meters which is approximately 9,687.51 square feet, then it is possible when considering the average square footage of a house according to the U.S. census is 2,422 square feet then it is possible that a Landsat image would be capable of capturing at least three houses in one pixel. Furthermore, this is problematic when using a support vector machine learning algorithm because image classification would result in the training classifier identifying three domiciles as one.

Since lower resolution images have been shown to be ineffective, we turn to looking at more tenable images that can provide a more realistic outcome in training a machine learning algorithm to find infrastructure. Tribes have an invaluable resource in using the Enhanced View program (EV) from Digital Globe as part of the National

Geospatial-Intelligence Agency (NGA). These images provide an excellent reference when analyzing the accuracy of machine learning.

As described above, dimensionality can be a curse or it can be a blessing. The results of this case study show that when a careful balance of image processing mixed with lower resolution images can produce very practical solutions.

Figure 4.4 - Limitations of Free or Lower Resolution Imagery versus Higher Resolution Imagery

The image in Figure 4.4 above is a 1-meter resolution image that tribes have access to through the NGA. The image in Figure 4.5 to the right was an image that was commissioned for this dissertation from a private drone contractor in order to do a small scale direct comparison with a .25-meter versus 1-meter comparison.

Figure 4.5 - .25-meter Drone Imagery

The next sections outline an extensive review of establishing a number of key ideas that are fundamental to understanding SVM methods. Although there is quite an extensive body of literature concerning this topic, this manuscript covers the basic building blocks from introducing the linearly separable and non-separable cases, kernel methods and finally the use of multi-class SVMs to achieve the outcomes at the end of the chapter.

SVM Case 1: The Linearly Separable Case

Izenman (2008) writes eloquently as to the theory behind the linear separable and non-separable SVM cases. In order to study more complex classification techniques, establishing these two concepts is crucial in handling non-linear transformations, kernels, and multi-class SVM's.

Assume we have available a learning set of data.

$$\mathbf{L} = \{(\mathbf{x}_i, y_i) : i = 1, 2, ..., n\}, \qquad 4.1$$

on the pair $(\mathbf{X}, Y)$, where $\mathbf{X} \in \Re^r$ and $Y \in \{-1, +1\}$. The binary classification problem is to use $\mathbf{L}$ to construct a function $f : \Re^r \to \Re$ so that

$$C(\mathbf{x}) = \text{sign}(f(\mathbf{x})), \ \mathbf{x} \in \Re^r, \qquad 4.2$$

is a classifier.

The separating function f then classifies each new point $\mathbf{x}$ in a test set T into one of two classes, $\prod_+$ or Π_-, depending upon whether $C(\mathbf{x})$ is +1 (if $f(x) \geq 0$) or -1 if ($f(x) < 0$)), respectively.

The goal is to have f assign all "positive" points in $\boldsymbol{\tau}$ (i.e., those with $y = +1$) to Π_+ and all negative points in $\boldsymbol{\tau}$ ($y = -1$) to Π_-. In practice, we recognize that 100% correct classification may not be possible.

First, consider the simplest situation: suppose the positive $(y_i = +1)$ and negative $(y_i = -1)$ data points from the learning set $\mathbf{L}$ can be separated by a hyperplane,

$$\left\{\mathbf{x}: f(\mathbf{x}) = \beta_0 + \mathbf{x}^\tau \boldsymbol{\beta} = 0\right\}, \qquad 4.3$$

where $\boldsymbol{\beta}$ is the *weight vector* with Euclidean norm $\|\boldsymbol{\beta}\|$, and β_0 is the bias. (Note: $b = -\beta_0$ is the *threshold*.) If this hyperplane can separate the learning set into the two given classes without error, the hyperplane is termed a *separating hyperplane*. Clearly, there is an infinite number of such separating hyperplanes. How do we determine which one is the best?

Consider any separating hyperplane. Let d_- be the shortest distance from the separating hyperplane to the nearest negative data point, and let d_+ be the shortest distance from the same hyperplane to the nearest positive data point. Then, the *margin* of the separating hyperplane is defined as $d = d_- + d_+$. If, in addition, the distance between the hyperplane and its closest observation is maximized, we say that the hyperplane is an optimal separating hyperplane (also known as a maximal margin classifier).

If the learning data from the two classes are linearly separable, there exists β_0 and $\boldsymbol{\beta}$ such that

$$\beta_0 + \mathbf{x}^\tau \boldsymbol{\beta} \geq +1 \text{ if } y_i = +1 \qquad 4.4$$

$$\beta_0 + \mathbf{x}^\tau \boldsymbol{\beta} \leq +1 \text{ if } y_i = -1 \qquad 4.5$$

If there are data vectors in $\mathbf{L}$ such that equality holds in (4.4), then these data vectors lie on the hyperplane $H_+ : (\beta_0 - 1) + \mathbf{x}^\tau \boldsymbol{\beta} = 0$; similarly, if there are data vectors in L such that equality holds in (4.5), then these data vectors lie on the hyperplane $H_- : (\beta_0 - 1) + \mathbf{x}^\tau \boldsymbol{\beta} = 0$. Points in $\mathbf{L}$ that lie on either one of the hyperplanes H_{-1} or H_{+1}, are said to be *support vectors*. See Figure 4.6.

The support vectors typically consist of a small percentage of the total number of sample points.

If $\mathbf{x}_{-1}$ lies on the hyperplane H_{-1}, and if $\mathbf{x}_{+1}$ lies on the hyperplane H_{+1}, then,

$$\beta_0 + \mathbf{x}_{-1}^\tau \boldsymbol{\beta} = -1, \quad \beta_0 + \mathbf{x}^\tau{}_{+1} \boldsymbol{\beta} = +1 \qquad 4.6$$

The difference of these two equations is $\mathbf{x}_{+1}^\tau \boldsymbol{\beta} - \mathbf{x}_{-1}^\tau \boldsymbol{\beta} = 2,$ and their sum is $\beta_0 = -\frac{1}{2}\left\{\mathbf{x}_{+1}^\tau \boldsymbol{\beta} + \mathbf{x}_{-1}^\tau \boldsymbol{\beta}\right\}.$ The perpendicular distances of the hyperplane $\beta_0 + \mathbf{x}^\tau \boldsymbol{\beta} = 0$ from the points $\mathbf{x}_{-1}$ and $\mathbf{x}_{+1}$ are

$$d_- = \frac{\left|\beta_0 + \mathbf{x}_{-1}^\tau \boldsymbol{\beta}\right|}{\|\boldsymbol{\beta}\|} = \frac{1}{\|\boldsymbol{\beta}\|}, \quad d_+ = \frac{\left|\beta_0 + \mathbf{x}_{+1}^\tau \boldsymbol{\beta}\right|}{\|\boldsymbol{\beta}\|} = \frac{1}{\|\boldsymbol{\beta}\|} \qquad 4.7$$

respectively. So, the margin of the separating hyperplane is $d = 2 / \|\boldsymbol{\beta}\|$.

The inequalities (4.4) and (4.5) can be combined into a single set of inequalities,

$$y_i(\beta_0 + \mathbf{x}_i^\tau \boldsymbol{\beta}) \geq +1, \quad i = 1, 2, ..., n. \qquad 4.8$$

The quantity $y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta})$ is called the margin of $(\mathbf{x}_i y_i)$ *with respect to the hyperplane* (4.3), $i = 1, 2, ..., n$. From (4.6), we see that $\mathbf{x}_i$ is a support vector with respect to the hyperplane (4.3) if its margin equals one; that is, if

$$y_i(\beta_0 + \mathbf{x}_i^T\boldsymbol{\beta}) = 1 \qquad 4.9$$

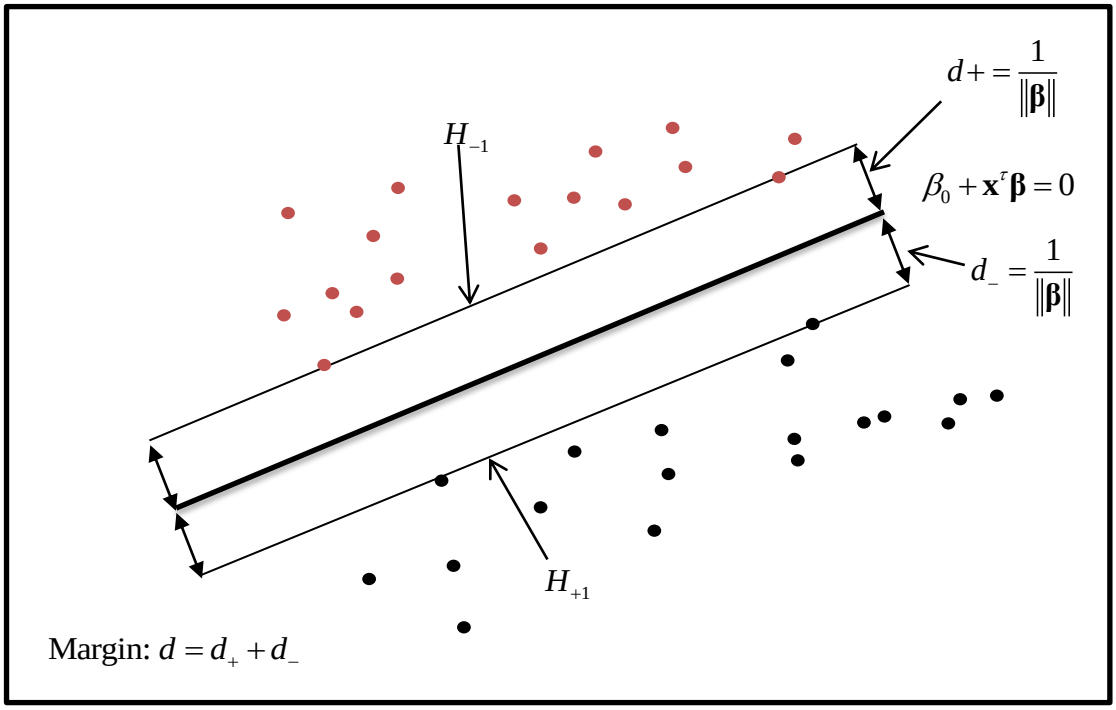


Figure 4.6 - Support Vectors in the Linearly Separable Case

The support vectors in Figure 4.6 are identified (points intersecting, H_+ and H_-). The empirical distribution of the margins of all the observations in $\mathbf{L}$ is called the *margin distribution of a hyperplane with respect to* $\mathbf{L}$. The minimum of the empirical margin distribution is the *margin of the hyperplane with respect to* $\mathbf{L}$. The problem is to find the optimal separating hyperplane; namely, find the

hyperplane that maximizes the margin, $2/\|\boldsymbol{\beta}\|$, subject to the conditions(4.8).

Equivalently, we wish to find β_0 and $\boldsymbol{\beta}$ to

$$\text{minimize } \frac{1}{2}\|\boldsymbol{\beta}\|^2, \qquad 4.10$$

$$\text{subject to } y_i(\beta_0 + \mathbf{x}_i^\tau\boldsymbol{\beta}) \geq 1,\ i = 1, 2, ..., n \qquad 4.11$$

This is a convex optimization problem: minimize a quadratic function subject to linear inequality constraints. Convexity ensures that we have a global minimum without local minima. The resulting optimal separating hyperplane is called the *maximal* (or *hard*) *margin solution*.

We solve this problem using Lagrangian multipliers. Because the constraints are $y_i(\beta_0 + \mathbf{x}_i^\tau\boldsymbol{\beta}) \geq -1\ i = 1, 2, ..., n$, we multiply the constraints by positive Lagrangian multipliers and subtract each such product from the objective function (4.10) to form the *primal functional*,

$$F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{n}\alpha_i\left\{y_i(\beta_0 + \mathbf{x}_i^\tau\boldsymbol{\beta}) \geq -1\right\} \qquad 4.12$$

where

$$\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_1)^\tau \geq \mathbf{0} \qquad 4.13$$

is the *n*-vector of (nonnegative) Lagrangian coefficients.

We need to minimize *F* with respect to the *primal variables* β_0 and $\boldsymbol{\beta}$, and then maximize the resulting minimum-*F* with respect to the *dual variables* $\boldsymbol{\alpha}$.

The Karush-Kuhn-Tucker conditions give necessary and sufficient conditions for a solution to a constrained optimization problem. For our primal problem, β_0, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ have to satisfy:

$$\frac{\partial F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0, \qquad 4.14$$

$$\frac{\partial F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0, \qquad 4.15$$

$$y_i(\beta_0 + \mathbf{x}_i^{\tau}\boldsymbol{\beta}) \geq 0, \qquad 4.16$$

$$\alpha_i \geq 0, \qquad 4.17$$

$$\alpha_i\{y_i(\beta_0 + \mathbf{x}_i^{\tau}\boldsymbol{\beta}) - 1\} \geq 0, \qquad 4.18$$

for $i = 1, 2, ..., n.$ The condition (4.18) is known as the *Karush–Kuhn–Tucker complementarity condition*. Solving equations (4.14) and (4.15) yields

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \qquad 4.19$$

$$\boldsymbol{\beta}^* = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \qquad 4.20$$

Substituting (4.19) and (4.20) into (4.12) yields the minimum value of $F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}),$ namely,

$$F_D(\boldsymbol{\alpha}) = \frac{1}{2}\left\|\boldsymbol{\beta}^*\right\|^2 - \sum_{i=1}^{n} \alpha_i \left\{ y_i \beta_0^* + \mathbf{x}_i^{\tau}\boldsymbol{\beta}^* - 1 \right\}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^{\tau}\mathbf{x}_j) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^{\tau}\mathbf{x}_i) + \sum_{i=`}^{n} \alpha_i$$

$$= \sum_{i=`}^{n} \alpha_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^{\tau}\mathbf{x}_i) \qquad 4.21$$

where we used (4.18) in the second line. Note that the primal variables have been removed from the problem. The expression (4.21) is usually referred to as the *dual functional* of the optimization problem. We next find the Lagrangian multipliers $\boldsymbol{\alpha}$ by maximizing the dual functional (4.21) subject to the constraints (4.17) and (4.19). The constrained maximization problem (the 'Wolfe dual') can be written in matrix notation as follows.

Find $\boldsymbol{\alpha}$ to

$$\text{maximize } F_D = \mathbf{1}_n^\tau \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\tau \mathbf{H}\boldsymbol{\alpha} \qquad 4.22$$

$$\text{subject to } \boldsymbol{\alpha} \geq 0,\ \boldsymbol{\alpha}^\tau \mathbf{y} = \mathbf{0} \qquad 4.23$$

where $\mathbf{y} = (y_i, ..., y_n)^\tau$ and $\mathbf{H} = (H_{ij})$ is a square $(n \times n)-$matrix with $H_{ij} = y_i y_j (\mathbf{x}_i^\tau \mathbf{x}_j)$. If $\hat{\boldsymbol{\alpha}}$ solves this optimization problem, then

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \mathbf{x}_i \qquad 4.24$$

yields the optimal weight vector. If $\hat{\alpha}_i > 0$, then, from (4.18), $y_i(\beta_0^* + \mathbf{x}^\tau \boldsymbol{\beta}^*) = 1$, and so $\mathbf{x}_i$ is a support vector; for all observations that are not support vectors, $\hat{\alpha}_i = 0$. Let $sv \subset \{1, 2, ..., n\}$ be the subset of indices that identify the support vectors (and also the nonzero Lagrangian multipliers). Then, the optimal $\boldsymbol{\beta}$ is given by (4.24), where the sum is taken only over the support vectors; that is,

$$\hat{\boldsymbol{\beta}} = \sum_{i \in sv} \hat{\alpha}_i y_i \mathbf{x}_i. \qquad 4.25$$

In other words, $\hat{\boldsymbol{\beta}}$ is a linear function only of the support vectors $\{\mathbf{x}_i, i \in sv\}$. In most applications, the number of support vectors will be small relative to the size of $\mathbf{L}$, yielding a *sparse* solution. In this case, the support vectors carry all the information necessary to determine the optimal hyperplane.

The primal and dual optimization problems yield the same solution, although the dual problem is simpler to compute and, as we shall see, is simpler to generalize to nonlinear classifiers. Finding the solution involves standard convex quadratic programming methods, and so any local minimum also turns out to be a global minimum. Although the optimal bias $\hat{\beta}_0$ is not determined explicitly by the optimization solution, we can estimate it by solving (4.18) for each support vector and then averaging the results. In other words, the estimated bias of the optimal hyperplane is given by

$$\hat{\beta}_0 = \frac{1}{|sv|} \sum \left(\frac{1 - y_i \mathbf{x}^\tau \hat{\beta}}{y_i} \right), \qquad 4.26$$

where $|sv|$ is the number of support vectors in $\mathbf{L}$.

It follows that the optimal hyperplane can be written as

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \hat{\beta}_0 + \mathbf{x}^\tau \hat{\boldsymbol{\beta}} \\ &= \hat{\beta}_0 + \sum_{i \in sv} \hat{\alpha}_i y_i (\mathbf{x}^\tau \mathbf{x}_i). \end{aligned} \qquad 4.27$$

Clearly, only support vectors are relevant in computing the optimal separating hyperplane; observations that are not support vectors play no role in determining

the hyperplane and are, thus, irrelevant to solving the optimization problem. The classification rule is given by

$$C(\mathbf{x}) = sign\left\{ \hat{f}(\mathbf{x}) \right\}. \qquad 4.28$$

If $j \in sv,$ then, from (4.27)

$$y_i \hat{f}(\mathbf{x}_j) = y_i \hat{\beta}_0 + \sum_{i \in sv} \hat{\alpha}_i y_i y_j (\mathbf{x}_j^\tau x_i) = 1 \qquad 4.29$$

Hence, the squared-norm of the weight vector $\hat{\boldsymbol{\beta}}$ of the optimal hyperplane is

$$\begin{aligned} \left\| \hat{\boldsymbol{\beta}} \right\|^2 &= \sum_{i \in sv}^{n} \sum_{j \in sv}^{n} \hat{\alpha}_i \hat{\alpha}_j y_i y_j (\mathbf{x}_i^\tau \mathbf{x}_j) \\ &= \sum_{j \in sv}^{n} \hat{\alpha}_j y_j \sum_{i \in sv}^{n} \hat{\alpha}_i y_i (\mathbf{x}_i^\tau \mathbf{x}_j) \\ &= \sum_{j \in sv}^{n} \hat{\alpha}_i (1 - y_j \hat{\beta}_0) \\ &= \sum_{j \in sv}^{n} \hat{\alpha}_j . \end{aligned} \qquad 4.30$$

The third line used (4.29) and the fourth line used (4.19).

It follows from (4.30) that the optimal hyperplane has maximum margin $2 / \left\| \hat{\boldsymbol{\beta}} \right\|$, where

$$\frac{1}{\left\| \hat{\boldsymbol{\beta}} \right\|} = \left(\sum_{j \in sv} \hat{\alpha}_j \right)^{-1/2} \qquad 4.31$$

Case Summary

In most real world applications, it will be unlikely that data will be conveniently separated in this context, but this does form the fundamental basis for understanding how to deal with more complex tasks a SVM is capable of handling such as linearly non-separating cases, non-linear SVM's, and the development of multi-class support vector machines over K-classes.

In addition, there are additional techniques such as support vector regression which defines a function that is used to track the points in a space, rather than separating them. This case study is specifically designed with classification of individual housing infrastructure using raster and jpg maps, thus regression will not be covered.

The next section establishes techniques that deal with cases when two classes are separable, but not linearly or there is no clear separability exists either linearly or non-linearly. When classes overlap, the result is one or more constraints will be violated due to high noise (i.e. large variance).

SVM Case 2: The Linearly Non-Separable Case

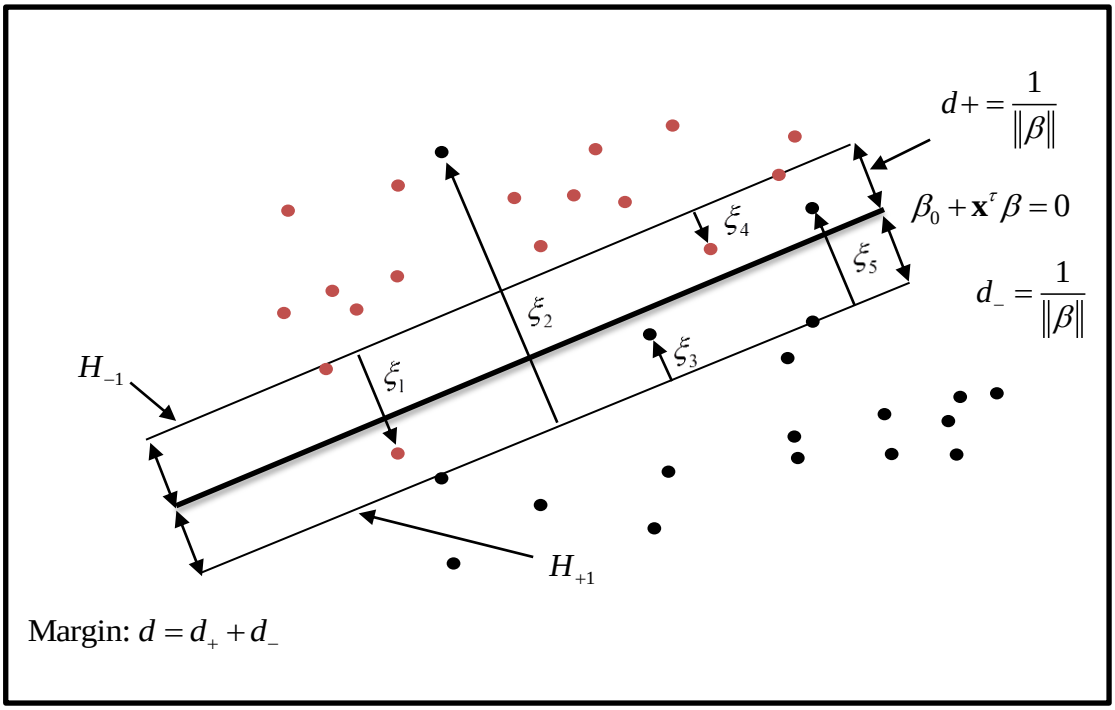


Figure 4.7 - Support Vectors in the Linearly Non-Separable Case

Steinwart and Christmann (2006) assert that in presence of noise, it commonplace for training points to be misclassified in order to avoid over fitting. The non-separable case above sets the stage for more complex problems. In satellite imagery it is often the case that many pixels will overlap that are also part of the target classification group. To handle overlapping data, in a more flexible way, the formulation of the margin in the linearly separable case must be defined as a *soft-margin solution* as in Figure 4.7.

The red points are defined as $y_i = -1$ and the black points correspond to data points with $y_i = +1$. Like before the thick black line represents the *separating* hyperplane,

$\beta_0 + \mathbf{x}^\tau \beta = 0.$ The support vectors are the points lying on the hyperplanes, H_+ and H_-. The points that violate the margin between the support vectors and the *separating* hyperplanes are referred to a nonnegative *slack variables* ξ_i for each data point, $(\mathbf{x}_i, y_i)$, in the learning set, $i = 1, 2, ..., n$.

Let

$$\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\tau \geq \mathbf{0} \qquad 4.32$$

The constraints (4.11) now become $y_i(\beta_0 + \mathbf{x}_i^\tau \boldsymbol{\beta}) + \xi_i \geq 1, \quad i = 1, 2, ..., n$. Data points that obey these constraints have $\xi_i = 0$. The classifier now has to find the optimal hyperplane that controls both the margin, $\frac{2}{\|\boldsymbol{\beta}\|}$ and some computationally simple function of the slack variables, such as

$$g_\sigma(\boldsymbol{\xi}) = \sum_{i=1}^{n} \xi_i^\sigma \qquad 4.33$$

subject to certain constraints. The usual values of σ are 1 ("1-norm") or 2 ("2-norm"). Here, we discuss the case of $\sigma = 1$ only.

The *1-norm soft-margin optimization problem* is to find to $\beta_0, \boldsymbol{\beta},$ and $\boldsymbol{\xi}$ to

$$\text{minimize } \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n} \xi_i^\sigma, \qquad 4.34$$

$$\text{subject to } \xi_i \geq 0,\ y_i(\beta_0 + \mathbf{x}_i^\tau \boldsymbol{\beta}) \geq 1 - \xi_i,\ \text{i=1,2,...,n,} \qquad 4.35$$

where $C > 0$ is a *regularization parameter*. C takes the form of a tuning

constant that controls the size of the slack variables and balances the two terms in the minimizing function.

Form the primal functional, $F_P = F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\eta})$, where

$$F_P = \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(\beta_0 + \mathbf{x}_i^\tau\boldsymbol{\beta}) - (1-\xi_i)\} - \sum_{i=1}^{n}\eta_i\xi_i, \qquad 4.36$$

with $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_n)^\tau \geq \mathbf{0}$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\tau \geq \mathbf{0}$. Fix $\boldsymbol{\alpha}$ and $\boldsymbol{\eta}$, and differentiate F_P with respect to $\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}$:

$$\frac{\partial F_P}{\partial \beta_0} = -\sum_{i=1}^{n}\alpha_i y_i, \qquad 4.37$$

$$\frac{\partial F_P}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \qquad 4.38$$

$$\frac{\partial F_P}{\partial \xi_i} = C - \alpha_i - \eta_i, \text{ i=1,2,...,n.} \qquad 4.39$$

Setting these derivatives equal to zero and solving yields

$$\sum_{i=1}^{n}\alpha_i y_i = 0, \ \boldsymbol{\beta}^* - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i, \ \alpha_i = C - \eta_i \qquad 4.40$$

Substituting (4.37) into (4.33) gives the dual functional,

$$F_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i^\tau \mathbf{x}_j) \qquad 4.41$$

which, remarkably, is the same as (4.18) for the linearly separable case. From the constraints $C-\alpha_i-\eta_i=0$ and $\eta_i \geq 0,$ we have that $0 \leq \alpha_i \leq C$ .In addition, we have the Karush–Kuhn–Tucker conditions:

$$y_i(\beta_0+\mathbf{x}_i^\tau\boldsymbol{\beta})-(1-\xi_i) \geq 0, \quad 4.42$$

$$\xi_i \geq 0, \quad 4.43$$

$$\alpha_i \geq 0, \quad 4.44$$

$$\eta_i \geq 0, \quad 4.45$$

$$\alpha_i\{y_i(\beta_0+\mathbf{x}_i^\tau\boldsymbol{\beta})-(1-\xi_i)\}=0, \quad 4.46$$

$$\xi_i(\alpha_i-C)=0, \quad 4.47$$

for $i=1,2,...,n.$ From (4.47), a slack variable, ξ_i , can be nonzero only if $\alpha_i=C.$ The Karush–Kuhn–Tucker complementarity conditions, (4.46) and (4.47), can be used to find the optimal bias β_0.

We can write the dual maximization problem in matrix notation as follows.

Find $\boldsymbol{\alpha}$ to

$$\text{maximize } F_D(\boldsymbol{\alpha})=\mathbf{1}_n^\tau\boldsymbol{\alpha}-\frac{1}{2}\boldsymbol{\alpha}^\tau\mathbf{H}\boldsymbol{\alpha} \quad 4.48$$

$$\text{subject to } \boldsymbol{\alpha}^\tau\mathbf{y} \geq 0,\ 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}_n \quad 4.49$$

The only difference between this optimization problem and that for the linearly separable case, (4.22) and (4.23), is that, here, the Lagrangian coefficients α_i, i=1,2,...,n, are bounded above by C; this upper bound restricts the influence of each observation in determining the solution. This type of constraint is referred to as a *box constraint* because $\boldsymbol{\alpha}$ is constrained by the box of side C in the positive orthant. From (4.49), we see that the *feasible region* for the solution to this convex optimization problem is the intersection of

the hyperplane $\boldsymbol{\alpha}^{\tau}\mathbf{y} = 0$ .with box constraint $\mathbf{0} \leq \boldsymbol{\alpha} < C\mathbf{1}_n$. if $C = \infty,$ then the problem reduces to the hard-margin case separable case.

If $\hat{\boldsymbol{\alpha}}$ solves this optimization problem, then,

$$\hat{\beta} = \sum_{i=\epsilon sv} \hat{\alpha}_i \mathbf{y}_i \mathbf{x}_i \qquad 4.50$$

yields the optimal weight vector, where the set *sv* of support vectors contains those observations in *L* which satisfy the constraint (4.42).

Case Summary

Constructing the natural order of linear, separable and non-separable cases is the building blocks for more complex ideas in classification theory. There would be no way to fully understand the next sections if there was not a fundamental theory of hyperplanes, margin, soft-margin solutions, slack variables, as well as defining the cost function in these terms. This review will further build as to how to handle situations where classification cannot be separated linearly.

SVM Case 3: Defining Non-Linear Support Vector Machines

Nonlinear Support Vector Machines

So far, we have discussed methods for constructing a linear SVM classifier. But what if a linear classifier is not appropriate for the data set in question? Can we extend the idea of linear SVM to the nonlinear case? The key to constructing a nonlinear SVM is to observe that the observations in $\mathbf{L}$ only enter the dual optimization problem through the inner products $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^\tau \mathbf{x}_j,\ i, j = 1, 2, ..., n.$

Nonlinear Transformations

Suppose we transform each observation, $\mathbf{x}_i \in \Re^r$ in $\mathbf{L}$ using some nonlinear mapping $\mathbf{\Phi} : \Re^r \to \mathbf{H}$, $\Phi : r \to H$, where $\mathbf{H}$ is an $N_{\mathbf{H}}$-dimensional feature space.

The nonlinear map $\mathbf{\Phi}$ is generally called the *feature map* and the space H is called the *feature space*. The space $\mathbf{H}$ may be very high-dimensional, possibly even infinite dimensional. We will generally assume that $\mathbf{H}$ is a Hilbert space of real-valued functions on $\Re$ *with* inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$.

Let

$$\mathbf{\Phi}(\mathbf{x}_i) = (\varphi_1(\mathbf{x}_i), ..., \varphi_{N_{\mathbf{H}}}(\mathbf{x}_i))^\tau \in \mathbf{H},\ i = 1, 2, ..., n. \qquad 4.51$$

The transformed sample is then $\{\mathbf{\Phi}(\mathbf{x}_i), y_i\}$, where $y_i \in \{-1, +1\}$ identifies the two classes. If we substitute $\mathbf{\Phi}(\mathbf{x}_i)$ for $\mathbf{x}_i$ in the development of the linear SVM, then data would only enter the optimization problem by way of the inner product $\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle = \mathbf{\Phi}(\mathbf{x}_i)^\tau \mathbf{\Phi}(\mathbf{x}_j)$. The difficulty in using nonlinear transformations in this way is computing such inner products in high-dimensional space $\mathbf{H}$.

The next section outlines a way of overcoming this computational constraint called the *kernel trick.*

Kernels and the Kernel Trick

A *kernel* is a way of computing the dot product of two vectors x and y in some (possibly very high dimensional) feature space, which is why kernel functions are sometimes called "generalized dot product".

The idea behind nonlinear SVM is to find an optimal separating hyperplane (with or without slack variables, as appropriate) in high-dimensional feature space $\mathbf{H}$ just as we did for the linear SVM in input space. Of course, we would expect the dimensionality of $\mathbf{H}$ to be a huge impediment to constructing an optimal separating hyperplane (and classification rule) because of the curse of dimensionality.

The fact that this does not become a problem in practice is due to the "kernel trick" which was first applied to SVMs by Cortes and Vapnik (1995). The so-called kernel trick is a wonderful idea that is widely used in algorithms for computing inner products of the form $\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle$ in feature space $\mathbf{H}$.

The trick is that instead of computing these inner products in $\mathbf{H}$, we compute them using a nonlinear kernel function, $K(\mathbf{x}_i, \mathbf{x}_i) = \langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \rangle$ in input space, which helps speed up the computations. Then, we just compute a *linear* SVM, but where the computations are carried out in some other space. Remember, that it is assumed to be a linear separable set of training data. Nevertheless, this is only the case in very few real-world applications. Now the kernel function comes to handy as a remedy, as an implicit mapping of the input space into a linear separable feature space, where our linear classifiers are again applicable.

Izenman (2008) defines:

> A kernel K is a function K: $K : \Re^r \times \Re^r \to \Re$ such that, for all $\mathbf{x}, \mathbf{y} \in \Re^r$,
>
> $$K(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle. \qquad 4.52$$
>
> The kernel function is designed to compute inner-products in $\mathbf{H}$ by using only the original input data. Thus, wherever we see the inner product $\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle$, we substitute the kernel function $K(\mathbf{x}, \mathbf{y})$. The choice of K implicitly determines both Φ and $\mathbf{H}$. The big advantage to using kernels as inner products is that if we are given a kernel function *K*, then we do not need to know the explicit form of Φ. We require that the kernel function be symmetric, $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$, and satisfy an inequality, $[K(\mathbf{x}, \mathbf{y})]^2 \le K(\mathbf{x}, \mathbf{x}) K(\mathbf{y}, \mathbf{y})$, derived from the Cauchy–Schwarz inequality. If $K(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \Re^r$, this implies that $\|\Phi(\mathbf{x})\|_{\mathbf{H}} = 1$.

A kernel *K* is said to have the *reproducing property* if, for any $f \in \mathbf{H}$

$$\langle f(\cdot), K(\mathbf{x}, \cdot) \rangle = f(\mathbf{x}) \qquad 4.53$$

If *K* has this property, we say it is a *reproducing kernel*. *K* is also called the *representer of evaluation*.

In particular, if $f(\cdot) = K(\cdot, \mathbf{x})$, then,

$$\langle K(\mathbf{x}, \cdot) K(\mathbf{y}, \cdot) \rangle = K(\mathbf{x}, \mathbf{y}). \qquad 4.54$$

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be any set of n points in $\Re^r$. Then, the ($n \times n$)-matrix

$\mathbf{K} = (K_{ij})$, where $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, 2, ..., n$, is called the *Gram* (or *kernel*) matrix of *K* with respect to $\mathbf{x}_1, ..., \mathbf{x}_n$. If the Gram matrix K satisfies $\mathbf{u}^\tau \mathbf{K} \mathbf{u} \geq 0$, for any *n*-vector u, then it is said to be *nonnegativedefinite* with nonnegative eigenvalues, in which case we say that *K* is a nonnegative-definite kernel (or *Mercer kernel*).

If *K* is a specific Mercer kernel on $\Re^r \times \Re^r$, we can always construct a unique Hilbert space $\mathbf{H}_K$, say, of real-valued functions for which *K* is its reproducing kernel. We call $\mathbf{H}_K$ a (real) *reproducing kernel Hilbert space* (*rkhs*). We write the inner-product and norm of $\mathbf{H}_K$ (or just $\langle \cdot, \cdot \rangle$ when *K* is understood) and $\| \cdot \|_{\mathbf{H}_K}$ respectively.

The Radial Basis Function

This case study's objective was to test the accuracy of how well an image classifier would perform with various resolutions of data. The geoprocessing tool in the spatial analyst toolbox in ESRI ArcMap has a support vector machine image classifier built into the platform. However, like most 'black box' functions, there was no accompanying literature defining the parameters of the machine learning algorithm. I submitted an inquiry to ESRI to research the algorithm and this is the statement I got back from the researcher:

"To clarify linear versus non-linear and the kernel used; the kernels used are non-linear radial basis function (RBF) kernels, and we do a 2-D grid search for the best parameter pair $[C, \gamma]$, as outlined in the paper Hsu, Chang, and Lin (2010)" (ESRI, personal communication).

Hsu, Chang, and Lin (2010) outlines the process of a SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

Given a training set of instance-label pairs $(\mathbf{x}_i, y_i),\ i = 1,...,l$ where $\mathbf{x}_i \in \Re^n$ and $\mathbf{y} \in \{1,-1\}^l$, the support vector machines (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) require the solution of the following optimization problem:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^\tau \mathbf{w} + C\sum_{i=1}^{l}\xi_i$$
$$\text{subject to } y_i(\mathbf{w}^\tau \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0.$$

Proof

To verify this result from the paper, recall Liebler (2003) defines the transpose of a column vector is a row vector such that $u \cdot v = u^{\tau} v = v^{\tau} u$ which $\mathbf{w}^{\tau}\mathbf{w}$ corresponds to equation 4.34 which asserts $\|\boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^{\tau}\boldsymbol{\beta} = \boldsymbol{\beta}\boldsymbol{\beta}^{\tau}$ and the *1-norm soft-margin optimization problem* in equations 4.34 and 4.35 is to find to $\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}$ $\min \frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{n} \xi_i^{\sigma}$, where $\sigma = 1$ and by mapping a non-linear function $\phi(\mathbf{x}_i)$ subject to the kernel trick:

$$\xi_i \geq 0,\ y_i(\beta_0 + \varphi(\mathbf{x}_i)^{\tau}\boldsymbol{\beta}) \geq 1 - \xi_i,\ \text{i=1,2,...,n,}$$

thus, we have proven the research given by ESRI is in the same solution provided in this literature review maps the function in accordance with parameters outlined in this manuscript.

Selection of the RBF Kernel

In general, the RBF kernel is a reasonable first choice. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. The RBF kernel consists of two tuning parameters (C, γ).

Since the number of hyper parameters influences the complexity any given kernel, the RBF kernel provides a sensible choice because it has fewer numerical difficulties over other kernels. One key point is in contrast to other kernel values that may go to infinity; the Radial Basis Function is subject to the constraint, $0 < K_{ij} \leq 1$.

There are some situations where the RBF kernel is not suitable; in particular, when the number of features is very large, one may just use the linear kernel. Thus, the radial basis kernel function is defined as $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\| x_i - x_j \right\|^2 \right)$, where $\gamma > 0.$

Here training vectors $\mathbf{x}_i$ are mapped into a higher and perhaps infinite dimensional space by this function as described on page 168 via, data would only enter the optimization problem by way of the inner product,

$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \right\rangle = \mathbf{\Phi}(\mathbf{x}_i)^r \mathbf{\Phi}(\mathbf{x}_j)$ by the function Φ. In addition, recall where $C > 0$ is a *regularization parameter*. C takes the form of a tuning constant that controls the size of the slack variables and balances the two terms in the minimizing function. Given this process; the next procedure is how the algorithm, performs a search for the parameters (C, γ). (Hsu et al., 2010)

Optimizing a Grid Search for Parameters in a Radial Basis Function

Recall on page 165, C is a type of constraint is referred to as a *box constraint* because $\boldsymbol{\alpha}$ is constrained by the box of side C in the positive orthant. Since C and γ are not known beforehand, a model selection is necessary so that the classifier can accurately predict the unknown or testing data.

Hsu et al., (2010) explain:

> Note that it may not be useful to achieve high training accuracy (i.e. a classifier which accurately predicts training data whose class labels are indeed known). As discussed above, a common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the

“unknown” set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation.

We recommend a “Grid-search" on C and using cross-validation. Various pairs of (C,γ) values are tried and the one with the best cross-validation accuracy is picked. The grid-search is straightforward but seems naive. In fact, there are several advanced methods which can save computational cost by, for example, approximating the cross-validation rate. There are two motivations why we prefer the simple grid-search approach.

The amount of computational work involved in the grid search for the SVM solution is much greater and, hence, a lot more expensive. One is that, psychologically, we may not feel safe to use methods which avoid doing an exhaustive parameter search by approximations or heuristics. The other reason is that the computational time required to find good parameters by grid search is not much more than that by advanced methods since there are only two parameters, (C,γ). Furthermore, the grid-search can be easily parallelized because each (C,γ) is independent. Many of advanced methods are iterative processes, e.g. walking along a path, which can be hard to parallelize. Since doing a complete grid-search may still be time-consuming, we recommend using a coarse grid first. After identifying a “better" region on the grid, a finer grid search on that region can be conducted. The grid search first searches for optimal parameters C and γ

Figure 4.8 - Training Set for the Support Vector Machine Using 3 Classes: Houses, Vegetation, and Pavement for the Grid Search

on an independent plane before generating the final classifier used in Figure 4.8 (pp. 5-7).

Figure 4.8 was created as a training set for simplicity. If you look closely, the only three features chosen were housing, roads, and green space. The algorithm performs the grid search for the optimal parameters and the final classifier was used to validate the actual image. The final part of this literature review, covers how to construct a multi-class support vector machine which produces the final images for examination.

SVM Case 5: Multiclass Support Vector Machines

Izenman (2008) continues:

To construct a true multiclass SVM classifier, we need to consider all K classes, $\Pi_1, \Pi_2, ..., \Pi_K$, simultaneously, and the classifier has to reduce the binary SVM classifier if $K = 2.$ Here we describe the construction due to Lee, Lin, and Wahba (2004).

Let $v_1, ...v_K$ be a sequence of K-vectors, where v_K has a 1 in the kth position and whose elements sum to zero, $k = 1, 2, ..., K;$ that is, let

$$v_1 = \left(1, -\frac{1}{K-1}, ..., -\frac{1}{K-1}\right)^\tau$$
$$v_1 = \left(-\frac{1}{K-1}, 1, ..., -\frac{1}{K-1}\right)^\tau$$
$$\vdots$$
$$v_K = \left(-\frac{1}{K-1}, -\frac{1}{K-1}, ..., 1\right)^\tau$$

Note that if $K = 2$, then the vector $v_1 = (1, -1)^\tau$ and $v_2 = (-1, 1)^\tau.$ Every $\mathbf{x}_i$ can be labeled as one of these K vectors; that is $\mathbf{x}_i$ has a label $\mathbf{y}_i = \mathbf{v}_K$ if $\mathbf{x}_i \in \Pi_k$, $i = 1, 2, ..., n, k = 1, 2, ..., K.$

Next, we generalize the separating function $f(\mathbf{x})$ to a K-vector of separating functions,

$$\mathbf{f}(\mathbf{x}) = \left(f_1(\mathbf{x}), ..., f_K(\mathbf{x})\right)^\tau \qquad 4.55$$

where

$$f_k(\mathbf{x}) = \beta_{0k} + h_k(\mathbf{x}),\ \mathrm{h}_k \in \mathrm{H}_K,\ \mathrm{k=1,2,...,K.} \qquad 4.56$$

In (4.56), H_K is a reproducing-kernel Hilbert space (rkhs) spanned by the $\{K(\mathbf{x}_i, \cdot), i = 1, 2, ..., n\}$. For example, in the linear case, $h_k(\mathbf{x}) = \mathbf{x}^\tau \boldsymbol{\beta}_k$, for some vector of coefficients $\boldsymbol{\beta}_k$.

We also assume, for uniqueness, that

$$\sum_{k=1}^{K} f_k(\mathbf{x}) = 0 \qquad 4.57$$

Let $\mathbf{L}(\mathbf{y}_i)$ be a K-vector with 0 in the kth position if $\mathbf{x}_i \in \Pi_2$, and 1 in all other positions; this vector represents the cost of misclassifying $\mathbf{x}_i$ (and allows for an unequal misclassification cost structure if appropriate). If $K = 2$ and $\mathbf{x}_i \in \Pi_1$, then $\mathbf{L}(\mathbf{y}_i) = (0,1)^\tau$, while if $\mathbf{x}_i \in \Pi_2$, then $\mathbf{L}(\mathbf{y}_i) = (1,0)^\tau$.

The multiclass generalization of the optimization problem, is to find functions $\mathbf{f}(\mathbf{x}) = \left(f_1(\mathbf{x}), ..., f_K(\mathbf{x}) \right)^\tau$ satisfying 4.57 which

$$\text{minimize } I_\lambda(\mathbf{f}, \Upsilon) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{L}(\mathbf{y}_i) \right]^\tau \left(\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i \right)_+ + \frac{\lambda}{2} \sum_{i=1}^{n} \left\| h_k \right\|^2, \qquad 4.58$$

where $\left(\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i \right)_+ = \left(\left(\mathbf{f}(\mathbf{x}_1) - \mathbf{y}_{i1} \right)_+, ..., \left(\mathbf{f}_K(\mathbf{x}_i) - \mathbf{y}_{iK} \right)_+ \right)^\tau$ and $\Upsilon = (\mathbf{y}_1, ..., \mathbf{y}_n)$ is a $K \times n -$ matrix.

By setting $K = 2$, if $x_i \in \Pi_1$, then $\mathbf{y}_i = \mathbf{v}_1 = (1,-1)^\tau$, and

$$\begin{aligned}\left[\mathbf{L}(y_i)\right]^\tau = (\mathbf{f}(x_i) - \mathbf{y}_i)_+ &= (0,1)((f_1(x_i)-1)_+, (f_2(x_i)+1)_+)^\tau \\ &= (f_2(x_i)+1)_+ \\ &= (1 - f_1(x_i))_+ \end{aligned} \qquad 4.59$$

while if $x_i \in \Pi_2$, then $\mathbf{y}_i = \mathbf{v}_2 = (1,-1)$, and

$$\left[\mathbf{L}(y_i)\right]^\tau = (\mathbf{f}(x_i) - \mathbf{y}_i)_+ = (f_1(\mathbf{x}_i)+1)_+. \qquad 4.60$$

If we set $K = 2$ in the second term of (4.58), we have that

$$\sum_{k=1}^{2} \|h_k\|^2 = \|h_1\|^2 + \|-h_1\|^2 = 2\|h_1\|^2, \qquad 4.61$$

The function $h_k \in \mathrm{H}_K$ can be decomposed into two parts:

$$h_k(\cdot) = \sum_{\ell=1}^{n} \beta_{\ell k} K(\mathbf{x}_\ell, \cdot) + h_k^{\perp}(\cdot) \qquad 4.62$$

where the $\{\beta_{\ell k}\}$ are constants and $h_k^{\perp}(\cdot)$ is an element in the RKHS orthogonal to H_K. Substituting (4.57) into (4.58), then using (4.62), and rearranging terms, we have that

$$f_K(\cdot) = -\sum_{k=1}^{K-1} \beta_{0k} - \sum_{k=1}^{K-1}\sum_{i=1}^{n} \beta_{ik} K(x_i, \cdot) - \sum_{k=1}^{K=1} h_k^{\perp}(\cdot) \qquad 4.63$$

Because $K(\cdot,\cdot)$ is a reproducing kernel,

$$\left\langle h_k, K(x_i, \cdot)\right\rangle = h_k(x_i),\ i = 1,2,...,n \qquad 4.64$$

and so,

$$\begin{aligned} f_k(x_i) &= \beta_{0k} + h_k(x_i) \\ &= \beta_{0k} + \langle h_k, K(x_i, \cdot) \rangle \\ &= \beta_{0k} + \Big\langle \sum_{\ell=1}^{n} \beta_{\ell k} K(x_\ell, \cdot) + h_k^{\perp}(\cdot) \Big\rangle \\ &= \beta_{0k} + \sum_{\ell=1}^{n} \beta_{\ell k} K(x_\ell, x_i) \end{aligned} \qquad 4.65$$

Note that, for $k = 1, 2, ..., K-1$,

$$\begin{aligned} h_k(\cdot)^2 &= \left\| \sum_{\ell=1}^{n} \sum_{i=1}^{n} \beta_{\ell k} K(\mathbf{x}_\ell, \cdot) + h_k^{\perp}(\cdot) \right\|^2 \\ &= \sum_{\ell=1}^{n} \sum_{i=1}^{n} \beta_{\ell k} \beta_{ik} K(\mathbf{x}_\ell, \mathbf{x}_i) + \left\| h_k^{\perp}(\cdot) \right\|^2, \end{aligned} \qquad 4.66$$

and for $k = K$,

$$\left\| h_K^{\perp}(\cdot) \right\|^2 = \left\| \sum_{k=1}^{K-1} \sum_{i=1}^{n} \beta_{ik} K(\mathbf{x}_\ell, \cdot) \right\|^2 + \left\| \sum_{k=1}^{K-1} h_k^{\perp}(\cdot) \right\|^2. \qquad 4.67$$

Thus, to minimize (4.67), we set $h_k^{\perp}(\cdot) = 0$ for all k. From (4.65), the zero-sum constraint (4.57) becomes

$$\bar{\beta}_0 + \sum_{\ell=1}^{n} \beta_\ell K(\mathbf{x}_\ell, \cdot) = 0 \qquad 4.68$$

where $\bar{\beta}_0 = K^{-1} \sum_{k=1}^{K} \beta_{0k}$ and $\bar{\beta}_i = K^{-1} \sum_{k=1}^{K} \beta_{ik}$. At the n data points,

$\{\mathbf{x}_i, i = 1, 2, ..., n\}$, (4.68) in matrix notation is given by

$$\left(\sum_{k=1}^{K} \beta_{0k} \right) \mathbf{1}_n + \mathbf{K} \left(\sum_{k=1}^{K} \boldsymbol{\beta}_{\cdot k} \right) = 0 \qquad 4.69$$

where $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))$ is an (nxn) Gram matrix and $\boldsymbol{\beta}_{\cdot k} = (\beta_{1k}, ..., \beta_{nk})^\tau$. Let $\beta_{0k}^* = \beta_{0k} - \bar{\beta}_0$ and $\beta_{ik}^* = \beta_{ik} - \bar{\beta}_i$. Using (4.68), we see that the centered version of (4.65) is $f_k^*(\mathbf{x}_i) = \beta_{0k}^* + \sum_{\ell=1}^n \beta_{\ell k}^* K(\mathbf{x}_\ell, \mathbf{x}_i) = f_k(\mathbf{x}_i)$.

Then,

$$\sum_{k=1}^K \left\| h_k^*(\cdot) \right\|^2 = \sum_{k=1}^K \boldsymbol{\beta}_{\cdot k}^\tau \mathbf{K} \boldsymbol{\beta}_{\cdot k} - K \bar{\boldsymbol{\beta}}^\tau \mathbf{K} \bar{\boldsymbol{\beta}} \le \sum_{k=1}^K \boldsymbol{\beta}_{\cdot k}^\tau \mathbf{K} \boldsymbol{\beta}_{\cdot k} = \sum_{k=1}^K \left\| h_k(\cdot) \right\|^2, \qquad 4.70$$

where $\bar{\boldsymbol{\beta}} = (\bar{\beta}_1, ..., \bar{\beta}_n)^\tau$; if $\mathbf{K}\bar{\boldsymbol{\beta}} = 0,$ the inequality becomes an equality and so $\sum_{k=1}^K \beta_{0k} = 0.$ Thus,

$$0 = K^2 \bar{\boldsymbol{\beta}}^\tau \mathbf{K} \bar{\boldsymbol{\beta}} = \left\| \sum_{i=1}^n (\sum_{k=1}^K \beta_{ik}) K(\mathbf{x}_i, \cdot) \right\|^2 = \left\| \sum_{k=1}^K \sum_{i=1}^n \beta_{ik} K(\mathbf{x}_i, \cdot) \right\|^2, \qquad 4.71$$

whence, $\sum_{k=1}^K \sum_{i=1}^n \beta_{ik} K(\mathbf{x}_i, \mathbf{x}) = 0, \ \forall \mathbf{x}.$ Thus,

$$\sum_{k=1}^K \left\{ \beta_{0k} + \sum_{i=1}^n \beta_{ik} K(\mathbf{x}_i, \mathbf{x}) \right\} = 0 \qquad 4.72$$

for every $\mathbf{x}$. So, minimizing (4.58) under the zero-sum constraint (4.57) only at the n data points is equivalent to minimizing (4.58) under the same constraint for every $\mathbf{x}$.

We next construct a Lagrangian formulation of the optimization problem (4.58) using the following notation. Let $\boldsymbol{\xi}_i = (\xi_{i1}, ..., \xi_{iK})^\tau$ be K-vector of slack variables corresponding to $(f(\mathbf{x}_i) - y_i)_+, \ i = 1, 2, ...n,$ and let $(\boldsymbol{\xi}_{\cdot 1}, ..., \boldsymbol{\xi}_{\cdot K})^\tau = (\boldsymbol{\xi}_{\cdot 1}, ..., \boldsymbol{\xi}_{\cdot K})^\tau$ be

the $(n\times K)-$matrix whose kth column is $\xi_{\cdot k}$ and whose ith row is $\boldsymbol{\xi}_i$. Let $(\mathbf{L}_1,...,\mathbf{L}_K)=(\mathbf{L}(\mathbf{y}_1),...,\mathbf{L}(\mathbf{y}_n))^\tau$ be the $(n\times K)-$matrix whose kth column is $\mathbf{L}_k$ and whose ith row is $\mathbf{L}(\mathbf{y}_i)=(L_{i1},...,L_{iK})$. Let $(\mathbf{y}_{\cdot i},...\mathbf{y}_{\cdot K})=(\mathbf{y}_1,...,\mathbf{y}_n)^\tau$ denote the $(n\times K)-$matrix whose kth column is $\mathbf{y}_{\cdot k}$ and whose ith is $\mathbf{y}_i$. The primal problem is to find $\{\beta_{0k}\},\{\boldsymbol{\beta}_{\cdot k}\}$, and $\{\xi_{\cdot k}\}$ to

$$\text{minimize } \sum_{k=1}^{K}\mathbf{L}_k^\tau\boldsymbol{\xi}_i+\frac{n\lambda}{2}\sum_{k=1}^{K}\boldsymbol{\beta}_{\cdot k}\mathbf{K}\boldsymbol{\beta}_{\cdot k} \qquad 4.73$$

subject to

$$\beta_{0k}\mathbf{1}_n+\mathbf{K}\boldsymbol{\beta}_{\cdot k}-\mathbf{y}_{\cdot k}\le\boldsymbol{\xi}_{\cdot k},\ k=1,2,...,K, \qquad 4.74$$

$$\boldsymbol{\xi}_{\cdot k}\ge\mathbf{0},\ \text{k=1,2,...,K,} \qquad 4.75$$

$$(\sum_{k=1}^{K}\beta_{0k})\mathbf{1}_n+\mathbf{K}(\sum_{k=1}^{K}\boldsymbol{\beta}_{\cdot k})=\mathbf{0} \qquad 4.76$$

Form the primal function $F_P=F_P(\{\beta_{0k}\},\{\boldsymbol{\beta}_{\cdot k}\},\{\boldsymbol{\xi}_{\cdot k}\})$, where

$$\begin{aligned}F_P=&\sum_{k=1}^{K}L_k^\perp\boldsymbol{\xi}_{\cdot k}+\frac{n\lambda}{2}\boldsymbol{\beta}_{\cdot k}^\tau\mathbf{K}\boldsymbol{\beta}_{\cdot k}\\&+\sum_{k=1}^{K}\boldsymbol{\alpha}_{\cdot k}^\perp(\beta_{0K}\mathbf{1}_n+\mathbf{K}\boldsymbol{\beta}_{\cdot k}-\mathbf{y}_{\cdot k}-\boldsymbol{\xi}_{\cdot k}\\&-\sum_{k=1}^{K}\gamma_k^\perp\boldsymbol{\xi}_{\cdot k}+\boldsymbol{\delta}^\tau\left(\sum_{k=1}^{K}\beta_{0K}\mathbf{1}_n+\mathbf{K}(\sum_{k=1}^{K}\boldsymbol{\beta}_{\cdot k})\right)\end{aligned} \qquad 4.77$$

In (4.77), $\boldsymbol{\alpha}_{\cdot k}=(\alpha_{1k},...,\alpha_{nk})^\tau$ and $\boldsymbol{\gamma}_k$ are n-vectors of nonnegative Lagrange multipliers for the inequality constraints (4.74) and (4.75), respectively, and $\boldsymbol{\delta}$ is

an n-vector of unconstrained Lagrange multipliers for the equality constraint (4.76).

Differentiating (4.77) with respect to $\beta_{0k}, \boldsymbol{\beta}_{\cdot k},$ and $\boldsymbol{\xi}_{\cdot k}$), yields

$$\frac{\partial F_P}{\partial \beta_{0k}} = (\boldsymbol{\alpha}_{\cdot k} + \boldsymbol{\delta})^{\tau} \mathbf{1}_n, \qquad 4.78$$

$$\frac{\partial F_P}{\partial \boldsymbol{\beta}_{0k}} = n\lambda \mathbf{K}\beta_{\cdot k} + \mathbf{K}\boldsymbol{\alpha}_{\cdot k} + \mathbf{K}, \qquad 4.79$$

$$\frac{\partial F_P}{\partial \boldsymbol{\xi}_{\cdot k}} = \mathbf{L}_k - \boldsymbol{\alpha}_{\cdot k} - \boldsymbol{\gamma}_k, \qquad 4.80$$

$$\boldsymbol{\alpha}_{\cdot k} \geq \mathbf{0} \qquad 4.81$$

$$\boldsymbol{\gamma}_k \geq \mathbf{0} \qquad 4.82$$

The Karush–Kuhn–Tucker complementarity conditions are

$$\boldsymbol{\alpha}_{\cdot k}(\beta_{0k}\mathbf{1}_n + \mathbf{K}\beta_{\cdot k} - \mathbf{y}_{\cdot k} - \boldsymbol{\xi}_{\cdot k})^{\tau} = 0, \; k = 1, 2, \ldots K, \qquad 4.83$$

$$\boldsymbol{\gamma}_k \boldsymbol{\xi}_{\cdot k}^{\tau} = 0, \; k = 1, 2, \ldots K, \qquad 4.84$$

where, from (4.80), $\boldsymbol{\gamma}_k = \mathbf{L}_k - \boldsymbol{\alpha}_{\cdot k}$. Note that (4.83) and (4.84) are outer products of two column vectors, meaning that each of the n^2 element-wise products of those vectors are zero. From (4.80) and (4.82), we have that $0 \leq \boldsymbol{\alpha}_{\cdot k} \leq \mathbf{L}_k$, k=1,2,...,K. Suppose for some i $0 < \alpha_{ik} < L_{ik}$; then, $\gamma_{ik} > 0$, and, from (4.84), $\xi_{ik} = 0$ whence from (4.83), $y_{ik} = \beta_{0k} + \sum_{\ell=1}^{n} \beta_{\ell k} K(\mathbf{x}_{\ell} \mathbf{x}_i)$.

Setting the derivatives equal to zero for $k = 1, 2, ..., K$ yields

$\boldsymbol{\delta} = -\bar{\boldsymbol{\alpha}} = -K^{-1}\sum_{k=1}^{K}\boldsymbol{\alpha}_{\cdot k}$ from (4.97), whence, $(\boldsymbol{\alpha}_{\cdot k} - \bar{\boldsymbol{\alpha}})^{\tau}\mathbf{1}_n = 0,$ and, from (4.79),

$\boldsymbol{\beta}_{\cdot k} = -(n\lambda)^{-1}(\boldsymbol{\alpha}_{\cdot k} - \bar{\boldsymbol{\alpha}}),$ assuming that $\mathbf{K}$ is positive definite. If $\mathbf{K}$ is not positive-definite, then $\boldsymbol{\beta}_{\cdot k}$ is not uniquely determined. Because (4.78), (4.79), and (4.80) are each zero, we construct the dual functional F_D by using them to remove a number of the terms of F_P.

The resulting dual problem is to find $\{\boldsymbol{\alpha}_{\cdot k}\}$ to

$$\text{minimize } F_D = \frac{1}{2}\sum_{k=1}^{K}(\boldsymbol{\alpha}_{\cdot k} - \bar{\boldsymbol{\alpha}})^{\tau}\mathbf{K}(\boldsymbol{\alpha}_{\cdot k} - \bar{\boldsymbol{\alpha}}) + n\lambda\sum_{k=1}^{K}\boldsymbol{\alpha}_{\cdot k}^{\tau}\mathbf{y}_{\cdot k} \qquad 4.85$$

subject to

$$\mathbf{0} \le \boldsymbol{\alpha}_{\cdot k} \le \mathbf{L}_k, \qquad k = 1, 2, ..., K, \qquad 4.86$$

$$(\boldsymbol{\alpha}_{\cdot k} - \bar{\boldsymbol{\alpha}})^{\tau}\mathbf{1}_n = 0, \quad k = 1, 2, ..., K \qquad 4.87$$

From the solution, $\{\hat{\boldsymbol{\alpha}}_{\cdot k}\}$,to this quadratic programming problem, we set

$$\hat{\boldsymbol{\beta}}_{\cdot k} = -(n\lambda)^{-`}(\hat{\boldsymbol{\alpha}}_{\cdot k}\hat{\bar{\boldsymbol{\alpha}}}), \qquad 4.88$$

where $\hat{\bar{\boldsymbol{\alpha}}} = K^{-1}\sum_{k=1}^{K}\hat{\boldsymbol{\alpha}}_{\cdot k}$.

The multiclass classification solution for a new $\mathbf{x}$ is given by

$$C_k(\mathbf{x}) = \underset{k}{\arg\max}\left\{\hat{f}_k(\mathbf{x})\right\}, \qquad 4.89$$

where

$$\hat{f}_k(\mathbf{x}) = \hat{\beta}_{0k} + \sum_{\ell=1}^{n} \hat{\beta}_{\ell k} K(\mathbf{x}_\ell, \mathbf{x}),\ k = 1, 2, ..., K. \qquad 4.90$$

Suppose the row vector $\hat{\boldsymbol{\alpha}}_i = (\hat{\alpha}_{i1}, ..., \hat{\alpha}_{iK}) = \mathbf{0}$ for $(\mathbf{x}_i, \mathbf{y}_i)$; then, from (4.88), $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, ..., \hat{\beta}_{iK}) = \mathbf{0}$. It follows that the term $\hat{\beta}_{ik} K(\mathbf{x}_i, \mathbf{x}) = 0,\ k = 1, 2, ...K$. Thus, any term involving $(\mathbf{x}_i, \mathbf{y}_i)$ does not appear in (4.90); in other words, it does not matter whether $(\mathbf{x}_i, \mathbf{y}_i)$ is or is not included in the learning set $\mathbf{L}$ because it has no effect on the solution. This result leads us to a definition of support vectors: an observation $(\mathbf{x}_i, \mathbf{y}_i)$ is called a support vector $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, ..., \hat{\beta}_{iK}) \neq \mathbf{0}$. As in the binary SVM solution, it is in our computational best interests for there to be relatively few support vectors for any given application. (pp. 391-397)

Case Summary

When constructing a multi-class SVM, there is an extensive body of literature that attempts to address how to handle $K > 2$ classes. In ArcMap, the machine learning algorithm allows for classifying multiple features and classes for the training set. This case study is a mix of understanding the power of machine learning through simplicity to create SMART solutions that are practical. The multi-class SVM strategies currently address two types of scenarios:

- One versus the rest, where a K-class problem is divided into K binary classification sub-problems of the type “kth class” versus “not kth class”.
- One versus one, where the K-class problem is divided into $\binom{k}{2}$ comparisons of all pairs of classes.

As we have established all of the technical aspects of the machine learning process, more importantly these concepts translate into a way to find a practical SMART solution, namely: Can we establish an infrastructure that maps key buildings to use for a multitude of geospatial tasks. The next section examines the results of machine learning process as it relates to comparing the resolution of two images that compare the tradeoffs of dimensionality versus computational limits.

Results of the Machine Learning Procedure

This case study was a proof of concept in designing a SMART solution for tribal communities. The data sovereignty framework is the strategic planning behind how to implement a designed data domain. In the first case study, the objective was to develop the framework as a proof of concept working with Claremont Graduate University and the Road Safety Institute. The data domain, *Tribal Transportation Safety* was developed theoretically to obtain tribal stakeholder feedback as to strength of the concepts. In addition, *Case Study 1* was part of a larger study in *Using GIS to Improve Tribal Traffic Safety*.

In this case, we could define the data domain to be *GIS Infrastructure Point Patterns*. In the simplest terms, developing a critical analysis will showcase the expertise I bring to moving from theory into practice. The next step is developing this smart solution into a practical way for stakeholders with less statistical background to obtain results that can be used in a real life setting.

In training the classifier described in the grid search, the original objective was to choose three simple classes and attempt to isolate housing infrastructure so a polygon could be constructed, and ultimately obtain a spatial point pattern. As we will see, the process illustrated here would be how you can design a “white paper”, where after reviewing the complex nature of the problem, we move to solve the problem, and use this as strategic primer for decision-making.

Drone .25-meter Digital Globe 1-meter

Figure 4.9 - Training Sets for the Support Vector Machine for Image Comparison

The first step was to train the classifier. Figure 4.8 on page 175 is a visualization of the three defined classes: housing, vegetation, and pavement. This simple design was intentional. I ran similar classifications for two raster images the Digital Globe 1-meter resolution and a commissioned drone image .25-meter as shown in Figure 4.9.

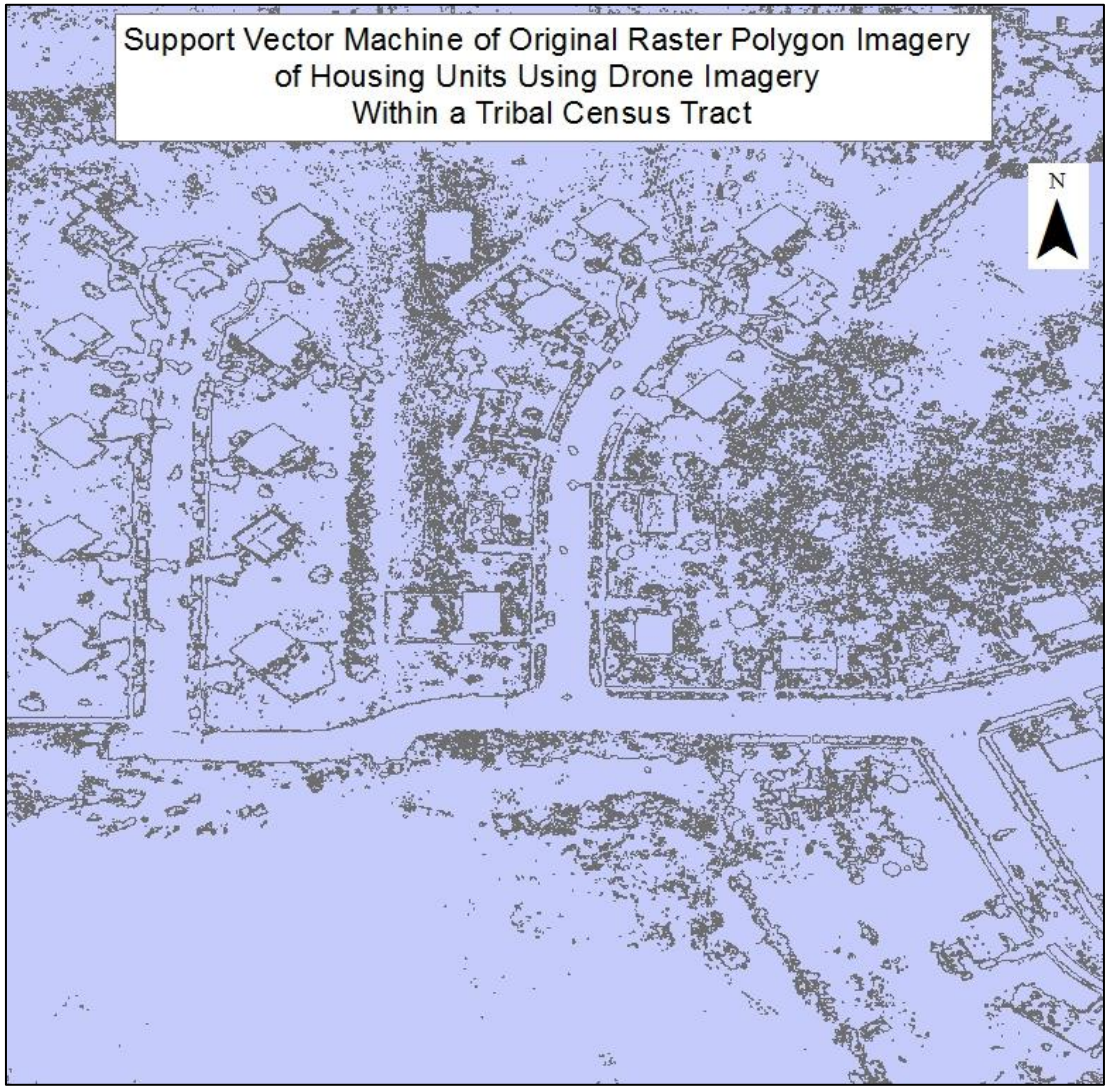


Figure 4.10 - Support Vector Machine

Once the polygons were drawn, the support vector machine geoprocessing tool classified the images into nearly 65,000 predictions. Figure 4.10 is an example of the resulting classification. Although it may seem to be a daunting task to differentiate this many predictions; this problem can be solved rather easily.

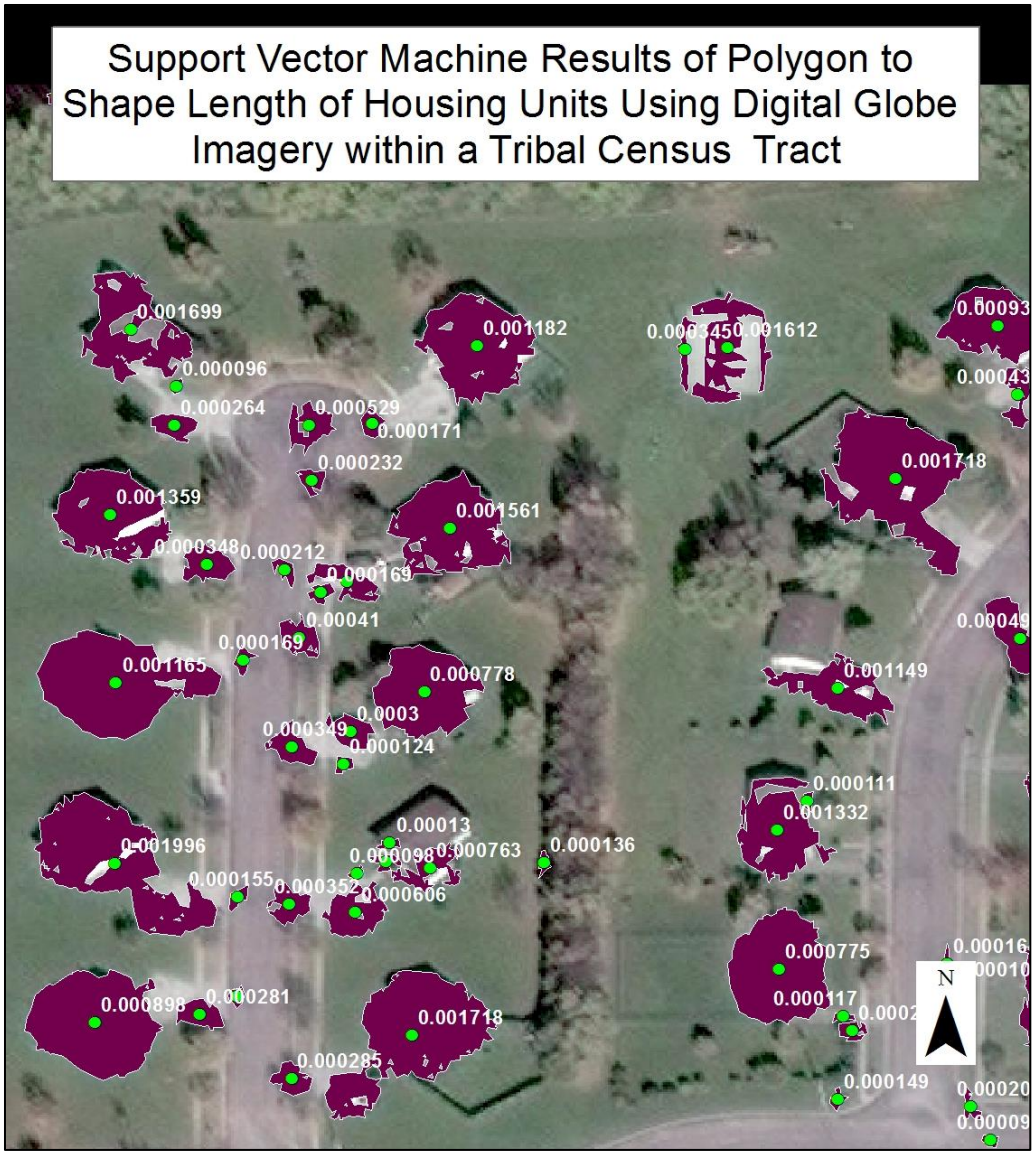


Figure 4.11 - Using Shape Length of the Classified Images to Remove Noise

The key output of the classification raster is the calculated shape length of the objects attempting to be classified. After running the summary statistics of the shape of each classified object and since we are looking for housing infrastructure; a quick identification of house shape length yielded a threshold of approximately .0001 or larger.

The same procedure was done for pavement and vegetation and a more refined selection was done in seconds. Figure 4.12 is the result of this selection.

As you can see, the support vector machine's further refinement of the infrastructure has become more apparent. To further study this refinement, I was interested in how effective at the computational cost, a high resolution drone image would be compared to a 1-meter satellite image.

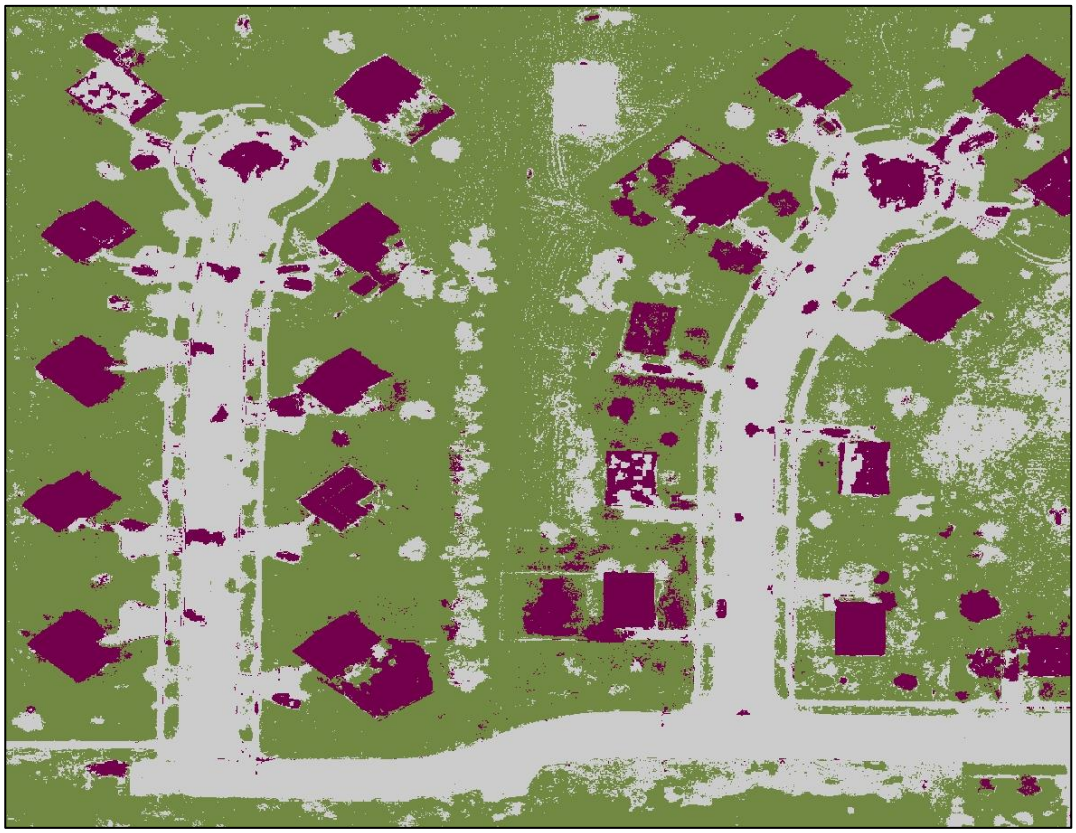


Figure 4.12 - Further Refinement of the Original Raster Classification

Figure 4.13 on the next page clearly shows the performance of a high resolution image is superior to a lower resolution raster image that is one meter or greater. So the question becomes: How can this analysis provide insight into economic development?

Drone .25-meter Digital Globe 1-meter

Figure 4.13 - Support Vector Polygons Created with .25m and 1m Resolutions

On one hand, tribes can obtain digital globe images and thus obtain images on the right in Figure 4.13. On the other hand, if tribes use the data sovereignty framework where a strategic plan can be developed around an indigenously centered data science white paper; then the case can involve using private drone image companies that capture this data under equitable interaction.

As described in the literature review, data obtained that is in direct control of the tribe is not only beneficial, but adds a dimension of data ownership that, when negotiated,

provides accurate infrastructure data for the use in community and economic development. Once these images have been created, the data sovereignty framework seeks to protect this data as a matter of sovereignty.

Figure 4.14 - A Point Pattern Created with Geospatial Points Based on the Polygon Construction

Now that we have examined these images in detail, the final digital infrastructure can be created for study. As you can see in Figure 4.14, an actualized point pattern can be created using a tool that creates a polygon of all classified images. In turn, this polygon can also display the centroid of the pattern. Figure 4.14 shows the initial calculated point pattern of the refined drone image.

Figure 4.15 - Candidate Point Pattern Selection Through Object Identification

Next, a simple search of polygon object ID, can refine this search quite easily. Clearly in this image, there is a finite number of housing units. For this image, there are twenty -two dwellings, and a quick check of polygon ID can refine the final infrastructure point pattern. The next image is the realized infrastructure point pattern.

Figure 4.16 - Final Candidate Point Pattern Selection

As we can see, the point pattern obtained is incredibly accurate given the simple selection of classes for the original grid search. This was done in hours rather than weeks or months. The power of machine learning is clearly the way to move forward when advocating for tribal communities by developing digital infrastructure in data ownership, privacy, and economic development.

This point pattern is now a *Master Address File* and has the power to attach any number of covariates that can model any topic that is of interest spatially. Tribes will undoubtedly have the associated postal addresses associated with any tribal housing under their respective tribal housing authority, so having the geospatial point pattern can

create an innumerable number of geospatial outcomes to the tribes benefit. The last piece of this case study is to examine the practicality of classifier accuracy versus machine learning accuracy regarding a confusion matrix.

What is a Confusion Matrix?

A confusion matrix summarizes the classification performance of a classier with respect to some test or training data. It is an n-dimensional matrix, indexed in one dimension by the true class of an object and in the other by the classes that the classifier assigns.

The matrix is essentially a type of contingency table. In the binary case, there are four cells of the matrix and the rows correspond to the actual class and the columns predicted or assigned class. This allows the classifier to count the number of times an object has been misclassified. The true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are metrics defined by the predicted versus actual values. Figure 4.17 demonstrates this concept (Ting as cited in Sammut & Webb, 2011). This can also be expanded to any $n \times n$ class.

		Assigned Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

Table 4.1 - An Example of a 2x2 Confusion Matrix

Two measures that are of interest are accuracy and the kappa value:

$$\text{Accuracy} = p_0 = \frac{\sum diagonal\ entries}{n} \qquad 4.91$$

$$kappa = \frac{p_0 - p_e}{1 - p_e} \qquad 4.92$$

where p_0 is the relative observed agreement among raters (identical to accuracy), and p_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category (Cohen, 1960).

Class	Housing	Vegatation	Pavement	Total	User Accuracy	Kappa
Housing	14	13	6	33	42.42%	
Vegetation	0	32	1	33	96.97%	
Pavement	2	17	14	33	42.42%	
Total	16	62	21	99		
Producer Accuracy					60.61%	
Kappa						40.91%

Table 4.2 - The Confusion Matrix Results from the Second to Last Refinement Using a Random Equalized Stratified Sample

The producer accuracy in Table 4.2 is the fraction of pixels classified correctly per total classifications and the kappa values are based on the total user's and producer's accuracies; it gives an overall assessment of the classification's accuracy. Overall, these accuracies are not entirely important. The reason this is problematic is because from a

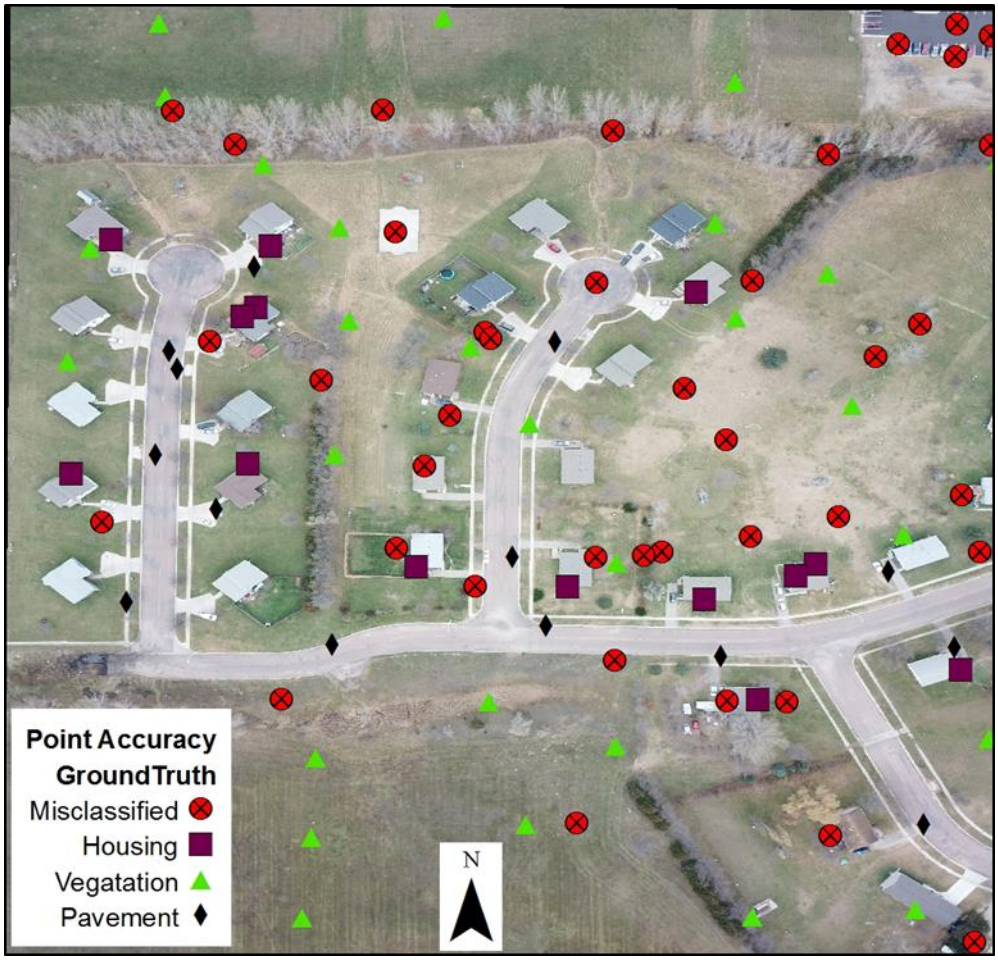


Figure 4.17 - The Confusion Matrix Results with Visual Assessment by Icon for 99 Equally Stratified Classes of 33 Randomly Points for Each Class

statistical point of view, we are testing the machine learning accuracy in accordance with a stratified set of equal random points when we are really only interested in how accurate the point pattern relative to the housing infrastructure is. Figure 4.17 provides a visual representation of the confusion matrix in Table 4.2. Again, how is this accuracy related to figure we obtained in the final point pattern in Figure 4.16? This consideration is only based on a rigorous evaluation of statistical techniques that weigh the importance of theoretical rigor versus practical application. In this case, we have shown that achieving

parity between these two schools of thought has be carefully balanced to achieve what I have deemed to be a SMART solution.

Scientific Implications

As this case study came to life, I was reminded how powerful technology has become. Since the speed at which these technologies continue to accelerate, it is equally hard to comprehend the number of capabilities these tools have in solving problems. The theoretical foundation of this concept can allow for further refinement of the classifier for more accurate results. In the case where drone images are not available, it is possible to use up to a one meter resolution; however the classifier will need finer grid searches and additional training classifier images. Either way, this method is not only efficient, but practical as well.

The point pattern obtained in this case study as a MAF has many useful properties. Attaching covariates whether categorical or continuous, can produce man y field oriented projects an organization can undertake. Field projects such as mapping utilities, 911 infrastructure, housing analysis, census, or even spatial modeling and simulation are all possibilities when a critical point pattern related to economic planning and development can be obtained for studying more complex point processes.

Framework Implications

The data domain created from this case study also has a number interesting implications to the data sovereignty framework. As this manuscript has outlined, pairing this outcome with the three key indicators creates multidimensionality. A further examination of the key indicators will drive a more refined strategic plan of how this data

domain can be used in practice. Since the scientific implications are independent of this framework, the key is to integrate this SMART solution for further study. The key indicators in the framework will allow for a more complex review of how this outcome can serve tribal governance and tribal communities once this is integrated into the scientific implications of this case.

Cultural Implications

Although this case study is just one step in a greater design I have constructed; nonetheless it is important to consider who these SMART solutions have been designed for: stakeholders with citizen science in mind. The citizens of a tribal nation have specific needs and priorities when considering such an ambitious design. They need to consider the culture, tribal government, and political implications that accompany a data driven decision such as this. The citizen science aspect aims to pair with the key indicators tribal community and culture to create dialogue and strategies to govern this process by community input and reflection.

Final Thoughts

As I have shown, the steps in using a machine learning algorithm for image classification are an achievable goal, even for people with no experience in GIS. Now that the smart solution has been created, a white paper prototype on how to perform the steps can be written for a specific client. The data sovereignty framework key indicator allows for an expansion of tasks that revolve around the data domain we have created, *GIS Infrastructure Point Pattern.* In the near future, this data domain and many others will serve communities as a matter of data sovereignty moving forward.

Chapter 5

Conclusions and Recommendations

This manuscript has been a huge undertaking. As I have described throughout this dissertation, there is a fundamental need to create smart, complex, and well-designed systems to promote nation building. I would like to outline some key issues I learned in the construction of this manuscript and outline a set of guidelines regarding some important topics of discussion moving forward.

I have always been a privacy advocate and I have always believed in open source philosophies. Throughout the last decade, many people have freely shared their personal and private information in exchange for services hosted by huge technology companies such as Google, Facebook, and Apple to name a few.

While some of these systems have been designed as open source, many of them are not. The rise of information technology and data science is in an interesting state of flux at the moment. The explosion of data has fundamentally changed the way humans conduct nearly everything in their lives. Most would argue that this is a good thing. My position is that I fundamentally embrace what technology represents, but again not without *context*.

Because I was forced to critically examine what the potential outcomes this framework would represent in a real life situation; I began thinking about how critical data has become in defining an organization's digital footprint. This footprint has an incredible potential to provide representation; however it also has the potential to sow division.

In March of 2018, it was revealed that a British political consulting firm named Cambridge Analytica, had improperly obtained the personal information from over 87 million Facebook users by acquiring this data from a researcher who 'claimed' to be conducting research for academic purposes. Needless to say, there has been a firestorm of any number of ethical, moral, and privacy issues involving data.

This is the fundamental issue data science faces moving forward. I created the *Data Sovereignty Initiative* as fundamental check to this type of data collection and practice. In the introduction, I outlined the fundamental reason I have pursued this area of study is to provide an ethical, cultural, and community based consultancy that is designed by an American Indian, for nation building to assist tribal communities with economic development, strategic planning, and data driven decision-making. I would like to comment on this further.

Privacy and ethics are the topics I would like to cover to complete the scope of the Data Sovereignty Initiative because as the Cambridge Analytica example demonstrates; it is not enough to simply assume that people or organizations will do the right thing. A careful plan must be devised to achieve a high ethical standard to assure tribes are in the best negotiating position and the data is used for what it was intended for.

Shanley (2015) outlined an extremely thorough analysis of spatial data sovereignty and privacy in Indian country. Her manuscript covered many concerns that tribes have with the use of data to undermine their sovereign status as a nation.

Shanley (2015) writes,

> The Supreme Court in Klamath Water Users used the application of Exemption 5 to tribal information shared with the federal government. The Court also explicitly rejected an "Indian trust responsibility" exemption to FOIA. While this is in keeping with the Court's history of narrowly interpreting FOIA to encourage public disclosure, it has had a deleterious effect on the federal-tribe trust relationship. Tribal information also may become public under discovery or judicial review of agency actions.
>
> Fundamental issues are at stake – Indian tribes' rights and interests in their natural resources and federal agency's decision-making processes that affect these resources. The incorporation of tribal expertise and information into environmental planning and policy formulation, however, is critical if Indian tribes' rights and interests are to be protected. If information is withheld, federal agency decision makers may deduce incorrectly that natural and cultural resources are insignificant. Yet, once these communications become part of federal agency record, they are at risk for disclosure under the FOIA. (p. 251)
>
> "We must be careful that the incorporation of indigenous knowledge in the planning process in fact leads to empowerment, and is not merely a repackaging and legitimization of state and corporate domination" (Shanley, 2015, p.251).

These statements are an example of a more complex set of issues that tribes face with storing, collecting and owning their data. Tribes can fall into peril if the agreements they enter into like Kessler-Mata (2014) asserted is not through equitable interaction.

Shanley (2015) continues:

> The United States does not have comprehensive federal privacy law. With the exception of a set of narrowly focused federal privacy laws, the private sector is largely unregulated. Although Fourth Amendment law may provide some insight into how the courts will characterize satellite surveillance by commercial operators, the courts will not introduce Fourth Amendment law when considering commercial aerial or satellite surveillance. Privacy is traditionally protected by common law torts. Craig (2007) suggests that 'potential causes of action associated with online satellite and aerial images include: (1) trespass; (2) nuisance; (3) invasion of privacy; (4) strict products liability; (5) violation of 42 U.S.C. Sec. 1983; (6) patent infringement; and (7) other miscellaneous actions.' Satellite imaging does not create a nuisance nor trespass private property, he contends, but it may present an "unreasonable intrusion upon the seclusion of another. (pp.274-275)

As Shanley (2015) has shown, the tribes have an uphill battle in achieving parity between their rights to sovereignty, data sovereignty, and how that affects governance and economic development. In the process of writing this manuscript, I have taken great care to try and understand these issues by designing key indicators in the framework to bring these issues to forefront so when making data-driven decisions; tribes have a better negotiating position.

Lastly, I would like to discuss the ethics of data and the use of data. Throughout my research, I have been closely watching the technology sector and have started to see

certain patterns emerge. One pattern is the idea of socially engineering systems to create social credit systems that penalize a person or persons whose behavioral data is in disagreement with a deterministically server client side set of digital rules. China is currently using this system to assign a human being a social credit score. For instance, if your friend does not have a favorable credit history, and you being friends with them, will jeopardize your ability to get a loan. This is what I call the biggest technological crisis that will overtake data science in next decade.

The tech sector has been unchecked in how the use data has created many unforeseen consequences we have yet to truly understand its scope. Stopping bad actors or users with no interest in privacy, or psychographic profiles that could be used to swing elections are merely just a symptom of what I would argue as lack of ethical standards in regard to data, collection and practice.

In addition, the lack of diversity in tech companies have created extensive networks of artificial intelligence research and development that are creating systems of machine learning that have a high risk of cultural bias if these machines are used to make decisions that affect people's quality of life. These are the issues of our times.

These observations are not meant to forecast the end of the world; rather it is to celebrate the beginning of a new age: The Age of the American Indian Data Scientist. *The Data Sovereignty Initiative: Creating SMART Solutions for Tribal Communities* represents a school thought as to how to use data and machine learning in the highest ethical standard for the community's benefit.

Moving forward beyond this manuscript lies a whole new unexplored world. To honor this new world and to honor the past, I have founded a company named Mato Ohitika Analytics. My late father Creighton's given Dakota name is Mato Ohitika, which translates as Brave Bear. In order to celebrate the historical sacrifice of those who came before me; we again must understand the *context* of that sacrifice.

Thus, the entire premise of the manuscript was not to have it sit on a shelf; but rather serve as the cornerstone of my company's strategic plans to develop smart solutions for tribal communities. In the coming months I will be developing a user's digital privacy bill of rights so the mission of my company adheres to high ethical standards; and seeks to promote those values through higher education and nation building.

I am reminded by what Wildcat (2001) said that provides context to these future mission goals:

> The question of self-determination from a standpoint of American Indian practice in education is essentially a question of the degree to which individuals and communities are actively engaged in the future - not in the abstract but what Dewey called the "live-in" present. For we are all involved in a living process - some merely less conscious or, I prefer to say, less aware than others about the future they enabling through their present activity. (p. 144)

The Data Sovereignty Initiative is actively engaged in writing a new future for American Indian people through data science. And if in the future, Mato Ohitika Analytics can be the cornerstone of the dialogue; then I am confident the future looks bright for not only American Indian Communities, but humanity as well.

APPENDIX

The Foundations of the Data Sovereignty Initiative: A Biographical Sketch

The conceptual footprint of this manuscript began many years ago when during my undergraduate studies at Colorado State University, the director of the Native American Cultural Center and I began working together on strategies to better understand how to retain American Indian students through their collegiate studies. To begin to address this issue, we developed mentoring and tutoring programs, and put the Native students attending the university in these positions (including myself) to take ownership of their fellow students' well-being and studies. This created a community based support system that over a decade later has resulted in an explosive growth of Native students enrolling and succeeding at Colorado State.

As I moved into teaching mathematics and school improvement at a number of tribal schools and colleges in my home state of South Dakota, I had firsthand experience understanding some of the actual reasons why many Native students who come from reservations have a very low probability of succeeding. Issues of historical trauma, poverty, lack of quality teachers, and many students have never had any of their family members attend college. This fundamentally changed my view of what I was resolved to do in service of my community.

What was most striking of these experiences was not that education isn't important; rather it was not a fundamental priority at any given time. The acceptance of these outcomes when you witness them first hand is beyond humbling. The thought had crossed my mind that I was fighting a battle that could never be won, and that my efforts

would succumb to massive failure because I had not realized the scope and context of the task at hand.

Through my experiences, a lot of doubt had been cast in my once altruistic view of how higher education would be the gatekeeper of the future of Indian communities. When I was pursuing my state teaching credentials to be a highly qualified mathematics teacher, I had written about my experiences in a required human relations course in enacting social justice through education.

I described my experiences in the tribal education system as:

> ...The fact remains that my high level of education and my spiritual connection to my people served to be two forces that contradicted each other due to historical trauma...It was always explained to me in the acculturation process that my education would be the cornerstone of modeling success. Ironically, my "white" education has been the demise of any attempt at social justice. Clearly the stagnation of social justice has affected me directly. The decisions of people in power whether Indian or non-Indian has forced me to question whether what I represent is an achievable goal at this state in the education system. Social justice as it relates to education is far more complex than it I had ever expected. So the question remains, in attempting to enact social justice what are the inherent pitfalls to promoting values that are met with extreme resistance?
>
> Though I have been told that our communities need someone like me; it is clear that that is only merely a starting point to truly understand what is at stake and what actions of any can be used to mitigate the circumstance. On a personal level,

> it is extremely devastating psychologically to be put off by society as an unequal unless you understand the context of stratification, prejudice, and racism. This class has offered me insight that although education is a personal choice, it is not a means to an end.
>
> Despite all the difficulties, I understand that the opportunity afforded to me will perhaps never be an opportunity afforded to some of my relatives that continue to live on the reservation. Furthermore, the children are greatly affected by irresponsible school decisions and as such, it is my responsibility despite the seemingly overwhelming challenge to overcome; I will continue to model the importance of education. The pursuit of social justice means that if it is this difficult for me, we must respect the extreme difficulties in letting the process evolve by continuing to work piece-by-piece to find a solution.

My thoughts from that candid reflection guide me today. Piece-by-piece is the operative word. When understanding the experiential nature of American Indian communities, it is imperative to understand the shared history of what Indian people had to endure and how incalculably, our culture remains intact. That is why this dissertation has become so important. The forward thinking of my committee chair and my department head have made this examination possible.

As we will see, the scholarship of our elders has somehow already foreseen my fate. And although I had not even realized this until I was constructing this manuscript, it is suffice to say that when I wrote years ago about the need to continue on piece-by-piece, I will be forever be reminded that in order for you to take your place in history,

requires honoring the sacrifices that were made for me to even have such a unique and special opportunity like this.

This initial introduction is crucial to establishing that when we ignore history, culture, and context, then we will never be able to achieve opportunities that are a real actualization of the power of higher education. The key is to define education praxis as Friere (1974) describes as “the important thing is to help men (and nations) help themselves, to place them in consciously critical confrontation with their problems, to make them agents of their own recuperation”.

As I transition from a worldview of context in this introduction it fundamentally builds the foundation of the data sovereignty framework I have conceived that can be used in real life. Ideas of self-determination, education, and nation building have their place because what builds educational success through praxis is context.

Self-Determination and Education

In 2014, when I began constructing my initial ideas about what this dissertation would represent a number of topics I was interested in came to mind: higher education, American Indian history and culture, and the ideas of what actually self-determination means in practice. My initial sense actualizing this in the SDSU Mathematics and Statistics Department seemed to be untenable.

I felt the extensive research I would need to undertake to realize a vision had to be centrally positioned around this self-determination construct in some way. The key question at hand was how I could convince my department to support such an unorthodox exploration in computational statistics that represented an absolute foreign way of

thinking given the United States euro-centric system of education which stresses examining the *parts* rather than the *whole*.

American Indian learning outcomes have rarely been acknowledged to be the antithesis to this style of teaching and learning. Tharp, (as cited in Roppollo and Crow, 2007) "described Native American cognition as the 'anchor example' for holistic thought, in which the pieces derived their meaning from the pattern of the whole, rather than the whole being revealed through the analysis of each of its sections. Neither of us had heard nor read anything professionally that contradicted this visual and holistic learning style for American Indian students" (p.3).

As an American Indian, this could not be closer to the truth. The fact I had to not only understand but overcome this obstacle in my higher education pursuits made this manuscript an important representation of redefining the paradigm of higher education.

I was struck when Vine Deloria Jr. and Daniel Wildcat in their collection of essays, *Power and Place* had written about self-determination and education and the pursuit I present today. "Native people navigating American systems of higher education must absorb a great deal of factual content, and they must also place that knowledge into the context of their own tribal and community traditions. For the American Indian students the scientific method and the Western worldview coexist with Native spirituality and a deep connection with the earth". The essays they present on Indian education covers topics of praxis that are philosophic, practical, and visionary in nature.

My understanding of self-determination began evolving as I began to understand that the complexity of the topic had already been covered by the Native scholars above in the supposition of my assessment with my doctoral dissertation in mind.

One thing to note: Deloria uses the term *Indian* in his writings and as such, he is referring to American Indians. I use the terms Native American or American Indian whenever possible to avoid confusion with other cultural groups.

Deloria (2001) writes:

> Self-determination inevitably had to take on different meanings when applied to Indian Tribes and reservations. And as to the original goal of the Kennedy and Johnson administrations was to delay the termination of federal services until such time as tribes achieved some measure of economic parity with their white neighbors, self-determination in the Indian context basically has meant Indians can administer their own programs in lieu of federal bureaucrats. Education was conceived as the handmaiden of development...
>
> While Indians have been penetrating the institutions of higher learning, the substance and procedures of these institutions have also been affecting Indians. Indians have found even the most sophisticated academic disciplines and professional schools woefully inadequate. This is because the fragmentation of knowledge that is represented by today's modern university does not allow for a complete understanding of a problem or a phenomenon...
>
> In the past four decades, while the movement for self-determination was proceeding, we have witnessed a drastic decline in politeness and civility in Indian communities. Indian meetings are many times difficult to attend because they consist of little more than people clamoring for attention and people busy impressing each other. The outstanding characteristic of Indian students today is the emergence of politeness as a personality trait. Science and engineering

> students more than others now seem to possess this most precious of all the old traditional personality traits.
>
> Here we may have an indication that the current generation of Indian youth is moving beyond the boundaries established for non-Indian self-determination, and now this generation stands ready to bring something entirely new to the process of applying Western scientific knowledge to Indian problems. If this observation is correct, then we will witness some very unusual things happening in Indian communities in the future. Indians who are now working at the professional level, particularly in science and engineering, will work their way through corporate and academic institutions and begin appearing as independent consultants and owners of small, technologically oriented businesses working in ecological restoration and conservation areas.
>
> Research institutes headed by Indians will begin to appear on certain college and university campuses doing complex research projects. One or two of these people will write extremely sophisticated papers and books that will be highly regarded in their professions.
>
> Indian students in colleges and universities will begin to combine majors, putting together unlikely and unpredictable fields. They will have some degree of difficulty doing so because of the department's inability to reconcile the students' interests within traditional Western disciplinary relationships. In the increasing number of Indian students will choose very specific new majors that represent non-Indian converts to do interdisciplinary work and that are almost entirely outside the fields being chosen by present Indian students.
>
> Indian graduate students will be doing very sophisticated dissertations, and in hard sciences, highly innovative research projects. (pp. 124-130)

This statement is profound. This statement is a truth of such high magnitude that it in essence it explains exactly the struggles, the search for maintaining culture identity, and the inherent risks I had to take to not only get approval for this dissertation, but

achieving parity with the balance between Western scientific principles and cultural tradition.

This journey has had as many ups and downs as I could not have imagined, from the experience of institutional racism, discrimination, sabotage, and professional envy. This is to be expected. No one said this would ever be easy, and more importantly no one said how incredibly important this task would become. Again, I am fortunate to have a department head who was forward thinking in approving the central theme of this manuscript. Because of this, I am able to explain why we need to ascribe "context" to the unique perspective of what this dissertation represents.

The reason education and higher education form the cornerstone of the analysis, is because the historical context and quite possibly the precedence has to be understood so that a shift in perception is possible. This discussion is crucial in framing the impact this work has in demonstrating to tribal youth what indigenous scholarship looks like and how it is perceived.

Education in its purest form ascribes a process of discovery. This process prescribes opportunity as well as endows the transfer of knowledge into wisdom. However, American Indian communities' continued struggle with understanding the value of education is not because it is not perceived to be important; rather the Western view of education has been the cornerstone of indoctrination.

Take for instance; the model for American Indian boarding schools was the Carlisle Indian School. Founded in 1879 in an abandoned army post in Pennsylvania, the goal of Carlisle was to strip all vestiges of Indian culture from the Indian students: they

were to speak only English, they were to dress in the American style, they were to eat American foods, they were to worship the Christian gods, and they were to live in American-style houses. More importantly, Indian students were forced to give up their names in exchange for a more "proper" English name (Adams, 1995).

The school was headed by Captain Richard H. Pratt, the former commandant of the Fort Marion Prison in Florida, which served as an Indian prison. While Pratt liked individual Indians, he had no use for Indian cultures and felt that these cultures would have to be destroyed if Indian people were to survive. Like many other Americans, Pratt felt that Indian ways were inferior in all respects to those of non-Indians. Thus the slogan for Carlisle was “Kill the Indian, Save the Man.” The students at the Carlisle Indian School were told in 1893 (The Carlisle Boarding School, 2013):

“You are a race thrown by the Providence of God in the pathway of a mighty and resistless tide of civilization, flowing Westward around you. So mighty is the flood that resistance is fruitless, and the only choice is between submission and destruction on the one hand, or joining the flood and floating with it, on the other.”

This is why the fundamental view of education needs to have context. Again, this manuscript is much more than a doctoral dissertation; it is a fundamental examination of educational praxis from an indigenous point of view. This provides evidence as to the challenges American Indian students have faced generationally in the pursuit of their studies. It has been only a few generations removed from the boarding school era, and these perceptions of higher education related to assimilation are very real.

Deloria (2001) continues:

> Education has generally been misunderstood by its practitioners. It is defined as both process and content, and it is exceedingly difficult to tell from educational behavior and philosophy whether or not the educator is making the proper distinctions…From the Miriam report of 1928 until the present we have been living in the age of process – which is to say, we have been more concerned about *how* children learn than with *what* they learn.
>
> During the past forty years we have been exclusively concerned with how they learn and have almost studiously avoided asking what it is they are learning. This situation is particularly difficult for students who are studying science because, in most respects, science is content and not process. Consequently, after educating Indian young people in schools that stress learning experiences, we suddenly place upon them the demand that they accommodate themselves to the scientific enterprise – which is to say, build scientific expertise on a secondary education that has very little content. The student has no choice except to attempt to learn the scientific curriculum as well as gain background in the mass of conflicting ideas that now passes for Western civilization.
>
> When the social adjustment from Indian community-based culture to non-Indian urban networking culture has to be made at the same time, many students adopt a very rigid posture concerning personal, group, and community values.
>
> Too often they model themselves after the professionals in their academic field or their institutional situation. This adjustment then forces them outside their Indian circle and greatly inhibits their ability to draw from their own tribal traditions but lessons that could be profitably learned regarding both science and the social world in which they live. That we are producing any Indians in science at all is a tribute to the perseverance of this generation of Indian young people. (pp. 81-83)

This is the fundamental obstacle that many Native students face in western euro-centric education science praxis: The disconnect with culture does not provide a

reasonable expectation to achieve parity with what I am establishing in this manuscript: an integrated cultural examination of how science connects to cultural identity which then informs a new educational praxis.

Finally, Deloria (2001) concludes:

> Self-determination will not be an issue because people will be doing it in forms that even they will not recognize. Although it appears easy to make vague predictions concerning the future of Indians and education, none of these ideas is an ad hoc concept. Rather, everything flows from the original idea of education acting as the motivational force in self-determination. The policymakers four decades ago assumed education would radically change Indian young people while also assuming that they would hold, as a constant, the value of returning to their tribes to take the lead in development projects. Higher education really was thought to be higher than the knowledge and experiences that Indians brought from their homes and communities.
>
> Higher education might have been more complicated, but it was too departmentalized, and consequently the chinks in the armor were all too apparent and left most Indian students with the feeling of having an incomplete knowledge. Unable to bring academic knowledge to its proper unity, more and more students are now supplementing the shortcomings of Western thought by placing it in the context of their own tribal traditions.
>
> Once the process of supplementation began, it would naturally follow that individuals would begin to compare specific items of Western knowledge with similar beliefs derived wholly from the traditions of their tribes. We see this process now emerging as an identifiable intellectual position of this generation of Indians. It will take a considerable period of time for new theoretical postures to be developed by this generation, but some individuals are well on their way to doing so.

As a new perspective is formed individual Indians will move completely through the institutional structures will take all conceptions of Indians beyond the ability of Western ideas to compete, and this conceptual shift will focus attention on cultural knowledge.

In a previous essay I discussed the fact that much of American education is really just training and indoctrination into the Western view of the world. Basically this view is held together by the sincerity of its followers. It does not have internal consistency of its own except in general methodological patterns whereby information is classified.

Indians over the long run are exceedingly hard to train because they easily get bored with the routine of things. Once they have understood and mastered a task it seems like a waste of time to simply repeat an activity. So for an increasing number of Indians the training received at institutions of higher learning only raises fundamental questions that are never answered to their satisfaction.

We can visualize the effects of education on Indians as follows, non-Indians lived within a worldview that separates and isolates and mistakes labeling and identification for knowledge. Indians were presumed to be within this condition except they were slower on the uptake and not nearly as bright as non-Indians.

In truth Indians were completely outside the system and within their own worldview. Initiating an accelerated education system for Indians was intended to bring Indians up to the parity of middle-class non-Indians. In fact this system has pulled Indians into the Western worldview, and some of the brighter ones are now emerging on the other side, having traversed the Western body of knowledge completely.

Once this path has been established, it is almost a certainty that the rest of the Indian community will walk right on through the Western worldview and emerge on the other side also. And it is imperative that we do so. Only in that way can we transcend the half millennium of culture shock brought about by the confrontation with Western civilization. When we leave the culture shock behind we will be

> masters of our own fate and able to determine for ourselves what kind of lives we will lead. (pp. 132-133)

The story map I created as a primer to this manuscript, *The Impact of Data Sovereignty on American Indian Self-Determination* was the first of many ideas that demonstrate my inherent connection to defining data sovereignty as a matter of context. Not only has Deloria and Wildcat laid the foundations of what the state of Indian education was, is, and will be; I am living proof as an exculpable measure of what traditional passing down of knowledge means and how we interpret the science and technology from a purely indigenous perspective. In the simplest terms, and what could be an implied by this insight, is what this dissertation has ultimately become in the context of historical sacrifice:

It asserts that the Age of the American Indian Data Scientist has begun.

REFERENCES

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Akabe, A., & Sugihara, K. (2012). *Spatial Analysis along Networks*. West Sussex, U.K.: John Wiley & Sons, Ltd.

American Statistical Association. (2015). *ASA Statement on The Role of Statistics in Data Science*. [PDF Document]. Retrieved from: https://ww2.amstat.org/misc/DataScience Statement.pdf

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Anselin, L. & Bera, A. K. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullau, A. and Giles, D. E. A., Eds., *Handbook of Applied Economic Statistics*. Marcel Dekker, New York, pp. 237–289.

Anthony, G., Greg, H., Tshilidzi, M. (2007). *Classification of Images Using Support Vector Machines* [PDF Document]. Retrieved from: https://arxiv.org/abs/0709.3967.

Baddeley, A., Rubak, E., & Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton, FL: Chapman and Hall/CRC Press.

Bardach, E. (2012). *A Practical Guide to Policy Analysis*. Los Angeles, CA: CQ Press.

Begay M., Cornell S., Jorgensen M., & Kalt, J., (2007) *Rebuilding Native Nations: Strategies for Governance and Development*. M. Jorgensen (Eds.). Tucson: The University of Arizona Press.

Cabrera, J., & McDougall, A. (2002) *Statistical Consulting*. New York, NY: Springer-Verlag New York Inc.

Chainey, S. (2010). *Spatial significance hotspot mapping using the Gi* statistic*. [PDF Document]. Retrieved from: http://www.popcenter.org/conference/conference papers /2010/Chainey-Gi-hotSpots.pdf.

Chainey, S. & Ratcliffe, J. (2005). *GIS and Crime Mapping*. Hoboken, NJ: John Wiley & Sons, Inc.

Chang, C-C. and C-J. Lin. 2011. LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2(27), pp. 1-27.

Chikio, H. (1998). What is Data Science? Fundamental Concepts and a Heuristic Example. *Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization.* Springer Japan. pp. 40–51.

Chiou, Y-C., Fu, C. (2015). Modeling crash frequency and severity with spatiotemporal dependence. *Analytic Methods in Accident Research*. 5-6, 43–58

Cohen, F. (2012). *Cohen's Handbook of Federal Indian Law* (2012 ed.). New Providence, NJ: LexisNexis.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 20(1): 37–46.

Cornell, S., & Kalt, J. P. (2007). Two approaches to development of native nations. In M. Jorgensen (Eds.). *Rebuilding Native Nations: Strategies for Governance and Development*. Tucson: The University of Arizona Press.

Corntassel, J., & Witmer II, R. (2008). *Forced Federalism: Contemporary Challenges to Indigenous Nationhood.* Norman: The University of Oklahoma Press.

Cortes, C. and Vapnik, V.N. (1995). Support vector networks, *Machine Learning*, 20, 273–297.

Cressie, N., & Wikle, C.K. (2011). *Statistics for Spatial-Temporal Data*. Hoboken, NJ: John Wiley & Sons, Inc.

Cressie, N. (2011). *Statistics for Spatial Data*. Hoboken, NJ: John Wiley & Sons, Inc.

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*. New York, NY: Springer Science + Business Media LLC.

Data Science. (n.d.). Retrieved March 11, 2018 from wiki: https://en.wikipedia.org/wiki/Data_science.

Dean, A., Morris, M., Stufken, J., & Bingham, D (Eds.). (2015). *Handbook of Design and Analysis of Experiments*. Boca Raton, FL: Taylor and Francis Group LLC.

Deloria Jr., V. (1976). A Better Day for Indians. New York: Field Foundation.

Deloria Jr., V. & Wildcat, D. R. (2001). *Power and Place. Indian Education in America*. Golden, CO: Fulcrum Resources.

Derr, J. (2000) Statistical Consulting: An Effective Guide to Effective Communication. Pacific Grove, CA: Brook/Cole.

Dhar, V. (2013). *"Data science and prediction". Communications of the ACM. 56(12): 64. doi:10.1145/2500499.*

Diggle, P., Ribeiro Jr., P. J. (2007). *Model-based Geostatistics*. New York, NY: Springer Science + Business Media LLC.

Donoho, D. (2000). *Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*. Department of Statistics, Stanford University.

Doran, G. T. (1981). "There's a S.M.A.R.T. Way to Write Management's Goals and Objectives", *Management Review*, Vol. 70, Issue 11, pp. 35-36.

Eiksund, S. 2009. A geographical perspective on driving attitudes and behavior among young adults in urban and rural Norway. *Safety Science* 47: 529–536.

European Citizen Science Association. (2015). *Ten Principles of Citizen Science.* [PDF Document]. Retrieved from: https://ecsa.citizen-science.net/documents.

Farnsworth, J. (2013). *Hot Spot Identification and Analysis Methodology*. Brigham Young University: All Theses and Dissertations Paper 3878.

Fitzmaurice, G., Laird, N., & Ware, J. (2004). *Applied Longitudinal Analysis.* Hoboken, New Jersey: John Wiley & Sons Inc.

Foody, G.M and A. Mathur. 2004. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6), pp. 1335-1343.

Flury, B. (1997). *A First Course in Multivariate Statistics*. New York, NY: Springer-Verlag New York Inc.

Freire, P., (2008). *Education for Critical Consciousness* (pp. 39-40). New York, NY: Sheed & Ward, Limited. (Original work published 1974).

Gelfad, A., Diggle, P., Fuentes, M., & Guttorp, P. (Eds.). (2010). *Handbook of Spatial Statistics*. Boca Raton, FL: Taylor and Francis Group LLC.

Getis, A., & Ord, J.K. (1992). The analysis of spatial association by use of distance statistics. *Geographic Analysis*, 24, 189-206.

Gidudu A., Hulley G., & Marwala, T. (n.d.). *Classification of Images Using Support Vector Machines*. [PDF Document]. Retrieved from: https://arxiv.org/pdf/0709.3967.pdf.

Healy, J. F. (2009). *Race, Ethnicity, Gender, and Class: The Sociology of Group Conflict of Change* (5th ed.). Los Angeles, CA: Pine Forge Press.

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. Luxembourg: Office for Official Publications of the European Communities.

Hilbe, J.M., (2009). *Logistic Regression Models*. Boca Raton, FL: Taylor and Francis Group, LLC.

Hofmann , M. (2006) Support Vector Machines-Kernels and the Kernel Trick. Retrieved from: http://www.cogsys.wiai.unibamberg.de/teaching/ss06/hs_svm/slides/SVM_ Seminarbericht_Hofmann.pdf.

Horan, T, Botts, N., Burkhard, R. (2010). A Multidimensional View of Personal Health Systems for Underserved Population. *Journal of Medical Internet Research*. 12(3), e32. doi: 10.2196/jmir.1355.

Horan, T. and Hilton, B. (2017). *State of Minnesota tribal areas crash analysis*. Using GIS to Improve Tribal Traffic Safety. Claremont Graduate University, Road Safety Institute.

Huang, C., L.S. Davis, and J.R.G. Townshend. 2002. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4), pp. 725-749.

Izenman, A. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning.* New York, NY: Springer.

Jimenez, J., (2010). *Social Policy and Social Change: Toward the Creation of Social and Economic Justice*. Thousand Oaks, CA: SAGE Publishing, Incorporated.

Jorgensen, M. (Eds.). (2007). *Rebuilding Native Nations: Strategies for Governance and Development*. Tucson: The University of Arizona Press.

Kalt, J., Cornell, S, Henson, E., Grant II, K., Taylor, J., Jorgensen, M., ... Nelson, H. (2008) *The State of Native Nations: Conditions under US Policies of Self-Determination*. New York, NY: Oxford University Press.

Kessler-Mata, K. (2014). *A Constitutive Theory of Tribal Sovereignty: The Possibilities of Federalism.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3638603)

Keuhl, R. (2000). Design of Experiments: *Statistical Principles of Research Design and Analysis* (2nd ed.).Pacific Grove, CA: Brook/Cole.

Lavenda, R., Schultz, E. (2003). *Core Concepts in Cultural Anthropology* (2nd ed.). Boston, MA: McGraw Hill Companies, Incorporated.

Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data, *Journal of the American Statistical Association*, 99, 67–81.

Leek, J. (2013). *"The key word in "Data Science" is not Data, it is Science". Simply Statistics.*

Lemont, E. D. (Eds.). (2006*). American Indian Constitutional Reform and the Rebuilding of Native Nations.* Austin: University of Texas Press.

Levine, N., Kim, K.E., Nitz, L.H., 1995a. Spatial analysis of Honolulu motor vehicle crashes: I. Spatial Patterns. *Accident Analysis and Prevention* 27 (5), 663–674.

Liebler, R. A. (2003). *Basic Algebra with Algorithms and Applications*. Boca Raton, FL: Chapman & Hall/CRC.

Loo, B., & Anderson, T. (2016). *Spatial Analysis Methods of Road Traffic Collisions*. Boca Raton, FL: Taylor & Francis Group.

Lu, Y., & Chen, X. (2007). On the false alarm of planar K-function when analyzing urban crime distributed long streets. *Social Science Research,* 36, 611-632.

MacDonald, H., & Peters, A. (2011). *Urban Policy and the Census*. Redlands, CA: ESRI Press.

McKinley Jones Brayboy, B., Fann, A., Castagno, A., Solyom, J. (2012). *Postsecondary education for American Indians and Alaska Natives*. Hoboken, NJ: Wiley Periodicals, Inc.

McGuigan, D. R. D. (1981) "The Use of Relationships between Road Accidents and Traffic Flow in 'Black-Spot' Identification," Traffic Engineering and Control, Vol. 22, No. 8-9, 448-453.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis. Volume 1: Geographic Patterns & Relationships*. Redlands, CA: ESRI Press.

Mitchell, A. (1999). *The ESRI Guide to GIS Analysis. Volume2: Spatial Measurements & Statistics*. Redlands, CA: ESRI Press.

Mueller, B. A., F. P. Rivara, and A. B. Bergman. 1988. Urban–rural location and the risk of dying in a pedestrian-vehicle collision. *Journal of Trauma & Acute Care Surgery* 28 (1): 91–94.

National Highway Traffic Safety Administration. 2008. Traffic safety facts, 2006 data: Rural/urban comparison. NHTSA, U.S. Department of Transportation, Washington, DC.

Native Nations Institute (n.d.). *What is Nation Building?* Retrieved February 15, 2018 from http://nni.arizona.edu/ programs-projects/what-native-nation-building.

Necula, E. (2015). Analyzing traffic patterns on street segments based on GPS data using R. Transportation Research Procedia. 10, 276-285. doi:10.1016/j.trpro.2015.09.077.

OECD. 2002. Safety on roads: What's the vision? Paris, France: Organization for Economic Co-operation and Development.

Ojibwa (2013). *The Carlisle Boarding School*. Retrieved from: http://nativeamerican netroots.net/diary/1497.

Okabe, A., and Sugihara, K. (2012). *Spatial Analysis along Networks*. West Sussex, U.K.: John Wiley & Sons, Ltd.

O'Neil, C & Schutt, R. (2014). *Doing Data Science*. Sebastopol, CA: O'Reilly Media, Inc.

O'Sullivan, D., & Unwin, D. J. (2010). *Geographic Information Analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Peters, A., MacDonald, H. (2004).*Unlocking the Census*. Redlands, CA:ESRI Press.

Plant, R. (2012). *Spatial Data Analysis in Ecology and Agriculture using R*. Boca Raton, FL: CRC Press.

Rainie, S., Rodriguez-LoneBear, D., & Martinez, A. (2017). *Policy Brief: Data Governance for Native Nation Building.* The University of Arizona Native Nations Institute. Retrieved from: http://usindigenousdata.arizona.edu/spotlight/policy-brief-data-governance-native-nation-rebuilding-0.

Rothaermel, F. T., (2017). *Strategic Management* (3rd ed.). New York, NY: McGraw Hill Education.

Ropollo, K. & Crow, C. L. (2007). Native American Education vs. Indian Learning. Still Battling Pratt After All These Years. *Studies in American Indian Literatures*. 19, pp. 3-8.

Serrano, F., Sans, F., Silva, C, & Keislinger, B., (n.d.). *Citizen Science White Paper.* [PDF Document]. Retrieved from: http://www.socientize.eu/sites/default/files/white-paper_0.pdf.

Shanley, L. A. (2015). *Spatial Data Sovereignty and Privacy in Indian Country: A Policy Analysis.* (Doctoral Dissertation). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. 3707906)

Sharda, R., Delen, D., & Turban, E. (2018). Business Intelligence, Analytics, and Data Science: A Managerial Perspective (4th ed.). Boston, MA: Pearson.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data. Analysis, Vol. 26 of Monographs on Statistics and Applied Probability*, Chapman & Hall. London.

South Dakota State University Office of Engineering Research. (2016). Engineering Research Review 2016. South Dakota State University: University Marketing and Communication, (pp. 10-13). Retrieved from: https://www.sdstate.edu/sites/default/files/2017-05/engineering_review_2016_2.pdf

Steinwart, I. & Christmann, A. (2008). *Support Vector Machines*. New York, NY: Springer Science + Business Media LLC.

Tawansi, K. (2012). Smart devices – the fastest technology adoption in history. Retrieved from: https://www.solentive.com.au/smart-devices-the-fastest-technology-adoption-in-history/.

The Harvard Project on American Indian Economic Development. (2008). *The state of Native Nations. Conditions under U.S. policies of self-determination*. New York, NY: Oxford University Press Inc.

Ting, K. (2011). Confusion Matrix. C. Sammut & G. Webb (Eds.). Encyclopedia of Machine Learning. New York, NY: Springer Science + Business Media LLC.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234–240.

US Indigenous Data Sovereignty Network. (n.d.). *About us*. Retrieved January 21,2018 from http://usindigenousdata.arizona.edu/about-us-0.

U.S. Department of Justice. (2005). Mapping Crime: Understanding Hot Spots (NCJ 209393). Washington, DC: U.S. Government Printing Office.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. New York, NY: Cambridge University Press.

Wu, T-F., Lin, C-J, & Weng, R.C. (2004). Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5, pp. 975-1005.

Xie, Z., & Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. Computers, *Environment and Urban Systems*, 32, 396-406.

Yamada, I., & Thill, J. (2004). Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography*, 12, 149-158.

Ying, V. (2013) *Methods for Spatial Analysis on a Network*. (Master's Thesis) Retrieved from: http://escholarship.org/uc/item/1t01p61g.

Zhang, L. and M. Milanova. (2013). An effective multi-feature fusion object-based classification method on ArcGIS platform using very high-resolution remote sensing images. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3(11), pp. 10-23.