SCIENCE IN THE DIGITAL AGE: OVERCOMING UNCERTAINTY AND THE ADOPTION OF VOLUNTEERED GEOGRAPHIC INFORMATION FOR SCIENCE

By

Shaun Arthur Langley

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Geography—Doctor of Philosophy

2014

ABSTRACT

SCIENCE IN THE DIGITAL AGE: OVERCOMING UNCERTAINTY AND THE ADOPTION OF VOLUNTEERED INFORMATION FOR SCIENCE

By

Shaun Arthur Langley

With the advent of Web 2.0, the public is becoming increasingly interested in spatial data exploration. The potential for Volunteered Geographic Information (VGI) to be adopted for Science through collaborations between researchers and non-scientists is of special interest to me. In particular, mobile devices and wireless communication permit the public to be more involved in research to a greater degree. Furthermore, the accuracy of these devices is rapidly improving, allowing me to address questions of uncertainty and error in data collections. Cooperation between researchers and the public integrates themes common to VGI and PGIS (Participatory Geographic Information) to bring about a new paradigm in GIScience. This dissertation discusses VGI in the context of a new paradigm, eScience, and the broader framework of Neogeography. I discuss current issues with data quality and uncertainty regarding VGI and detail one approach to quality credibility of the data. Finally, the dissertation outlines the framework for utilizing VGI in the context of case study in disease ecology for the purpose of surveillance of tsetse flies, the primary vector of African Trypanosomiasis. My system allows for two-way communication between researchers and the public for data collection, analysis, and the ultimate dissemination of results. Enhancing the role of the public to participate in these types of projects can improve both the efficacy of disease surveillance as well as stimulating greater interest in science.

This work is dedicated to Sam, my little bundle of joy and the motivation to push myself, and to Courtney whom I love and admire greatly. You inspire me to be a better person, and your never-ending support for me gave me the strength to push on, even when it seemed difficult.

ACKNOWLEDGEMENTS

I would like to acknowledge all of those individuals who helped me over the course of my tenure as a PhD student. To the staff in Geography, particularly Sharon Ruggles, I am grateful for all you have done to help me. I would have been perpetually lost without you.

I am grateful for the help I received while in Kenya. In particular, I want to acknowledge Dr. Maitima for his guidance in pulling everything together while in Nguruman. I also want to acknowledge the assistance of Joel Meiyponi in the collection of my flytrap data, and for assisting with the translation of interviews that both he and I conducted in Nguruman. I'm also very grateful for the healthcare workers that tended to me following my accident. I also want to acknowledge the Elders in Ol'Kirmatian for their willingness to allow me to operate in the area.

I am immensely grateful for my committee members – Dr. Ashton Shortridge, Dr. Sue Grady, Dr. Edward Walker, and my advisor Dr. Joseph Messina. Your support, guidance, feedback, and collaborations were instrumental in getting me to this point.

I am grateful for the support of all my friends, in particular Dr. Hamm for pushing me to finish this dissertation. The support of friends is critical to making it through Graduate School and you all provided that support without precondition.

I am so thankful for the support of my Parents and family for being patient with me as I am concluding my 14th year in college. Only sometimes did you comment that it was taking an awful long time to finish. I hope that you feel it has been worth it!

Finally, and most importantly, I want to thank Courtney for your unwavering love and faith in me. You have taught me more than you know; and to top it off, you've given me the most wonderful gift of all, Sam. It is now my mission in life, over the next 60 years or so, to try and repay you all the kindness and patience you've shown me.

The work in this dissertation was supported in part through a grant from National Institutes of Health, Office of the Director, Roadmap Initiative, NIGMS Award No. RGM084704A, from three Graduate Office Fellowship awards, and from the US Government's Student Loan Program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Conceptual Framework

What does it mean to be a scientist? What does it mean to be a Geographer? These questions have existed and been debated for as long as the discipline has existed. Geography is ripe with tradition, having evolved many times to suit the consciousness of the day (Livingstone, 1992). Early Geographers like Francis Bacon and John Locke were Empiricists, advocating the scientific method as the means to explain the world. Plato was a proponent of realism, the philosophy that states that truths are universal and have an objective or absolute existence. René Descartes, arguably a geographer in name only, was a proponent of subjectivism, perceiving knowledge as truly subjective. Under Descartes' philosophy, there were no absolute, external, or objective truths; rather, reality existed uniquely for each individual in their own way. Thomas Kuhn transformed the discipline with a relativism paradigm, arguing that knowledge, truth, and morality exist in relation to culture, society, or in a historical context; however, truths were not absolute and thus conflicting theories weren't necessarily the result of being right or wrong, rather having arrived at a conclusion from different perspectives. Finally Karl Popper proposed the critical rationalist philosophy, arguing that the falsifiability of science was paramount to the generation of knowledge, and that all scientific theories should be rationally criticized and subjected to tests of falsifiability.

As researchers began to acknowledge the importance of context in describing the Geography of a place, citizens were engaged directly by researchers seeking to capture the "local

knowledge" that only a citizen can convey (Elwood, 2006a; Robbins, 2003). Participatory science emerged in the 1990s as a context within which we can incorporate quantitative, qualitative, and cartographic forms of data into a GIS environment (Elwood, 2006b; Warren, 1991). The term was first used by Warren (1991) and Pickles (1995) to define the method of engaging and empowering communities to address political, social, and environmental questions that directly concern and impact them (Robbins, 2003). Warren (1991) notes, "Local knowledge is increasingly acknowledged to be scientific in the sense that it evolves from experimental techniques of trial-and-error conditions not entirely distinct from scientific practices". He argues that this knowledge should not be too quickly dismissed as it can inform us of the social, political, and economic structure of communities. Furthermore, the use of cartographic and GIS tools can empower local communities to use their own knowledge and understanding of their environment to better manage their resources, even resolve standing disputes. The latter applications are commonly associated with Participatory GIS (PGIS) in the work of Sarah Elwood (e.g. 2006a, 2006b) and Michael McCall (2005). Turner and Hiernaux (2002) demonstrate the ability of cartographic methods, informed by local knowledge, to develop more effective local management strategies than can be achieved through a spatial modeling experiment alone. Finally, Robbins (2003) extends the application of PGIS by Turner and Hiernaux (2002) indirectly in his notion of "indigenous GIS". It is the goal of PGIS to generate knowledge in collaboration with communities to enable them to independently critique spatial data for their own purposes and benefits (Elwood, 2006a). To quote Elwood (2006b),

> *"the existing literature demonstrates that the nature of the participatory knowledge production in PPGIS (public participation GIS) is the result of complex interactions of technological and social factors, but there have been*

relatively few attempts to detail how these intersecting factors play out in and are shaped by the daily choices and negotiations of PPGIS projects."

Universal throughout the evolution of Geography and perceptions of science was the distinction between citizens and scientists. There are those who can do science – experts in their field, educated and trained, often residing in academic institutions, and producers of knowledge; and there are those who cannot – citizens, laypeople, and consumers of knowledge. It is difficult to identify a specific shift in consciousness; rather there has been a subtle recognition of the role citizens can, and do, play in science. Beginning in the early 1900s, the Audubon Society began enlisting the help of citizen scientists and volunteers in their annual Christmas Bird Watch. Still today, these amateur ornithologists participate in the annual bird count; to-date they have amassed more than 31 million records since its inception (Haklay, 2012; Silvertown, 2009). The volunteers who participate do not do so at their own leisure though. They are specifically trained in data collection methods and their data are subjected to at least a minimal amount of quality control.

Haklay (2011) identifies four types of citizen science, which he distinguishes based on an individual's training and level of participation. The first type is "crowdsourcing". Howe (2006) defines it as engagement in the collection of data with a minimum level of effort. At this stage, citizens are perceived as nothing more than data collection tools, or sensors, off which we can glean information about their environment.

The second level is termed "Distributed intelligence". Here, citizens are tasked with slightly more complex tasks. They are usually given some amount of training in data collection methods, and may be asked to perform simple interpretations of their environment. Citizens thus

become basic interpreters of the data they collect, rather than simply sensors of information (Haklay, 2011).

The third level "participatory science" is marked by significant participation on the part of citizens who collaborate with scientists directly at each stage in the project from inception to execution and analysis (Haklay, 2011; Irwin, 1995). This mode of citizen science has dominated the Geography literature and has garnered a great deal of attention for its ability to grant empowerment to disenfranchised groups.

Finally, the fourth level of citizen science is "extreme". It is at this stage that science becomes truly collaborative as the citizen is fully integrated into the scientific process (Haklay, 2011). This can potentially allow for citizens to be scientists without the presence of the professional or expert.

The last decade has seen dramatic technological and communications advances. With the advent of Web 2.0 and the proliferation of new technologies and modes of communication, exposure to geographical knowledge is increasingly pervasive in our society (O'Reilly, 2006, 2007) (Goodchild, 2007a, 2007b). Utilizing web-based applications, citizens have developed an interest in becoming active participants in describing their environment by volunteering geographic information.

Whereas citizen science defines the actions of the individual in participating in the scientific process, Neogeography reframes the context in which it occurs. Neogeography defines the erosion of traditional roles and the blurring of the distinction between citizens and scientists (Goodchild, 2009). Central to the idea is a reframing of the space in which knowledge generation occurs. The literature points to the emergence of the Geoweb, the online environment where

geographical information is mashed with abstract data (Haklay, 2011). However, the activities of individuals in the Geoweb have arisen organically, lacking the traditional oversight and training in methodology given by scientists. Today there are literally thousands of web applications in which individuals can actively volunteer information, often geographical in nature (Goodchild, 2009).

Volunteered geographic information (VGI) is a term that was coined by Michael Goodchild to describe the phenomenon of citizens contributing, volunteering, and consuming geographical information outside the purview of academia (Goodchild, 2007a). Perhaps the most commonly cited examples of this phenomenon are Wikimapia[1] and OpenStreetMap[2]; in these examples, the citizens contribute or volunteer information and descriptions about the built environment around them (Goodchild, 2007a; Haklay & Weber, 2008). However, the past decade has seen a tremendous increase in the utilization of VGI (Elwood et al., 2011).

In this dissertation, I use the term Volunteered GIS (VGIS) to refer to a conceptualization of GIS that emphasizes the interest of non-scientists to be active participants in the generation of spatial knowledge (Flanagin & Metzger, 2008; Goodchild, 2007c, 2010b). A tradition previously reserved for "professional" scientists, VGIS represents a shift from viewing science as having a single authority (the scientist) to a model where authority is relative and expressed contextually. Information abundance, repetition, and the information collective conveys credibility to itself (Craglia et al., 2007). However, the issue of data quality has been debated extensively in the literature, and represents a critical hurdle to the adoption of VGIS for Science.

[1] http://wikimapia.org
[2] http://www.openstreetmap.org

Data Quality

The quality of spatial data is the foundation for its utility in the scientific process. Quantifying error and related uncertainty is the most critical evaluation of data that can be made. The credibility of any information or data is directly related to an assessment of data quality. The issue of data quality has a long tradition in the literature, particularly as it pertains to spatial data. One of the most comprehensive lists of data quality metrics was outlined by van Oort (2005). In his dissertation, he identifies 11 components of data quality that pertain to spatial data (Table 1.1). The assessment of data quality has objective and subjective components, which collectively convey the reliability we have as to the truthfulness of the information. Although there are a multitude of metrics, arguably the most important, and objective, components of spatial data quality are: positional accuracy, attribute accuracy, and completeness (Haklay, 2010). Within the context of traditional spatial data infrastructures (SDIs) these metrics facilitate an assessment of the variability of the data with respect to the population being measured and allow for an assessment of uncertainty. Assessing data quality using objective metrics allows for us to make statements with regard to confidence in our results.

In the age of Neogeography, data are routinely volunteered without the corresponding information needed to make an objective assessment of data quality (e.g. positional error, spatial resolution, etc.). A subjective assessment of data quality, can serve as a means to determine a dataset's credibility and as such make a determination of its fitness-for-use (Grira et al., 2009).

As the determination of data quality is itself a subjective determination on the part of the consumer, the communication of the metrics is critical. To this extent, the communication of quality metrics is critical. Within metadata of a dataset, the user is provided with the information

needed to make a quality assessment. However, too often the communication to users (and sometimes scientists) is done without regard for whether or not the receiver is capable of understanding or perceiving it as is intended by the communicator (Grira et al., 2009). Reporting data quality through metadata is therefore an ineffective means of communication and can result in potential misuse of data (Comber et al., 2006; Grira et al., 2009). Subjective assessments are a matter of perception. With regard to citizens and non-experts, the perception of quality is less a matter of quality metrics, but rather relates to their perception of the communicator themselves (Metzger et al., 2003).

The emergence of Citizen Science has generated volumes of data that are distinct from traditional datasets, primarily because of the manner in which they were collected. There has been a great deal of debate as to the best way to assess the credibility of volunteered data (Elwood, 2006b; Flanagin & Metzger, 2008; Metzger et al., 2003). Many illustrations of this assessment rely on a comparison of VGI against datasets of known quality (Koukoletsos et al., 2012). But perhaps the most promising approach is to correlate the credibility of the communicator with the quality of the data (Corbett, 2012).

The uncertainty concerning data quality for volunteered geographic information (VGI) and crowdsourced data has resulted in poor adoption of citizen science initiatives for science, whether experimental or analytical (Craglia et al., 2007; Elwood et al., 2013; Goodchild, 2009). A 2009 survey conducted by Elwood et al. (2011) looked at a range of projects identified to be utilizing VGI in some form. Of the 99 studies they evaluated, only 3% were sponsored by an academic institution, 7% were sponsored by a Government program, and 7% were affiliated with an NGO. The majority of VGI initiatives are associated with for-profit institutions (63%). The

overwhelming element of for-profit companies utilizing VGI runs counter to the notion of VGI as an expression of citizen science (Goodchild, 2007a). However, there is significant academic interest in adopting VGI initiatives to enhance collaborations between researchers and the communities (in which they operate), and to develop framework data (Craglia et al., 2007; Elwood et al., 2011; Haklay, 2012). Framework data constitute the core of a spatial data infrastructure (SDI) as it aims to represent core phenomena; it contains data on geodetic control, orthoimagery, elevation, transportation, hydrography, governmental units, and cadaster (Craglia et al., 2007; Elwood et al., 2011).

For VGI and crowdsourced data to be valuable for science in an academic context, we must address the lingering questions of credibility (of the data) and uncertainty regarding data quality (Craglia et al., 2007; Flanagin & Metzger, 2008; Haklay et al., 2010).

Spatial data management

The development of new communications and technology in the early 1990s created a need for new data management strategies, and a fundamentally different standard data model — a set of abstractions that define object, relation, and attribute types and their use (Devogele et al., 1998; R. Groot & McLaughlin, 2000). Such data models form the basis for a spatial data infrastructure (SDI) for the purpose of "facilitating coordination, production, access, and use of spatial data" (Budhathoki et al., 2008; R. Groot & McLaughlin, 2000; I Masser, 2005). In 1994, President Clinton signed an executive order establishing a national SDI (Coordinating geographic data acquisition and access: The national spatial data infrastructure 1994). For more than a decade, the standard SDI was sufficient for the data types available. However, with the emergence of the eScience paradigm (Gray & Szalay, 2006) and Web 2.0 (O'Reilly, 2006, 2007), the types of spatial

information no longer fit neatly within in the constructs of the traditional data model (Rouse et al., 2007). Thus, there were calls to modernize the standard SDI.

Next generation SDIs have implemented web services and broadened the types of data supported; however the majority still perceive the user role as passive (Budhathoki et al., 2008), thus rendering them ineffective for citizen science initiatives. This is a large part of the reason that current SDIs are so vastly underutilized, resulting in widespread inconsistencies and a lack of interoperability among systems (Budhathoki et al., 2008; Ian Masser, 2005).

It is not sufficient to simply define an alternate data model for use with VGI; such an approach reduces the status of the information to a second tier, perpetuating the inequality of roles and utility of the data. Users of SDIs tend to be expert organizations (Budhathoki et al., 2008), while Neogeographers operate in a patchwork system of inconsistent authorities. Bringing forward a unified data model will not only elevate the authority of VGI, but will improve overall data quality by standardizing the communication of quality metrics.

Concluding Thoughts

The acquisition of data for science is arguably the single largest hurdle researchers face. Information must be obtained in a controlled manner so as to assure the interpretation under the prescribed theoretical model. Any uncertainty surrounding the quality or integrity of the information obtained can have a crippling effect on the ability of the researcher assert significance for any observed patterns (in a traditional statistical framework). Crowdsourced data or VGI can provide significant contributions to science if questions of quality and integrity can be addressed. Some have suggested that these uncertainties render such sources of information useless; however I think we are approaching a new era in science that will dramatically change

our perception of such sources and bring about a new paradigm. I think the technological advances of the past century and the advent of Web 2.0, combined with the proliferation of mobile technologies and the increasing incorporation of smart sensors into these devices will provide new opportunities for ordinary citizens to be active participants in science through the collection, dissemination, and even analysis of information. Such opportunities are not only beneficial for researchers to access new kinds of information, but also present a unique opportunity for citizens to be more involved in science – reversing the disturbing trend of scientific illiteracy we are witnessing in society today. In this regard, addressing the questions posed in this dissertation become critical for ushering in a new era of scientific discovery and opportunity for future generations.

Specific Aims

Objective 1: Address three recurring problems with spatial data management: scalability, reliability, and security by:

1. Communicating a conceptual model for a comprehensive open-source computing environment that promotes the efficient organization, storage and retrieval of disparate data.
2. Extending the discussion of spatial databases by presenting a model framework for a spatial DBMS that rigorously and consistently manages both spatial and nonspatial data.

Objective 2: Demonstrate the utility of VGI by:

1. Describing a prototype for the utilization of VGI to enhance disease surveillance programs.

2. Articulating an approach for integrating VGI into a traditional species distribution model.

Objective 3: Address lingering concerns of credibility and data quality in VGI by:

1. Illustrating how to dynamically assess the reliability of reporters of VGI.
2. Assessing the impact of incorporating VGI of varying quality into a traditional species distribution model.

Outline of the dissertation

Chapter 2 of the dissertation addresses specific aim 1 — issues of spatial data management that are exacerbated by the volume of data generated through crowdsourcing and VGI. In this paper I outline the framework for a spatial database management system (sDBMS) that facilitates the efficient storage, query, and retrieval of spatially explicit data. The paper was published in 2011 in the Journal of Map & Geography Libraries and appears in this dissertation in its published form.

Chapter 3 of the dissertation answers specific aim 2 in illustrating the utility of VGI in science. I utilize a case study in disease ecology to demonstrate how VGI can be incorporated into a traditional species distribution model to enhance the output and relevance for managing disease vectors more effectively. This chapter was published in 2013 in the International Journal of Applied Geospatial Research, and appears in its published form.

Chapter 4 addresses specific aim 3 in which I directly discuss issues of credibility and quality for VGI. I outline an approach in which the reliability of volunteers (or reporters) of VGI can be computed, and how this measure can serve as a surrogate metric of data quality.

Table 1.1: Components of data quality for spatial data

Metric	Definition	Expression
Attribute accuracy	"an assessment of the accuracy of the identification of entities and assignment of attribute values in the data set."	*Khat or Cohen's Kappa*
Positional accuracy	"… to the degree to which the digital representation of a real-world entity agrees with its true position on the earth's surface"	$RMSE = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} \delta x_i^2 + \sum_{i=1}^{n} \delta y_i^2\right)}$
Temporal accuracy	"… the agreement between encoded and actual temporal coordinates."	Textual
Semantic accuracy		$\frac{\sum_{v \in V} w(v) \cdot d\left(\gamma(v), \gamma'(v)\right)}{\sum_{v \in V} w(v)}$
Completeness	"measurable error of omission observed between the database and the specification."	**Omission:** % of data missing relative to the specification **Commission:** % of data present that is not in the current specification of dataset or extract **Coverage ratio:** occurrences of one variable per unit of another
Logical consistency	"refers to the absence of apparent contradictions in the database"	textual

Continued on next page

Table 1.1 (cont'd)

Metric	Definition	Expression
Spatial resolution	"... the fineness of detail that can be observed."	Integer
Temporal resolution	"… the minimum duration of an event that is discernible."	Integer
Thematic resolution	**Categorical data:** "resolution is defined in terms of the fitness of category definitions." **Quantitative data:** "resolution is determined by the precision of the measurement device."	Integer
Fitness-for-use	"ability to use the dataset for a particular purpose or situation."	textual

Continued on next page

Table 1.1 (cont'd)

Metric	Definition	Expression
Lineage	"[Lineage] refers to source materials, methods of derivation and transformations applied to the database. This includes temporal information (data that the information refers to on the ground), and is intended to be precise enough to identify the sources of individual objects (i.e. if the database was derived from different sources, lineage information is to be assigned as an additional action viewing objects in a spatial overlay)."	textual
Meta-quality	"a measurement of the collective quality of the data."	
Variability	"… the difference between expected measures and actual values."	$\epsilon = y - \mu$

CHAPTER 2

EMBRACING THE OPEN-SOURCE MOVEMENT FOR THE MANAGEMENT OF SPATIAL DATA: A CASE STUDY OF AFRICAN TRYPANOSOMIASIS[3]

Abstract

The past decade has seen an explosion in the availability of spatial data not only for researchers, but the public alike. As the quantity of data increases, the ability to effectively navigate and understand the data becomes more challenging. Here we detail a conceptual model for a spatially explicit database management system that addresses the issues raised with the growing data management problem. We demonstrate utility with a case study in disease ecology: to develop a multi-scale predictive model of African Trypanosomiasis in Kenya. International collaborations and varying technical expertise necessitate a modular open-source software solution. Finally, we address three recurring problems with data management: scalability, reliability, and security.

Introduction

The trans-disciplinary nature of modern research in disease ecology often requires and generates vast quantities of data varying thematically and in structure. The data management challenge is considerable and often cost prohibitively so. Data arise from a multitude of sources and occur in

[3] This chapter appears in its published form. Minor adjustments have been made to the text to address invalid references and/or grammatical errors. I address more substantially changes to the code in Appendix C.

spatially explicit or aspatial forms with concomitant structures. Rarely are these data accompanied by the ontologically coherent metadata necessary to facilitate cooperation and collaboration. Recent discourse in the studies of infectious disease ecology have suggested a need to emphasize the role of space and land cover change dynamics in describing the interactions of diseases with environmental processes (Ostfeld et al., 2008). Therefore, effective engagement in disease ecology research requires the ability to access, correctly interpret and integrate these highly diverse data (Longley et al., 2005; Shekhar & Chawla, 2003; Watson et al., 2004).

In the early 1990's there was a great deal of research concerning the expansion of traditional database management systems (DBMS) to incorporate functionality for handling spatially explicit data types (Shekhar & Chawla, 2003; Michael Stonebraker & Moore, 1995). Traditional DBMSs are self-describing in that the definitions of data are stored in a catalog, along with the raw data, without the need to store separate descriptive files. This is an extremely efficient means for managing and accessing large quantities of data. Databases improve our abilities to interact with data through the construction and querying of indices, which serve as a data roadmap stored on the physical media. Early database frameworks rarely interacted directly with spatial data (Egenhofer, 1994; Michael Stonebraker & Kemnitz, 1991). Yet with technological advances, particularly the development of satellite platforms for remote sensing, great quantities of data were being generated, necessitating the development of DBMS capabilities specifically to facilitate handling of spatial data.

SQL (standard query language) was formally adopted in 1986 by the American National Standards Institute, as SQL-86, and is the most widely used database querying language (Lans,

2007). The SQL language framework provided a logical structure for querying information stored in DBMSs. The first versions of SQL however, could not explicitly handle spatial data structures. Several extensions of SQL were proposed including GEOQL, SAND, GEO-Kernel, and PSQL; however the extension that proved most influential, and eventually adopted was Egenhofer's Spatial SQL (Adam & Gangopadhyay, 1997; Egenhofer, 1994). Spatial SQL extended the domain to include spatial operators and attributes. Egenhofer (1994) further defined the Graphical Presentation Language (GPL), a set of tools in which the results of a spatial query could be manipulated. The Open GIS consortium (OGIS) promoted spatial functionality by recommending a set of critical reforms in SQL, mainly the adoption of Egenhofer's spatial abstraction model, which introduced GEOMETRY as a base-class for spatial objects (Egenhofer, 1994; OGIS, 1999; Shekhar & Chawla, 2003). These recommendations were fully adopted in 1999 with the release of SQL3 (Lans, 2007; OGIS, 1999).

Data sharing is a critical consideration for our research group as we maintain collaborations with institutions and researchers throughout much of the United States and East Africa; yet the biggest problem we face as a group is the ability to share data and analysis. We require a new medium that facilitates this flow of information without the need to physically carry the data between institutions. Internet-based GIS and data servers are one solution we employ to allow for simultaneous interaction and analysis of the data. Internet-based GIS emerged with the release of two projects, GeoChange (Drew & Ying, 1996) and the Alexandria Digital Library (Smith & Frew, 1995). GeoChange in particular set the standard in terms of functionality and usability with an interface supporting batch processing, import/export, and automatic metadata generation (Adam & Gangopadhyay, 1997). The Alexandria Data Library,

though spare in relation to GeoChange, became widely adopted by the University of California system for in-house sharing of proprietary spatial data. Today, despite the improvements in functionality, most spatial databases still lack the ability to efficiently handle the raster data structure. Two ongoing projects promise to bring this functionality into spatial DBMSs (SDBMSs) in 2010. Of particular interest to us is the WKTRaster project, a community supported open-source project by the Open Source Geospatial Foundation (OSGEO). WKTRaster extends the functionality of the PostGIS library to implement the RASTER type class in the same way GEOMETRY was done in order to support spatially explicit handling of vector data (Michael Stonebraker & Moore, 1995). WKTRaster, when fully implemented, will facilitate consistency in data handling of all spatial and non-spatial data types, similar to the manner in which spatial data are managed with the georaster loader included in the Oracle Spatial 11g package.

While advances in scientific understanding and new technologies have helped to curb the spread of infectious disease in the developed world, the same cannot be said about the developing world. As the incidence of infectious diseases decreased in the developed world, fewer resources and attention were devoted to combating those diseases in the developing world, contributing to an increase in the prevalence of many illnesses there (Cohen, 2000). In order to develop better strategies for combating disease, we need to enhance our understanding of the underlying ecological conditions that contribute to the emergence of diseases and deliver solutions practicable in developing world contexts (Ostfeld et al., 2008). Disease ecology is generally not considered a discipline in itself but rather seeks to understand the relationship between disease epidemiology and the landscape (climate, physical, and human) (Johnson & Thieltges, 2010; Keesing et al., 2006; Sutherst, 2004; Tatem et al., 2006). The trans-disciplinary

nature of the field creates a unique set of problems, many of which pertain to the use of data while maintaining the rigorous standards mandated by Institutional Review Boards, HIPAA, and international research and privacy standards. Very little has been published that directly explores these types of management problems within the disease ecology literature. Routinely, issues of scalability, reliability, and security emerge that hinder the effective dissemination of federally funded data and models. Storage of large quantities of data must at a minimum facilitate the range of applications necessitated by the questions posed in disease ecology. Data ***scalability*** speaks to the ability of researchers to address questions at multiple scales of spatial or temporal resolution, depending upon the question being asked; the storage of such data must facilitate the rapid, concurrent access and integration of the data across varying resolutions (Shekhar & Chawla, 2003). ***Reliability*** requires mechanisms to ensure that data mismatches or inappropriate analytical methods are identified or prevented (Devillers & Jeansoulin, 2006; Shi et al., 2002). Furthermore, data reliability, particularly with concurrent usage and modification of the data, necessitates mechanisms for ensuring the integrity of the underlying data over time (Olson, 2003; Shi et al., 2002). Finally, ***security*** issues arise when interacting with individually identifiable human data or sensitive community data stored or generated within the DBMS (Olson, 2003). While institutional guidelines and privacy laws may restrict access of the data to pre-approved users, the limitations should not preclude non-privileged users from asking broader questions, which may interact with the underlying data when aggregated to remove identifiable data or other information that may be restricted by institutional guidelines or laws (e.g. ethnic identity at low densities in census block group data). Finally, privacy restrictions should be scalable, changing dynamically with the user and scale of resolution requested.

In collaboration with the International Livestock Research Institute in Kenya (ILRI), we have accumulated an extraordinary volume of data for Kenya. Irrespective of theme, international collaborations often present unique problems in terms of the management, sharing, and dissemination of data necessary to carry out analyses. Our framework for a data management system is a novel solution for spatial modeling in disease ecology, and the use of open-source software exclusively makes this a cost effective solution for sharing with international collaborators and organizations with limited budgets. The entire suite of data and models is designed to be packaged electronically or on a portable drive to facilitate electronic transfer or physical transportation.

A framework for data management cognizant of these issues and flexible in the use of restrictions is an ideal solution for working with data types characteristic of research in disease ecology. As part of the National Institutes of Health "Roadmap" program and with the National Institutes of Health General Medical Sciences support, we are developing a multi-scale predictive model that defines the relationship between climate change, land use and land cover change, social systems, and the distribution of tsetse flies and sleeping sickness across Kenya (Makido et al., 2007). Here we present a case study for the implementation of a generalizable disease ecology DBMS framework that provides scalability, reliability, and security to optimize interactions between users and the data.

Case Study: A Model for African Trypanosomiasis in Kenya

African Trypanosomiasis (AT), or sleeping sickness, is a major threat to human health across Africa, particularly among impoverished peoples (Brun et al., 2010; Gyapong et al., 2010). Typically considered a disease of the past, its prevalence has increased in recent years,

particularly in East Africa, due to the declining emphasis on trapping and control, climate, and anthropogenic factors (Batchelor et al., 2009; Bauer et al., 1992; WHO, 2005). Although monitoring has improved, the extent to which AT impacts East Africa is largely unknown. Recent contributions by foreign countries and aid organizations directed towards addressing AT have declined dramatically in contrast to the increased attention towards AIDS, malaria, and other diseases (Siringi, 2003; WHO, 2001). The WHO has responded by designating AT a neglected tropical disease (Brun et al., 2010; Kennedy, 2005; WHO, 2006). Not fully understanding the ecological processes that contribute to the spread of AT may result in the inefficient application of control regimes and misallocation of resources; thus retarding the efforts of the African Union to combat and control AT (Cox, 2004). As Trypanosomiasis has increased in prevalence, the impact on human and animal populations has been considerable, resulting in severe economic hardship for rural families throughout East Africa (Campbell et al., 2000; Campbell et al., 2004).

Tsetse flies, Glossinidae family, are the primary vectors for the cyclical transmission of African Trypanosomiasis. The general distribution of tsetse has been demonstrated in terms of the biophysical extent and the presence of suitable hosts (Cecchi et al., 2008; KETRI, 1996). However, the precise limits, historical and contemporary, have not been formalized experimentally (Wint, 2001). Furthermore, the current distribution belts reflect outdated data and methodologies (Joint WHO Expert Committee and FAO Expert Consultation on the African Trypanosomiases (1976: Rome), 1979; Muriuki et al., 2005; Wint, 2001). Through prior studies our research group has been able to describe the inaccuracies of these distribution limits, particularly in terms of seasonal changes (DeVisser & Messina, 2009; Moore & Messina, 2010).

Furthermore, global climate change is shifting tsetse habitats, though the degree to which this is occurring is unknown (Sutherst, 2004).

To adequately understand the mechanisms behind the increasing incidence of AT, it is important to consider a diverse range of inputs from social, physical, climatic, and even political dimensions. The range of scientific disciplines and methodologies required necessitates an extensive volume of data be created, collected, and maintained. The management of the volumes and types of data, physically and logistically, has proven to be a significant challenge and the one in which we address in this paper. Thus we present a conceptual model for a comprehensive open-source, computing environment that promotes efficient organization, storage, and retrieval of disparate data. Furthermore, we extend the discussion of spatial databases by presenting a model framework for a spatial DBMS that rigorously and consistently manages both spatial and non-spatial data.

Data Holdings and Acquisitions

We have collected all publicly available data and a significant portion of the known privately held relevant AT disease ecology spatial data for Kenya. The data fall into a classification scheme defined by topography, soils, vegetation, climate, ecological diversity, water resources, and anthropogenic factors known to control or influence the ecological processes driving tsetse distributions over time and space (see **Error! Reference source not found.** for a summary). Non-spatial data consist primarily of governmental and intra-agency reports obtained through private libraries in Kenya. The reports collected focus on policies and governmental/community control and eradication programs. Currently, these reports are neither catalogued nor indexed, limiting efficient use of any information they may contain.

Table 2.1: A summary of our data library grouped by major theme

Biophysical	Social	Geographical
Precipitation totals	Population Density	Land Use / Land Cover
Monthly Temperatures	Population Predictions	Land Use Projections
Evapotranspiration	Historical Population	Satellite imagery
Temperature Scenarios	Town Locations	Lakes
Precipitation Scenarios	District Census Data	Rivers
Historical Climate Data	Livestock	River Basins
Agro-Climatic Zones	Wildlife	Roads
Agro-Ecological Zones	Agricultural Production	National Parks
Lithology	Poverty	Railroads
Soils	Various Cadastral data	National Boundaries
Landforms	Entomological	Administrative Districts
Forest Range	Tsetse Distribution	Elevation
Wetlands	Tsetse Habitat Suitability	Topographic Maps (Countrywide)

Remotely sensed image data comprise the majority by physical file size of our data collection. We possess the majority of the known publicly available aerial imagery for Kenya from the Landsat, MODIS, PALSAR, and ASTER platforms, including a number of image products summarizing land use classification (MODIS types 1-5 for 2001 to 2005 and Landsat MSS derived 1 km for 1980), land surface temperature (MODIS LST), precipitation (WorldClim 30yr average at 1 km), and vegetation indices[4] (NDVI for 2001 to 2008 every 16 days, 250 m resolution). Additional sources of land use and land cover information are provided with Africover as vector or raster data types, GLC2000, CLIP cover1, and UMD Global Land Cover. We possess elevation data from ASTER, SRTM, and digitized topographic maps (30 m, 90 m, and 250 m spatial resolutions, respectively). With regards to the distribution of tsetse, we possess vector and raster

[4] Data are publicly available

digitized distributions of fly belts for Kenya for 1967, 1973, 1996, and 2000[5]. Finally, there are rasterized estimates of livestock densities (# per km^2) for 2007 (ILRI[6]). Utilizing these data, DeVisser et al. (2010) developed the TED model to predict the distribution of tsetse in Kenya. Figure 2.1 shows the predicted minimum range (the areas where tsetse persist all year) between 2002 and 2009. To effectively and efficiently recover and maintain the value of the data, we require a solution that not only provides for efficient storage and retrieval of the data, but which also allows for automated metadata generation.

We possess 103 spatial datasets of socio-economic and demographic assessments of Kenya between 1971 and 2008. A portion of these data were provided by the Integrated Public Use Microdata Series (IPUMS) International dataset for Kenya and originate from the Kenya 1989 and 1999 census collected by the Kenya National Bureau of Statistics. The IPUMS data are a systematic sample of every twentieth household, which represented a sampling fraction of 5% and expansion factor equal to 20; a long form questionnaire was implemented surveying individuals within households. Location data for each individual is limited to respondent's province and district, the first and second administrative levels respectively, of five possible levels each at an increasingly finer spatial scale of resolution. Data at finer spatial scales are not available as part of the Kenya Bureau of Statistics' effort to maintain privacy. A further complication was a change in the number of districts in Kenya from 42 in 1989 to 69 districts in 1999; however, this change was made by subdividing existing districts allowing rough comparisons between associated regions. The IPUMS sample provides 97 household and

[5] Not all maps are publicly available
[6] International Livestock Research Institute (Nairobi, Kenya)

individual variables, most importantly geographic information (urban-rural status, province, district), utility (electricity, water supply, sewage type, and type of cooking fuel), and dwelling (number of rooms, toilet type, floor, wall, and roof material). Other relevant data include

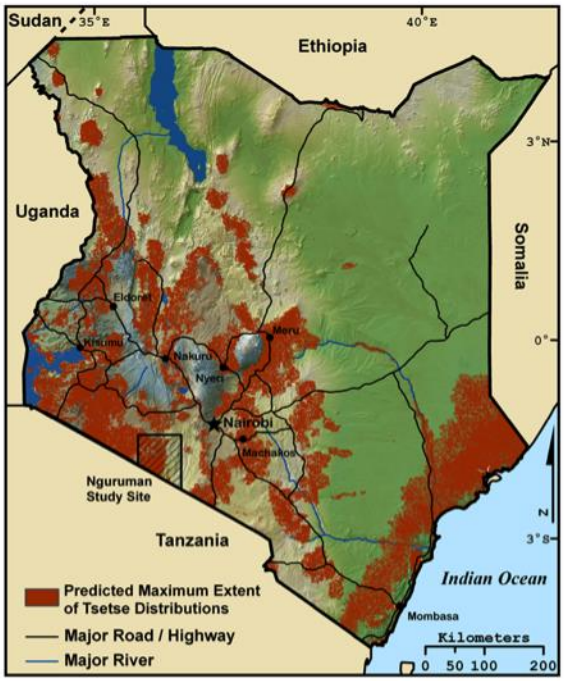

Figure 2.1: Maximum extent of tsetse distribution predicted by the TED Model between the beginning of 2002 and the end of 2009. As described by DeVisser et al. (2010), the TED Model uses five scenes of MODIS 1km annual land cover, 207 scenes of MODIS 250 m Normalized Difference Vegetation Index (NDVI), and 207 scenes of MODIS 1km day/night Land Surface Temperature (LST) products to predict the fundamental niche of tsetse in Kenya, and a fly movement model to predict tsetse distributions or realized niche of the tsetse species of interest. The TED Model is written in Python scripting language and is run within ArcGIS 9.2 (or later versions of ArcGIS). Other data sets used to construct the map include shapefiles of Kenyan major roads, highways, cities, rivers, Kenyan water bodies, lakes, and African country political boundaries. To construct the background topographic relief map, the Shuttle RADAR Topographic Mission (SRTM) 90 m Digital Elevation Model (DEM) was used to create a gridded hillshade product, and a dry season NDVI scene was combined with the SRTM DEM to create an elevation/vegetation color scheme.

individual variables describing household position, demographic characteristics, education, employment, migration, and disability (see Table 2.2 for a sample). There are a total of 1,074,048 individual entries for the 1989 census and 1,407,597 for the 1999 census samples.

Decisions on data management often conflict during collaborative research, resulting in the lack of a cohesive strategy for data management and the inability to share such data effectively. While our colleagues in Kenya have the technological skills to work with spatial data, the telecommunication network is insufficient to provide the necessary bandwidth or reliability to acquire the data via direct transfer over FTP or other similar protocol. In order to overcome this problem in the short term, it is necessary for us to carry the data into the country on physical media, and to have it accompanied by a data management system that can facilitate interaction with the large quantity of data. Thus, efficiency and portability are significant concerns.

Development of a Spatial Database System

Our solution for a spatial DBMS involves bridging a variety of software packages following the basic framework as described by Câmara et al. (1996) for the development of the TerraLib GIS Library and the integration considerations posed for the MurMur project (Parent et al., 2006). First, we outline the conceptual framework, and second, the implementation of the design. Third, we describe the development of routines, batch or other preconfigured shell scripts that can be selected to run either from the command line or through an interactive GUI prompt, whereby a user can add and recall raw data files, or query the database to return a mash up of spatial data files or metadata. Finally, we develop a set of SQL triggers (code set to run when activated by a defined action), to enforce data integrity.

Conceptual Model

Figure 2.2 demonstrates our conceptual model for a spatial DBMS. In contrast to previous implementations of MySQL, Postgres, and other common spatial databases, modern DMBS models facilitate raw data and metadata to be stored together, embedded in the database (Elmasri & Navathe, 2004; Watson et al., 2004). The proposed spatial computing environment uses open source, community supported software and standards, providing a solution to the data management problem that is temporally extensible. The database is portable in so much that we can copy the database and software binaries to a portable drive that can be carried to Kenya. Of critical concern to us in the selection of software components was the interoperability of the system, the ability of components to interface and work together. Our selected DBMS is PostgreSQL (Postgres), an advanced, readily available, open source, object-relational database management system. Utilizing standard SQL syntax, Postgres allows for complex query capabilities, including spatial queries, and facilities strict rule and primary key enforcement. Postgres is also extensible, allowing for the addition of new functionality (Michael Stonebraker & Kemnitz, 1991; M Stonebraker & Rowe, 1986).

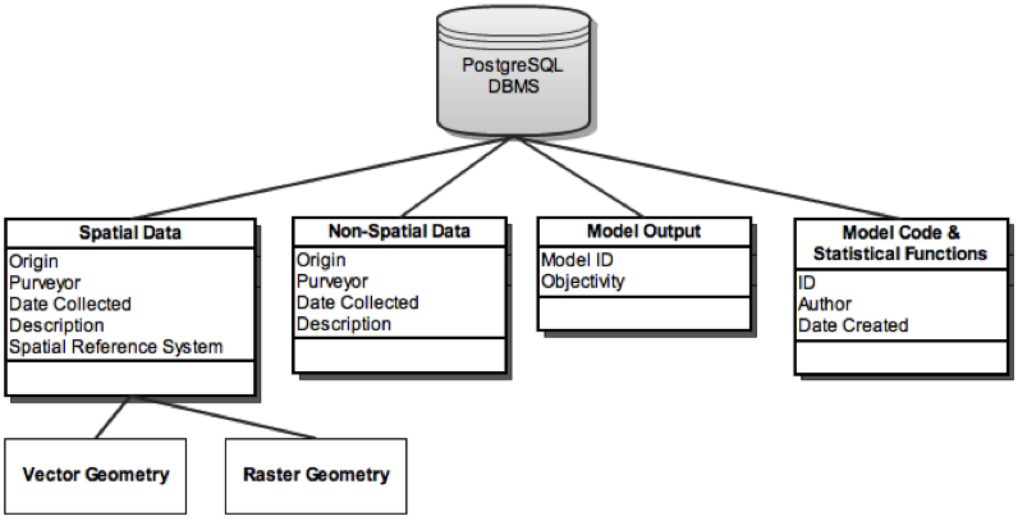

Figure 2.2: A conceptual model framework for the spatial DBMS.

PostGIS ("PostGIS," 2008) is an extension to the Postgres language that adds functionality for the storage and retrieval of spatial data files. PostGIS is, at its core, a suite of tools that serve as the backend for spatial functionality in Postgres. Of particular interest is the WKTRaster (Beta 0.1.6) project, which extends the ability of a Postgres database to store and index raster data, a first of its kind. The project mirrors the inherent vector-based functions (of GEOMETRY type) for raster data. The result is a single set of SQL functions that handle both spatial data types. This extension has the potential to greatly enhance and facilitate the utilization of raster data by end users. Figure 2.3 presents the conceptual model for the user interface to the Postgres DBMS. We incorporate a variety of software packages, explained later, each of which provides the user with

statistical, visual, or geoprocessing capabilities; the user can interact with these packages through a GUI or through a command line interface.

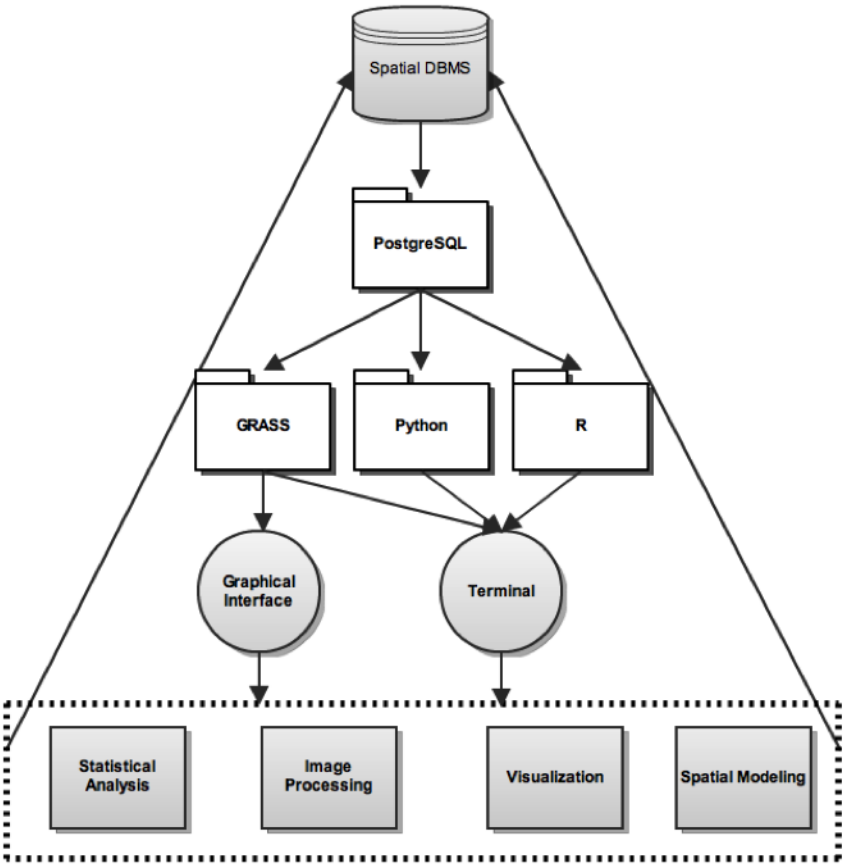

Figure 2.3: Flow of data through the model implementation as users interact with the system. It should be noted that flows are one-directional, meaning that although users can interact with the data to generate analysis, the results must be stored as a separate entity in the database. This ensures that the underlying data cannot be changed.

GRASS (Geographic Resources Analysis Support System) is one of the few open source options for the handling of raster GIS data. Developed by the Open Source Geospatial Foundation, GRASS is an increasingly popular solution for the use and analysis of spatial data in

academic research. Like the many other components selected for the system presented here, GRASS is interoperable with Python and Postgres and is extensible, allowing users to create and add new functionality.

Although there are a plethora of statistical packages available, R is our preferred statistical package in large part because it interfaces easily with GRASS and Postgres. We can do so without needing to install any additional software components. R is an open-source solution, developed by an international community of users, to create an alternative to the often expensive and restrictive programs offered by statistical companies. Although R lacks a graphical interface, it uses far less computer memory than most other comparable statistical packages. This allows us to perform complex analyses with fewer hardware demands, an important consideration for maintaining portability of the project.

Finally, Python ("Python," 2010) is an object-oriented programming language developed by Guido van Rossum in 1991 and maintained, in large part, by the open source community. Python is an efficient scripting language that facilitates interoperability between our database (Postgres), statistical analysis software (R), and GIS (Grass) (Neteler et al., 2008). Python is a highly intuitive, object-oriented programming language, easy enough to learn and use that it makes for a good solution to bridge our project components. Furthermore, the open-source nature of the language fits well with the other components, enabling us to incorporate additional extensions written by the community.

The most critical consideration for long-term data storage is persistence and security (Elmasri & Navathe, 2004; Watson et al., 2004). Digital data are inherently ephemeral in that over time physical storage media will fail or degrade, thus requiring continuous rewriting on digital

media to ensure persistence. Though drive technology has progressed significantly in the past decade, drive failure is not uncommon; the volume of data and frequency with which the data are accessed puts immense stress on the physical mechanisms. Therefore, it is necessary to employ a strategy that ensures the long-term viability of the data. To this extent, we employ a RAID array (Redundant Array of Independent Disks) as our secondary storage medium, mirroring data between groups of drives. This enables corrupted data resulting from disk failure to be recovered in real-time, without the need to create tertiary backup regimes. Furthermore, we enforce constraints to access of the data in making the data read-only, reducing the chance a write error will occur or values inadvertently be changed. Finally, all files will have MD5 checksums (a cryptographic hash code generated from a file's binary data) included as an attribute of the file allowing us to verify the integrity of any data file (Rivest, 1992). The strategies employed here will also enable us to protect against accidental user error, which may result in inadvertent modification or deletion of data. Security restrictions, though a good first line of defense, are frequently circumventable. The ability to rollback changes within Postgres, as well as being able to restore data from the RAID, will ensure the long-term integrity of the data library.

Database Standards

Generic frameworks for database development, such as the one we outline here, should use established ontologies in their component descriptions, which we satisfy by conforming to the formal ontology described by the Open Geospatial Consortium (OGC). The OGC is an independent group tasked with the purpose of developing and maintaining a set of standards for the management of spatial data to promote both consistency and interoperability across GIS platforms. GDAL (the Geographic Data Abstraction Library) stands as a single standard for

interoperability of raster data within the GIS community. As a library it facilitates the conversion between data products. However, as is the case when working with proprietary data, conversion between formats requires a commonly understood intermediate. The GDAL standards and abstraction libraries utilized in our project facilitate this conversion between data formats by providing a commonly understood intermediate. Custom GDAL libraries are largely used in our implementation of GRASS as a mechanism to add support for a range of occasionally unsupported data formats.

There are an inordinate number of disparate standards for metadata generation and inclusion, none of which are remotely universal. Regardless, it is necessary for our own data management that we accept and enforce a single standard for metadata management. Since our DBMS facilitates the storage of raw data within the DBMS, it precludes the need to store metadata separately as this information is included within the database as discrete variables (the raw data are technically also included as an attribute value). However, it is foreseeable that we may need, on occasion, to transfer data outside the realm of the database. Under this scenario, it is necessary to have a mechanism to recreate the metadata files discarded earlier. Thus, we incorporate custom scripts that can be called to assemble the necessary metadata precursor information from the data table. The resulting metadata report meets the formatting and content criteria specified by the OGC and ISO19115 specifications.

SQL-Rule-Based Interactions and Scripting

In an effort to ensure the integrity of the raw data and prevent accidental deletion or modification, all users are restricted in the ways they can interact with the DBMS. In the majority of cases, raw data are restricted to read-only access by users at the database level. One way we

achieve this is by restricting UPDATE and DELETE privileges to the database administrator account only. This ensures that raw data cannot be changed or deleted through unverified scripting. It further facilitates the simultaneous access and use of raw data by multiple processes. Figure 2.3 symbolizes the flow of data from the DBMS to the user. While the raw data are read-only, it is useful to permit users to save the output of models to the database with the option to link to the source data. Indexing these data sets together is useful when new users explore the data. We will discuss this functionality further in the later section "A User Perspective: Interfacing with the DBMS". Metadata for these model outputs will vary, but will at a minimum include the user identity and model code, as well as a summary of the model objective. A set of predefined rules or triggers is loaded into the DBMS and provides enforcement of the desired constraints (e.g. spatial/temporal mismatch encountered during scripted analysis). These rules serve to provide a minimal standard of validity and consistency for model output and statistical analysis (Devillers & Jeansoulin, 2006; Shi et al., 2002).

The storage of raw data within the DBMS precludes the need to regularly interact with different data formats. Nevertheless, it may be useful to incorporate a mechanism whereby we can convert data among data formats or export data into a range of formats. While the majority of these tasks can be accomplished within the GRASS interface, we include in our library a set of python scripts to extend our ability to move among data formats (particularly useful for the range of raster imagery available). For example, our "data bank" includes raster imagery (*.img, .tif, .hdf, among others), vector data (*.shp and others), text documents (*.doc, .docx, .txt, .pdf, and others), metadata files corresponding to acquired imagery (*.xml, .pdf, .txt, and others), as well as an assortment of other types not specifically mentioned here.

DBMS Interface – A Manager Perspective

Implicit in the design of our database are SQL rules (and triggers) that constrain the ways that users can work with data in an effort to prevent common mistakes. Since the database holds data at varying resolutions and extents, we constructed rules to check for common errors committed by users in selecting data layers. In the event that data layers are deemed incompatible, the user is alerted to the mismatch and encouraged, though not required, to restate their request. These rule sets operate at the point of data retrieval and storage. When importing data, we check and request the user define the relationship between spatial data and metadata. If metadata are not available, the user is prompted to input known characteristics of the data file in an attempt to force this paired relation. Common examples of information prompted include the timestamp of data acquisition or generation, solar or view angle (for satellite derived imagery), or a definition of codes that may occur within the data table (often the case with census data). Some of these data can be retrieved from the raw data, such as extent, summary statistics, file type, and others. Enforcing these relations at the time of storage will greatly reduce the number of unpaired spatial data files and corresponding descriptors.

These rules sets also operate on a variety of demographic or other similar data types stored in our database. In our specific implementation of the database, we have sample volumes of census data collected by the Kenyan Government and acquired from IPUMS. As data is input into the database, users are prompted as to the type of data being input and plain English definitions of the variables (a long form description of the purpose of the data) included in the dataset. We developed a means to query the data file for variables and return this list in memory

to the rule set. In the event the data file does not return the correct list of variables, the user is prompted to specify the range of variables or create them.

DBMS Interface – A User Perspective

While many incantations of interfaces are possible, one means by which users can interact with data is through a web browser, which connects to the database via an application developed for Mac OS. Through the application, users have access to all the tools needed to query, utilize, and analyze data from the DBMS. From a set of menus, the user is prompted to select a subset of data. Next, they are given the opportunity of either selecting from a set of precompiled models or statistical scripts that can analyze the data, or the data can be brought forward and immediately visualized in the browser window. Finally, the user is provided with command line tools that can augment their interaction with the data. This approach to data interaction lowers the learning curve for new users and gets them instantly connected with the data.

Statistical and Analytical Analysis

We extend the functionality of our DBMS to assist users in browsing the library of data by using the R package to automatically calculate descriptive statistics. These summaries are potentially most valuable to users not involved with the production of the data, or who may not be familiar with a region of interest. Furthermore, providing a means to compute descriptive statistics automatically enforces consistency between files that is often a symptom of user error when files are independently managed.

Perhaps the most common challenge users face in interacting with databases is retrieving data both correctly formatted and appropriate for use in a particular analysis (Longley et al.,

2005). The scripted analysis tools help to bridge the gap in understanding for users not familiar with the R statistical package. With a simple menu prompt, the user is able to specify which data files and statistical methods are to be employed by the program. Though not a comprehensive selection, it provides a means to conduct a range of simple statistical comparisons. Further analyses can be performed though the R package directly linked to the database.

Implementation

To enable the reader to implement our DBMS model, we provide a general outline of the steps needed to install and configure the software packages. We make no assumptions as to the hardware on which the model is implemented. Our solution requires a number of software components, including PostgreSQL, PostGIS (with GDAL and WKT Raster extensions), Python, GRASS, and the R statistical package. In this section we will detail the required steps for the implementation of each component within open source Debian distributions of Linux.

Software Packages

Debian distributions of Linux take advantage of the aptitude system for software distribution, and therefore you can install all packages from the terminal. To install:

```
sudo apt-get install python postgresql
sudo apt-get install gdal-bin postgresql-8.4-postgis postgis grass r-
     base
```

You may also be interested in installing the Quantum GIS (QGIS) program, another convenient GUI that interfaces with GRASS, which can also be installed with aptitude:

```
sudo apt-get install qgis
```

In order to allow interaction between GRASS and the GDAL libraries, you need to build and configure the GRASS plugin for GDAL. Download and follow the installation instructions provided by:

http://trac.osgeo.org/gdal/wiki/GRASS

Finally, you need to install optional extensions to the R package. Start R from a terminal by typing R; then install packages with:

```
install.packages("ctv")
library(ctv)
install.views("Spatial")
```

Initializing Postgres

You will need to login to Postgres for the first time using the default "postgres" login. From the terminal window you will then be able to create user accounts, set access restrictions, and the database for the project data. First start Postgres from a terminal by typing:

```
psql -U postgres
```

You can now create your user accounts. In our implementation of the project, all users are granted limited permissions by default. Additional rights can be given later depending on the needs of the user.

```
CREATE USER <username>;
CREATEDB <dbname>;
GRANT ALL to <username> ON <dbname>;
```

It is necessary to configure the database to enable spatial functionality with PostGIS. The instructions given are also available from the PostGIS documentation at:

http://postgis.refractions.net/documentation/

First, type:

```
CREATELANG plpgsql <dbname>
```

Now, copy the PostGIS spatial definition files into your new database and navigate to the PostGIS installation directory specified during installation and type:

```
psql -d <dbname> -f postgis.sql
\q  #Quits the Postgres session
```

For additional security options that are available, you may refer to the Postgres documentation[7].

You may now login to your new database from a terminal window by typing:

```
psql -U <username> <dbname>
```

Use caution when performing upgrades to PostGIS as you will likely have to rebuild support for your database.

WKT Raster extension for PostGIS

WKT Raster is not yet included in the PostGIS pre-compiled binary as it is (as of 09/15/2010) still in beta testing. Therefore, you will need to download and compile the extension. The following instructions are available from the WKT Raster project documentation[8]. Depending on your specific system configuration, it may be necessary to build the PostGIS libraries from source as well as the required dependencies. Download the source code from:

http://www.postgis.org/download/

Unpack the tar file and navigate to the WKTRaster directory. Generate a configuration profile by typing:

```
./autogen.sh
```

[7] http://www.postgresql.org/docs
[8] http://trac.osgeo.org/postgis/wiki/WKTRaster/Documentation01

Now run a configuration script. You will need to locate the installation directory for PostGIS:

```
./configure --with-postgis-sources=/thesrc/postgis-version
```

If no errors were generated, you can install WKTRaster. Since you are already in the installation directory, simply type:

```
make & make install
```

Configuring GRASS

The first time you start GRASS, you will have to specify a location and path for your data. Choose the data path that you setup earlier (it should already be the default entry). In order to create a new location, it is easiest to have a georeferenced data file that spans the region of interest. Although not necessarily required by the software, it greatly simplifies the process of adding a spatial reference system to the data library to specify it right away. Start GRASS by navigating to the icon or from a terminal:

```
grass64 –gui
```

Click on the Location wizard icon on the right hand side of the window. Follow the on-screen instructions, selecting your georeferenced data file when prompted. Alternatively, you can create a new location by entering grass with the default example “spearfish60”. From the GRASS terminal navigate to the directory where your data is stored and type:

```
r.in.gdal i=<yourimage> -o <mapname> location=<newlocation>
```

The function `r.in.gdal` will parse the image file and automatically define the region based on the extent of the selected image.

Loading Data and interfacing GRASS with Postgres

There are a great many possibilities for importing data and interfacing with GRASS. Here we demonstrate several examples for importing data into Postgres and accessing those data within GRASS. Landsat 7 data are downloadable from the USGS as georeferenced TIFF for individual bands. To import them into a sample database "KENYA" utilize a python loader script[9]:

```
gdal2wktraster.py -r *.tif -t Landsat -s 4326 -k 100x100 -I >
      Landsatloader.sql
```

The script does not directly load the data into the database; rather it creates the necessary SQL script to do so. In this example, the -r option enables multiple files (selected with the asterisk) to be imported simultaneously into a single table. The -t option specifies the table name the data will be imported to. With the -s option, we specify the spatial reference system using SRID numbers (Spatial Reference Identifier, OGC specifications), WGS84 in this example (Herring 2006). The -k option splits each raster into tiles that are 100x100 pixels. Finally, the -I option requests that a spatial index file be created for each raster tile. Next pass the SQL loader for processing with:

```
psql -U <username> -f Landsatloader.sql <yourdatabase>
```

If you do not want to create a spatial index or forgot to do so in the first step, you can easily create one at the Postgres prompt by:

```
CREATE INDEX Landsat_SI ON Landsat USING GIST (ST_ConvexHull (rast));
```

In this example, the GIST (Generalized Search Trees) index is used, which has a balanced tree index structure similar to a B-tree (Hellerstein, 1999; Kornacker, 1999). An example of a common

[9] The loader is now called "raster2postgresql" and is installed by default with PostGIS.

format vector data type is the ESRI shapefile and an example is a digitized map of fly belt regions FLY.shp. Postgres facilitates the importing of vector data from the terminal with:

```
shp2pgsql -s 4326 -I -D FLY.shp <yourdatabase>.flybelts > flybelts.sql
```

Here the -s flag specifies the spatial reference system using SRID codes. The -I option requests the script initialize a GIST spatial index on the geometry column of the data. Finally, the -D option creates a dump file (SQL loader) that can be imported into Postgres from the terminal, a faster means of adding data to the database. Now pass the SQL loader to Postgres with:

```
psql -U <username> -f flybelts.sql <yourdatabase>
```

Alternatively, import vector data using a graphical interface by loading:

```
shp2pgsql –gui
```

To load the data into GRASS, initialize the Postgres driver from within the GRASS terminal. Start GRASS specifying a work location (see previous instructions for the generation of LOCATION). Now, load the driver defining the connection between Postgres and GRASS:

```
db.connect driver=pg database="host=localhost, dbname=<yourdatabase>"
db.login user=<username>
db.connect -p
db.tables –p
```

After initializing the connection, it is now possible to query the database; for example, retrieving the flybelts data described earlier:

```
v.in.ogr dsn="PG:host=localhost dbname=<yourdatabase> user=<username>"
     layer????? output=flybelts type=boundary,centroid
v.db.select flybelts
v.info -t flybelts
d.vect flybelts
```

Data in the HDF formats can only be read with additional GDAL libraries, which are not included with the standard distribution. However, the version of GDAL made available through the Ubuntu repositories appears to support some limited functionality. If further interaction with HDF is required, you will need to compile GDAL manually, inputting development files downloadable from the HDF Group[10]. HDF formats are containers, and thus may hold multiple data sets. Header data are accessed with:

```
gdalinfo sample.hdf
```

Sub dataset names are formatted as:

```
HDF4_SDS:subdataset_type:file_name:subdataset_index
```

A portion of the output reads:

```
SUBDATASET_8_NAME=HDF4_SDS:MODIS_L1B:GSUB1.A2001124.0855.003.200219309
      451.hdf:7   SUBDATASET_8_DESC=[408x271] Range (16-bit unsigned
      integer)
```

Detailed headers for this sub dataset can be viewed with:

```
gdalinfo
      SUBDATASET_8_NAME=HDF4_SDS:MODIS_L1B:GSUB1.A2001124.0855.003.2002
      19309451.hdf:7
```

Since the GRASS plugin was compiled with HDF support, image data, by individual band, are directly imported into GRASS with `r.in.gdal`.

```
r.in.gdal
      HDF4_SDS:MODIS_L1B:GSUB1.A2001124.0855.003.200219309451.hdf:7
      out=hdfexample
```

[10] http://www.hdfgroup.org

Once the data are loaded into the GRASS interface, they can be imported into Postgres either directly with db.connect or by exporting the image as .TIF and importing with the gdal2wktraster.py script.

Limitations and Future Expansion

The initial phase of our project is limited to the objectives addressing the misuse, misrepresentation, and effective archiving of our data library. As advances and improvements are made to the telecommunications infrastructure in Kenya, we will be able to share a common easily accessible repository for data with our colleagues outside the United States. The DBMS framework, including necessary software packages and data, can be packaged together and distributed via portable hard drive. Future updates to our software model will include a web-based interface that will allow users to interact with the DBMS, including the suite of analysis and visualization tools (R and GRASS), without the need to install and configure these programs locally. This reduces the hardware requirements for working with the data, potentially allowing a broader base of users to share in data access. This approach has been extensively applied in many of the institutional projects referenced commonly in the literature (Câmara et al., 1996; Parent et al., 2006).

Summary

Disease ecology is a trans-disciplinary field, exploring the complex interactions between diseases and the environment. Computational and collaborative barriers inhibit meaningful advances in the field. Major problems include data management schemas to facilitate the scalability (data are re-sampled dynamically to avoid redundant storage), reliability (concurrent access to data

permitted while ensuring the raw data cannot be changed), and data security (the database allows for a dynamic security access policy while meeting the HIPAA and IRB requirements). As data become aggregated with decreasing spatial resolution, many of the privacy concerns disappear but tracking and management problems proliferate. The DBMS must dynamically alter the restriction rule-set to account for the aggregation and application challenges. Previous implementations of data management systems required that multiple instances of the data be stored, creating a problem of exponentially increasing data storage demands. Currently, no framework for data management that addresses this set of **integrated** concerns exists.

Over the course of our research on African Trypanosomiasis, we have accumulated a large library of data. Our database model utilizes open-source software so as to allow for flexibility and extendibility to the model implementation. We take advantage of the new PostGIS extension, WKTRaster, to allow for the storage of raster imagery within the database. This allows us to enforce a single standard in the way all data formats, irrespective of the contents, are managed. With this development, we are finally able to store the entirety of our data library, spatial and non-spatial, explicitly within a single database implementation, and the restrictions and rule-sets coded into the DMBS should ensure the long-term integrity and security of the data.

The overarching goal of this project is to create a multi-scale predictive model for the tsetse and African Trypanosomiasis to provide a means whereby governments, communities, and NGOs can make informed decisions for disease control or suppression that are spatially and temporally aware. Given the variable background and technical expertise of the different groups, our solution should be simple enough for the most basic user, yet powerful enough to be useful

for complex analyses by the most skilled. To maximize the utility of this system, we will utilize a participatory design framework to develop Mac, iPhone, and Web applications for interfacing with the models and data.

Open-source software is uniquely capable of rapid adoption of new technologies and functionality due to the base of developers working on modular extensions to the software base. Spatial databases are increasingly common. Mobile data applications are becoming increasingly location-aware (e.g. Facebook, Twitter, Loopt, Turn-by-turn Navigation (Waze), among others), with the purpose of providing users with context-specific information and opportunities. These GIS technologies increasingly promote web-based interfaces to a spatial database. Perhaps the most exciting vision for GIS is the adoption of mobile technologies, now incorporating GPS technology, to provide for context-aware interaction with spatial database systems. The next step in our project is to implement an efficient, two-way web-portal for the spatial DBMS that will allow us to interact with data and model results from mobile devices in the field. Combined with automated scripting and model execution, this would have the potential to dramatically increase the amount of information shareable with local communities. While not yet possible due to barriers in Kenya, we work towards the vision of achieving synchronicity between science and practice.

Table 2.2: A sample of the subset of Kenya census data we hold from the 1990 National Census

	cntry	year	sample	serial	persons	wthh	subsamp	gq	unrel	urban	provke	distke	ownrshpd	electrc
1	404	1989	4041	1000	4	20	26	10	0	2	1	1010	216	2
2	404	1989	4041	1000	4	20	26	10	0	2	1	1010	216	2
3	404	1989	4041	1000	4	20	26	10	0	2	1	1010	216	2
4	404	1989	4041	1000	4	20	26	10	0	2	1	1010	216	2
5	404	1989	4041	2000	1	20	76	10	0	2	1	1010	216	1
6	404	1989	4041	3000	4	20	2	10	0	2	1	1010	216	1
7	404	1989	4041	3000	4	20	2	10	0	2	1	1010	216	1
8	404	1989	4041	3000	4	20	2	10	0	2	1	1010	216	1
9	404	1989	4041	3000	4	20	2	10	0	2	1	1010	216	1
10	404	1989	4041	4000	1	20	92	10	0	2	1	1010	140	2
11	404	1989	4041	5000	1	20	81	10	0	2	1	1010	216	1
12	404	1989	4041	6000	12	20	5	10	0	2	1	1010	216	1
13	404	1989	4041	6000	12	20	5	10	0	2	1	1010	216	1
14	404	1989	4041	6000	12	20	5	10	0	2	1	1010	216	1

CHAPTER 3

UTILIZING VOLUNTEERED INFORMATION FOR INFECTIOUS DISEASE SURVEILLANCE

Abstract

With the advent of Web 2.0, the public is becoming increasingly interested in spatial data exploration. The potential for Volunteered Geographic Information (VGI) to be adopted for passive disease surveillance and mediated through an enhanced relationship between researchers and non-scientists is of special interest to the authors. In particular, mobile devices and wireless communication permit the public to be more involved in research to a greater degree. Furthermore, the accuracy of these devices is rapidly improving, allowing the authors to address questions of uncertainty and error in data collections. Cooperation between researchers and the public integrates themes common to VGI and PGIS (Participatory Geographic Information), to bring about a new paradigm in GIScience. This paper outlines the prototype for a VGI system that incorporates the traditional role of researchers in spatial data analysis and exploration and the willingness of the public, through traditional PGIS, to be engaged in data collection for the purpose of surveillance of tsetse flies, the primary vector of African Trypanosomiasis. This system allows for two-way communication between researchers and the public for data collection, analysis, and the ultimate dissemination of results. Enhancing the role of the public to participate in these types of projects can improve both the efficacy of disease surveillance as well as stimulating greater interest in science.

Introduction

Recent publications surrounding Volunteered Geographic Information (VGI) broadly represent the belief among some in the academic community that non-scientists can be engaged in and benefit from spatial data analysis (Connors et al., 2012; Flanagin & Metzger, 2008; Goodchild, 2007a, 2010a), a field previously reserved exclusively for academics. Focus on VGI represents a paradigm shift from viewing science as having a single authority (the scientist) to a model where authority is relative and expressed contextually. Abundance, repetition, and the collective assessment of data (as well as the ability to correct) convey credibility to information that would not necessarily exist otherwise (Connors et al., 2012). In this sense, a non-scientist plays a role in validating data collected by others, and collectively assessing data quality (Connors et al., 2012; Craglia, 2007).

The concept of Web 2.0 incorporates bi-directional collaborations in which users collectively collate spatial data, stored in a central cloud repository and accessible by anyone for whatever purpose deemed worthy. The Web 2.0 paradigm is represented widely through web projects such as Wikimapia, OpenStreetMap, and even Google Earth. Within the context of these volunteered GISystems (VGIS), users contribute information to develop a collective knowledge base. Recent advances in mobile technology have furthered the applicability of Web 2.0 projects, enabling easier access to the information, and even allowing for novel uses of crowd-sourced information (Rosenberg, 2011). Sui (2008) extends the paradigm to include “the wikification of GIS”, a notion which he defines as being the shift in perception that only people who are specifically trained to “do GIS” should interact with spatial data and perform analysis. It is upon

this notion, specifically, that VGIS endeavors to enhance the role of the user in the collection and analysis of spatial data.

The use of volunteered information for disease surveillance draws upon themes in the participatory GIS (PGIS) literature in suggesting that GIS technologies can operate in concert with volunteered information and local knowledge (S Boroushaki & Jacek Malczewski, 2010; Connors et al., 2012; Elwood, 2010; Flanagin & Metzger, 2008). The key distinction between classical PGIS methods and VGIS involves the role of the scientist. We refer here to McCall's (2005) discussion of good governance through improving dialogue, legitimizing and using local knowledge, the redistribution of resources, access rights, and new skills training in geospatial methods. These concepts support the idea that a PGIS or VGIS approach can contribute to the adoption of new technologies for disease surveillance.

Background

Traditional Paradigm

The traditional paradigm in GIScience partitions individuals into experts versus non-experts. In an academic context, this treats scientists as the experts and citizens as non-experts. Under this traditional paradigm, public participation in the research process is hindered by a number of factors. Most importantly, the traditional roles of experts (scientists) versus the public leaves little room to consider alternative knowledge bases (i.e., local knowledge). Furthermore, there is limited opportunity for citizens to become informed, equal participants, thereby limiting the potential applicability of any results/understanding gleaned from the research process (Soheil Boroushaki & Jacek Malczewski, 2010).

Under the traditional GIS model, technology and software are not readily accessible, requiring either a specific skill set or simply being priced beyond the consumer market. Therefore, citizens are relegated to operating as consumers of information exclusively, or as indirect producers, mediated by communication to researchers in small group projects. Their interaction with the data in this regard is strictly as a provider of information, not as producers of spatial data products. Finally, the traditional GIS model treats data validation as achieved largely through reputation (Flanagin & Metzger, 2008). Scientists and researchers are perceived as producers of reliable data due to past training in data collection and analysis. Furthermore, the peer review process adds credibility by requiring outside researchers to assess quality. Broadly though, data collected by researchers is assumed to be reputable because it is collected within the context of academic endeavors, and done by trained individuals. Information of this sort is generally accepted to be true until shown to be otherwise. With few exceptions, the vast majority of GIS data products are produced under the traditional GIS model. Citizens are largely excluded from the process of data collection and analysis (Connors et al., 2012). We do not, however, suggest that the traditional model must be replaced. Instead, we propose the standard model be extended to facilitate collaboration between citizens and researchers.

VGIS Paradigm

Volunteered GIS represents a paradigm shift from viewing science as having a single authority (the scientist) to a model where authority is relative and can be expressed contextually. Information abundance, repetition, and the collective assessment of data convey plausibility to data that would not otherwise necessarily exist. In this sense, a non-scientist plays a role in validating data collected by others; collectively assessing data quality (Oreskes et al., 1994). This

concept is explored further by Craglia (2007) in his assessment of individuals as geosensors, empowering them to validate global models using their own perceptions or impressions of the data. Volunteered GIS therefore represents the broad interest by non-scientists to be engaged in and to benefit from spatial data analysis.

Although Goodchild coined the term "volunteered GIS", the movement towards a new paradigm really began a decade earlier with the desire, on the part of scientists, to engage citizens directly in the research process. Sara Elwood, through her work with PGIS, exemplifies this desire and her work has been instrumental in the evolution of the VGIS paradigm (Elwood, 2006b). Other contributions have included work by Elmes et al. (2005) with their description of a "community integrated GIS", Turner's "Neogeography" (2006), Balram and Dragicevic's "collaborative GIS" (2006), and Sieber's "public- participation GIS" (2006). Collectively the work of these individuals demonstrates the broader goal of direct community engagement in the research process.

However, researchers have also made significant strides towards integrating components of a VGIS into their own projects, including studies in environmental sensing, decision-making, resource management, and community risk assessment. Project GLOBE, OakMapper, and Audubon's Christmas Bird Count (Connors et al., 2012; Goodchild, 2007a; House et al., 2001; Yaukey, 2010) are long running projects for the purpose of monitoring spatial and temporal distributions of resources and phenomena. By employing citizens to collect data, researchers are able to more effectually analyze spatial processes by generating much larger quantities of data. While data quality remains a concern, the large quantity of data collected diminishes the influence of inaccurate data (Flanagin & Metzger, 2008).

The use of VGIS to answer questions of decision making draw upon the PGIS literature in suggesting that GIS technologies and implementations can assist in conflict resolution and multiple-criteria decision making (Soheil Boroushaki & Jacek Malczewski, 2010). Flanagin and Metzger (2008) make reference to these types of questions in using GIS for collective community efforts. The key distinction between classical PGIS methods and VGIS involves the role of the scientist. PGIS seeks to improve dialogue between actors for the purpose of legitimizing and using local knowledge, the redistribution of resources access and rights, and new skills training in geospatial methods (McCall & Minang, 2005). However, the researcher plays a limited role as teacher. VGIS builds on this by leveling the authority between actors; scientists and non-scientists are viewed as having [almost] equal authority, allowing both actors to communicate more freely with each other and to share expertise. Our prototype supports the idea that a PGIS or VGIS can contribute to addressing questions of decision-making, and later resource management and conflict resolution.

Volunteered GIS alters the standard GIS paradigm by substituting a producer-user model instead of the traditional expert-user archetype. Under this framework, researchers and citizens can act as either producers or consumers of spatial data, depending on the context within which they are interacting with the data. Producers in this case need not necessarily be experts in all areas of GIS or the broader research context. Rather the producer's role is given to any individual with information to contribute to the aggregate knowledge base (Soheil Boroushaki & Jacek Malczewski, 2010; Flanagin & Metzger, 2008). User roles are given to individuals who consume a spatial data product for any purpose. Under the new paradigm, roles are not fixed and not exclusive.

Finally, under the traditional GIS model, data were perceived to be trustworthy because of the perceived authority of the scientist. However, with changing roles, we need a new model for data error assessment (Flanagin & Metzger, 2008; Goodchild, 2007a, 2007c). Possibly the single largest barrier to the utilization of volunteered information is the uncertainty surrounding its credibility (Connors et al., 2012; Flanagin & Metzger, 2008; McKnight et al., 2011). Under the VGIS model, the credibility of volunteered information is achieved through volume. Intuitively, we understand that if multiple individuals report similar information, the reports are likely credible representations of the truth. The larger volume of data collected through a VGIS, albeit repetitive, can achieve the same threshold for credibility as data collected under the traditional model (Flanagin & Metzger, 2008).

Related, there's a significant degree of uncertainty as to the nature of volunteered information with respect to the types of error. McKnight et al. (2011) explore the relation of volunteered information to assess spatial distribution of West Nile virus in Michigan. In their analysis, they raised the issue of uncertainty with regards to types of error that influence the data. For example, users may reliably report positive observations (i.e. in this case, observations made of dead birds), but reports are likely to indicate the absence of data. Therefore, volunteered information is heavily biased towards the observation of an outcome, and should not be interpreted as a metric of prevalence. Utilization of volunteer information must be cognizant of the nature of uncertainty.

The prevalence of mobile devices that have GPS capabilities, including cell phones, tablets and laptop computers, has increased the accessibility of spatial data. Hardware is no longer priced outside the realm of ownership for many people in the world, meaning that users can now

directly engage with spatial data in ways that they simply could not do before. The interaction of the public with spatial data is now so prevalent that most users have developed sufficient technological skills and spatial cognition (through interaction with online mapping tools) to enable them to interact with spatial data in an intelligent manner, precluding the need for training prior to participation in GIScience research. Paradoxically, it would appear that spatial cognition is unrelated to global geographic awareness. Although people are able to position themselves abstractly on the landscape, they remain illiterate as to the broader geographic context in which they live.

Software interfaces fall into two broad classes: traditional desktop products and web-based applications. For the purposes of inter-acting with a VGIS, citizens are most likely to use a web application since this does not require a specific platform or license to run. Desktop applications, on the other hand, can be distributed to certain groups, allowing for a more targeted interaction with the spatial data. The open-source software movement is most directly credited with making GIS software accessible to the public, removing financial and hardware restrictions for many GIS products. Most notable among these are GRASS and Quantum GIS, free GIS packages modestly equivalent to ESRI's ArcGIS®. Interfaces for spatial data analysis have been developed with R and Python, interacting directly with Grass and Quantum GIS. Increasing familiarity on behalf of the public in spatial tools, geospatial technologies, and mapping increases the likelihood that they will be able to act as producers of high quality information. While the increasing availability of mobile technologies has spurred public interested in GIS, the cost of adopting new technologies remains a principle challenge, particularly in developing countries.

Disease Surveillance

Disease surveillance systems are established for the purpose of collation, analysis, and dissemination of information so as to facilitate the allocation of resources in handling disease outbreaks (Thacker et al., 1983). Broadly, surveillance programs are categorized as either passive or active. Passive disease surveillance programs rely on reporting by healthcare providers to public health authorities when specific signs and symptoms are observed, a diagnosis is made and/or a diagnostic test is confirmed. Public health authorities collate these reports and assess the need for a coordinated response. Upon making a determination, the authorities communicate back to the local health care providers, and the necessary recommendations are set forth to address the disease outbreak. An example of passive surveillance in the United States involves the reporting of certain communicable diseases by health care workers after a diagnosis is made. These reports are received by public health officials who are tasked with ensuring the disease does not pose a threat to the welfare of the public.

Passive disease surveillance systems are hierarchical in nature with space and time important factors. Knowledge of highly infectious diseases may be reported up and information on the control of those diseases may be reported down the hierarchy very rapidly, whereas more common diseases may be reported and intervened on slower schedules such as monthly or semiannually. The management and coordination of communication within a passive surveillance system therefore, needs to be agreed upon by all parties (i.e., levels within the hierarchy in order to ensure the protection of population health). However, passive surveillance systems are widely criticized for underreporting diseases (Thacker et al., 1983).

When passive surveillance systems break down and mandated reportable disease(s) are not communicated from the local to central levels, there is a need to respond by implementing an active surveillance program.

Active surveillance programs directly address the underreporting of disease by utilizing teams to assess local conditions. Such programs begin with the recognition at the central level that the expected communication in space and/or time has not been received and in response actively reach out to that location for the information. During these visits, retrospective data are collected and the management of the passive surveillance system is revived (e.g., manpower, technology). The operations of disease surveillance systems are therefore highly dependent upon the cooperation of all participants at level of the hierarchy. One well-cited example of active surveillance is Snyder and Merson's (1982) meta-analysis of diarrheal disease prevalence and mortality throughout the developing world. Here they review 24 studies where data was actively collected (either through home visits or other means) by trained personnel. In contrast to a passive surveillance program, workers were employed for the sole purpose of collecting disease prevalence data.

Case Study

Purpose

African Trypanosomiasis (AT) is a zoonotic disease transmitted by the tsetse fly. In Kenya, the two most common forms of AT are *Trypanosoma brucei* (Nagana), the form of the disease that affects cattle, and *Trypanosoma rhodesiense* (Sleeping Sickness) that affects humans. While sleeping sickness is relatively rare in Kenya, Nagana is widespread and represents a major threat to the livelihood of pastoralists (Baird et al., 2009; Tarimo-Nesbitt et al., 1999; Waller, 1990). The

prevalence of Nagana has increased in recent decades due to a decline in control regimes, climate change, and anthropogenic factors (Batchelor et al., 2009; Bauer et al., 1992; WHO, 2005). Our case study describes the prototyping of a VGIS for the purpose of surveillance of an infectious disease vector.

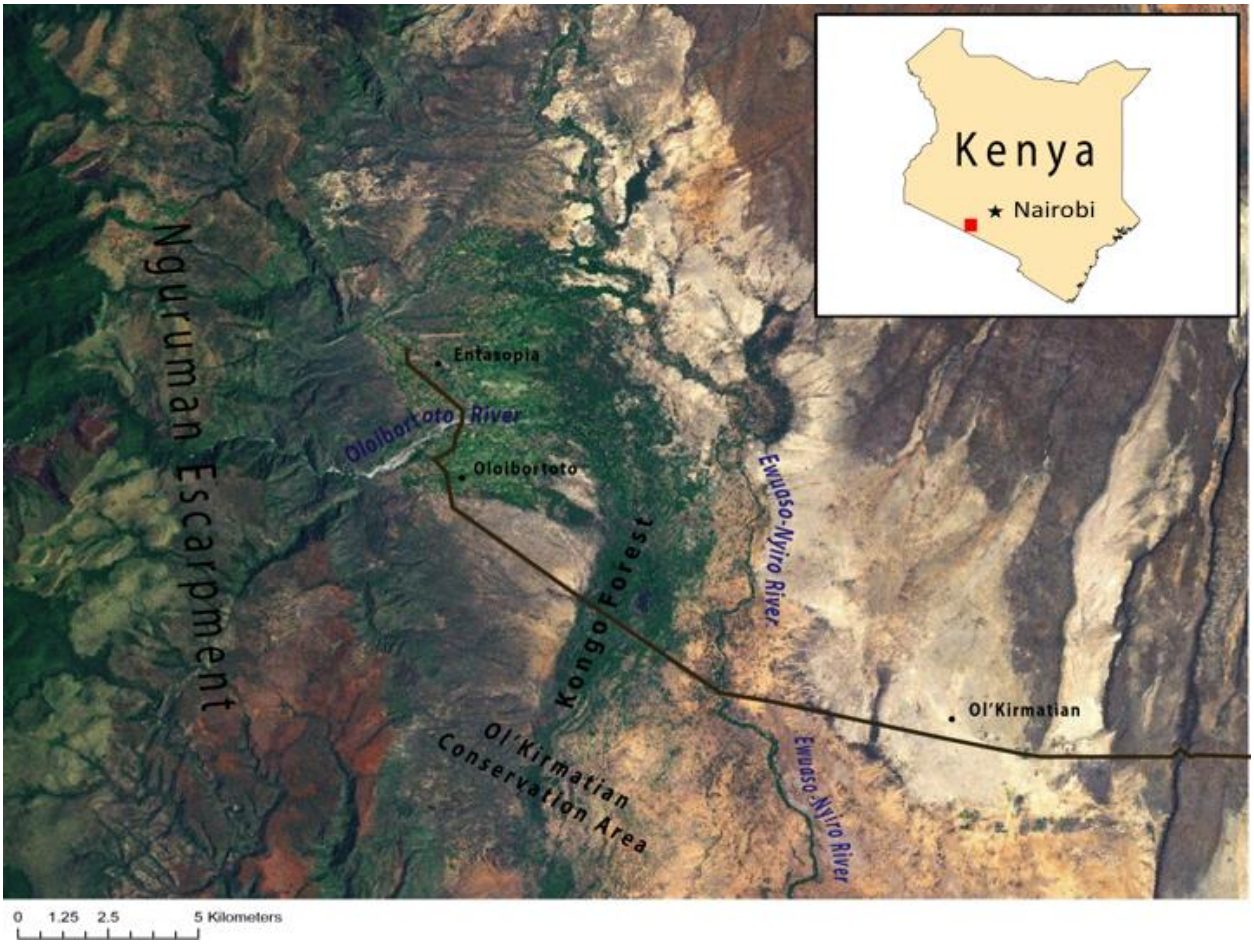

Figure 3.1: Study Area

Site Description

Nguruman (Figure 3.1) is located at the base of the Rift Valley in southern Kenya, just east of the Nguruman Escarpment. Formally, Nguruman is the local Maasai name for the settlement, which

occupies the area west of the Ewuaso-Nyiro River and the Kongo Forest to the Oloibortoto water intake; it is bounded to the south by the Ol'Kirmatian Conservation Area and to the north by the Oloibortoto River. Nguruman is also referred to locally as Oloibortoto. North of the Oloibortoto River, and broadly included in our study area, is Entasopia, the largest settlement in the area. The political "capital" of the Nguruman area is Ol'Kirmatian, a settlement 6.5km west of the Ewuaso-Nyiro River, and home to the District office for the Kenya government as well as the office of the local governor for the Ol'Kirmatian group ranch, the political arm of the Maasai in this area.

From the base of the Rift to Oloibortoto, the predominant land use is smallholder agriculture. Streams dissect the region and are maintained by the community as means to irrigate their farms. Dominant agricultural crops throughout the region include tomatoes, vegetables destined for South Asian markets, and fruit trees (e.g., bananas, mangos) (Langley, 2010). Southeast along the road from Oloibortoto to the Kongo forest, vegetation density rapidly increases. The area is extremely rocky and dominated by herbaceous and woody shrub vegetation, most abundant of which are *Acacia tortillis*, *Salvandora persca* (toothbrush tree), *Grewia tembensis*, and *Cordia sinensis* (Maitima; Morris et al., 2009). Dominant grasses throughout the region include *Sporobolus spp.*, *Setaria spp.*, and *Cynodon dactylon* (Morris et al., 2009). The Kongo forest is the area of densely vegetated land between Oloibortoto and the Ewuaso-Nyiro River. It is within this zone that we find abundant tsetse; unfortunately this zone is often the only option available to the community for grazing their animals during the dry season. Moving east from the Ewuaso-Nyiro River, the landscape dries quickly, resulting in a rapid decrease in vegetation density. During the dry season, the area is devoid of most vegetation; however after a short period of rains, the grasses in the area of Ol'Kirmatian re-emerge with

vigor. Across the entire region, these eco-zones are highly dynamic and respond rapidly to local climatic shifts and the occurrence of precipitation.

Global climate change is dramatically influencing the local environment within our study area (Moore et al., 2012). In past decades, annual precipitation in southern Kenya has remained relatively constant despite significant increases in annual mean temperatures, however the variance of the magnitude of precipitation events have increased and the seasonality of total precipitation has become less predictable (Altmann et al., 2002; Messina et al., 2012; Moore & Messina, 2010). The observed climate change and uncertainty in precipitation will undoubtedly threaten the livelihood of farmers and pastoralists (Fischer et al., 2005). Indeed, these concerns were conveyed to us in the course of our work; many farmers have already found it difficult to determine the right time for planting due to changes in local weather and precipitation events (Langley, 2010).

Trypanosomiasis (Nagana) in cattle is a major threat to the livelihood of Maasai pastoralists in Nguruman. The risk of infection is chief among their concerns to the health and well-being of their cattle herds. An important consideration for the community is the management of grazing for cattle herds among the members of the group ranch. A committee of elders, whose chief aim is to maximize utilization of the limited resources (while advocating sustainability) for the benefit of the community, manages the patchwork of grazing areas. Of particular interest to the grazing committee (as expressed through interviews) is the ability to work with our research lab to incorporate predictions of the spatial and temporal trends in tsetse populations and models of risk aversion.

DeVisser et al. (2010) developed a species distribution model for tsetse (TED) that predicts tsetse presence/absence every 16 days based on the habitat requirements and movement rates of the fly. The precision of the model predictions is limited spatially by the resolution of the inputs (250m), and temporally by the availability of MODIS LST and NDVI data products (8 and 16 days respectively). It is well established that tsetse are highly responsive to microclimatic conditions supported by local variations in vegetation (Terblanche et al., 2008). The spatial resolution of the TED model predictions limits consideration of such local configurations, thereby increasing the likelihood of errors of omission. The TED model was designed to identify endemic tsetse and does not model transient tsetse populations. By incorporating volunteered information from citizen reporters, TED could better illustrate the distribution of the flies over space and time by reducing errors of omission and reporting transient populations. Volunteered information may also be used (to an extent) to confirm (Oreskes et al., 1994) the TED model predictions by giving us a means to estimate model uncertainty.

Conceptual Model

Here we elaborate on the previously published framework for a VGIS (Langley & Messina, 2011) by illustrating the construction and deployment of a working prototype. Furthermore, we discuss potential challenges and limitations of our implementation and propose strategies to address these issues. Figure 3.2 outlines the basic implementation of the proposed VGI and the methods by which users will be able to interact with the spatial database (sDBMS), specifically through mobile devices. The core of the proposed implementation is a Postgres database server that stores both the spatial data as well as scripts to compute predictions of tsetse distributions, process volunteered data (submitted first to a reliability assessment), and automatically retrieve

and process remotely sensed imagery as it becomes available. Users interact with the database through an Apache server and an HTML interface. A MapServer implementation provides functionality for visualization of spatial data. All components are open-source and platform independent so as to convey maximum portability.

Our implementation of a VGIS seeks to achieve three goals: 1) facilitate user interaction with the VGIS and model results so as to allow for the reporting of information that may correct otherwise inaccurate data (defined as those predictions that contradict ground-level reports); 2) assess the reliability of volunteered information; and 3) incorporate volunteered information to calibrate a model of tsetse distribution and reduce errors of omission. To assess the functionality of the VGIS to achieve these goals, we have developed a working prototype of the system to illustrate our approach.

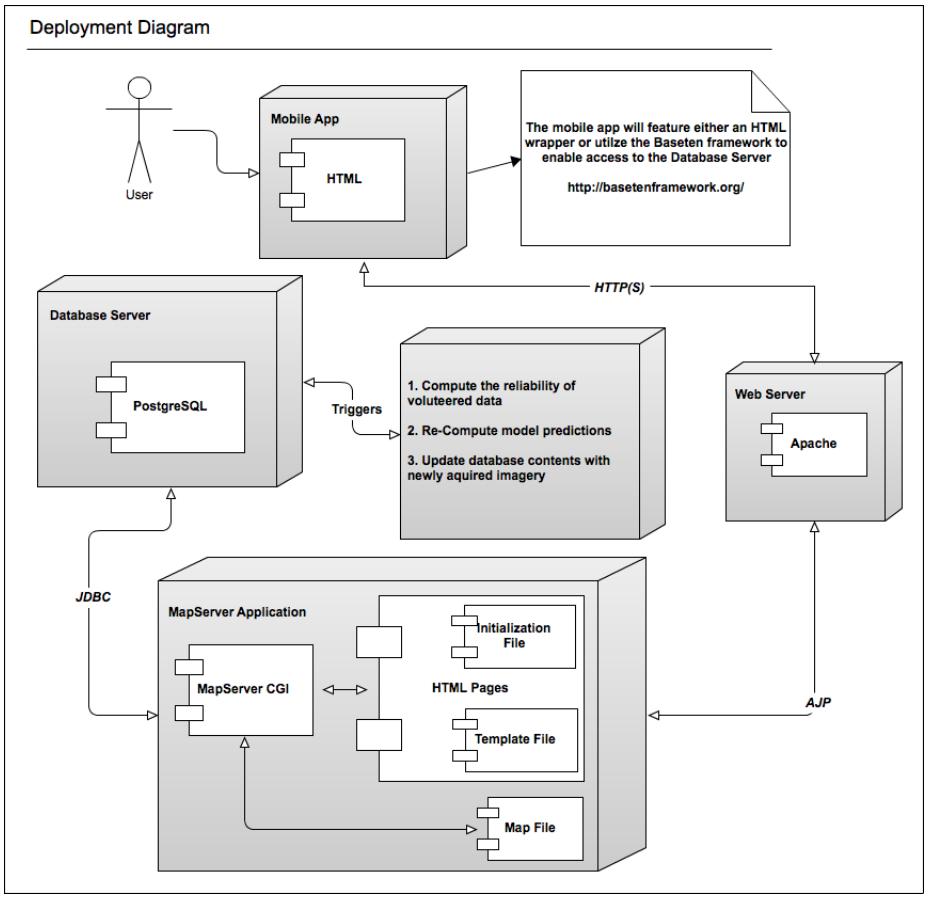

Figure 3.2: This deployment diagram illustrates the interaction of the separate components of the VGIS and the flow of information between each component

We incorporate a variety of software packages, including GRASS and QGIS (for visual GIS support), Python and R (for statistical and modeling tasks), each of which provides the user with statistical, visual, and geoprocessing capabilities; the user can interact with these packages through a GUI or through a command line interface. Our selected DBMS is PostgreSQL 9.1 (Postgres), an advanced, readily available, open-source, object-relational database management system. Using standard SQL syntax, Postgres allows for complex query capabilities, including spatial queries, and facilitates strict rule and primary key enforcement. Postgres is also

extensible, allowing for the addition of new functionality (Michael Stonebraker & Kemnitz, 1991; M Stonebraker & Rowe, 1986). In contrast to previous implementations of MySQL, Postgres, and other common spatial databases, modern DMBS models facilitate the combined storage of spatially explicit data and corresponding metadata together in the database (Elmasri & Navathe, 2004; Watson et al., 2004). The proposed spatial computing environment uses open-source, community-supported software and standards, providing a solution to the data-management problem that is temporally extensible. Of critical importance to us is the improved functionality available in PostGIS 2.0, which adds support for raster data types. PostGIS is an extension to the Postgres language that adds functionality for the storage and retrieval of spatial data. PostGIS is, at its core, a suite of tools that serves as the back end for spatial functionality in Postgres.

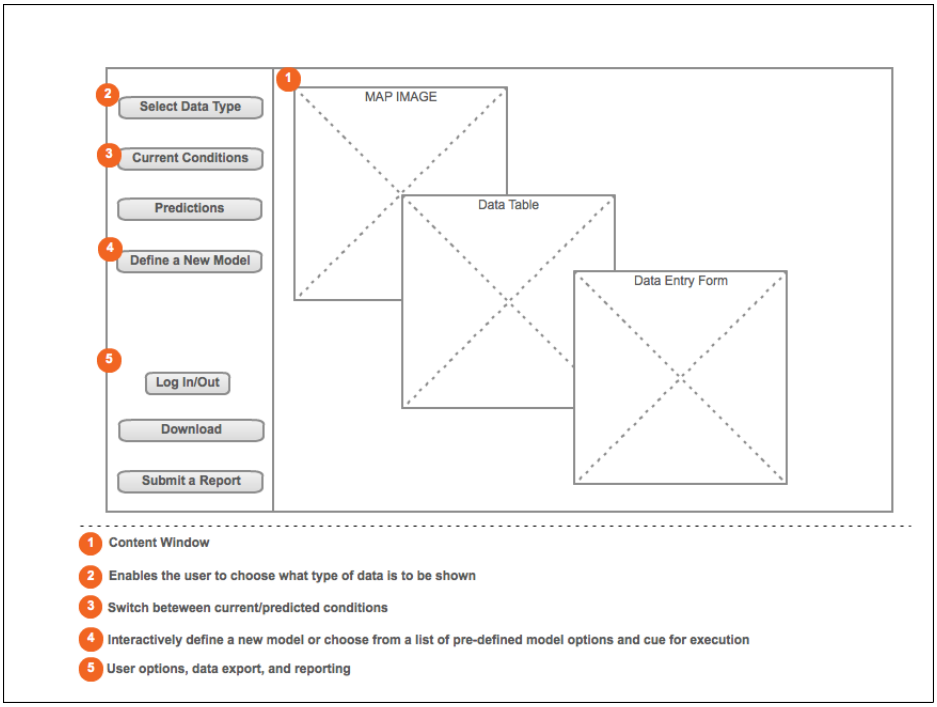

Figure 3.3: We propose an iOS application (for iPhone or iPad) that allows for users to interact with the VGIS, explore model predictions, volunteer data, or to contribute

Data Collection and Interface

Users and producers alike are able to interact with the VGIS in many ways, each according to their own skills, interests, and available hardware. In our case study, we outline the mechanisms whereby participants (either researchers or community members) can interact with the VGIS.

Figure 3.3 illustrates the broad deployment strategy for mobile device interaction. Related to the deployment of a mobile interface, the web interface sports a comprehensive suite of tools available to all participants from any web browser, with functionality dependent upon the credentials supplied to the system.

Most participants will find themselves interacting with the VGIS primarily through their mobile devices. For this purpose, we propose an iOS application that, for the most part, simply employs an HTML wrapper allowing the user to interact with a MapServer application. Users are able to query the database for specific, albeit limited, types of data, even define a specific range of times over which to aggregate the data; the application is geographically aware, so it is able to return information for a user's specific coordinates by passing the current longitude and latitude to the server. Most importantly, a user is able to volunteer information, with regards to the distribution of tsetse, through the application. Simply, a user can use this function to report that tsetse flies are present at their current location. A user's unique device ID is logged with the report and serves as a surrogate measure to distinguish between users.

Users are able to interact with the system in different environments, including a web browser, a desktop application, and on a mobile device (Figure 3.3). Users ultimately will be able to conduct a range of operations, such as obtaining spatially contextual information and model predictions, defining new model runs, exporting data, and submitting volunteered information

or reporting map/model errors; however for the purpose of our prototype, functionality is limited to the data querying, visualization, and reporting of tsetse occurrences.

Information Reliability

The traditional model of data reliability emphasizes the authority of the researcher and our belief that trained individuals will generate reliable, trustworthy data (Craglia, 2007). Within the context of a VGIS, we relax this assumption, instead qualifying reliability through data volume, the idea being that credible information will tend to be generated independently by more than one user (Flanagin & Metzger, 2008).

There are two fundamental approaches to assessing the reliability of crowd-sourced information. In the simplest case, information is assumed either credible or not until confirmed or rejected by a subsequent report. Under this model, all participants are treated equally with respect to their prior knowledge/skills; reliability is assessed by their peers through the creation of informal social networks [of trust] (M Bishr & Kuhn, 2007; Mohamed Bishr & Mantelas, 2008; Flanagin & Metzger, 2008; Metcalf & Paich, 2005).

The second model takes a more nuanced perspective of the user, taking into account the skill set of the person filing a report and their prior credibility. This approach is best approximated as a Bayesian model of data quality where the reliability of a report is dependent on the prior assessment of the user and previous reports made to the system (Crosetto & Rodriguez, 2001). If a number of prior reports are rejected, the individual is given a low reliability score that may lead to automatically rejecting any subsequent reports made (unless of course those reports are later confirmed independently). However, if the user has a history of high quality submissions that are routinely confirmed, they may be given a high credibility score, leading to automatic

accepting of the report into the database. Prior experience with the user is the crux of this model approach to data quality. Conati (2004) demonstrated this approach in evaluation of models of user affect; their study required that they be able to assess the reliability of self-reporting of emotional states.

$$User\ Rating = \alpha + \Delta \qquad (3\text{-}1)$$

α = prior score

Δ = change in score output from Figure 3.4

In our case study, we employ a simple decision model (Figure 3.4), that integrates social trust networks (e.g. Mohamed Bishr & Mantelas, 2008; Metcalf & Paich, 2005) and Bayesian methods (e.g. Conati, 2004; Crosetto & Rodriguez, 2001) to assess the accuracy of data and the reputability of volunteers who report on the presence of tsetse flies. To demonstrate our approach, we evaluate the reliability of volunteered information under two scenarios. In the first, a single reporter volunteers on multiple occasions. In this scenario, the reliability of the information is determined and the rating (Equation 3-1) can be associated with the reporter's ID. Subsequent reports are evaluated on the merits of the information as well as the reliability score of the reporter. In short, a reliable reporter is likely to submit reliable information. A record can also be approved if a user is deemed trustworthy under the model. This value is calculated over time as a measure of the number of reports that are confirmed versus contradicted.

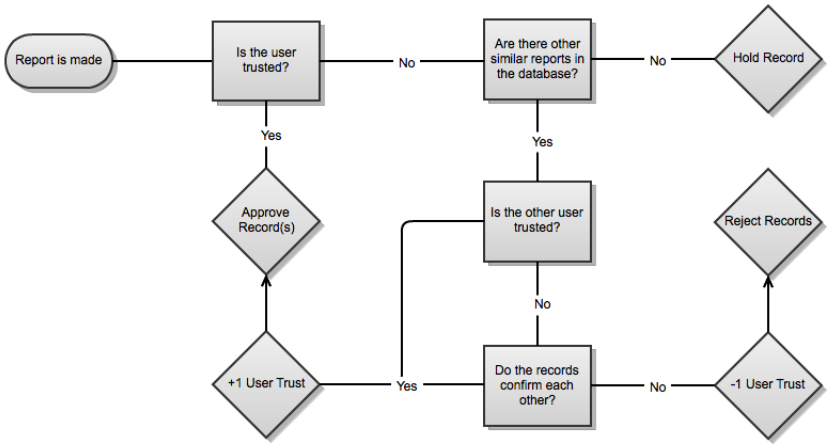

Figure 3.4: To assess the reliability of volunteered information, a report is evaluated in the context of a set of conditions. This figure presents a logical thought diagram for the application of the computation of reliability (3-2)

In the second scenario, several reporters each volunteer information only once. In this case, a reliability score cannot be computed or used to evaluate the reliability of the information; a report made under this scenario must be evaluated solely on the merits of the content. There are two components in the report (in addition to the information itself) that are used to assess reliability, context and authorship. In this scenario, authorship is of limited value since each reporter submits only once; we cannot conceptualize an author profile. However, we can evaluate the content of the information in the context of current predictions (of tsetse distribution) as well as prior years' predictions for the same period. A reliability score (Equation 3-2) is computed as a cumulative product of a user's rating, the number of times a cell is occupied in the previous time step in the current year, the number of times the cell is occupied on the same date in previous years, and the number of neighbors occupied in the previous time step. A

report is deemed credible if the score exceeds a certain threshold. This threshold is initially set to 5, but should be re-evaluated periodically to ensure data quality is maintained.

$$Reliability = \theta + \rho + \frac{\kappa}{4} + \gamma \qquad (3\text{-}2)$$

θ = user score

ρ = the number of times the cell was occupied previously, including the previous time step in the same year and the same time step in the previous year (max = 2)

κ = number of neighboring cells that are occupied (max = 8)

γ = number of supporting reports

Volunteered information under both scenarios can also be evaluated in the context of TED model predictions. Model predictions that take into account volunteered reports are compared to predictions made 16 days prior as well as to the distribution of tsetse at the same time in the previous year. We can reasonably assume that pockets of tsetse should maintain connectivity. If a report is made of tsetse occurrence in an isolated area (as measured by number of neighbors, κ) where no tsetse are predicted to occur, the probability of this report being accurate is low. If we were to incorporate these data into the model, the resulting predictions might dramatically impact the local reliability of the model outputs. By incorporating volunteered information into our prediction of tsetse distribution, we can better represent fine scale variability, particularly with regards to our ability to represent real-time distributions.

Utilizing Volunteered Information to Reduce Model Error and Uncertainty

Previously, we detailed our approach to assessing the reliability of volunteered information in the context of our case study. To illustrate the performance of the VGIS in making this determination, we simulate the reporting of tsetse occurrences across the study area. The simulated reports are generated for each iteration of TED model prediction. Additionally, we can illustrate reliability assessment under each of the two scenarios we detailed earlier; multiple reports from a single user or single reports made from multiple users.

The TED model outputs a binary raster at 250 m pixels which represents the minimum mapping unit for the predicted distribution of tsetse on the date the latest MODIS data product was captured; the predictions are not real-time estimates (always 30-45 days past) of tsetse distribution and are designed to underestimate the maximum distribution. Incorporating volunteered data allows us to fill in the gap, providing more up-to-date predictions (Figure 3.5). If the data reports are deemed reliable and differ from TED predictions, the cell represented in the binary raster for the previous time step is updated to reflect tsetse presence. The next iteration of TED will build on the 'corrected' raster.

Tsetse distributions expand and contract with seasonal climate. They achieve a minimum distribution at the peak of the dry season; these regions of minimum tsetse distribution are termed 'reservoirs' (DeVisser et al., 2010). Of relevance to our case study, we can use these minimum distributions as opportunities to 'reset' the model predictions so as to reduce any errors of omission that may have resulted over the previous season from incorporating volunteered information. In doing so, we can ensure that TED model predictions are reliable estimates of the minimum distribution of tsetse.

To test the functionality of the VGIS, we will simulate the implementation of the system to test the evaluation of volunteered information and the integration of this information with the DeVisser's tsetse distribution model. These simulations will primarily explore the assessment of volunteered reports of tsetse presence, under three scenarios. The first illustrates the case a report fills in a gap in the predicted distribution of tsetse (Figure 3.5a). Presumably, this is a product of error in the estimation of safety distribution. As stated previously, the TED model is designed to minimize errors of commission at the expense of added errors of omission. The assumption here is that the gap observed in the distribution is a product of this process. When a report is made, occurring within the bounds of this, the probability of that report being accurate is reasonably high. Therefore, our model should assign a high reliability score to that report.

The second case involves a report that connects two clusters of tsetse. In this case, we assume that patches of tsetse distribution should, for the most part, maintain connectivity in some way. If a report is made that establishes connectivity between patches (Figure 3.5b), there's a high likelihood that this report is reliable. Therefore, our model should assign a score that reflects this likelihood.

Finally, we simulate the case in which a report is made which places tsetse in a region that is isolated from predicted patches of tsetse distribution (Figure 3.5c). Since we have no prior reason to believe tsetse occur in this region based on model projections, there is a low likelihood that this report is true. Therefore, our model should assign a reliability score that emphasizes the extreme nature of this report. Through these simulations, we can identify the effective threshold for reliability.

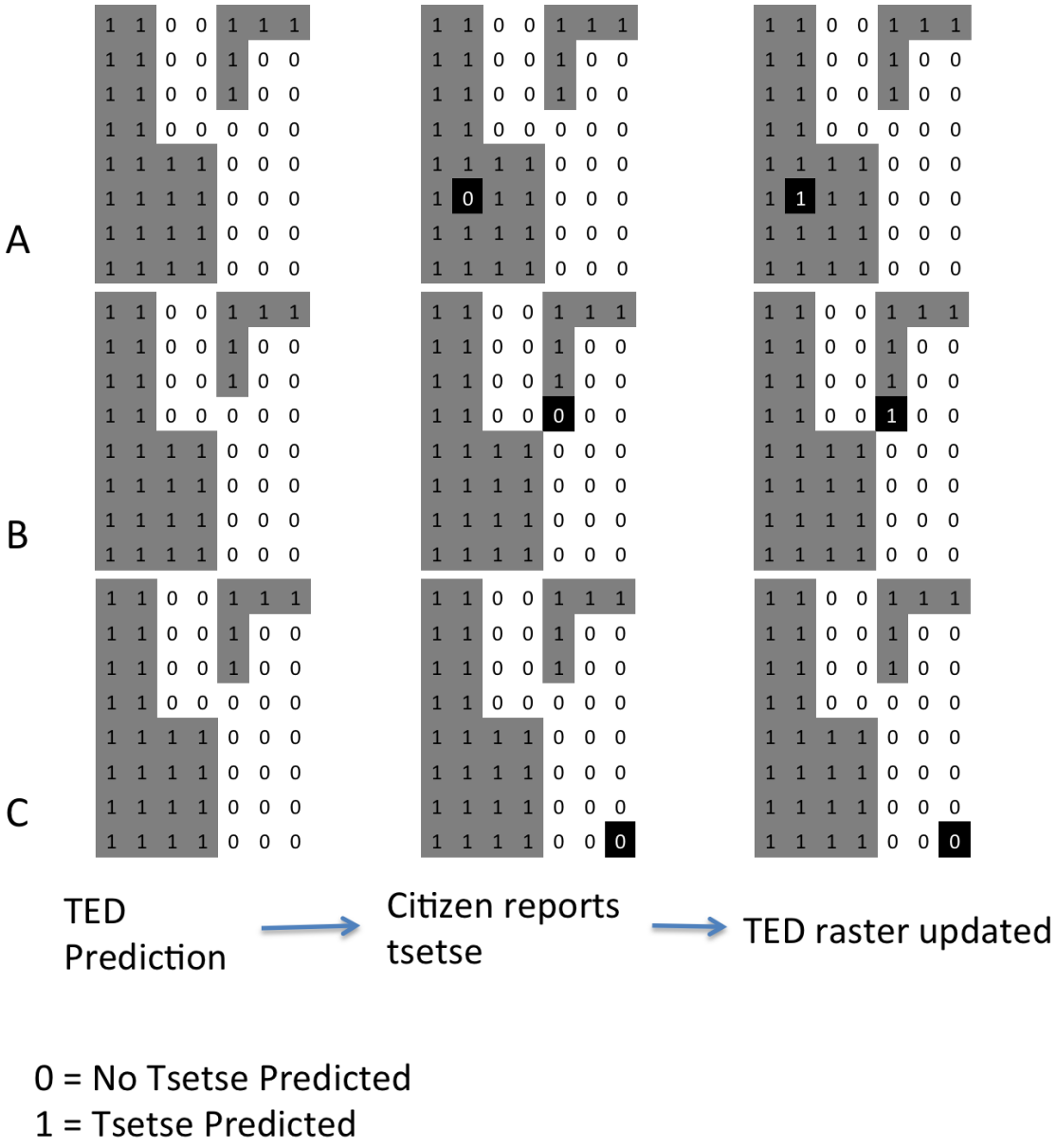

Figure 3.5: Users may volunteer reports of tsetse presence under a range of scenarios. (A) Illustrates the case where a report fills in a gap in a patch of tsetse, likely correcting an error in TED model predictions. (B) Illustrates the case where a report establishes connectivity between two isolated patches of tsetse. (C) Illustrates the case where a report of tsetse presence is spatially isolated from the predicted distribution of tsetse. In each case, a user is presented with a prediction of tsetse distribution from the TED model (Column 1). Users identify an error in the model, observing tsetse in an area where they are not predicted to occur, and submit a report (Column 2 - black box). The report is submitted for reliability assessment; if deemed reliable, TED model predictions are updated to reflect the new information (Column 3 – black box).

Conclusion and Limitations

Communication barriers present one of the most prominent barriers to disease surveillance programs. When communications between health care providers and regional health authorities break down under passive surveillance systems, there is a need to make attempts to directly collect disease incidence data directly. Yet, there are significant hurdles to implementing active surveillance programs (e.g., costs and logistics). By adopting concepts of crowdsourcing, public participation, and volunteered GIS, we can open the door for an intermediate solution for disease surveillance. Such an intermediate solution employs citizens to collect surveillance information, increasing the manpower available to collate the information. It may not then be necessary to dispatch health professionals to procure the data directly. Targeted campaigns can also be utilized to solicit participation on behalf of the public to assist in collecting surveillance data. Finally, our approach to assessing the credibility of volunteered information increases the utility and reliability of data obtained from these campaigns.

However, there are significant limitations to a full implementation of the VGI in our case study. The region in Kenya in which we are working is remote; there are significant challenges in terms of communication connectivity, reliable electricity, and necessary hardware; AMREF (American Medical and Research Foundation) and the African Conservation Centre (AAC) have made significant improvements to local infrastructure, but much more is required. Cost remains the most substantial hurdle for regional implementation. Our utilization of open source solutions mitigates, but does not eliminate this challenge. Absent assistance from international partners,

the likelihood of full implementation of the disease surveillance system will certainly remain in the domain of our scientific and development collaborators.

Connors et al. (2012) draw attention to the potential value of incorporating additional sources of information (e.g., Twitter, Flickr), aside from direct volunteering through a VGIS, to allow for increased participation; however, in doing so we would be introducing new types of uncertainty to the models. Our current design attempts to limit error exclusively to those of omission (i.e., we have tried to ensure that TED model predictions are estimations of the minimum area tsetse are distributed). In this way, we have greater confidence over the areas TED predicts tsetse to occur. When users volunteer reports of tsetse occurrence, they do so by providing GPS coordinates of their location (this is done in the background through the iOS application). Incorporating Twitter feeds, geo-tagged Flickr photos, among others, would on the one hand provide us with more information; however the cone of location uncertainty of that information is much greater and far less tractable. These sources of volunteered information represent important avenues for future development, particularly in the broader field of VGI, but at this time are beyond the scope of what we believe to be possible to include in our project.

Critical to the success of VGIS for disease surveillance is adequate public participation. Too few reporters can make it difficult to assess credibility and limits the conclusions that can be drawn from the information collected; however if incentives for participation are carefully considered, there can be a drive for individuals to accurately and reliably contribute to the system. Integration of volunteered information for disease surveillance, especially in low-income countries, can be used as an alternative to the high costs of active surveillance programs, which are often implemented in rural areas to learn more about disease prevalence. The prototype for

a VGIS outlined in this study demonstrates how technology and participatory science can advance passive disease programs to improve public health in needed parts of the world.

CHAPTER 4

USING META-QUALITY TO ASSESS THE UTILITY OF VOLUNTEERED GEOGRAPHIC INFORMATION FOR SCIENCE

Introduction

The scientific paradigm has evolved many times over the last millennium --- empirical, theoretical, and computational paradigms have dominated our identity as scientists. However, we are standing on the apex of another transition as technological and communications barriers are toppled (Elwood et al., 2013; Gray & Szalay, 2006), and the distinction between amateur and professional scientist is eroded. Neogeography characterizes the "blurring of the distinctions between producer, communicator, and consumer of geographic information"; the separation of scientist and layperson, expert and novice, is obscured as citizens engage in the generation of new knowledge (Goodchild, 2009). As citizens engage in Science, we need to reconsider our traditional notions of authority, expertise, and purpose.

Neogeography, a type of citizen science, has garnered a great deal of attention in the literature as we struggle to conceptualize the nature of "geographic expertise"; however, the involvement of citizens in science has long been established (Goodchild, 2009; A. Turner, 2006). Participatory science has sought to involve citizens directly in academic research and related exploits (Elwood, 2006b; Haklay et al., 2008; Tulloch, 2008) on the premise that citizens are more informed actors with respect to their local environment than researchers operating externally. Citizens are perceived to hold authority through experience and status, and are acknowledged

for their capacity to convey unique understanding, or indigenous knowledge (Elwood, 2006b; Elwood et al., 2013).

With the advent of Web 2.0 (O'Reilly, 2005, 2006) and the widespread availability of new technologies (Corbett, 2012; Haklay et al., 2008), citizens are increasingly exposed to geographical information. Citizens also increasingly volunteer spatially explicit (geographical) information that is of relevance or interest to them, often integrating this information with existing datasets, or mashups, utilizing it for their own gain (Miller, 2006; A. Turner, 2006). Goodchild coined the term "volunteered geographic information" (VGI) to refer to spatial data that is contributed by ordinary citizens, irrespective of their training in scientific methods (Goodchild, 2007a). The notion of VGI grew out of recognition of the limitations of traditional methodologies for adequately mapping and assembling spatial information around the world that provided both good coverage and fine temporal resolution (Elwood, 2008b; Elwood et al., 2011; Goodchild, 2008b). Goodchild further articulated the issue, drawing from the broader social science literature, postulating that the problem of data coverage can be mitigated were we to harness the "six billion sensors" on the earth (Goodchild, 2007a). He is of course referring to the earth's population of citizens, whom he notes begin to acquire spatial knowledge at a young age. Combined with Web 2.0 technologies, he asserts citizens can use the tools available to them to "volunteer" spatial knowledge of the world around them. While traditional spatial data infrastructures (SDIs) represent one-way communication models (where users only receive information from experts), VGI represents a model where communication flows in both directions without consideration of the role of the individual (Goodchild, 2009). As a framework, VGI encompasses citizen participation from a range of social classes and computing practices with

the express purpose of harnessing the collective intelligence (Connors et al., 2012; Elwood, 2006b); it builds on the notion that data can be shaped by social and political processes and an individual's expertise, context, and spatial awareness (Elwood, 2008a, 2008b; Elwood & Leitner, 2012; Harvey, 2012). Local knowledge is crucial to an accurate geographic description of communities and social groups, involving the citizen in the process of data collection; this understanding enables science to more accurately explain geographic phenomenon.

VGI in practice is now commonplace. Arguably one of the most successful, if not the most widely cited, outlet for VGI has been Wikimapia (Elwood et al., 2011; Goodchild, 2007a). Here individuals contribute knowledge of the physical, built environment around them in order to create as accurate a representation as possible. Other prominent examples include OpenStreetMap and Google Maps (utilizing the Google's API or Maps Maker) (Haklay et al., 2008). Recent events have also demonstrated the potential for VGI to assist in disaster response (Goodchild, 2010a).

However, the utility of VGI remains limited. In the context of the broader GIS literature, data quality has always been a concern (Elwood et al., 2011; Flanagin & Metzger, 2008). In the case of VGI, this concern is exacerbated due to the lack of expertise, or credibility, of the individual (Flanagin & Metzger, 2008). Given that VGI is user-generated information by non-experts, there is no quality assurance of the data (Craglia, 2007). It is viewed as non-authoritative data; as such, some have argued it should be informative, but not relied upon (Corbett, 2012). Others have raised concerns over the motivations of the individual, whether data is volunteered with an intent to inform or mislead, an act of digital vandalism (Tulloch, 2007).

Many approaches have been taken to assess the quality and reliability of VGI (e.g. Corbett, 2012; Elwood & Leitner, 2012; Flanagin & Metzger, 2008; Langley & Messina, 2013). The majority of these approaches thus far have been conceptual in nature, with few implementations of reliability assessments for VGI. The most common of these methods involves social trust networks and reputation models (Corbett, 2012; Maué, 2007). Under this approach, data quality is checked by other project participants for errors and inconsistencies. In this model, no single expert is tasked with reviewing each volunteered report. Another approach recommended has been to use existing data sets (collected using more authoritative methods) to check for inconsistencies in data. However, quality is not absolute; a datasets fitness-for-use is contextual and may have varying degrees of suitability for different users (Goodchild, 2008a). No single metric can be used to determine whether a data set is suitable across all ranges of potential uses. Thus, the context of a user's participation and interaction with VGI must be taken into account when considering accuracy/quality of VGI.

Given the concerns raised over the uncertainty of data quality in VGI, there is significant debate as to the utility of VGI for science. There are surprisingly few examples of academic initiatives utilizing VGI. (2011) inventoried 99 projects utilizing VGI and found only 3% to have academic affiliations. Perhaps one of the most prominent examples of VGI in science is the Audubon Society's Christmas Bird Count. This project has amassed a significant volume of volunteered data; however despite attempts to train volunteers in data collection, lingering questions of data quality, of reliability, have limited any analytical value and integration potential with authoritative datasets (Wiersma, 2010).

Flanagin and Metzger (2008) frame the issue of data quality of VGI in terms of credibility, and the primary components of trustworthiness and expertise (of the data source). The credibility (or believability) of VGI can be described objectively by traditional measures of data quality – the degree to which the information can be considered accurate, or as the subjective perception on the part of the consumer (Flanagin & Metzger, 2008). Credibility is related to notions of trust, reliability, accuracy, reputation, authority, and competence. However, for VGI to be useful for science, it is the traditional, objective "credibility-as-accuracy" measure demanded (Flanagin & Metzger, 2008). To fully quantify error in data, it is necessary to have a measure or to make assumptions as to the nature of the population being measured, to compare the distribution of data against the population as a whole. It is in this way we measure attribute accuracy, completeness, thematic resolution, and variability, to name only a few. Other measurements rely on feedback from measurement equipment, such as positional accuracy, temporal accuracy, spatial and temporal resolution, among others. As is often the case with VGI, the individual either operates without measurement equipment, or does not volunteer the additional metadata with their report. Participatory science and VGI Science (VGIS) often involve datasets for which the nature of the population is not immediately known. Therefore, a direct quantification of the error of VGI is only possible in a post-hoc analysis. However, it is the immediate benefit VGI can provide us that is of interest here and so we must develop a mechanism to evaluate the merits of VGI in real time (as it is contributed). In the absence of an ability to directly measure error and uncertainty parameters of volunteered data, we can use a surrogate measure, ***meta-quality***, a measurement of the collective quality of the data (van Oort, 2005).

The objective of our work here is to improve the perceived value of VGI for science by demonstrating a methodology for VGI data quality assessment. We accomplish this through a mechanism to explicitly assess the reliability of reporters based upon their respective VGI contributions.

To better illustrate our approach, we apply the methodology to a case study in disease ecology where we model the distribution of the tsetse fly, the principle vector of African Trypanosomiasis in sub-Saharan Africa. The "tsetse ecological distribution model" or TED, is based on an assessment of environmental characteristics critical for the persistence of the fly (DeVisser et al., 2010). The model simulates a 16 day cycle of the expansion and contraction of the fly population as response to the changing fundamental niche given the movement potential of the fly population (DeVisser et al., 2010). The model is a conservative estimation of the population distribution specifically minimizing errors of commission; therefore, the TED model is an estimation of the minimum extent of tsetse at each point in time. However, the model is reliant on a static land cover classification and makes no adjustment for error intrinsic to the model (DeVisser et al., 2010). The TED model produces estimates of the spatial distribution as binary outputs indicating presence/absence of the fly for each time period.

Potentially the most important contribution to incorporating VGI into a species distribution model of the kind here is the fact that we can explicitly address one component of model error (omission) without contributing additional error. TED was developed as a conservative model of the minimum expected distribution of tsetse. By incorporating VGI into the model results, we can effectively facilitate the population expanding over gaps of unsuitable habitat, either due to actual conditions or poor input data. It is known that microclimates provide

refuge for tsetse in areas where the habitat would be otherwise unsuitable (Ford, 1971; Moore & Messina, 2010). The spatial resolution of the underlying MODIS data miss these microsites and therefore omit these cells in the estimated distribution. Allowing the distribution to be updated based on the VGI, would allow us to more accurately reflect conditions as they exist reflecting sub-pixel dynamic that otherwise would not be possible. Incorporating VGI into the model results to expand the distribution can therefore reduce errors of omission without contributing additionally to errors of commission, thereby reducing total error, and thus improving data quality. Incorporating VGI into TED requires two distinct steps: 1) determine the reliability of the reporter to assess whether the VGI meets the threshold for acceptance, and 2) update the tsetse distributions by changing the binary tsetse presence/absence value for the cell (in which the datum is located) to 1 – indicating presence of the fly. In cases where VGI reflects the predicted distribution, no change is made.

Methodology

Here we undertake a series of experiments to illustrate the integration of VGI into a traditional analytical model. First, we explore the characteristics of VGI and its impact on model results. Second, we evaluate the sensitivity of the model to three types of error common to crowdsourced data. Finally, we explore the importance of reliability, as measured by a reputation score (Frew, 2007; Langley & Messina, 2013; Maué, 2007), in determining the threshold for accepting the data for inclusion in the model, under both static (a pre-defined score) or dynamic (a varying score) conditions.

To simulate the generation of VGI, we first consider the different kinds of reporters and the characteristics of the data they might contribute (Table 4.1). We identify four basic types of

reporters: 1) “always right”, 2) “always, intentionally wrong”, 3) “random”, and 4) “normal”. The “always right” reporter represents individuals who are judged, post hoc, to be highly reliable and the data they contribute are of high quality, often promoted to the role of moderator in online forums (Maué, 2007); there is no (or minimal) spatial or temporal error component to the data they contribute. The “always, intentionally wrong” reporter represents individuals who consistently, and/or intentionally provide erroneous data (M Bishr & Kuhn, 2007; van den Berg et al., 2011); these reporters are unreliable and the data they contribute should always be rejected. The “random” reporter represents individuals who generate data, falling on a random distribution, reporting tsetse, for example, at apparently random locations across the landscape (whether or not they are actually present) ignorant of underlying habitat conditions (Chow, 2012; D. Coleman & Sabone, 2010); due to the random nature of the reports, the data are therefore unreliable. Finally, the “normal” reporter represents the typical individual who volunteers information; the individuals have a high degree of credibility and the data are usually high quality (Flanagin & Metzger, 2008), but there is a spatial and temporal error component to the data they contribute. It is this type of reporter that we are most interested in evaluating reliability.

In the context of our case study, the simulated data for each reporter are based on habitat suitability criteria. In a real scenario, it is not possible to assess the accuracy of any report by itself; rather we can only assess the fitness-for-use of the data by placing it in application context and asking whether it is plausible (Grira et al., 2009; R. T. A. d. Groot, 2012). We simulate this by evaluating the data based on the likelihood of the data being correct given the underlying habitat conditions. To simulate the data, we identify a set of conditions that would be consistent with reports made for each reporter type, and use these conditions to identify points that can be used

in our sample data set. Table 4.1 fully describes the types of reporters and the set of conditions used to simulate data. For completeness, we explore the impact on the predicted occurrence of tsetse by simulating data, not only from the four reporter types but also from data generated from all combinations of habitat suitability criteria. It is based, in part, on these simulations that we identified the specific combination of criteria that would be used to render simulated VGI.

The simulated data are based on the underlying conditions present at each time step in the model, but not necessarily on the predicted occurrence for that simulation. For each set of criteria and combination thereof, we ran 100 simulations, identifying 100 points in each time step to serve as mock reports. Pooling these data points together results in 10,000 potential locations (some locations are represented more than once in the pool due to random selection in the simulations) for reports for each time step from which we randomly draw from when simulating reporters. This allows us to incorporate a minimum amount of stochasticity that would exist with reporters in a real-world scenario.

The basic TED model was implemented in GRASS based on the methods outlined by DeVisser et al. (2010) (see Appendix A for code). Building on our implementation of the TED model, we model the predicted distribution of tsetse, incorporating VGI, and evaluate the magnitude of the difference. Appendix B presents sample code for one simulation run (simulation 11). Each model was written in BASH, a UNIX shell-scripting language. The models were run on the HPCC cluster at Michigan State University for a total of 9,321 simulations representing an estimated 13,981 hours of computing time.

The normal reporter is defined as an individual who usually provides credible data, but has the potential to submit erroneous data. Incorporating these inaccuracies into the data

stream produces some degree of error in the model output. In reality, it is not possible determine the truthfulness of the data; therefore we must be able to determine the influence of error on the model output. The standard "normal" reporter is assigned an error rate of 10% (an arbitrary assignment); we measure the effects of this error by evaluating the impact on the resulting distribution when the "normal" reporter is assigned an error rate of 50%. As the data are constructed based on the combination of habitat suitability criteria, we evaluate introducing error into the model in different ways. Erroneous data are simulated by selecting points in areas of unsuitable habitat by shifting the location of the point (simulating positional error), or by holding the data until the following time step (simulating temporal error). A z-score is computed comparing each set of criteria against a simulation where points are selected at random, as well as a test of significance against the output from the TED model alone (no VGI data incorporated).

An assessment of the reliability of the VGI requires us to first generate a dynamic history for each reporter that reflects the plausibility of the data as determined by habitat suitability criteria. Each reporter is assigned a score, a measurement of their reputation, which is a product of these criteria (slightly modified from Langley and Messina 2013 to allow for negative changes in reputation). The index returns an ordinal measurement of reliability; it is not constraint to a particular range, rather is structured such that positive scores convey reliability. It is computed as:

$$Reliability = \theta + \rho + \frac{\kappa}{4} + \gamma \qquad (4\text{-}1)$$

$\theta =$ reporter's score

$\rho =$ the number of times a cell was previously occupied (-1 if 0)

$\kappa =$ the number of occupied cells in 4-cell neighborhood (-1 if 0)

γ = the number of supporting reports (-1 if 0)

We arbitrarily selected threshold scores of 5 and 8 for incorporation of the VGI into the TED model results. A paired t-test is used to measure the significance of adjusting the threshold and the potential importance the specific selection has on the resulting predicted occurrence. An alternative approach to the arbitrary assignment of scores is to determine the threshold at which reporter types can be distinguished from each other. We subject the history of reporter scores to a k-means test; this analysis tries to iteratively place each reporter into one of two clusters (we define these clusters to mean reporters of "plausible" or "erroneous" data). Cluster centers were defined at random from the set of scores for each test. As reporter scores increase over time, we expect it will take a certain number of model time steps before they will group properly. The average reporter score (for the plausible group) from 100 iterations can be interpreted as a reasonable threshold score under a static model.

Over time, the scores for reporters quickly exceed the small thresholds we set (reaching values > 100 at the end of the simulation), which results in unqualified acceptance of the VGI into the model. As such, we cannot detect or respond (within a reasonable time) to changing behavior among reporters, reflecting the inability of arbitrary, static thresholds to capture potential declining reliability and reputation of reporters over time. In the final set of simulations, we explore the possibility of using a dynamic score model, where the threshold for acceptance is drawn from the distribution of all reporter scores at each time step. For each simulation, we set a threshold equal to the 1st quartile score, mean, or 3rd quartile score from the distribution of all reporters' scores at that time. This allows us to include only the most reliable reporters from our total pool of participants, and the longer the model operates over time, the more reliable

our output becomes. The net benefit to the model should thus improve over time. Sets of paired t-tests are used to measure the significance of the difference in predictions from the three threshold models.

Results

In our case, the likelihood that tsetse are present in an area, the subject of the VGI in question, is correlated with the habitat suitability as measured by land cover, land-surface temperature, and NDVI. A reporter's score is a measurement of their reputation, akin to eBay's ratings system, which quantifies the history of the individual to perform in a manner that is perceived positively by their peers (Maué, 2007). We assume that if a reliable reporter contributes information that confirms another's data, the likelihood that datum being accurate is improved. However, this method of confirmation by peers necessitates a set of reporters who have attained a data history. Until a reporter attains a certain reputation, we do not have enough information to assess data quality; however, we have seen that different reporters themselves quickly separate from each other, allowing us to partition out individuals who are either reporting randomly (and thus frequently inaccurately) or are simply providing erroneous data intentionally. Partitioning out these two types of reporters alone immediately improves the quality of the contributed data.

Varying the criteria for spatially locating VGI greatly influences the overall impact on the predicted occurrence of tsetse, however the impact varies markedly from year to year due to environmental conditions and shifts in the habitat suitability (Table 4.2). Randomly locating points results in an overall 9.81% (4.23% − 13.66% for individual model years) increase in the number of cells in which tsetse are predicted to occupy over the time period in the model (recall that incorporating VGI into the TED model can only increase the prevalence of tsetse). However

targeting specific locations where habitat is suitable and at least one neighbor is predicted to be occupied (the criteria we assign to our normal reporter), yields an overall 0.03% (0.02% – 0.05%) increase in occupied cells. Notably, selecting suitable habitat alone as our criteria influenced the results the most, with an overall 14.06% (7.22% – 17.94%) increase in predicted occurrence. Likely this speaks to the design goal of the TED model to minimize errors of commission. Predictably, constraining report locations to only those cells in which tsetse are predicted to occur (the condition for our "always right" reporter) yields no increase in the predicted occurrence of tsetse over the base model. Selecting locations in which tsetse are not predicted to occur or where habitat is unsuitable (conditions for the "wrong" reporter or a component of error in the normal reporter, respectively) yields an overall 10.59% and 8.23% increase in the predicted occurrence. All criteria tested yielded significantly different results over the random model ($p < 0.001$ in each case).

In the static threshold score model, there was no significant difference in the overall predicted occurrence of tsetse ($p > 0.4$). However, utilizing a dynamic threshold score model resulted in significant differences between all three models (1st quartile, mean, and 3rd quartile) with p-values < 0.001 in each case. The overall increase in predicted occurrence was 0.8%, 0.43%, and 0.12% respectively; however, the results varied widely from year to year for both static and dynamic threshold models (see Table 4.3). [Note: simulations 8 through 12 in the table consider the cases for only normal reporters].

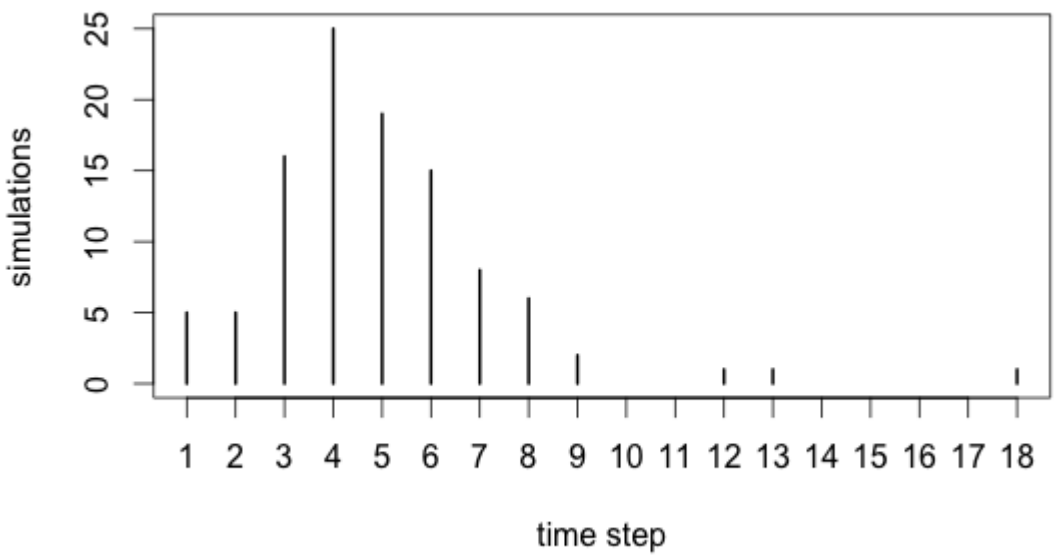

Figure 4.1: A frequency plot representing the time-step in which reporters cluster into two groups, for 100 replications of simulation 13

The four types of reporters cluster into two groups — see simulations 13 and 14 (Table 4.4) for the cases where all reporter types are considered. The four reporters are not fully distinguishable from each other at any time in our models (k-means with four clusters). Figure 4.1 presents the distribution curve (for all 100 replications) for the time step, at which point the reporters can be distinguished using a k-means clustering approach. For simulation 13, where a threshold score of 5 is used, the reporters can be separated, on average, in the 5th time step (mean = 4.93, median = 5). The average reputation score in the 5th time step is 10.87 for the "plausible" group. Reporters in simulation 14 (50% error rate) do not consistently cluster together into two groups.

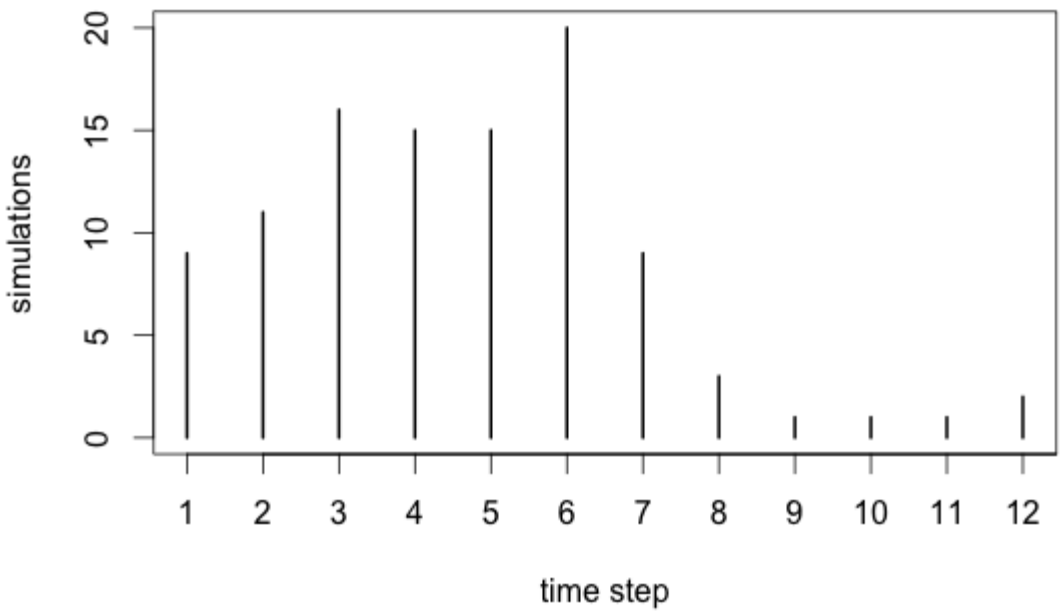

Figure 4.2: A frequency plot representing the time-step in which reporters cluster into two groups, for 100 replications of simulation 10

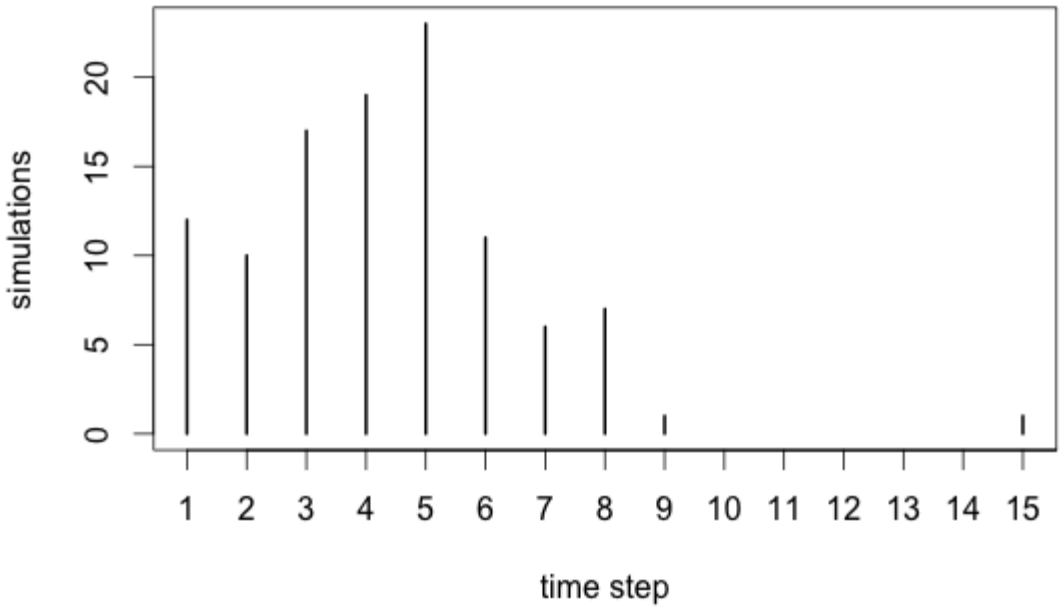

Figure 4.3: A frequency plot representing the time-step in which reporters cluster into two groups, for 100 replications of simulation 11.

Considering the dynamic score models, there were no significant differences in the time needed for reporters to group together. For the 1st quartile threshold score (simulation 10), reporters clustered into two groups, on average, in the 5th time step (mean = 4.61, median = 5). The average score for the "correct" reporters in the 5th time step was 18.87 (Figure 4.2). In the mean threshold score models (simulation 11), reporters clustered together in the 4th time step (mean = 4.32, median = 4). The average reputation score for reporters in this time step was 15.06 (Figure 4.3). Finally, for the 3rd quartile threshold score model, reporters clustered together in the 4th time step (mean = 4.21, median = 4) with an average reputation of 15.14 (Figure 4.4).

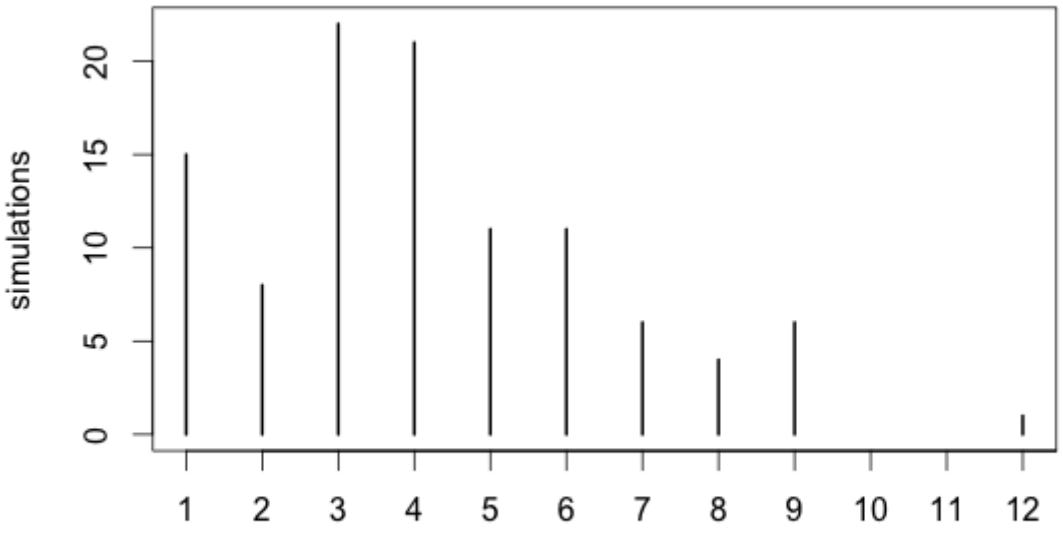

Figure 4.4: A frequency plot representing the time-step in which reporters cluster into two groups, for 100 replications of simulation 12

The nature of error (positional vs. temporal) introduced into our models through incorporating VGI did not appear to change the magnitude of the impact on predicted occurrence. This was also true when varying the magnitude of the error, at least for the range tested (5 – 25%). We did observe a significant increase in the predicted occurrence of tsetse

when the magnitude of the error introduced was 50% (where each reporter had a 50% chance of contributing erroneous data); introducing error of any type, though, results in a significant increase in the predicted occurrence compared to the case where no error is considered (simulation 4). Therefore, at least in our case study, the error introduced from VGI is not expected to a statistically significant effect on the prevalence of tsetse. This suggests that our models are resilient to the introduction of some erroneous data. Adaptations of our model to different studies will nevertheless necessitate an exploration of the role of introduced error from VGI to assess the resiliency of scientific models.

While the analysis reveals significant differences in the predicted tsetse occurrence from incorporating VGI into the TED model, global metrics are difficult to interpret given the importance of spatial structure in the dataset. To this extent, visualizing the structure of tsetse distribution patterns can lead to novel interpretations of the influence of VGI. Figure 4.5 - Figure 4.7 present the predicted distribution of tsetse over our study area (for simulations 10, 11, and 12 respectively); cell values indicate the proportion of time steps in the model (every 16 days between 2004-2006) where tsetse are predicted to occur, averaged across 100 replications. The distributions incorporating VGI closely mirror the base TED model with marked differences between core tsetse areas. These maps illustrate specific areas where VGI is particularly influential, likely due to the ability of tsetse populations to “jump” patches of unsuitable habitat.

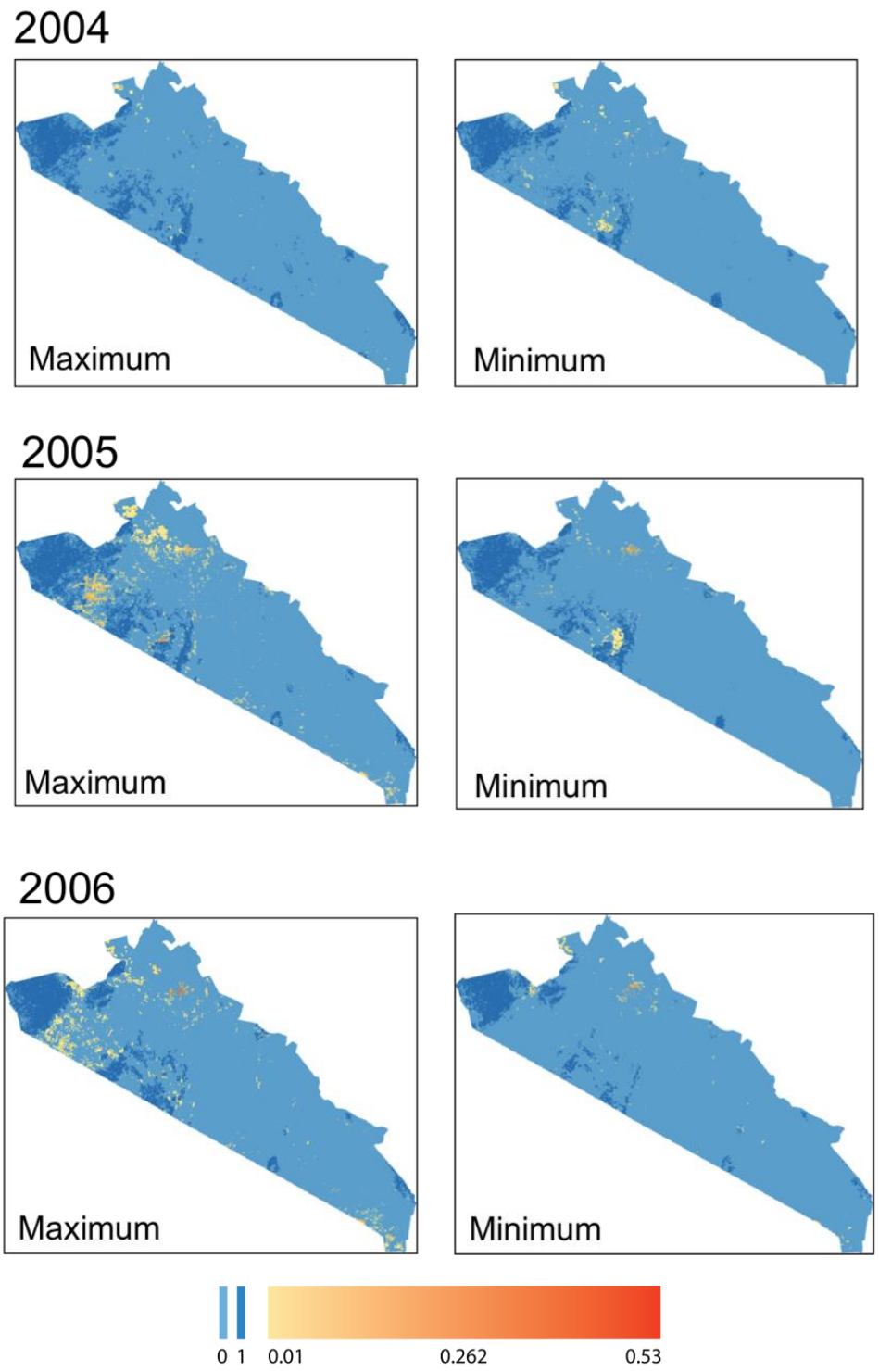

Figure 4.5: The theoretical maximum and minimum extent (respectively) for the distribution of tsetse for simulation 10. Values represent the proportion of time-steps in the model where tsetse were present; this is a rough approximation of the probability of tsetse occurrence.

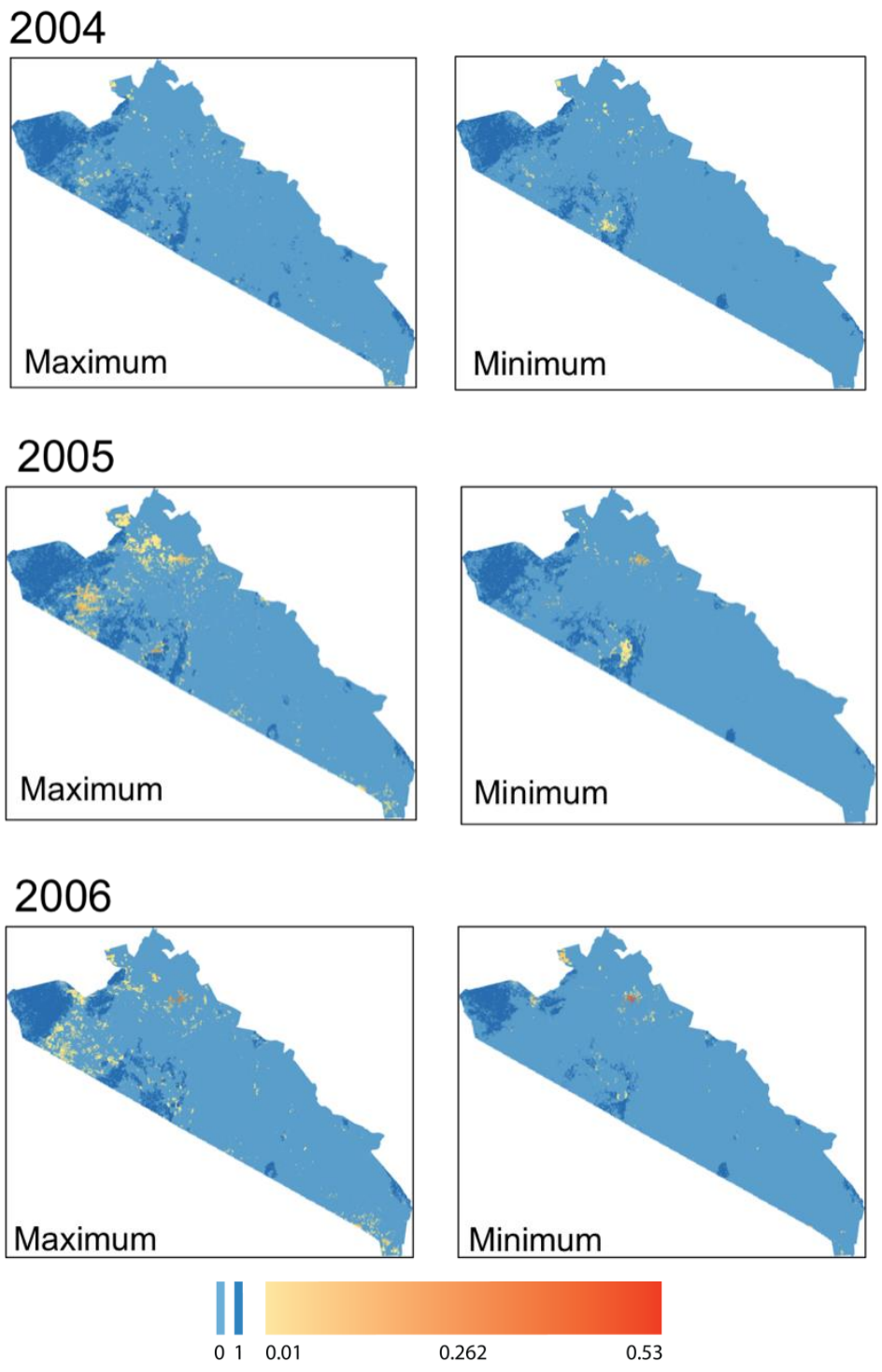

Figure 4.6: The theoretical maximum and minimum extent (respectively) for the distribution of tsetse for simulation 11. Values represent the proportion of time-steps in the model where tsetse were present; this is a rough approximation of the probability of tsetse occurrence.

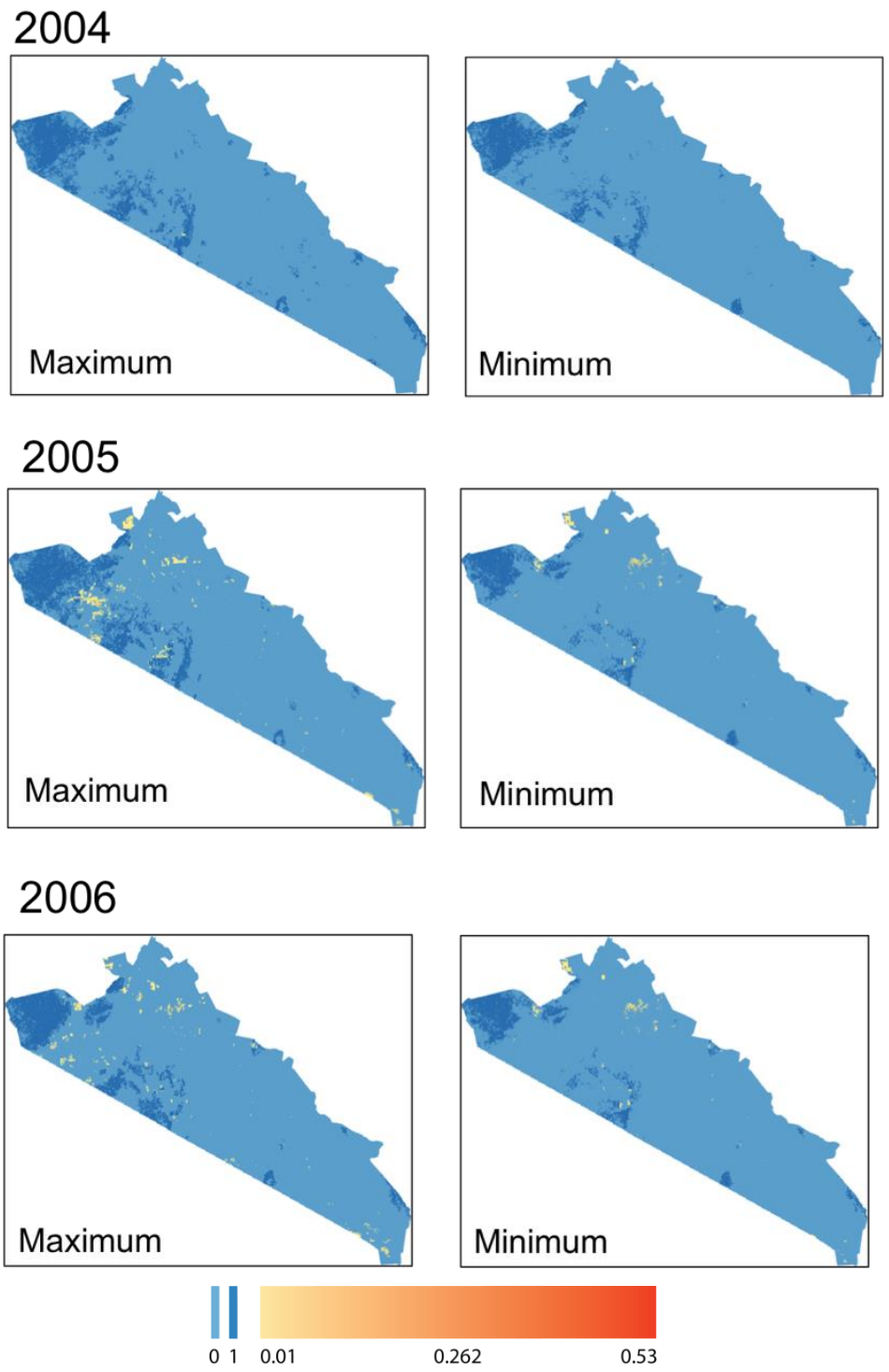

Figure 4.7: The theoretical maximum and minimum extent (respectively) for the distribution of tsetse for simulation 12. Values represent the proportion of time-steps in the model where tsetse were present; this is a rough approximation of the probability of tsetse occurrence.

Time is a significant factor to consider when evaluating the results of our models. In describing the output of TED model predictions, DeVisser et al. (2010) noted that tsetse populations tended to reach their maximum extent at the end of the long rains (ending the beginning of June). Populations tended to reach their minimum extent at the end of the cool dry season (mid- to late-October). This interpretation of tsetse population distributions comports with what is observed in my simulations, and is grounded in an ecological understanding of tsetse population dynamics.

Discussion

Volunteered geographic information can make valuable contributions to science, enhancing datasets from more authoritative sources. However, integrating VGI data necessitates assessing the error and uncertainty of those data. Direct quantification of data quality in this context is difficult; the traditional components (e.g. accuracy, precision, and variance) typically cannot be ascertained for VGI. It is critical for us to at least be able to qualify data quality, as it serves as the foundation from which we assess fitness-for-use. We have proposed using reputation or reliability (of the reporter) as a surrogate measure of meta-quality. As an initial assessment, meta-quality allows us to begin to break through the cloud of uncertainty inherent with VGI.

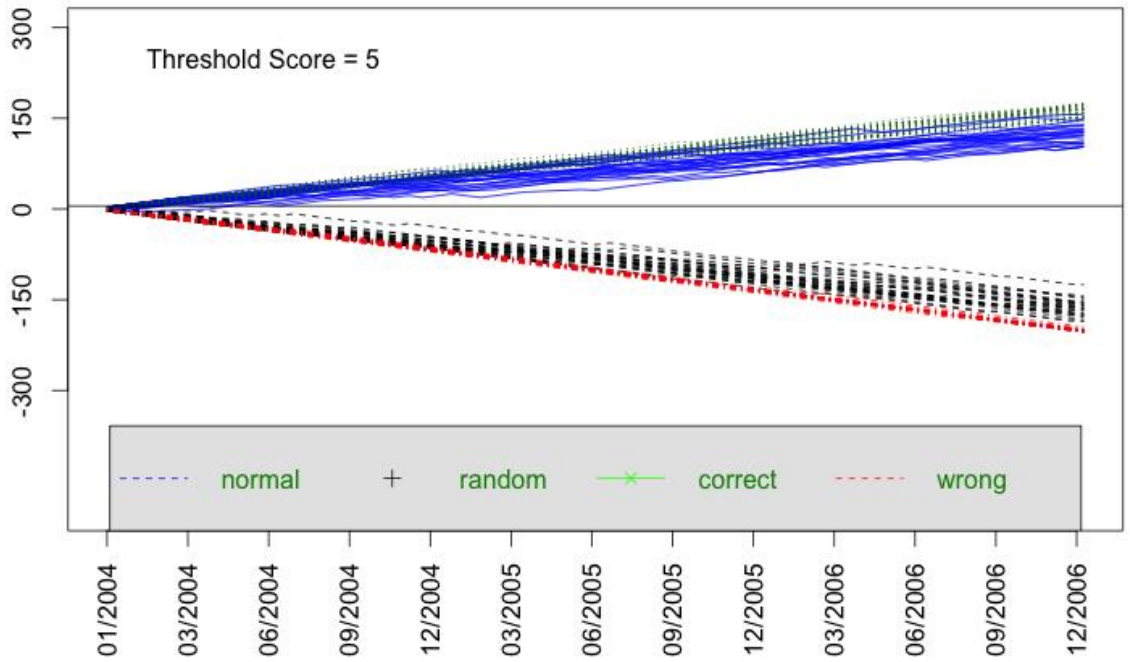

Figure 4.8: This figure overlays the scores of 100 reporters for simulation 8

We build on the power of the reliability/reputation assessment by considering a dynamic threshold scoring model. While we considered three different criteria for establishing a threshold (defined as the 1st quartile, mean, and 3rd quartile values in the distribution of reporter scores in each time step), we did not find a significant difference between them – as measured by an overall increase in the prevalence of tsetse in our models. In considering only those individuals whose reliability exceeds the mean score for all reporters, we only incorporate VGI from a subset of reporters we deem the most reliable. As scores improves for all individuals (regardless whether we have incorporated their data into our models), the threshold for acceptance/inclusion in our models also increases (approximately linearly in our models – Figure 4.8 shows the trend for one simulation). Over time, the quality of VGI data that we incorporate will improve, and the impact

of any erroneous data we have included should decrease. Most importantly, a dynamic threshold model facilitates detection of declining performance (of a reporter) and a rapid response to limit the acceptance of poor quality data.

The potential value of a means to assess data quality of VGI is immense. The strongest hurdle to fully utilizing VGI has been our inability to measure data quality and uncertainty. In demonstrating a valuation system for VGI (based on the reputation of reporters themselves), we have, in part, overcome this hurdle. To date, the utilization of VGI for science has been reserved for those cases only where the performance of reporters is controlled through training and guidance while closely monitoring the entire process from data collection to communication (Elwood & Leitner, 2012; Tulloch, 2008; Wiersma, 2010). But this runs contrary to many of the perceived strengths of VGI, the dissolution of traditional roles (Elwood et al., 2013; Goodchild, 2007c, 2009; Haklay et al.), and the establishment of a two-way communication model for geographical information (Goodchild, 2007a). Projects that have tried to embrace VGI have done so under the old model of participatory science, and thus are subject to all the perceived and actual limitations (Elwood, 2006b; Miller, 2006). Many factors influencing quality remain difficult to measure, including rates of participation and motivation to participate; the value of VGI cannot be fully appreciated until we can reliably assess these factors and the role they play in determining data quality.

It is our position that incorporating VGI into standard scientific models, particularly those where available data are sparse, can significantly improve the performance of the models and the predictive or explanatory power of the results. Consider the case of “Digital Earth”; first conceived by then US Vice-President Al Gore, it represented a push to represent the planet in

high-resolution, multi-dimensional space for the primary purpose of improving our predictive capabilities of Earth's ecosystems (Craglia, 2007; Craglia et al., 2012). Twelve years later, significant gaps still exist, particularly in terms of our capacity to collect certain types of data of sufficient quality and resolution (Craglia et al., 2012). Harnessing the collective power of earth's citizens, the aggregate power of "six billion sensors", we can make significant strides to improving the predictive capacity of our models through incorporating new types of information (Goodchild, 2007a). Therefore, it is critical we continue to explore ways to assess the credibility of VGI, to embrace the new geographical traditions, while respecting the scientific paradigms of the past.

Table 4.1: Reporter types and the criteria used to simulate their behavior

Reporter	Type	Model Criteria
1	Always right	Tsetse predicted
2	Always, intentionally wrong	Tsetse not predicted, habitat unsuitable
3	Random	Spatially random
4	Normal	Suitable habitat + one occupied neighbor

Table 4.2: Simulation results for simulated conditions. Values represent percent increase over the base TED model

		% Gain				Variance			
Sim	**Criteria**	**Overall**	**2004**	**2005**	**2006**	**Overall**	**2004**	**2005**	**2005**
1	Random	9.81	4.23	13.66	11.85	144.02	76.83	105.73	106.62
2	Suitable habitat	14.06	7.22	17.94	17.58	108.26	73.41	80.88	77.11
3	One neighbor	0.29	0.17	0.39	0.32	93.37	27.62	66.28	65.48
4	Suitable habitat + one neighbor	0.03	0.02	0.04	0.05	19.65	10.86	16.52	13.97
5	Tsetse present	0	0	0	0	0.01	0.01	0.01	0.01
6	Tsetse not present	10.59	4.78	14.57	12.71	128.16	75.51	97.03	91.45
7	Habitat unsuitable	8.23	3.46	11.75	9.66	138.43	79.15	112.02	109.58

Table 4.3: The percentage increase in the prevalence of tsetse over the base TED model for simulations 8-12

		% Gain				Variance			
Sim	**Score**	**Overall**	**2004**	**2005**	**2006**	**Overall**	**2004**	**2005**	**2006**
8	5	1.28	0.27	1.94	1.68	139.56	46.8	113.56	100.52
9	8	1.22	0.23	1.88	1.6	138.6	44.09	112.57	99.1
10	1st quartile	0.8	0.13	1.15	1.2	120.62	37.76	92.9	95.07
11	Mean	0.43	0.05	0.6	0.68	109.46	28.92	83.06	83.27
12	3rd quartile	0.12	0	0.14	0.24	77.36	10.47	55.49	66.11

Table 4.4: The percentage increase in the prevalence of tsetse over the base TED model for simulations 13-20

		% Gain				Variance			
Sim	**Error type**	**Overall**	**2004**	**2005**	**2006**	**Overall**	**2004**	**2005**	**2006**
13	10%	1.38	0.39	2.07	1.74	139.44	48.86	116.26	95.18
14	50%	5.23	1.88	7.94	5.95	144.9	70.59	124.32	99.84
15	Spatial shift 5%	1.39	0.39	2.13	1.7	137.33	49.43	112.58	92.83
16	Spatial shift 10%	1.39	0.44	2.22	1.52	142.05	54.59	118.36	97.97
17	Spatial shift 25%	1.41	0.44	2.18	1.65	131.66	50.61	108.22	95.31
18	Temporal shift 5%	1.46	0.43	2.21	1.79	135.71	50.68	112.63	97.89
19	Temporal shift 10%	1.54	0.45	2.4	1.81	149.88	52.54	124.28	97.73
20	Temporal shift 25%	1.61	0.5	2.47	1.9	148.95	55.65	119.68	97.55

CHAPTER 5

SUMMARY AND CONCLUSION

Introduction

With the emergence of new technologies, the availability of new web based tools, and the prevalence of geographic information, the traditional distinctions between citizen and scientist, consumer and producer of knowledge are eroding and the line between them is blurred. In this era of eScience and Neogeography, citizens are embracing new freedoms to not only combine datasets through mashups, but to engage in the generation of new knowledge. These activities, traditionally the exclusive purview of the academy, necessitate new approaches for management. The sheer volume of information being generated requires new approaches for storage and retrieval. Furthermore, we need to devise new methods to assess the quality of these data, particularly as traditional objective measurements of error are no longer communicated through metadata, as has been the case with traditionally collected spatial information. Data quality metrics are necessary for us to qualify the credibility of the information before us to determine a dataset's fitness-for-use in a particular analysis.

Summary of Main Findings

Objective 1: Address three recurring problems with spatial data management: scalability, reliability, and security by:

1. Communicating a conceptual model for a comprehensive open-source computing environment that promotes the efficient organization, storage and retrieval of disparate data.

2. Extending the discussion of spatial databases by presenting a model framework for a spatial DBMS that rigorously and consistently manages both spatial and nonspatial data.

Effective data management in the age of Neogeography is an emerging problem. Today's data management strategies not only need to account for the volume of information, but also the modes by which it is acquired and disseminated. The rigorous standards to which we have traditionally subjected data in terms of quality and completeness are often unfeasible. As citizen science initiatives expand, volumes of spatial data are generated, much of which would traditionally be considered incomplete, either because it is not accompanied with comprehensive metadata or does not consistently report all attributes. Traditional DBMS afford a great deal of flexibility, but they often do so through a lack of built-in checks for consistency and quality. Modern spatial data infrastructures (SDIs) must address this changing nature of data.

Three recurring problems with data management are routinely cited in the literature. *Scalability* refers to the ability of SDIs to manage interaction with data at multiple spatial, temporal, and thematic resolutions (Shekhar & Chawla, 2003). *Reliability* speaks to the need for mechanisms to protect against loss of data either through inadvertent changes or malicious behavior; however, it must do so without impeding interaction with the data (Devillers & Jeansoulin, 2006; Shi et al., 2002). Finally, *security* refers to the need to strictly control the dissemination of various types of data (some of which may contain personally identifiable, private information) without impeding the ability to use the information. Furthermore, it must facilitate management of ownership; in particular, ownership of VGI is still a commonly cited concern in the literature. Such safeguards are commonly required in academic or government

institutions. The implementation of the DBMS proposed directly addresses each of these problems.

The aim of chapter 2 was twofold: first, outline the problem of spatial data management in the age of Neogeography and the volume of data generated by citizens as VGI; second, demonstrate an approach for the handling of spatial data of varying types in a manner that facilitates efficient querying, storage, and retrieval. The ways in which we interact with spatial data have changed. Thus, the goal here was to develop a DBMS that afforded flexibility and extendibility to allow for the integration of future advances in technology.

It was important to utilize only open-source software solutions so as to aid its adoption by communities and organizations seeking to incorporate VGI into their projects. The open-source software is also extendible, fulfilling one of the primary objectives. The DBMS facilitates efficiency in the storage of spatial data by incorporating the WKTRaster module to the PostGIS libraries[11]. WKTRaster extends the functionality of PostGIS to allow for the raster-based imagery to be stored in the database just as has been possible for vector-based data. This makes it possible to query imagery directly at multiple levels, and return it at the desired resolution and extent. The decision to utilize PostgreSQL as the core of our DBMS also addresses the second and third concerns raised: reliability and security (although the specific safeguards were not discussed in detail until Chapter 3).

[11] Since publication (Langley & Messina, 2011), WKTRaster has been fully implemented in PostGIS 2.0. It is now referred to in the documentation simply as PostGIS Raster.

Objective 2: Demonstrate the utility of VGI by:

1. Describing a prototype for the utilization of VGI to enhance disease surveillance programs.
2. Articulating an approach for integrating VGI into a traditional species distribution model.

The last decade has seen a dramatic rise in the availability of spatially explicit geographical information and the frequency with which non-scientists are exposed to and utilize it. This is in large part a result of new modes of communication and interaction via web based applications, a phenomenon often referred to as Web 2.0 (Corbett, 2012; Haklay et al., 2008; O'Reilly, 2006, 2007). In Chapter 3, I made reference to two prominent and often cited examples of GIS in a Web 2.0 environment; these include Wikimapia and OpenStreetMap (Goodchild, 2007a; Haklay & Weber, 2008). Both sites allow for users to volunteer information about their environment. These initiatives are distinctly spatial in nature and generate volumes of new geographical knowledge, yet operate entirely outside the purview of the academy. I refer to these types of initiatives (those devoted to the volunteering of geographic information) as a type of volunteered GIS or VGIS.

The potential for scientists to tap into the VGI data collective, and to make use of new tactics for data collection is attractive; Goodchild (2007c) makes reference to harnessing the power of earth's billions of sensors as a means of knowledge production. However, there are significant concerns that have been raised with regards to the credibility of VGI (Flanagin & Metzger, 2008). As such, there have been very few illustrations of VGI being utilized in scientific research (Elwood et al., 2011).

The aim of Chapter 3 was to elucidate the value of VGI when incorporated into a traditional species distribution model. Building off Chapter 2 (which outline the framework for the SDI) we demonstrated the framework for incorporating the VGI. Our primary objective in this chapter was to propose an approach to assessing the reliability of reporters of VGI which can serve as an indicator of the quality of the data itself (Maué, 2007). We propose a quantitative metric, Reliability, that is computed as the product of underlying model conditions; it is based in part on the notion that VGI can be assessed by comparing it against datasets of known quality (Koukoletsos et al., 2012). This approach has been cited in the literature as a possible method, and implemented in part in several studies as a post-hoc analysis (D. J. Coleman & Georgiadou, 2009; Corbett, 2012; Frew, 2007; Maué, 2007). We propose a threshold be established below which a reporter's data are excluded/rejected.

Finally, we outlined our approach to integrating VGI into a traditional species distribution model. In our case study, we build on a model developed by DeVisser et al. (2010) that models the distribution of tsetse in Kenya on the basis of remotely sensed imagery as inputs. The implementation in our study is such that the assessment of reliability is made at the moment data is submitted; therefore immediate decisions can be made as to incorporating or excluding the data from the communicated results.

Objective 3: Address lingering concerns of credibility and data quality in VGI by:

1. Proposing a method for dynamically assessing the reliability of reporters of VGI.
2. Assessing the impact of incorporating VGI of varying quality into a traditional species distribution model.

In order to firmly establish the utility of VGI for science, it's critical we be able to judge the credibility of the information. The intent of Chapter 4 was to elucidate an approach for the measurement of data quality for VGI. In reviewing the traditional metrics used to assess data quality, it is noted that objective assessments are typically not possible for VGI (Flanagin & Metzger, 2008); therefore we propose a subjective measure that uses the reliability of the reporter (to produce credible data) as a surrogate measure to communicate data quality (Maué, 2007). This is drawn, in part, from the eBay model that references reputation as an indicator of the probability of a positive future transaction. We reference meta-quality (an officially recognized metric) as the means by which we can communicate credibility of the VGI.

In the second part of this paper, we illustrate the impact VGI can have on the output of a traditional species distribution model. Since all data are prone to include error, and it is our assertion that credibility of the data communicates an estimation of error, we demonstrated the impact different types of error can have and at varying magnitudes. Finally, we demonstrate that the model outlined in Chapter 3 for assessing the reliability of reporters does in fact allow us to partition out reporters perceived as credible and those who are not. We introduced the notion of a threshold score model, which represents the point at which an individual is deemed reliable, and have their data incorporated. We noted that the static model (one set threshold) did not allow us to effectively account for changes in reporter behavior, leaving the system vulnerable to "digital vandalism" (Tulloch, 2007). However, a dynamic model that adjusts the threshold to include only the most reliable of reporters, allows the system to respond to changing behaviors, mitigating the impact of malice and the incorporation of erroneous data.

Theoretical Implications of this Dissertation

As summarized in the main findings, this research has contributed to the discussion of uncertainty of data quality for VGI. Numerous studies have articulated the issue, raising concerns for utilizing VGI in the absence of effective data quality assessments (Corbett, 2012; Flanagin & Metzger, 2008; Maué, 2007; Tulloch, 2007). In this dissertation I have demonstrated one approach to evaluating VGI utilizing subjective measures of credibility. Numerous theoretical studies have laid the framework for this approach, specifically correlating a reporter's reputation - their reliability to contribute credible information - with the quality of VGI and crowdsourced data (D. J. Coleman & Georgiadou, 2009; Elwood et al., 2013; Maué, 2007). Other studies have suggested that the credibility of individual datum is irrelevant as the collective contributions of reporters, averaged together, can produce information that is highly credible (Elwood & Leitner, 2012; Haklay et al., 2010). An extension of Linus' Law for Neogeography states that the number of participants can be correlated with the credibility of the resulting information set (Haklay et al., 2010). A popular expression of this law by Raymond Raymond (1999) says, "Given enough eyeballs, all bugs are shallow". The point here is that errors are less influential as the number of participants in a project increases.

This dissertation specifically addresses issues raised by Flanagin and Metzger (2008), Tulloch (2007), and Craglia (2007) regarding the lack of quality assurance available for VGI. I further propose that as the reliability of VGI data quality metrics is recognized, Corbett's (2012) critique of VGI as non-authoritative should no longer be a barrier to utilizing VGI for science. It is my assertion that the credibility of any information or data is directly related to an assessment of data quality. While the data quality of VGI is most commonly communicated subjectively,

evaluating quality using objective methods allows for us to more directly make statements with regard to confidence, particularly when evaluated in the context of more traditional data collection methodologies.

Recommendations for Future Research

Future research should focus on a number of specific objectives that have been conveyed in this dissertation. We must further the development of web-based mapping capabilities to incorporate new tools for efficient querying of spatially explicit information. National SDIs have, to a limited extend, implemented web-based technologies; however their purpose is tailored to the expert scientist for data analysis and visualization. The same functionality has not been translated to the web-based GIS portals used by non-scientists (Craglia et al., 2012). As citizen scientists engage in more advanced geospatial analyses, they demand the same functionality afforded to academics. Craglia (2012) describes a vision in which farmers are able to use GIS tools to monitor crop status and yield, perform long-term risk analysis on market prices and trends, receive early warning of extreme weather and other environmental dangers, and overall improve the management of their resources. However these capabilities require enhanced communication and technological improvements not currently available to the general public. The Digital Earth project is building off of the achievements in citizen mapping technologies to improve access of non-scientists to spatial data and their ability to engage in more advanced spatial tasks (Goodchild et al., 2012).

There has been significant discussion in the literature surrounding the lack of credibility inherent in VGI (e.g. Elwood et al., 2013; Flanagin & Metzger, 2008; Goodchild & Li, 2012). Numerous approaches have been proposed for assessing data quality, but most are still in the

theoretical stage (Goodchild & Li, 2012). Few studies have demonstrated the implementation of quality assessments for VGI, and to my knowledge, all are post-hoc assessments. Linus' Law is often cited as justification for leniency in credibility assessments of VGI based on the correlation observed between the number of participants in a project (or volunteers) and the aggregate quality of the information (Goodchild & Li, 2012; Haklay, 2010; Raymond, 1999). However, this approach precludes individual assessments of quality; if integration of VGI (such as was observed in our case study) necessitates evaluating uncertainty of each datum, the approach is not applicable.

Current approaches demonstrated in the literature assess VGI, not on its qualities, but rather in comparison against datasets of known quality or on the bases of tertiary qualities (Cipeluch et al., 2010; Girres & Touya, 2010; Koukoletsos et al., 2012). We need a methodology that expedites an assessment of VGI objectively on its own merits, and in real-time. Despite its limitations, VGI has the potential to serve as a means of acquiring high quality data for little cost by harnessing the collective power of the crowdsourcing community. Therefore, we must develop new approaches to objectively evaluate VGI on its own merits so we can quantify error and uncertainty in the data.

The method proposed in this dissertation (Chapter 4) facilitates a credibility assessment using a surrogate metric, meta-quality, which correlates the reporter's reputation with the credibility of the information they contribute. However, the approach has only been demonstrated theoretically; it must now be subjected to experimentation in real-world applications to quantify the correlation between reporter reputation/reliability and the

credibility of the data they volunteer on an individual basis and not necessarily as a part of the collective.

Finally, there is a need to explore the application of subjective valuations of VGI on existing crowdsourced datasets so that the data can be subjected to the same scientific analyses available to more authoritative data sets. This has been undertaken, to a limited extend, on the Wikimapia and OpenStreetMap datasets (Haklay, 2010), but we do not have established methodologies for post hoc valuations.

There is significant interest in the literature for methodologies that will allow incorporating and utilizing VGI and other crowdsourced data for science (e.g. Craglia, 2007; Goodchild, 2009; Haklay, 2010). In this dissertation, I broadly set out to advance the utility of VGI for science. I articulated a data model for an SDI (Chapter 2) that is tailored to and supports the integration of VGI. The model I have proposed takes steps to improve the communication of data quality metrics that are commonly missing with crowdsourced data. In Chapter 3, I demonstrated how VGI could be integrated with a more traditionally authoritative dataset. Here I expounded on the potential value this additional data stream can provide in the context of the limitations with the existing datasets. Finally in Chapter 4, I responded to those who have raised concerns with a lack of quality assurance in VGI by proposing a method to assess the credibility of volunteered data. I did this by evaluating the context of the information and related this back to the reliability of the reporter. Subsequent contributions by this reporter are then assessed in a Bayesian model where the credibility of the data is a function of its context as well as the prior performance of the reporter. Aggregately, the methods introduced and the points made with

regard the specific advancements (both theoretical and practical), all advance the affirmed goals set out in this dissertation.

APPENDICES

Appendix A

BASE CODE SIMULATION

```
# Base TED simulation model
# Author: Shaun Langley
# Last Modified 3/25/2013

years="2003 2004 2005 2006"

times="001  017  033  049  065  081  097  113  129  145  161  177  193
     209  225  241  257  273  289  305  321  337  353"

# The model will loop through each year and time value noted here.

g.mapset -c baseSim1 # create a seperate mapset for each simulation to
     isolate output from each other to prevent possible data overwrite
g.mremove rast=* -f
g.region Kenya
r.mapcalc "distrib.tmp=initDistrib"

for year in $years; do
     for i in $times; do
          NDVImap=`g.mlist type=rast pat="bin${year}_${i}_NDVI"`
          LSTDaymap=`g.mlist type=rast
     pat="bin${year}_${i}_Day_LST_250m_Terra_16day"`
          LSTNightmap=`g.mlist type=rast
     pat="bin${year}_${i}_Night_LST_250m_Aqua_16day"`
          LULCmap=`g.mlist type=rast
     pat="bin${year}_LULC_Type_1_250m"`
          r.mapcalc "suitable.$year.$i=($NDVImap * $LSTDaymap *
     $LSTNightmap * $LULCmap)"
          r.neighbors input=distrib.tmp output=distrib.grown.tmp
     size=5 method=maximum --o
          r.mapcalc "distrib.tmp=(distrib.grown.tmp *
     suitable.$year.$i)"
          g.copy distrib.tmp,distrib.$year.$i
     done
done

g.remove distrib.tmp
g.remove distrib.grown.tmp

g.region zoom=studyarea
r.mask studyarea

r.series input=`g.mlist pat=distrib.200[4-6].* sep=,`
     output=baseSim_average method=average
r.series input=`g.mlist pat=distrib.200[4-6].* sep=,`
     output=baseSim_sum method=sum
```

```
r.series input=`g.mlist pat=distrib.2004.* sep=,`
     output=baseSim_2004_average method=average
r.series input=`g.mlist pat=distrib.2004.* sep=,`
     output=baseSim_2004_sum method=sum
r.series input=`g.mlist pat=distrib.2005.* sep=,`
     output=baseSim_2005_average method=average
r.series input=`g.mlist pat=distrib.2005.* sep=,`
     output=baseSim_2005_sum method=sum
r.series input=`g.mlist pat=distrib.2006.* sep=,`
     output=baseSim_2006_average method=average
r.series input=`g.mlist pat=distrib.2006.* sep=,`
     output=baseSim_2006_sum method=sum

r.mask -r

curl https://prowlapp.com/publicapi/add -F
     apikey=2daeeaf780b2e94281d1089752c6698da81434a9 -F
     application="GRASS" -F description="TED base simulation complete"
```

Appendix B

SIMULATION 11 CODE

```
#########################
# simulation 37 - all reporters with score tracking -- threshold=mean
# Author: Shaun Langley
# Last Modified 3/25/2013

g.mapset -c expSim37_$rep
g.mapsets addmapset=baseSim1
g.mremove rast=* -f
g.mremove vect=* -f
g.copy studyarea@PERMANENT,studyarea
g.region Kenya
r.mapcalc "distrib.tmp=initDistrib"

years="2004 2005 2006"

times=`seq -w 1 16 365`

ptslasttime=2003_353
lasttime=2003.353
lastyear=2003

for year in $years; do
for i in $times; do

r.neighbors input=distrib.tmp output=distrib.grown.tmp size=5
     method=maximum --o
r.mapcalc "distrib.tmp=(distrib.grown.tmp * suitable.$year.$i)"

# Collect the first point
if [ `shuf -i 1-10 -n 1` -le 9 ]; then
v.extract input=sim4_pts_${year}_${i}@results output=pts_selected
     random=1 --quiet --o
else
v.extract input=sim17_pts_${year}_${i}@results output=pts_selected
     random=1 --quiet --o
fi

# each reporter has a 10% chance of being wrong.  So pick points
     individually until there are 100 points with each reporter
     possibly geting it wrong.

while [ `v.info pts_selected | grep -e "Number of points:" | awk '{
     print $5 }'` -lt 25 ]; do

if [ `shuf -i 1-10 -n 1` -le 9 ]; then
```

```
v.extract input=sim4_pts_${year}_${i}@results output=pts_selected_4
     random=1 --quiet --o
v.patch input=pts_selected_4 output=pts_selected -ae --o
else
v.extract input=sim17_pts_${year}_${i}@results output=pts_selected_17
     random=1 --quiet --o
v.patch input=pts_selected_17 output=pts_selected -ae --o
fi

done

g.remove vect=pts_selected_4
g.remove vect=pts_selected_17

# random reporter
v.extract input=sim1_pts_${year}_${i}@results output=pts_random
     random=25 --o

# correct reporter
v.extract input=sim15_pts_${year}_${i}@results output=pts_correct
     random=25 --o


# wrong reporter
v.extract input=sim17_pts_${year}_${i}@results output=pts_wrong
     random=25 --o

v.patch input=pts_random,pts_correct,pts_wrong output=pts_selected -ae
     --o

g.remove vect=pts_random
g.remove vect=pts_correct
g.remove vect=pts_wrong

# Add columns

v.db.addcol pts_selected columns="reporter integer, score double,
     neighbors double, suitable double, occ_ly double, occ_lt double,
     support double, delta double"

# Add reporter IDs

val=1
for v in `v.category pts_selected option=print`; do
v.db.update pts_selected column=reporter value=$val where="cat=$v"
val=$val+1
done

# Initialize score

if [ ${year}${i} -eq 2004001 ]; then
echo "loop1"
```

```
v.db.update pts_selected column=score value=0
else
echo "loop2"
for r in `seq 1 1 100`; do
v.db.update pts_selected column=score value=`v.db.select
     pts_selected_$ptslasttime columns=score where="reporter=$r" -c`
     where="reporter=$r"
done
fi

# Number of neighbors
# The default is to sum all 9-cells in the neighborhood.  The
     weights.txt file specifically excludes the middle cell from being
     included.  This is the way it was published, but I might want to
     include this in the computation of scores

r.neighbors input=distrib.tmp output=neighbors_count size=3 method=sum
     weight=weights.txt
r.mapcalc "neighbors_count=neighbors_count/4"
v.what.rast vector=pts_selected raster=neighbors_count
     column=neighbors
v.db.update pts_selected column=neighbors value=-1 where="neighbors=0"
v.db.update pts_selected column=neighbors value=0 where="neighbors is
     null"
g.remove neighbors_count

# Suitable Habitat -- not included initially in the published paper

v.what.rast vector=pts_selected raster=suitable.$year.$i
     column=suitable
v.db.update pts_selected column=suitable value=-1 where="suitable=0"
v.db.update pts_selected column=suitable value=0 where="suitable is
     null"

# Occupied in t-1

v.what.rast vector=pts_selected raster=distrib.$lasttime column=occ_lt
v.db.update pts_selected column=occ_lt value=-1 where="occ_lt = 0"
v.db.update pts_selected column=occ_lt value=0 where="occ_lt is null"

# Occupied in y-1

v.what.rast vector=pts_selected raster=distrib.$lastyear.$i
     column=occ_ly
v.db.update pts_selected column=occ_ly value=-1 where="occ_ly=0"
v.db.update pts_selected column=occ_ly value=0 where="occ_ly is null"

# Supporting reports

v.distance from=pts_selected to=pts_selected upload=dist col=support
     dmax=250 dmin=1
```

```
v.db.update pts_selected column=support value=-1 where="support is
     null"
v.db.update pts_selected column=support value=1 where="support > 0"

# compute delta score

v.db.update pts_selected column=delta value="neighbors + suitable +
     occ_lt + occ_ly"

# update the overall score

v.db.update pts_selected column=score value="score + delta"

# discard points that don't meet minimum standards

score=`v.univar pts_selected column=score type=point | grep -e "mean:"
     | awk '{ print $2 }'`
echo "mean score is ${score}"
v.db.update pts_selected column=value value=0 where="score <=
     ${score}"

# output to raster

v.to.rast input=pts_selected type=point output=pts_selected use=attr
     column=value
r.null pts_selected null=0

# update distribution

r.mapcalc "distrib.tmp=if(pts_selected == 1, 1, distrib.tmp)"

g.copy distrib.tmp,distrib.$year.$i
g.copy vect=pts_selected,pts_selected_${year}_${i}
g.remove rast=pts_selected
g.remove vect=pts_selected
g.remove distrib.grown.tmp

lasttime=$year.$i
ptslasttime=${year}_${i}
if [ $i -eq 353 ]; then
lastyear=$year
fi
done
done

g.remove distrib.tmp

g.region zoom=studyarea
r.mask studyarea
r.series input=`g.mlist pat=distrib.200[4-6]* sep=,
     mapset=expSim37_$rep` output=sim37_run$rep\_average
     method=average
```

```
r.series input=`g.mlist pat=distrib.200[4-6]* sep=,
     mapset=expSim37_$rep` output=sim37_run$rep\_sum method=sum
r.series input=`g.mlist pat=distrib.2004.* sep=, mapset=expSim37_$rep`
     output=sim37_2004_run$rep\_average method=average
r.series input=`g.mlist pat=distrib.2004.* sep=, mapset=expSim37_$rep`
     output=sim37_2004_run$rep\_sum method=sum
r.series input=`g.mlist pat=distrib.2005.* sep=, mapset=expSim37_$rep`
     output=sim37_2005_run$rep\_average method=average
r.series input=`g.mlist pat=distrib.2005.* sep=, mapset=expSim37_$rep`
     output=sim37_2005_run$rep\_sum method=sum
r.series input=`g.mlist pat=distrib.2006.* sep=, mapset=expSim37_$rep`
     output=sim37_2006_run$rep\_average method=average
r.series input=`g.mlist pat=distrib.2006.* sep=, mapset=expSim37_$rep`
     output=sim37_2006_run$rep\_sum method=sum

r.mask -r


curl https://prowlapp.com/publicapi/add -F
     apikey=2daeeaf780b2e94281d1089752c6698da81434a9 -F
     application="HPCC" -F description="TED simulation 37   , run $rep
     complete"
```

Appendix C

ADDENDUM TO CHAPTER 2

Since the publication of Chapter 2 (Langley & Messina, 2011), the process for standing up a postgres database with PostGIS Raster support has changed substantially. This appendix serves to provide updated code and instructions for installing the database in Ubuntu 13.10.

Overview of Requirements

Package Versions Required (May 2014):

Proj4 (4.8.0)	http://download.osgeo.org/proj/proj-4.8.0.tar.gz
GEOS (3.4.2)	http://download.osgeo.org/geos/geos-3.4.2.tar.bz2
GDAL (1.10.1)	http://download.osgeo.org/gdal/1.10.1/gdal1101.tar.gz
GRASS (6.4.3)	http://grass.osgeo.org/grass64/source/grass-6.4.3.tar.gz

Add postgres repository to the package manager (Ubuntu 13.10 Saucy)

```
sudo sh -c 'echo "deb http://apt.postgresql.org/pub/repos/apt/ saucy-
     pgdg main" >> /etc/apt/sources.list'
wget --quiet -O - http://apt.postgresql.org/pub/repos/apt/ACCC4CF8.asc
     | sudo apt-key add -
```

Install packages from repository

```
sudo apt-get install <package>
```

- build-essential
- postgresql-9.3
- postgresql-server-dev-9.3
- pgadmin3
- libxml2-dev
- libjson0-dev
- xsltproc
- docbook-xsl
- docbook-mathml

Compile and Install GEOS 3.4.X

PostGIS 2.1 is best used with GEOS >= 3.4 for several new features, however Ubuntu 13.10 only has GEOS 3.3.3 available in packages, so it needs to be built from source. If you don't need the new features, instead install the libgeos-dev package.

```
wget http://download.osgeo.org/geos/geos-3.4.2.tar.bz2
```

```
tar xfj geos-3.4.2.tar.bz2
cd geos-3.4.2
./configure
make
sudo make install
cd ..
```

Compile and Install PostGIS

```
wget http://download.osgeo.org/postgis/source/postgis-2.1.1.tar.gz
tar xfz postgis-2.1.1.tar.gz
cd postgis-2.1.1
```

Implement Basic configuration for PostGIS 2.1, with raster and topology support:

```
./configure
make
sudo make install
sudo ldconfig
sudo make comments-install
```

Lastly, enable the command line tools to work from shell:

```
sudo ln -sf /usr/share/postgresql-common/pg_wrapper
     /usr/local/bin/shp2pgsql
sudo ln -sf /usr/share/postgresql-common/pg_wrapper
     /usr/local/bin/pgsql2shp
sudo ln -sf /usr/share/postgresql-common/pg_wrapper
     /usr/local/bin/raster2pgsql
```

Spatially enabling the database

Connect to the database. To add raster support:

```
CREATE EXTENSION postgis;
```

To create topology support:

```
CREATE EXTENSION postgis_topology;
```

REFERENCES

REFERENCES

Adam, N. R. A., & Gangopadhyay, A. A. (1997). *Database Issues in Geographic Information Systems*. Boston: Kluwer Academic Publishers.

Altmann, J., Alberts, S. C., Altmann, S. A., & Roy, S. B. (2002). Dramatic change in local climate patterns in the Amboseli basin, Kenya. *African Journal of Ecology, 40*(3), 248-251. doi: 10.1046/j.1365-2028.2002.00366.x

Baird, T., Leslie, P., & McCabe, J. (2009). The Effect of Wildlife Conservation on Local Perceptions of Risk and Behavioral Response. *History Teacher, 37*, 463-474.

Balram, S., & Dragićević, S. (2006). *Collaborative Geographic Information Systems*. Hershey: Idea Group Publishing.

Batchelor, N., Atkinson, P., Gething, P., Picozzi, K., Fevre, E. M., Kakembo, A., & Welburn, S. C. (2009). Spatial Predictions of Rhodesian Human African Trypanosomiasis (Sleeping Sickness) Prevalence in Kaberamaido and Dokolo, Two Newly Affected Districts of Uganda. *PLoS neglected tropical diseases, 3*(12), e563. doi: 10.1371/journal.pntd.0000563

Bauer, B. O., Kabore, I., Liebisch, A., Meyer, F., & Petrich-Bauer, J. (1992). Simultaneous control of ticks and tsetse flies in Satiri, Burkina Faso, by the use of flumethrin pour on for cattle. *Tropical Medicine and Parasitology, 43*(1), 41-46.

Bishr, M., & Kuhn, W. (2007). Geospatial information bottom-up: A matter of trust and semantics. In S. I. Fabrikant & M. Wachowicz (Eds.), *The European Information Society: Leading the Way with Geo-information* (pp. 365-387). Springer.

Bishr, M., & Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth. *GeoJournal, 72*(3-4), 229-237. doi: 10.1007/s10708-008-9182-4

Boroushaki, S., & Malczewski, J. (2010). Measuring consensus for collaborative decision-making: A GIS-based approach. *Computers, Environment and Urban Systems, 34*(4), 322-332. doi: 10.1016/j.compenvurbsys.2010.02.006

Boroushaki, S., & Malczewski, J. (2010). Using the fuzzy majority approach for GIS-based multicriteria group decision-making. *Computers and Geosciences, 36*(3), 302-312.

Brun, R., Blum, J., Chappuis, F., & Burri, C. (2010). Human African trypanosomiasis. *The Lancet, 375*(9709), 148-159. doi: 10.1016/S0140-6736(09)60829-1

Budhathoki, N. R., Bruce, B. C., & Nedović-Budić, Z. (2008). Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal, 72*(3-4), 149-160. doi: 10.1007/s10708-008-9189-x

Câmara, G., Souza, R., Freitas, U. M., & Garrido, J. (1996). SPRING: Integrating remote sensing and GIS by object-oriented data modelling. *Computers & Graphics, 20*(3), 395-403.

Campbell, D. J., Gichohi, H., Mwangi, A., & Chege, L. (2000). Land use conflict in kajiado District, Kenya. *Land Use Policy, 17*(4), 337-348.

Campbell, D. J., Lusch, D., Smucker, T., & Wangui, E. E. (2004). Root causes of land use change in the Loitokitok Area, Kajiado District, Kenya. *Land Use Change Impacts and Dynamics (LUCID) Project Working Paper, 19*.

Cecchi, G., Mattioli, R., Slingenbergh, J., & de La Rocque, S. (2008). Land cover and tsetse fly distributions in sub-Saharan Africa. *Medical and Veterinary Entomology, 22*(4), 364-373. doi: 10.1111/j.1365-2915.2008.00747.x

Chow, T. E. (2012). "We Know Who You Are and We Know Where You Live": A Research Agenda for Web Demographics (pp. 265-285). Dordrecht: Springer Netherlands.

Cipeluch, B., Jacob, R., Winstanley, A., & Mooney, P. (2010, 20-23rd July 2010). *Comparison of the accuracy of OpenStreetMap for Ireland with Google maps and Bing maps.* Proceedings of the Proceedings, Ninth International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.

Cohen, M. L. (2000). Changing patterns of infectious disease. *Nature, 406*(6797), 762-767.

Coleman, D., & Sabone, B. (2010). Volunteering geographic information to authoritative databases: Linking contributor motivations to program characteristics. *Geomatica, 64*(1), 27-40.

Coleman, D. J., & Georgiadou, Y. (2009). Volunteered Geographic Information: the nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research, 4*(1), 332-358.

Comber, A. J., Fisher, P. F., Harvey, F., Gahegan, M., & Wadsworth, R. (2006). Using metadata to link uncertainty and data quality assessments. In A. Riedl, W. Kainz & G. A. Elmes (Eds.), *Progress in Spatial Data Handling: 12th International Symposium on Spatial Data Handling* (pp. 279-292). Springer Berlin Heidelberg.

Conati, C. (2004). How to Evaluate Models of User Affect? In T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, E. André, L. Dybkjær, W. Minker & P. Heisterkamp (Eds.), *Affective Dialogue Systems* (Vol. 3068, pp. 288-300). Berlin, Heidelberg: Springer Berlin Heidelberg.

Connors, J. P., Lei, S., & Kelly, M. (2012). Citizen Science in the Age of Neogeography: Utilizing Volunteered Geographic Information for Environmental Monitoring. *Annals of the Association of American Geographers, 102*(6), 1267-1289. doi: 10.1080/00045608.2011.627058

Corbett, J. (2012). "I Don't Come from Anywhere": Exploring the Role of the Geoweb and Volunteered Geographic Information in Rediscovering a Sense of Place in a Dispersed Aboriginal Community (pp. 223-241). Dordrecht: Springer Netherlands.

Cox, F. E. G. (2004). History of sleeping sickness (African trypanosomiasis). *Infect Dis Clin N Am, 18*, 231-245. doi: 10.1016/j.idc.2004.01.004

Craglia, M. (2007). Volunteered geographic information and spatial data infrastructures: When do parallel lines converge. Position paper for the VGI Specialist Meeting, Santa Barbara 13-14 December 2007.

Craglia, M., de Bie, K., Jackson, D., Pesaresi, M., Remetey-Fülöpp, G., Wang, C., . . . Woodgate, P. (2012). Digital Earth 2020: towards the vision for the next decade. *International Journal of Digital Earth, 5*(1), 4-21. doi: 10.1080/17538947.2011.638500

Craglia, M., Goodchild, M. F., Annoni, A., Câmara, G., Gould, M., Kuhn, W., . . . Parsons, E. (2007). Next-generation digital earth. *International Journal of Spatial Data Infrastructures Research, 3*, 146-167.

Crosetto, M., & Rodriguez, J. P. (2001). Uncertainty and sensitivity analysis: tools for GIS-based model implementation. *International Journal Of Geographical Information Science, 15*(5), 415-437. doi: 10.1080/13658810110053125

Devillers, R., & Jeansoulin, R. (2006). *Fundamentals of spatial data quality*. Newport Beach, CA: ISTE.

DeVisser, M., & Messina, J. P. (2009). Optimum land cover products for use in a Glossina-morsitans habitat model of Kenya. *Int J Health Geogr, 8*(39), 39. doi: 10.1186/1476-072X-8-39

DeVisser, M., Messina, J. P., Moore, N. J., & Lusch, D. (2010). A dynamic species distribution model of Glossina subgenus Morsitans: The identification of tsetse reservoirs and refugia. *Eco Soc America, 1*(1), 1-21.

Devogele, T., Parent, C., & Spaccapietra, S. (1998). On spatial database integration. *International Journal Of Geographical Information Science, 12*(4), 335-352.

Drew, P., & Ying, J. (1996). *GeoChange: an experiment in wide-area database services for geographic information exchange*. Paper presented at the ADL '96 Forum Forum on Research and Technology Advances in Digital Libraries. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=502512

Egenhofer, M. J. (1994). Spatial SQL: A Query and Presentation Language (PDF). *IEEE Transactions on Knowledge and Data Engineering, 6*(1), 86-95.

Elmasri, R., & Navathe, S. (2004). *Fundamentals of database systems* (4th ed.). Boston: Pearson/Addison Wesley.

Elmes, G. A., Dougherty, M., Challig, H., Karigomba, W., McCusker, B., & Weiner, D. (2005). Local knowledge doesn't grow on trees: Community-integrated geographic information systems and rural community self-definition. *Developments in Spatial Data Handling*, 29-39.

Elwood, S. A. (2006a). Critical issues in participatory GIS: deconstructions, reconstructions, and new research directions. *Transactions in GIS, 10*(5), 693-708.

Elwood, S. A. (2006b). Negotiating Knowledge Production: The Everyday Inclusions, Exclusions, and Contradictions of Participatory GIS Research. *The Professional Geographer, 58*(2), 197-208. doi: 10.1111/j.1467-9272.2006.00526.x

Elwood, S. A. (2008a). Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal, 72*(3-4), 173-183. doi: 10.1007/s10708-008-9186-0

Elwood, S. A. (2008b). Volunteered geographic information: key questions, concepts and methods to guide emerging research and practice. *GeoJournal, 72*(3-4), 133-135. doi: 10.1007/s10708-008-9187-z

Elwood, S. A. (2010). Geographic information science: emerging research on the societal implications of the geospatial web. *Progress in Human Geography, 34*(3), 349-357. doi: 10.1177/0309132509340711

Elwood, S. A. (2011). Geographic Information Science: Visualization, visual methods, and the geoweb. *Progress in Human Geography, 35*(3), 401-408. doi: 10.1177/0309132510374250

Elwood, S. A., Goodchild, M. F., & Sui, D. (2013). Prospects for VGI Research and the Emerging Fourth Paradigm. In D. Sui, S. A. Elwood & M. F. Goodchild (Eds.), *Crowdsourcing geographic knowledge* (pp. 361-375). Dordrecht: Springer Netherlands.

Elwood, S. A., Goodchild, M. F., & Sui, D. Z. (2011). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers, 102*(3), 110809092640007. doi: 10.1080/00045608.2011.595657

Elwood, S. A., & Leitner, H. (2012). GIS and Community-based Planning: Exploring the Diversity of Neighborhood Perspectives and Needs. *Cartography and Geographic Information Systems, 25*(2), 77.

Fischer, G., Shah, M., N Tubiello, F., & van Velhuizen, H. (2005). Socio-economic and climate change impacts on agriculture: an integrated assessment, 1990-2080. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1463), 2067-2083. doi: 10.1098/rstb.2005.1744

Flanagin, A. J., & Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal, 72*(3-4), 137-148. doi: 10.1007/s10708-008-9188-y

Ford, J. (1971). *The role of the trypanosomiases in African ecology. A study of the tsetse fly problem*. Oxford: Clarendon Pr.

Frew, J. (2007). *Provenance and volunteered geographic information*. Workshop on Volunteered Geographic Information. Retrieved June 30, 2013 from http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Frew_paper.pdf

Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS, 14*(4), 435-459.

Goodchild, M. F. (2007a). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221.

Goodchild, M. F. (2007b). Citizens as sensors: web 2.0 and the volunteering of geographic information. *Geofocus, 7*, 8-10.

Goodchild, M. F. (2007c). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*, 24-32.

Goodchild, M. F. (2008a). *Assertion and authority: the science of user-generated geographic content.* Proceedings of the Proceedings of the Colloquium for Andrew U Frank's 60th Birthday, Department of Geoinformation and Cartography, Vienna University of Technology, Vienna.

Goodchild, M. F. (2008b). Commentary: whither VGI? *GeoJournal, 72*(3), 239-244.

Goodchild, M. F. (2009). NeoGeography and the nature of geographic expertise. *Journal of Location Based Services, 3*(2), 82-96. doi: 10.1080/17489720902950374

Goodchild, M. F. (2010a). Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth, 3*(3), 231-241. doi: 10.1080/17538941003759255

Goodchild, M. F. (2010b). Twenty years of progress: GIScience in 2010. *Journal of Spatial Information Science, 1*(1), 3-20. doi: 10.5311/josis.v0i1.32

Goodchild, M. F., Guo, H., Annoni, A., Bian, L., de Bie, K., Campbell, F., . . . Woodgate, P. (2012). Next-generation Digital Earth. *Proceedings of the National Academy of Sciences, 109*(28), 11088-11094. doi: 10.1073/pnas.1202383109

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110-120. doi: 10.1016/j.spasta.2012.03.002

Gray, J., & Szalay, A. (May 29, 2006). *eScience: The Next Decade Will Be Exciting*. Lecture presented at ETH, Zurich. Retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/ETH_E_Science.ppt

Grira, J., Bédard, Y., & Roche, S. (2009). Spatial data uncertainty in the VGI world: Going from consumer to producer. *Geomatica, 64*(1), 61-71.

Groot, R., & McLaughlin, J. D. (2000). *Geospatial data infrastructure : concepts, cases, and good practice*. Oxford: Oxford University Press.

Groot, R. T. A. d. (2012). *Evaluation of a volunteered geographical information trust measure in the case of OpenStreetMap.* (MS), University of Münster, Germany. Retrieved from http://run.unl.pt/handle/10362/8301

Gyapong, J., Gyapong, M., Yellu, N., Anakwah, K., Amofah, G., Bockarie, M., & Adjei, S. (2010). Integration of control of neglected tropical diseases into health-care systems: challenges and opportunities. *The Lancet, 375*(9709), 160-165. doi: 10.1016/S0140-6736(09)61249-6

Haklay, M. M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design, 37*(4), 682-703. doi: 10.1068/b35097

Haklay, M. M. (2011, 27 November, 2011). Citizen Science as Participatory Science. Retrieved from http://povesham.wordpress.com/2011/11/27/citizen-science-as-participatory-science/

Haklay, M. M. (2012). Citizen Science and Volunteered Geographic Information: Overview and Typology of Participation. In D. Sui, S. A. Elwood & M. F. Goodchild (Eds.), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice* (pp. 105-122). Dordrecht: Springer Netherlands.

Haklay, M. M., Basiouka, S., Antoniou, V., & Ather, A. (2010). How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *Cartographic Journal, The, 47*(4), 315-322. doi: 10.1179/000870410X12911304958827

Haklay, M. M., Singleton, A., & Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass, 2*(6), 2011-2039. doi: 10.1111/j.1749-8198.2008.00167.x

Haklay, M. M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing, 7*(4), 12-18. doi: 10.1109/MPRV.2008.80

Harvey, F. (2012). To Volunteer or to Contribute Locational Information? Towards Truth in Labeling for Crowdsourced Geographic Information (pp. 31-42). Dordrecht: Springer Netherlands.

Hellerstein, J. (1999). The GIST Indexing Project. from http://gist.cs.berkely.edu

House, R., Javidan, M., & Dorfman, P. (2001). Project GLOBE: An Introduction. *Applied Psychology, 50*(4), 489-505. doi: 10.1111/1464-0597.00070

Howe, J. (2006, June 2006). The rise of crowdsourcing. *Wired magazine, 14*.

Irwin, A. (1995). *Citizen Science: A Study of People, Expertise and Sustainable Development*. New York, NY: Routledge.

Johnson, P. T. J., & Thieltges, D. W. (2010). Diversity, decoys and the dilution effect: how ecological communities affect disease risk. *The Journal of experimental biology, 213*(Pt 6), 961-970. doi: 10.1242/jeb.037721

Joint WHO Expert Committee and FAO Expert Consultation on the African Trypanosomiases (1976: Rome). (1979). *The African Trypanosomiases: report of a joint WHO expert committee and FAO expert consultation, Rome, 8-12 November, 1976*. Geneva. World Health Organization. Retrieved from http://www.who.int/iris/handle/10665/41349

Keesing, F., Holt, R. D., & Ostfeld, R. (2006). Effects of species diversity on disease risk. *Ecology Letters, 9*(4), 485-498. doi: 10.1111/j.1461-0248.2006.00885.x

Kennedy, P. (2005). Sleeping sickness-human African trypanosomiasis. *British Medical Journal, 5*(5), 260-267.

KETRI. (1996). Tsetse Distribution in Kenya Showing Tsetse Belts and Conservation Areas. Kenya Trypanosomiasis Research Institute.

Kornacker, M. (1999). *High-Performance Extensible Indexing*. Paper presented at the VLDB Conference, Edinburgh, Scotland. ftp://ftp.u-aizu.ac.jp/ftp/pub/dbms/gist/hiperf-gist.pdf

Koukoletsos, T., Haklay, M. M., & Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transactions in GIS, 16*(4), 477-498. doi: 10.1111/j.1467-9671.2012.01304.x

Langley, S. A. *Developing a participatory GIS network for the mapping of African trypanosomiasis in Kenya* [Unpublished Interview Data]. (2010).

Langley, S. A., & Messina, J. P. (2011). Embracing the Open-Source Movement for Managing Spatial Data: A Case Study of African Trypanosomiasis in Kenya. *Journal of Map and Geography Libraries, 7*(1), 87-113. doi: 10.1080/15420353.2011.534693

Langley, S. A., & Messina, J. P. (2013). Utilizing Volunteered Information for Infectious Disease Surveillance. *International Journal of Applied Geospatial Research, 4*(2), 54-70.

Lans, R. F. v. d. (2007). *Introduction To Sql: Mastering The Relational Database Language* (4th ed.). Upper Saddle River, NJ: Addison-Wesley.

Livingstone, D. N. (1992). *The Geographical Tradition: Episodes in the History of a Contested Enterprise*. Malden, MA: Blackwell Publishing Ltd.

Longley, P. A., Goodchild, M., Maguire, D. J., & Rhind, D. W. (2005). *Geographic Information Systems and Science* (2 ed.). Hoboken, NJ: John Wiley & Sons.

Maitima, J. M. (N.d.). *The variabilities in plant phenological activities in response to rainfall and temperature variability in Nguruman*. Unpublished manuscript.

Makido, Y., Shortridge, A. M., & Messina, J. P. (2007). Assessing Alternatives for Modeling the Spatial Distribution of Multiple Land-cover Classes at Sub-pixel Scales. *Photogrammetric Engineering and Remote Sensing, 73*(8), 935.

Masser, I. (2005, Oct. 14-16). *The future of spatial data infrastructures.* Proceedings of the ISPRS Workshop on Service and Application of Spatial Data Infrastructure, Hangzhou, China.

Masser, I. (2005). *GIS worlds: creating spatial data infrastructures* (1st ed.). ESRI Press.

Maué, P. (2007). *Reputation as tool to ensure validity of VGI*. Position Paper for Specialist Meeting on VGI. Santa Barbara, CA: Retrieved from http://www.ncgia.ucsb.edu/projects/vgi/docs/position/Maue_paper.pdf

McCall, M. K., & Minang, P. A. (2005). Assessing participatory GIS for community-based natural resource management: claiming community forests in Cameroon. *Geographical Journal, 171*(4), 340-356.

McKnight, K. P., Messina, J. P., & Shortridge, A. M. (2011). Using Volunteered Geographic Information to Assess the Spatial Distribution of West Nile Virus in Detroit, Michigan. *International Journal of Applied Geospatial Research, 2*(3), 72-85.

Messina, J. P., Moore, N. J., DeVisser, M., McCord, P. F., & Walker, E. D. (2012). Climate Change and Risk Projection: Dynamic Spatial Models of Tsetse and African Trypanosomiasis in Kenya. *Annals of the Association of American Geographers, 102*(5), 1038-1048. doi: 10.1080/00045608.2012.671134

Metcalf, S., & Paich, M. (2005). Spatial dynamics of social network evolution. *Vol, 51*, 61801.

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2003). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication yearbook, 27*, 293-336.

Miller, C. C. (2006). A Beast in the Field: The Google Maps Mashup as GIS/2. *Cartographica: The International Journal for Geographic Information and Geovisualization, 41*(3), 187-199. doi: 10.3138/J0L0-5301-2262-N779

Moore, N. J., Alagarswamy, G., Pijanowski, B., Thornton, P. K., Lofgren, B., Olson, J., . . . Qi, J. (2012). East African food security as influenced by future climate change and land use change at local to regional scales. *Climatic Change, 110*(3-4), 823-844. doi: 10.1007/s10584-011-0116-7

Moore, N. J., & Messina, J. P. (2010). A Landscape and Climate Data Logistic Model of Tsetse Distribution in Kenya. *PloS one, 5*(7), e11809. doi: 10.1371/journal.pone.0011809

Morris, D. L., Western, D., & Maitumo, D. (2009). Pastoralist's livestock and settlements influence game bird diversity and abundance in a savanna ecosystem of southern Kenya. *African Journal of Ecology, 47*(1), 48-55. doi: 10.1111/j.1365-2028.2007.00914.x

Muriuki, G. W., Njoka, T., Reid, R., & Nyariki, D. (2005). Tsetse control and land-use change in Lambwe valley, south-western Kenya. *Agriculture, Ecosystems and Environment, 106*(1), 99-107.

Neteler, M., Beaudette, D. E., Cavallini, P., Lami, L., & Cepicky, J. (2008). GRASS GIS. In G. B. Hall & M. G. Leahy (Eds.), (Vol. 2, pp. 171-199). Berlin, Heidelberg: Springer Berlin Heidelberg.

O'Reilly, T. (2005, September 30). What is Web 2.0. Retrieved June 30, 2013, from http://oreilly.com/web2/archive/what-is-web-20.html

O'Reilly, T. (2006, December 10). Web 2.0 Compact Definition: Trying Again. Retrieved June 30, 2013, from http://radar.oreilly.com/2006/12/web-20-compact-definition-tryi.html

O'Reilly, T. (2007). What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies, 1*(17).

OGIS. (1999). *Open GIS Simple Features Specification for SQL (Revision 1.1)*. Open Geospatial Consortium. Retrieved from www.opengeospatial.org

Olson, J. E. (2003). *Data Quality: The Accuracy Dimension*. San Francisco, CA: Morgan Kaufmann Publishers, Elsevier Science.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science (New York, NY), 263*(5147), 641-646. doi: 10.1126/science.263.5147.641

Ostfeld, R. S., Keesing, F., & Eviner, V. T. (2008). *Infectious Disease Ecology: Effects of Ecosystems on Disease and of Disease on Ecosystems*. Princeton, NJ: Princeton University Press.

Parent, C., Spaccapietra, S., & Zimányi, E. (2006). The MurMur project: Modeling and querying multi-representation spatio-temporal databases. *Information Systems, 31*, 733-769.

Pickles, J. (1995). *Ground truth: The social implications of geographic information systems*. The Guilford Press.

PostGIS (Version 1.3.2) (2008) [Computer Program]: Refractions Research Inc. Retrieved from http://postgis.net

Python (Version 2.6.5) (2010) [Computer Software]: Python Software Foundation.

Raymond, E. (1999). The cathedral and the bazaar. *Knowledge, Technology & Policy, 12*(3), 23-49.

Rivest, R. (1992). *The MD5 message-digest algorithm*. MIT Laboratory for Computer Science and RSA Data Security, Inc. Retrieved from http://www.faqs.org/rfcs/rfc1321.html#b

Robbins, P. (2003). Beyond ground truth: GIS and the environmental knowledge of herders, professional foresters, and other traditional communities. *History Teacher, 31*(2), 233-253.

Rosenberg, T. (2011, Apr 28). Crowdsourcing a Better World - NYTimes.com, Editorial. *New York Times*. Retrieved from http://opinionator.blogs.nytimes.com/2011/03/28/crowdsourcing-a-better-world

Rouse, L. J., Bergeron, S. J., & Harris, T. M. (2007). Participating in the Geospatial Web: Collaborative Mapping, Social Networks and Participatory GIS. In A. Scharl & K. Tochtermann (Eds.), *The Geospatial Web*. London: Springer.

Shekhar, S., & Chawla, S. (2003). *Spatial databases: A tour*. Prentice Hall.

Shi, W., Fisher, P., & Goodchild, M. F. (2002). *Spatial Data Quality*. New York, NY: Taylor & Francis.

Sieber, R. E. (2006). Public participation geographic information systems: A literature review and framework. *Annals of the Association of American Geographers, 96*(3), 491-507.

Silvertown, J. (2009). A new dawn for citizen science. *Trends in Ecology & Evolution, 24*(9), 467-471. doi: 10.1016/j.tree.2009.03.017

Siringi, S. (2003). Kenya rejects drugs deal. *The Lancet Infectious Diseases, 3*(6), 320-320. doi: doi: DOI: 10.1016/S1473-3099(03)00641-8

Smith, T., & Frew, J. (1995). Alexandria digital library. *Communications of the ACM, 38*(4), 62.

Snyder, J. D., & Merson, M. H. (1982). The magnitude of the global problem of acute diarrhoeal disease: a review of active surveillance data. *Bulletin of the World Health Organization, 60*(4), 605.

Stonebraker, M., & Kemnitz, G. (1991). The POSTGRES next generation database management system. *Communications of the ACM, 34*(10), 78-92. doi: 10.1145/125223.125262

Stonebraker, M., & Moore, D. (1995). *Object Relational DBMSs: The Next Great Wave*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Stonebraker, M., & Rowe, L. (1986). *The design of POSTGRES.* Proceedings of the SIGMOD Conference.

Sui, D. Z. (2008). The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. *Computers, Environment and Urban Systems, 32*.

Sutherst, R. (2004). Global change and human vulnerability to vector-borne diseases. *Clin Microbiol Rev, 17*(1), 136-173. doi: Doi 10.1128/Cmr.17.1.136-173.2004

Tarimo-Nesbitt, R., Golder, T. K., Dransfield, R., Chaudhury, M. F., & Brightwell, R. (1999). Trypanosome infection rate in cattle at Nguruman, Kenya. *Veterinary Parasitology, 81*(2), 107-117.

Tatem, A. J., Rogers, D. J., & Hay, S. I. (2006). Global Transport Networks and Infectious Disease Spread. *Advances in Parasitology, 62*, 293-343. doi: 10.1016/S0065-308X(05)62009-X

Terblanche, J. S., Clusella-Trullas, S., Deere, J. A., & Chown, S. L. (2008). Thermal tolerance in a south-east African population of the tsetse fly Glossina pallidipes (Diptera, Glossinidae): Implications for forecasting climate change impacts. *Journal of insect physiology, 54*(1), 114-127.

Thacker, S. B., Choi, K., & Brachman, P. S. (1983). The Surveillance of Infectious Diseases. *JAMA, 249*(9), 1181-1185. doi: 10.1001/jama.1983.03330330059036

Tulloch, D. L. (2007). Many, many maps: Empowerment and online participatory mapping. *First Monday, 12*(2).

Tulloch, D. L. (2008). Is VGI participation? From vernal pools to video games. *GeoJournal, 72*(3-4), 161-171. doi: 10.1007/s10708-008-9185-1

Turner, A. (2006). *Introduction to Neogeography*. O'Reilly.

Turner, M., & Hiernaux, P. (2002). The use of herders' accounts to map livestock activities across agropastoral landscapes in Semi-Arid Africa. *Landscape Ecology, 17*(5), 367-385. doi: 10.1023/A:1021238208019

van den Berg, H., Coetzee, S., & Cooper, A. K. (2011, May 31 - June 2). *Analysing commons to improve the design of volunteered geographic information repositories.* Proceedings of the AfricaGEO 2011, Cape Town, South Africa.

van Oort, P. (2005). *Spatial data quality: from description to application.* (PhD), Wageningen University. Retrieved from http://edepot.wur.nl/38987

Waller, R. D. (1990). Tsetse fly in western Narok, Kenya. *The Journal of African History, 31*(1), 81-101.

Warren, D. M. (1991). *Using indigenous knowledge in agricultural development*. World Bank.

Watson, H. J., Wixom, B. H., & Goodhue, D. L. (2004). Data warehousing: the 3M experience. In H. R. Nemati & C. D. Barko (Eds.), *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance* (pp. 202). Hershey, PA: Idea Group Pub.

WHO. (2001). *Report on African Trypanosomiasis (sleeping sickness)*. Paper presented at the Report of the Scientific Working Group meeting on African trypanosomiasis Geneva, 4-8 June, 2001. http://www.who.int/tdrold/cd_publications/pdf/aftryp_swg.pdf

WHO. (2005). Control of Human African Trypanosomiasis: A Strategy for the African Region. 1-9.

WHO. (2006). African Trypanosomiasis. Retrieved February 10, 2007, from http://www.who.int/media/centre/factsheets/fs259/en/

Wiersma, Y. F. (2010). Birding 2.0: Citizen Science and Effective Monitoring in the Web 2.0 World. *Avian Conservation and Ecology, 5*(2).

Wint, W. (2001). *Kilometre resolution Tsetse Fly distribution maps for the Lake Victoria Basin and West Africa*. Food and Agriculture Organization/International Atomic Energy Agency Joint Division. Retrieved from http://ergodd.zoo.ox.ac.uk/tseweb/iaea/tse1kmrep.doc

Yaukey, P. H. (2010). Citizen Science and Bird–Distribution Data: an Opportunity for Geographical Research. *Geographical Review, 100*(2), 263-273.