INCORPORATING GEOGRAPHIC AND PHYLOGENETIC INFORMATION INTO DENSITY-EQUALIZING MAPS

by

Alexander R. Keddy

Submitted in partial fulfillment of the
requirements for the degree of
Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

Table of Contents

List of Tables

List of Figures

Abstract

Phylogeography is the study of how geographic and environmental factors affect the evolution of organisms. Phylogeographic analysis combines evolutionary information, often represented using phylogenetic trees, with geographic representations of an observed data set. A key element of phylogeography is the use of branch lengths in a tree, which correspond to accumulated evolutionary differences among organisms, to generate an overall view of phylogenetic diversity in a region. Here we describe the Geographically Coupled Phylogenetic Distance (GCPD), a new method for associating phylogenetic diversity with location information. The GCPD uses location-based phylogeographic information to calculate a minimum spanning tree where locations are vertices. Branch weights of this graph are then substituted from geographic distance to the phylogenetic distance between sites, creating quantitative location-based representations of the phylogenetic difference between sites.

One application of the GCPD is in phylogeographic visualization. Density-equalizing maps, also known as cartograms, can preserve geographic relationships and attributes such as political divisions, but use map distortions to visualization quantitative data such as election results or population distributions. We have adapted the Gastner-Newman algorithm to create map distortions based on location rather than shape data, which allows enhanced visualization of phylogeographic data. Our approach is implemented in the GenGIS software package, and can be applied to all widely used digital elevation and image formats.

We used the GCPD to generate cartograms of two biological data sets: a 2010 pandemic of *Vibrio cholera* and the diversity of the Californian salamander *Aneides lugubris*. In the cholera data set, we were able to preserve the global context of the outbreak, while highlighting the crucial regional patterns in two countries, Nepal and Haiti, implicated in a crucial transmission event. GCPD highlighted the distribution of phylogenetic groups (clades) of salamanders and showed differences between major clades in terms of both geography and phylogeny. The implicit effects of restrictive geographic boundaries such as valleys and mountain ranges were inferred in the diffusion through the addition of phylogenetic information to the map. To accelerate the creation of cartogram visualizations, we developed methods to simplify construction of the density matrix used to build the cartogram, which yielded improvements in both run time and memory consumption. While interactive time calculations are still not feasible for high-density maps, we have achieved up to two-fold increases in running time.

List of Abbreviations Used

CT Cooley-Tukey.

DNA Deoxyribonucleic acid.

EBOV Ebola Virus.

EID emerging infectious diseases.

FFT Fast Fourier Transform.

GCPD Geographically Coupled Phylogenetic Distance.

GIS Geographic Information Systems.

HIV Human Immunodeficiency Virus.

IRIDA Integrated Rapid Infectious Disease Analysis.

LBI local branching index.

LGT Lateral Gene Transfer.

ML Maximum Likelihood.

MSP Minimal Skeletal Point.

MST Minimum Spanning Tree.

NCBI National Center for Biotechnology Information.

NGS Next-Generation DNA sequencing.

NN Neural Network.

PD Phylogenetic Distance.

PE Phylogenetic Endemism.

RP reduction percentage.

rRNA ribosomal ribonucleic acid.

SNP Single-Nucleotide Polymorphism.

SOM Self Organizing Maps.

WE Weigthed Endemism.

WGS whole-genome sequencing.

WSKIZ Weighted Skeletonization By Influence Zones.

Acknowledgements

Firstly I'd like to thank my supervisor Rob Beiko. Somehow he accepted the dishevelled computer scientist who spilled into our interview off of a motorcycle into his lab. It has given me the opportunity to perform research I enjoy, travel across the country, and present in front of international audiences. For these reasons and more, I'm unbelievably grateful.

Special thanks to everyone who's passed through the Beiko and Blouin (supervisors included) labs over the past four years. Whether it's been shooting paint at each other, being locked in tiny rooms, or just sharing the lab environment, my time here has been defined by you. You've all made my time here nothing less than excellent. Michael Hall and Emma Sylvester deserve a little extra credit for all the late nights. I wish you all nothing but the best.

Thanks to all the IRIDA folks for accepting me into the group. Especially Thomas Matthews, Franklin Bristow and Aaron Petkau for all the help debugging installs and service requests.

I'd also like to thank my family, Robert, Ann and Morgan. For their love and support, the drives to school, the delicious food, for pushing me to challenge myself, and always encouraging me to find the next step. I wouldn't be here without you.

Finally I'd like to thank Batman, for teaching me that all one needs to take on the world is an inquisitive mind and determination, and that when push comes to shove a big metal suit with a handful of space rocks doesn't hurt either.

Chapter 1

Introduction

Using geography in epidemiological studies dates back to the beginning of the profession itself. It was the late John Snow, credited as the father of modern epidemiology, who in 1854 discovered the cause of a *Vibrio cholerae* outbreak in Soho, London [69]. He discovered the source of the outbreak by generating large amounts of metadata through residential surveys combined with geography based analytics. A disproportionate amount of cases were found to exist around one particular water well, having found that the infected population either used this well solely, or were exposed to it in some way. Here an extensive survey of the area as well as geographic information was able to stop one of the deadliest cholera outbreaks London had ever seen, as well as identify cholera as a waterborne disease.

Endemism is an important concept in biogeography that describes the tendency of closely related individuals of a species or populations to cluster geographically. Endemism can highlight groups of special interest, and is an important component of conservation biology. Spatial relationships are important for many applications in epidemiology and ecology including identifying species ranges, source tracking in epidemics, and defining optimal conservation areas. These fields employ phylogenetic trees to express the relationships between sets of taxa. The summation of the branch lengths from the positions of both taxa of interest in the tree to the root provides a quantitative representation of their difference. This measure is known as Phylogenetic Distance (PD), and allows for the relationships between leaves of the tree to be quantitatively expressed. When PD is computed through the root of the tree it can be normalized by the total length of the tree to produce a PD between 0 and 1 which communicates the proportion of the variation within the tree two taxa represent. Precise calculations of endemism and other spatial attributes require complete

delineation of species' ranges, but resource limitations often constrain the ability to carry out comprehensive sampling. As such many studies are forced to rely on much sparser data collection. For example, the case of an outbreak the flow of information can be sporadic as cases are only known during treatment. Not all affected individuals will seek treatment, and some individuals may be misdiagnosed [32].

Water supplies have already been shown to be a source of population level infection, such as in the case of the London cholera epidemic. In 1991 Reif et al. conducted a study to investigate associations with negative health effects in Denver communities due to waste site runoff. A neurobehavioral study of 204 residents of the area identified by proximity to the waste site showed positive correlations with neurological disorders and reproductive health [61]. Geographic Information Systems (GIS) have also been applied to epidemiological studies, such as the investigation of causal relationships between landfill sites in the United Kingdom with birth anomalies [20]. This study drew upon location based data points for 19,196 landfill sites, though detailed information on boundaries are unavailable. This study found only a slight increase in likelihood for birth anomalies within these zones, and could not identify a causal mechanism for health defects.

For many epidemiologists and ecological researchers, GIS is an essential information display and analysis tool [43]. These systems can integrate natural and political features, population information, time, and other attributes of a data set. Overlays, geographic projections and interactive environments allow for deep exploratory analysis. One approach that is gaining widespread use is the density-equalizing map or *cartogram* [47][56]. Cartograms reproject physical maps based on a set of attributes, such that regions with large values for that attribute are emphasized. In the case of ecological or epidemiological analysis, small regions may contain a large number of observations that cannot be distinguished, while large areas of the map are devoid of observations. Cartograms offer an opportunity to adjust the visualization to preserve the global context of an analysis while emphasizing areas of particular focus.

To apply cartograms in a biodiversity setting, we have extended the Gastner-Newman diffusion-based cartogram algorithm [27] to deal with point observations

and incorporate phylogenetic information. The Gastner-Newman approach uses diffusion, the concept in physics that a gas released in a volume will disperse to a uniform distribution, to distort a geographic projection by a non-geographic property. Our implementation of the algorithm in the GenGIS software package [59] introduces several features not found in other popular GIS software such as ArcGIS [77], ScapeToad [3] or the R package ggplot2 [76]. Firstly, these applications rely on vector maps to create their projections. These vector maps use shape projections to represent geographic regions separated by borders, and each enclosed region of the map is uniformly encoded with a value of interest. In the GenGIS implementation any flat image can be used, from a map file to a simple flat image. This allows for non-geographic distortions to be made, such as diagrams of the human body or hospital floor plans. Secondly, GenGIS uses location-based information, so that metadata can be added and selected easily. Our approach allows new visual insights to be gained from visual analysis of phylogenetic and geographic data. Although epidemiological data is a primary driver of the work in this thesis, we also demonstrate the applicability of the method to ecological data sets more generally.

1.1 Epidemiology

Epidemiology is the study of the effects, causes and patterns of health and disease as they affect populations. Primarily here we will discuss epidemiology as it pertains to physical health through the study and identification of infectious diseases. Environmental epidemiology, which is concerned with the associations between environmental exposure and health outcomes [57], also plays a role in pathogen management, as outbreaks of disease are inevitably of a spatial nature, as can be seen by the 2014 West African Ebola outbreak [73]. One of the most important roles of epidemiologists is the identification and characterization of pathogenic organisms. Early successes with serological approaches, a technique of identifying antibodies in blood samples in response to viral attacks, provided early success despite the limited comparative power of phenotypic approaches. Due to the limited number of antigenic genes responsible for defining the serotype in a genome, as well as the potential for variance among strains to make it impossible to select defining loci, the need for new techniques was

apparent [75].

The advent of Next-Generation DNA sequencing (NGS) provided a solution to many of the problems of serotyping by allowing the complete sequencing of pathogen genomes. The advancement in NGS technologies has made whole-genome sequencing (WGS) widely accessible: Whereas in 2005 a single sequencing run could only produce one gigabase of data, by 2014 this rate had increased more than a thousand times, to 1.8 terabases in a single run. While in 2001 the first human genome required 15 years to sequence and cost nearly $3 billion (i.e., one billion nucleotides), NGS technology had brought the cost down to $1000 a genome, completing approximately two every hour [38]. With the introduction of NGS in epidemiology came genome sequencing and digitization of sequenced pathogens, and the necessity of new bioinformatics algorithms that can efficiently handle large-scale genomic data sets. Large databases of previously sequenced genomes, such as GenBank [7] are essential to comparative analyses of genetic content, allowing for genealogical estimates. These archived sequences are commonly used as models for sequence alignment and quality control. When large amounts of previously generated sequences are available they provide historical and/or geographical context to newly generated sequences.

1.1.1 Example Applications of Genomic Epidemiology

In 2014 Western Africa was hit with the largest outbreak of Ebola Virus (EBOV) ever seen. EBOV was one of five strains of the Ebola virus, which can have mortality rates as high as 90% [23]. Of concern to epidemiologists was the origins of the current Ebola strain. To analyse this outbreak 99 samples were sequenced from 78 individuals in Sierra Leone and combined with previously sequenced Ebola cases [30]. Due to the large amount of intrahost and interhost variation it was possible to characterize the transmission of the virus. Through phylogenetic analysis it was established that the most likely source of this Ebola outbreak was from a previous outbreak in Central Africa in 2004, and that there was no ongoing human-reservoir exposure reintroducing the disease into West Africa [29].

One of the interesting aspects of this outbreak was the ability to generate genomic

sequences as the outbreak developed. This allowed for the spread of the Ebola virus to be characterized both geographically and temporally using web based tools such as Nextflu(`ebola.nextflu.org`). Pressure to identify the path of infection also pushed the public release of this data, allowing different research groups to combine many data sources. The authors of [6] created a chronological phylogeographic visualization using 13 data sources. Phylogenies were creatued using augur(`https://github.com/blab/nextflu/tree/master/augur`) and displayed with respect to time, region or local branching index (LBI) [53]. Technologies which allow ad hoc sequence generation and analysis continue to improve; for example the Nanopore sequencer used by the European Mobile Laboratories to sample and release ongoing cases in only a few days [26]. To address the need for rapid, repeatable analysis of large data sets, several initiatives including Canada's Integrated Rapid Infectious Disease Analysis (IRIDA) network are being developed.

In 2010 an outbreak of *Vibrio cholerae* ravaged the already distressed nation of Haiti after it had been hit by a devastating earthquake. This epidemic of cholera, which had not been seen on the island nation previously resulted in approximately 600,000 cases and upwards of 7,500 deaths [4]. In order to investigate this outbreak 23 genomes were sequenced using the Illumina platform [41]. These sequences were taken over a variety of time points and geographic locations within Haiti. They were then compared against 85 different isolates of cholera retrieved from the National Center for Biotechnology Information (NCBI). These 108 genomes were compared to identify 566 positions of variance in the genome (known as Single Nucleotide Polymorphisms or SNPs) which were used to construct a phylogeny. After the phylogenies were compared it was found that the Hatian strains of cholera were most closely related with samples taken from Nepal, which itself grouped closely with samples from India and Bangladesh.

1.2 Phylogenetics

Phylogenetics is the study of evolutionary relationships comparing heritable traits across a set of biological entities. Deoxyribonucleic acid (DNA) and protein sequences

provide a great deal of information for phylogenetic analysis: as sequences mutate over time, the rates of these changes can be estimated through observation, and used to infer evolutionary history. While the size and content of genomes can vary widely among organisms, the use of core genetic components such as 16S ribosomal ribonucleic acid (rRNA) genes allow for distantly related organisms to be compared due to the universal presence of the 16S gene in all living organisms, as well as its stability within the genome. These exact properties however make rRNA unsuitable for very closely related organisms, such as strains of pathogens. This is because 16S genes have a high level of conservation, making them good for comparing distantly related organisms but a poorer choice for closely related organisms. A whole genome comparison approach is more suitable as it takes into account any changes which take place in the genome.

Several methods exist to produce phylogenies, one of the most popular being Maximum Likelihood (ML). This method was first applied to phylogeny creation by Felsenstein in 1981 [24] , and since its first introduction has been a favoured statistical approach to creating phylogenies. Popular implementations of the ML approach include RAxML [71] and PHYML [33] , which all seek to create more computationally efficient implementations using heuristics. A competing approach for phylogeny creation uses Bayesian statistics in order to create trees based upon observed rates of transition between genetic characters. MrBayes [65] is a piece of software commonly used to contrast different hypotheses about the relationships among taxa.

One pervasive problem of the microbial world is that microbial genomes are fluid in their genetic make-up. Lateral Gene Transfer (LGT) is a process which allows different bacteria to take up sequences from close and distant relatives, in addition to the genetic material they inherit directly from their ancestors [70]. This hampers attempts to place bacterial species in the evolutionary tree as one can never be sure if two species appear closely related because of a recent speciation event, or an acquired gene sequence. Problems also exist using gene content to compare organisms, as comparing different genes can yield wildly different tree topologies. The adoption of genome sequencing and more advanced methods of phylogenetic reconstruction can address these problems. In particular using whole genomes to build phylogenetic trees

does not rely on any one gene to construct the phylogeny, and can therefore produce an overall pattern of similarity that may be more reflective of parent-offspring inheritance [9].

Canada's IRIDA project is an example of a resource that aims to automate the entire genomic epidemiological pipeline, from pathogen isolation to phylogenetic analysis and interpretation. IRIDA is concerned with assembling and annotating pathogenic organisms for real time investigation. IRIDA seeks to facilitate the use of WGS in epidemiological studies by providing a centralized source of WGS compatible pipelines. A key component of the project is the SNVPhyl pipeline, which uses the presence of SNPs to create phylogenies. The whole process is automated as a Galaxy workflow [60]. The goal of the IRIDA system is to be able to identify close relatives of outbreaks as they happen in order to both identify what is the causative agent of an outbreak as well as its most likely geographic origin.

1.3 Phylogeography

Phylogeography is the application of phylogenetic models of diversity in conjunction with the geographical locations of samples of a set of organisms. It aims to combine these phylogenies and their geographic, environmental and temporal placement to identify possible drivers for their evolution and current geographic distribution. By addressing these questions phylogeography is able to address questions in public health [1], species divergence [50], conservation studies [55], and other areas.

Often phylogeographic tools are developed for use in particular domains such as conservation studies. Many of these tools were developed as extensions to existing visualization solutions such as R [49] or ArcGIS. Online tools for geographic visualization like CartoDB [17] focus on providing a robust suite of visualizations for a general audience via web based tools. GEOPHYLOBUILDER is a tool that extends ArcGIS to allow for phylogenetic information to be displayed, queried and analysed [42]. A key advantage of this package is its ability to visualize and compare multiple phylogenies. Tools like SPREAD [8] approach the problem of geographic diffusion

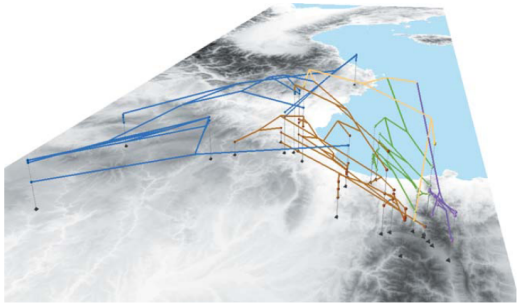
(a)  (b) 

Figure 1.1: (b) An example of the SPREAD program visualizing the geographic history of rabies in raccoons across the eastern seaboard under a continuous diffusion model. Image taken from [8]. (a) A visualization of *Ephippiger ephippiger*, a European Bushcricket, created with GEOPHYLOBUILDER such that individuals are the tips of the tree, and are connected to the triangular sample locations. The phylogeny is color coded by deep clades. Image taken from [42].

by using Bayesian models to predict the dispersion of species over time and geography. SPREAD also allows statistical analysis to be performed from ancestral nodes of the phylogenetic tree by accessing evolutionary information from all levels of the tree. GenGIS is another stand-alone application used for phylogenetic analysis and visualization [59]. Like SPREAD and GEOPHYLOBUILDER, GenGIS allows for phylogenies to be displayed on a geographic space. It also allows for analytic tools to access the weights of the phylogeny, as well as access to Python and R for complex statistical analysis. GenGIS also offers several 2D and 3D tree visualizations, including algorithms to optimize the layout of the tree relative to a set of geographic points.

One problem in phylogeography is that it is limited largely to a fixed geographic projection. This can become confusing for large amounts of dense geographic records, creating a large amount of clutter in one geographic area. This problem is compounded in phylogeography as location colouring is often used as an informative

display, which for large amounts of divergent or encoded locations can result in dense clustering of rainbow-like points, creating more confusion than explanation. GenGIS differentiates between samples and locations [59], which limits the number of overlapping points necessary on a map, but offers visualizations like pie charts to display the composition of each site.

1.4 Density-Equalizing Maps

Density-equalizing maps or *cartograms* are a form of map representation that adjust the traditional equal area map based on some quantitative variable of interest. These map distortions can be achieved in many ways, with the earliest ones being produced by hand [19]. Later advances in technology allowed for mechanical approaches to cartogram creation relying upon ball bearings to flex thin metal bands representing political divisions [68]. The advancement of computer systems allowed for automatic computational approaches to be developed. Waldo Tobler was the first researcher to develop such a method, which overlayed the map with a discrete grid. By slightly adjusting each grid cell with a linear function in concordance with the variable of interest the map can be distorted over many iterations, producing a cartogram [74].

Henriques et al. and Sagar both define the cartogram error as

$$\text{Relative Area Error} = \frac{|A_j^{Desired} - A_j^{Current}|}{A_j^{Desired} + A_j^{Current}}, \tag{1.1}$$

where $A_j^{Desired}$ is the expected position of geographic feature A_j under perfect distortion and $A_j^{Current}$ is its current position during cartogram iterations [67][37]. This provides the error for different political boundaries within the global context. Gastner and Newman provided a less specific error formula for each region, which can be generalized to

$$\text{Relative Error} = \frac{\text{Area of Cartogram} \times \text{Total Attribute of All Regions}}{\text{Total Area of All Regions} \times \text{Attribute of Region}} - 1, \tag{1.2}$$

[27]. Both methods of error calculation rely upon a vectored shape file representation in order to have perfect knowledge of the desired effect of distortion, as well as the area and attribute effects on each political border.

Here I will introduce several prominent approaches for creating density equalizing maps in overview, Self Organizing Maps (SOM), mathematical morphology and diffusion-based cartograms.

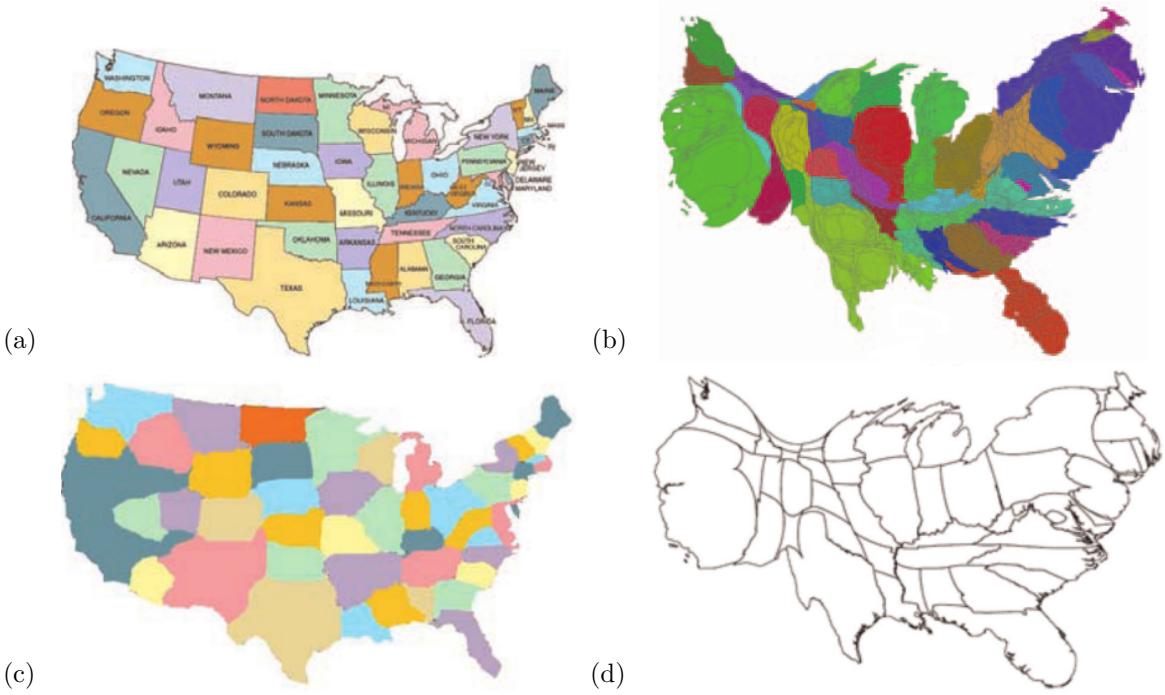


Figure 1.2: Examples of cartograms of the United States using population as the variable of interest. (a) is a standard projection map, (b) is generated using Self Organizing Maps, (c) is generated using mathematical morphometrics and (d) uses diffusion [67]. All images taken from [37].

1.4.1 Self-Organizing Maps

A recent method for cartogram creation was developed by Henriques et al. in 2009. This method known as CartoSOM uses SOMs in order to create accurate cartograms. SOM is a method developed by Kohonen in the 1980s [44] which employs a Neural Network (NN) in order to learn a mapping between an input space and an output space such that higher dimensional data may be displayed in a human readable space while retaining spatial relationships between individual points. Typically the output space of a SOM is two dimensional [37].

CartoSOM starts by dispersing points uniformly throughout each unique geographic region. Regions are generally defined as political borders in practice. These points are dispersed in accordance with a linear function, which represents the attribute of interest. For example, if one is creating a cartogram based on population, areas with larger populations will contain more points than those with smaller populations. After this step a uniform grid is laid over the map, which makes up the SOM. The training phase is then initialized such that the units of the SOM grid move in the map, attempting to mimic the distribution of the points placed earlier. The NN then applies a label to each node of the SOM based on the node's location in the map. Each node is then mapped back to its original position, morphing the area of the map to reflect the variable in question rather than geographic area.

This approach has several limitations. Performance improvements to this method have focused on parallelization alone, as opposed to algorithmic refinements, which means that the run time of this approach is directly proportional to the amount of threads it can be exposed to. SOMs also suffer from low-density magnification, whereby areas of low interest are artificially inflated in order to keep them from shrinking from the map. While this may or may not be of benefit to a given cartogram, it needs to be taken into consideration. Algorithms do exist which can relieve some of the bias applied by this magnification [5]. Figure 1.2 shows that SOM based methods are at least competitive with other methods in creating recognizable and representative cartograms.

1.4.2 Mathematical Morphology

One limitation of many traditional approaches to cartogram creation, such as those proposed in [74] or [28], is that the global shape of a map is rarely preserved after a transformation. As such it can become difficult for those not well versed in the geographic position and ordering of a locality to intuitively recognize the subject of a cartogram. Many methods rely on local shape preservation to convey this information, such that enough of the political divisions preserve enough of their shape that the global body is still identifiable. Mathematical morphology-based cartograms use

the opposite approach, allowing local regions to morph significantly while keeping the global shape intact. An example of this can be seen in Figure 1.2c where the global projections of the US remains accurate, but California has wrapped around surrounding states.

In practice the computation of mathematical morphology as related to map projections requires three steps: the definition of states or zones, the computation of centroids, and finally the calculation of the cartogram. The first step is analogous to the starting density in diffusion based cartograms, or the definition of a lattice in cartograms proposed by Tobler [28][74]. In effect political divisions are largely considered to be the most prudent choice, as they provide a human recognizable boundary. Computing the centroids for each of these zones is performed using the Minimal Skeletal Point (MSP) computation [67], which consists of creating a skeletal network of points for each region. Each network is then pruned of exterior points recursively until only one point, or one set of core points, is left. This area is defined as the MSP, and is considered the centroid of each state. From here each centroid is the seed location for Weighted Skeletonization By Influence Zones (WSKIZ). One way to imagine this application is that each centroid is a lake, and that flood water from each lake is allowed to propagate through the map according to the flow defined by distribution of a target variable [67]. This allows for more dense areas to flow into areas of low influence.

One of the significant advantages of this technique for cartogram creation is its preservation of the global geographic shape being distorted. Unfortunately in order to do this individual regions often lose recognizable shape. Conceptually this is an obvious side-effect of the morphological approach, as the distribution of area behaves like flood waters. Areas of high density will flow into areas of low density, conforming to the boundaries of whatever shapes are around them.

1.4.3 Diffusion

Diffusion based density-equalizing maps are currently one of the most popular cartogram algorithms in use. Developed by Gastner and Newman in 2004, this

physics based approach has been adopted by ArcGIS, and is also available as a C based package from Newman (`http://www-personal.umich.edu/~mejn/cart/`). Like most techniques for cartogram creation, the diffusion-based approach is rooted in the methodology created by Tobler [74], though improvements were made in both the segmentation of the areas to warp, and the technique used to warp these segments. The motivation for this approach in cartogram creation came from the limitations in the approaches of the time, such as Tobler's *Rubber Band Cartograms*, Gusein-Zade's *Line Integral Method* or Dougenik's *Contiguous Area Cartogram* [37], which in terms of both speed and accuracy were far from optimal solutions. By appropriating a calculation for spatial distribution from elementary physics, an approach was created which made significant advances to overcome these shortcomings [28].

Diffusion based cartograms are an application of the partial differential equation for diffusion employed in physics [28]. The main principle of this algorithm is that areas of high density are allowed to flow, or diffuse, into areas of low density. This change in density coincides with a change in area representation, e.g. the farther an areas density diffuses the larger it will be in the finished distortion. In practice the Fast Fourier Transform (FFT), as well as the backtransform are applied to solve both the population density function as well as the velocity function required to compute the diffusion of a given area. This is accomplished by treating the density function as a cosine Fourier transform, and solving the diffusion equation using the diagonal. By employing this technique, despite increased computational complexity compared to Tolber's method, run time is accelerated enough to offset the additional complexity. Despite the use of FFT, the algorithmic complexity of this implementation is at least no worse than most other methods of cartogram creation developed in the 1980s and 1990s, as complexity is heavily tied to the dimensions of the map, or the dimensions of a sub-sampled matrix representation [28]. It creates cartograms with levels of error that are typically smaller than other algorithms such as those achieved by methods proposed by Dorling, Sagar and Gausian-Zade [67][37].

The advantages of this approach to density-equalizing map creation are threefold: execution speed, readability and spatial accuracy. By allowing users to choose a trade off between density equalization and map readability via the scale of distortion,

agreeable geographic representations are nearly guaranteed within a few iterations of this technique. While the accuracy of a given projection relies heavily on the initial density, the commonly used solution of using political regions offers a good trade-off between readability and accuracy. By restarting the diffusion process on a fully diffused cartogram for a population-based distortion of the United States of America, Gastner and Newman managed to reduce relative errors in the worst performing political boundaries from 20% and greater down to at most 3.5% [27]. They argue that further iterations would continue to decrease the relative area accuracy of their approach, but that for most applications 3.5% will be indistinguishable by the human eye [27].

1.5 Overview

Here we present a method for reconciling geographic and phylogenetic information for location-based observations. Locations are defined as unique positions on a map or image, such as geographic coordinates. These locations correspond to one or more organism observations and must also have a position in the phylogenetic tree in order to have a non-zero score assigned to them. In cases where more than one sample exists at a location all samples are compared against their neighbours. Our approach differs from other phylogenetic measures at it incorporates a geographic backbone to complement phylogenetic relationships. This backbone is used to distribute phylogenetic weights among neighbours. We then use these location-based scores to apply a diffusion-based cartogram in order to emphasize areas with high or low scores depending on the goal of the visualization. These areas will expand in the projection, de-emphasizing low information portions of the map. Heuristic improvements, such as a density matrix reduction, were introduced into the cartogram process in order to decrease runtime and RAM consumption.

Chapter 2

Geographically Coupled Phylogenetic Distance and Map Distortions

2.1 Quantifying the Relationship Between Phylogeny and Geography

Species diversity measures are important tools to describe the variety and relative abundance of species [48]. Species richness, defined simply as the number of distinct species present in a site or region, is one of the most commonly used approaches to quantify diversity in ecology. Species richness is overly simplistic for many problems, and suffers from over sampling problems [12] and comparison problems [31]. There also exist a variety of statistical techniques to estimate richness in dense areas such as rainforests [15][13][10].

Many measures have been developed to quantify diversity in different ways; this is referred to as the diversity of diversity problem [48]. They are all primarily concerned with reconciling two kinds of information, species richness and *evenness*, which considers the relative abundance of species at different sites in addition to their presence or absence. Endemism is a concept frequently applied to conservation studies to account for species diversity based on environmental uniqueness. Weigthed Endemism (WE) is a common measure of endemism, which quantifies species richness in terms of geographic distribution.

$$WE = \sum_{t \in T} \frac{1}{R_t}. \tag{2.1}$$

WE works by partitioning a geographic space into a discrete grid. Abundance records are then calculated for each cell of the grid. A WE score is then calculated for cell. This score is calculated using the inverse summation for each taxon t's range R in the grid over all taxa T in the cell. In this way a taxon found only in one cell

will have a WE of 1, and a taxon found in two cells will have a WE of 0.5. The larger the WE the more unique a species is to the geographic region. WE was found to have a high level of association with species richness for some species, which identified the need for other methods to quantify endemism [16].

Phylogenetic Endemism (PE) is an extension of this idea and combines geographic spread with phylogenetic information. Phylogeny is incorporated to account for problems of taxonomic classification. This is done by using the WE of all the taxa in a cell to normalize their PD, thus quantifying how phylogenetically unique the organisms in a cell are, as well as how ubiquitous they are to the environment.

$$PE = \sum_{c \in C} \frac{L_c}{R_c}, \tag{2.2}$$

Where c is a single branch on the minimum spanning path C which joins the taxa to the root of the tree. R_c is the union of all ranges R_t for each taxon descendent from branch c [66]. By combining PD and WE, geographic sites can be compared based on the number of evolutionarily distinct species they have, and the extent of overlap between the most distinct species. This allows policy makers to create the most beneficial areas of conservation by identifying the regions that contain the most phylogenetically distinct species over a geographic range which contains a maximal number of such species.

Such measures can be of considerable use in epidemiology as well, as phylogeographic measures can assess whether pathogen variants tend to be narrowly or widely distributed. For example, avian influenza is a widely distributed strain of the influenza A virus which spreads largely through fecal contamination. Avian influenza affects global populations of birds, especially the young, and spikes seasonally with migration patterns [51]. In contrast Human Immunodeficiency Virus (HIV) is a much more locally concentrated pathogen, spreading more slowly in populations through sexual interaction. The disease is prevalent in Africa, affecting 20-30% of the residents in the worst afflicted urban centres, with rates steadily climbing with age [2].

Richness counts were used in a study of global emerging infectious diseases (EID) events, but no further phylogenetic information was incorporated [40]. Geographic

visualization was limited to the distribution and count of cases at specific regions. Patterns of infection were apparent by number of events, but not comparable. Location based data has been used to estimate species ranges in ancient fossil records. Species ranges were described by plotting each record on a map and creating a polygon around each observed record [72]. This differs from other approaches in species range calculations as it does not rely on any statistical inference of geographic diffusion [22]. These records were used to investigate ancient species invasions, and found such events correlated with sea-level depositional environment changes [64]. Phylogenetics and geography have been used to investigate patterns in plant distribution. Here Donoghue et al. attempted to reconcile several phylogenetic approaches to infer species deviation events. Initial clade sorting into geographic tracks was performed based on sample distribution in major areas of endemism. Hypothetical divergence events inferred from phylogenetic information were visualized using cladograms [18].

Our measure, the Geographically Coupled Phylogenetic Distance (GCPD), aims to reconcile geographic information with phylogenetic trees to quantify the spread and relatedness between samples using locations as proxies for geographic entities, visualizations that can account for the uneven clutter in GIS visualizations from biodiversity samples. This contrasts techniques such as a visual zoom by keeping the global geographic context of the projection.

2.2 GCPD Overview

The GCPD relies on a given location set containing two kinds of information: a phylogenetic tree with associated branch lengths that allow the calculation of PD and location describing the position of each taxon. The coordinate system will typically be geographic, but any two-dimensional coordinate system can be used. These two types of information are reconciled over geographic space using a network structure linking the taxa of interest as vertices. The edges between each pair of taxa are used to combine the phylogenetic information stored at each, using the phylogenetic information as weights. Through this process each location is assigned a value which expresses the similarity or dissimilarity of the samples at the location display with

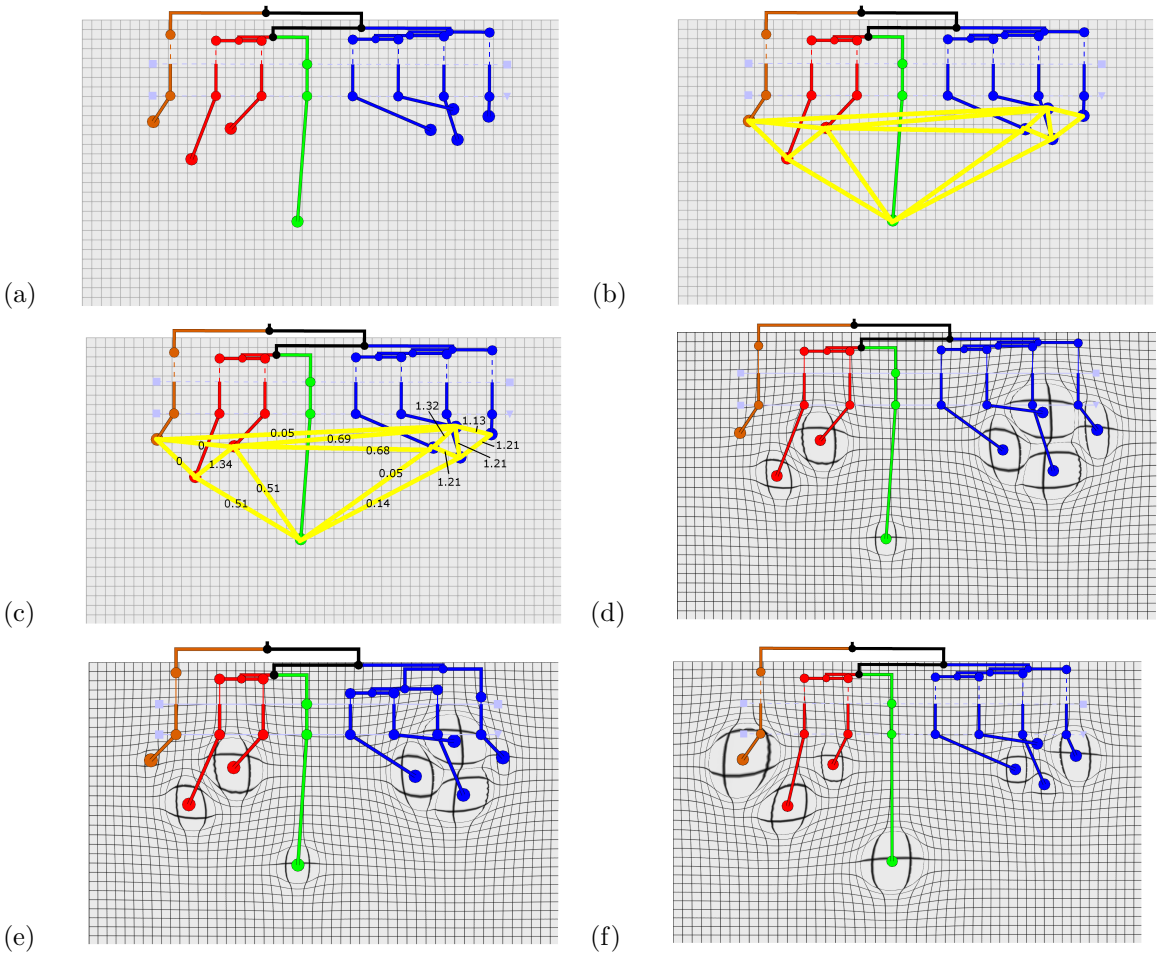


Figure 2.1: (a) A set of geographic locations representing the distribution of samples in a space with an overlaid phylogram. (b) Creation of the Delaunay triangulation. (c) Each branch in the triangulation is prescribed a weight based on the patristic distance between all samples at each connected location. (d) A density equalizing cartogram created using the Gastner-Newman method and GCPD. (e) outlines a similar cartogram as (d), with one branch of the phylogenetic tree has been changed to increase the distance between the taxa and the rest of the tree. (f) a cartogram created using the same branch weights as (d) and 1-GCPD.

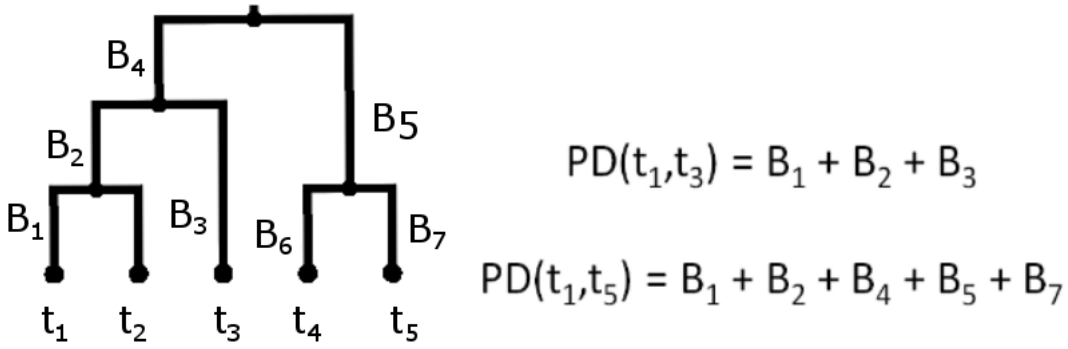


Figure 2.2: An example of the PD calculation on a five-taxon tree. Here PD is calculated between taxa t_1 and t_3 as well as t_1 and t_5. The PD of t_1 and t_c is calculated as the sum of all branches in the path between t_1 and t_3.

their geographic neighbours.

2.2.1 Definition of GCPD

In order to incorporate geographic and phylogenetic components into one measure a geographic scaffold which links all locations together must first be created (Figure 2.1b-c). This scaffold is used to represent the spatial relationships among all locations, which due to their discrete nature leave large swaths of uninformative area between them. This empty space presents problems to many methods, such as endemism calculations [66], as they expect continuous metadata over the landscape. All methods of cartogram creation presented above also expect continuous data, which is available in data formats such as vector map files through large scale ecological studies. Creating an all against all scaffold would be unwise as such a comparison, while producing a robust graph, would create a uniform value for all locations. Such a comparison would be likely to bury any phylogenetic relationship in noise. It is more pertinent to use a measure which introduces the shortest paths between locations, as they are more likely to be the true paths taken by species as their populations diffuse.

This assumption is of course not guaranteed to be true as migration paths are not guaranteed to be linear. For example glacial refugia[46] confine members of the same species to small, disjoint areas, which is often followed by outward expansion when the ice recedes. To identify shortest paths, Delaunay triangulations are deployed, which are a dual graph of the Voronoi Tesselation, a method used to subset the area of a plane into the largest possible fields from seeded centroids. A Voronoi Tesselation is defined as

$$R_k = x \in X | d(x, P_k) <= d(x, P_j) \forall j \neq k \tag{2.3}$$

where x is a position in metric space, d is some distance function (typical geographic examples employ Euclidean or Manhattan distances), K is the set of all indices, $(P_k)_{k \in K}$ is a tuple of all sites within X, and R_k is the Voronoi region associated with site P_k. Another way of explaining R_k would be to say that it is the set of all points in X whose distance to some site P_k is not greater that its distance to any other site P_j, where j is any non k index.

The Delaunay triangulation is a useful extension of this algorithm, which instead of finding maximal areas belonging to each point, finds optimal paths between points. This is accomplished by creating edges between three points, i.e. a triangle, such that the concentric circle whose circumference passes through all three points does not pass through or contain any other point. The Delaunay triangulation possesses the same properties as a Minimum Spanning Tree (MST), Gabriel Graph, and a Relative Neighbourhood Graph [58]. Here the Delaunay triangulation is used for its MST properties to create a scaffold over a set of locations. This undirected scaffold may be thought of as a set of possible shortest paths which represent candidate transmission or migration routes between sites. In the second step of GCPD, phylogenetic information is added to the geographic scaffold. This integration allows geographic distance to be balanced against phylogenetic distance, with regional effects dependent on the correlation between these two properties.

Evolutionary history is incorporated using the patristic distance definition of PD. The patristic distance of two samples is the sum of branch lengths in B (where B is

the set of all branch weights) between any two leaf nodes i,j in the phylogenetic tree taking the shortest path between each node. We then introduce an indicator variable $x_{ij,k}$ which is 1 if a branch b lies on the shortest path between nodes i and j, and 0 otherwise. The path between i and j does not need to pass through the root. This can be expressed more formally as:

$$PD_{i,j} = \sum_{k} x_{ij,k} b_k. \tag{2.4}$$

Depending on whether the intention is to highlight the similarity or dissimilarity between sites distance can be interpreted as 1-PD or PD respectively. The distinction here is important, as using just PD will highlight samples which have less in common in terms of sequence content, while 1-PD will assign larger values to two locations containing highly similar evolutionary content. Using PD is suitable for applications in conservation study such as endemism, while 1-PD is useful when trying to ascertain similarity such as the case of pathogen outbreaks.

A similar approach was taken by [66],who normalized the PD of samples by the geographic area covered by each taxon in order to identify the most important areas for conservation. Our approach differs from theirs as PD is not taken as the sum of all branch weights on a path through the root of the tree. We made this distinction as Rosauer et al. [66] wanted PD to be consistent across all samples as a proportion of the total length of the tree in a range between 0 and 1. In their case this also forced the PD and PE measures to be on the same scale. We found this method to be of little benefit as it forces paths between taxa which neighbour each other in a phylogenetic tree to first pass through the root of the tree, significantly increasing the amount of branches which need to be incorporated, and inflating the PD score of such a path.

To add phylogenetic information to the triangulation, every edge in the triangulation is given a weight corresponding to the PD of the taxa at each location, as seen in Figure 2.1c. If multiple taxa exist at each location they must be combined. It is possible here to select three options for combination: sum, local average and global average. Summation provides the easiest interpretation, but places bias on locations with more samples. The local average normalizes GCPD by the number of edges each

node in the triangulation has respectively, so that locations with more branches are down-weighted more heavily than those with fewer branches. This provides the best measure for how phylogenetically homogeneous a given location is in its local area. Alternatively global averaging normalizes the GCPD by the total number of edges in the triangulation, so that each location is divided by the same value. This allows for a representation of how important a given location is, and how strongly it affects the overall network signal.

Input: L as the set of all geographic locations
MST = the Delaunay triangulation between all locations.
Result: GCPD scores for all locations
for $Location_1$ *in* L **do**
  T_1 = the set of all taxa at $Location_1$
  neighbours = all neighbours of $Location_1$ in the MST;
  for $Location_2$ *in neighbours* **do**
    T_2 = the set of all taxa at $Location_2$
    for t_i *in* T_1 **do**
      for t_j *in* T_2 **do**
        GCPD($Location_1$) += PD($taxa_i$,$taxa_j$)/2;
      end
    end
  end
  GCPD($Location_1$) = Norm(GCPD($Location_1$));
end

Where PD is a function which computes the phylogenetic distance between any two taxa. Norm() is defined as the user selected normalization function, which is either a sum, local average or global average as previously defined. The PD between any two taxa in this approach is divided in half so that each location can have a portion of the effect of that relationship.

2.2.2 Geographic Projection and Visualization

As the GCPD measure is aimed at quantifying evolutionary distance over geographic space, it makes sense to visualize this measure within a GIS framework. Recent work in epidemiological visualization has shown that providing geographically contextualized health based information creates a very real benefit to conveying the effects of geography and transmission [57]. This same research found that by employing cartogram methods to present findings, policy decision makers were more likely to understand the findings. Density-equalizing maps can have a similar representational effect on GCPD to present the role different regions play in the phylogenetic spread (Figure 2.1d).

The advantages of the diffusion-based approach to cartogram construction (see 1.4) make it the best option for GCPD-based transformations. By allowing a system which is agnostic towards the projection of a given map this algorithm provides a solution which works equally well for geographic formats as it does for flat images. This allows for the production of distortions in various contexts, from traditional phylogeographic approaches to epidemiological studies on flat images of hospitals, to cross sectional environmental studies. This projection goal falls in line with design goals of the GenGIS platform in supporting many different projection systems.

Diffusion based cartograms have been shown in shape file maps to preserve geographic features more than other approaches to cartogram creation[37]. As all of the mentioned implementations expect data encoded shape files as inputs, the transition to a location based information system will compromise some of this accuracy, as geographic features are not encoded in a location based representation. Biodiversity information stored in a location based system is typically sparse by nature as well, only containing pertinent information in the points associated in the location file, while vector shape files contain continuous metadata throughout each shape. This means that empty areas of the map are a known quantity with a shape file, where as in location based observations every cell that does not contain a location is naively considered to be empty. When creating a cartogram from such information under a shape file each region is prescribed one consistent value, creating uniform diffusion

into the surrounding areas. Under a location based approach however no information is known about the area, only the local point. This means that diffusion can only occur outwards from this point, with no information about border placement. As such it makes sense to start with an approach which preserves global shape so warping does not jeopardize the geographic context when the distortion is location based.

Diffusion based cartograms more readily address this problem as they rely upon a density matrix interpretation of the map, where the number of rows in the matrix equates to the height of the map L_y and number of columns the map width L_x. This L_x by L_y matrix is the operational counterpart of the original map projection. The boundary limits are defined automatically as the dimensions of the map, and the number of divisions of the matrix upon each axis is based upon the underlying wire frame in the raster map projection.

As our geographic representation is location based, this density matrix will be filled in with the location points, such that the centrality of the points of interest can be represented in the matrix, converting latitude and longitude into the corresponding row and column in the density matrix. As these locations are discrete points they will not belong to more than one cell of the matrix, which will typically produce a sparse matrix. For the example of a 4050x2050 matrix (which is not an overly large pixel density for a high quality map) containing 100 distinct locations less than one thousandth of the cells will contain populated information. As the diffusion method of cartogram creation expects richly populated information from a vectored image, leaving the density matrix so sparsely populated would result in these points diffusing very quickly into an overwhelmingly uniform space and ultimately producing very little, if any, noticeable change in the map projection.

Due to this, the area around the informative cells should then be populated in accordance with some function $f(x)$ which may be used to disperse this value of interest in the density matrix. By increasing the number of cells holding values of interest the amount of diffusion is increased. The tuning parameter which controls the diameter of this dispersal is called location diameter. To provide further control over the amount and quality of the distortion a variable multiplier is provided, which increases

the values stored in each location. These act to further increase the amount of flow within the cartogram. Each parameter effects the shape of the completed distortion and is not informed by the geographic or phylogenetic information. Selections for these parameters will differ with sample distribution and geographic features.

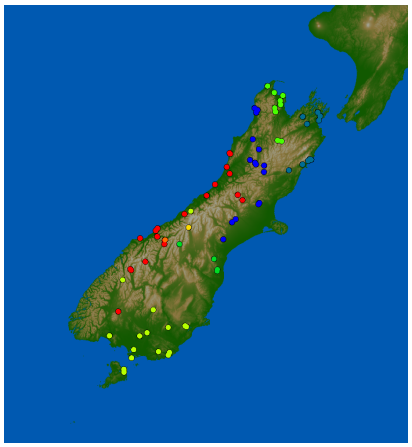
In order to account for the unpopulated portions of the density matrix the "neutral buoyancy" condition applied by Gastner and Newman is used [27]. The neutral buoyancy condition automatically places a boundary of uniform population density around the area of interest which stops the flow of diffusion, as its value is equivalent to that of a fully diffused matrix. The neutral buoyancy condition prevents the release of density into the areas beyond the borders of the map. This principle is similarly applied to account for the empty "sea" of density around the points of interest, creating large areas of low resistance to absorb the diffusion and allow the map to disperse.

2.2.3 Integrating Cartograms with GenGIS

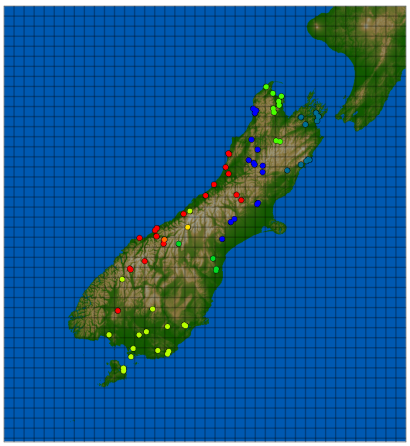
Applying the Gastner-Newman method of cartogram creation requires significant modifications to the data model in GenGIS. GenGIS typically integrates three types of data: the map (which can be a raster file or image file), the locations and a phylogenetic tree. The map can be either a raster projection or a vectored shape file, or can include both types of representation. Locations are made up of individual points on the map primarily consisting of geographic coordinates. Finally the phylogenetic tree is a Newick file representation of the phylogenetic relationships between the locations, which may or may not contain informative branch lengths.

The Gastner-Newman algorithm requires a density matrix representation of a map and its associated distortion values to create a cartogram. This density matrix is used to create a model of how much each area of the map needs to move to create uniform diffusion. This model is then fed individual (x,y) coordinates and returns their transformed positions. In order to make this process compatible with the information stored in GenGIS the grid representation which underpins the geographic coordinates of the raster map is used as the basis for the number of divisions in the

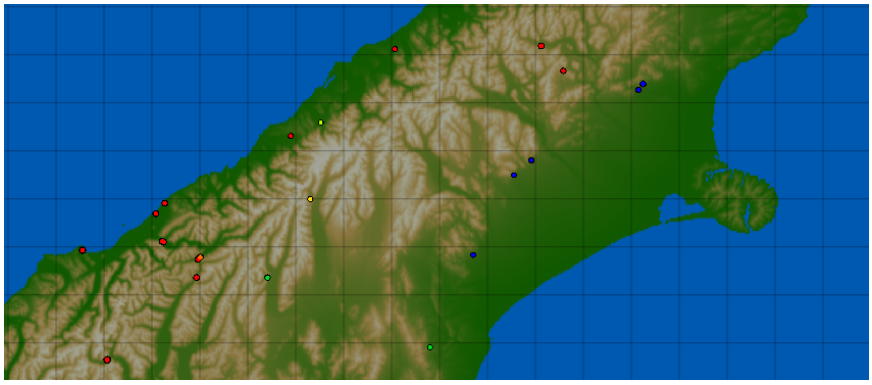
(a)



(b)



(c)



(d)

0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	2	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	1	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	0.1	2	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	3	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	1	0.1	3	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.1	0.1	1	0.1	0.1	0.1	0.1	0.1	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Figure 2.3: Creation of the density matrix. (a) A base map with sample locations indicated. In this example location counts will be used. (b) A coarse-grained empty matrix overlaid onto the map. (c) A zoomed in cross section of the original map. (d) Location count values added to the matrix using no location diameter or variable multiplier effects. All cells of the matrix which do not contain locations are given a value of neutral buoyancy.

density matrix. The locations on the map are then translated onto their corresponding positions in the matrix, and it is populated as described in Figure 2.2.2.

After the matrix is populated and the model of diffusion is created, each of the GenGIS layers needs to be interpolated from its original positions to its diffused state. The map layer is the most intuitive to interpolate, as the diffusion matrix is in its native coordinate system. The location and vector layers need to be translated into the map's coordinate system before they can be interpolated as they rely on latitude/longitude, UTM or pixel coordinates. The new positions need to be translated back to the location and vector coordinate systems respectively after interpolation.

2.2.4 Diffusion-based Density-Equalizing Maps

The method of Gastner-Newman diffusion based cartogram creation which we have adapted relies heavily upon Fick's two laws of diffusion [28]. These laws serve as a method in physics to model the flow of molecules in a confined container. The objective of this diffusion is for the container to be filled in such a way that the density of molecules is uniform throughout. Fick's first law of diffusion states that diffusion in two dimensions can be described as

$$J = -D\nabla\phi, \tag{2.5}$$

where J is the diffusive flux, D is the diffusion coefficient, ϕ is the concentration of the substance per volume and ∇ is the gradient of diffusion. The steeper ∇ is the faster the substance diffuses in time.

Fick's second law is a combination of conservation of mass and Fick's first law. It is used to predict how the concentration of a substance changes over time as it diffuses. It can be described as the partial differential equation

$$\frac{\partial\phi}{\partial t} = D\frac{\partial^2\phi}{\partial x^2}, \tag{2.6}$$

for a one dimensional space with a constant diffusion coefficient. Here t is time and x is position. In two dimensions it can be generalized using Laplacian $\Delta = \nabla^2$

$$\frac{\partial\phi}{\partial t} = D\Delta\phi. \tag{2.7}$$

In the general treatment of diffusion, the current state of flux can be described as

$$J = v(r,t)p(r,t), \tag{2.8}$$

where v is a function of velocity, p is a function of density, and both functions operate over position (r) and time (t).

As in the production of a cartogram we are not interested in any one state of the diffusion process, but rather in running the process to convergence, time can be said to approach infinity and D can be set to one without loss of generality. As diffusion follows a gradient of least resistance, meaning that it always flows from areas of high concentration to areas of low concentration, Fick's first law can be generalized as

$$J = -\nabla p \tag{2.9}$$

for the case of cartogram creation.

The diffusing population is also conserved locally so that

$$\nabla \cdot J + \frac{\partial p}{\partial t} = 0. \tag{2.10}$$

By combining diffusion equation 2.9 with 2.11 and 2.8 we arrive at

$$\nabla^2 - \frac{\partial p}{\partial t} = 0, \tag{2.11}$$

and

$$v(r,t) = -\frac{\nabla p}{p}, \tag{2.12}$$

respectively.

The cartogram is created by solving 2.11 for p(r,t) for every position in the density matrix starting from the original state, which is the initial observed population density matrix. After this step velocity is calculated for each position at that time slice using 2.12. In this way the displacement $r(t)$ of any point on the map at time t can be calculated over the integral

$$r(t) = r(0) + \int_0^t v(r,t')dt'. \tag{2.13}$$

Integrating this equation as $t \to \infty$ solves the cartogram.

2.2.5 Applying the Fast Fourier Transform to Diffusion

The previously outlined approach is sufficient to solve the diffusion problem and create a cartogram. The procedure is computationally demanding, and algorithmic improvements are necessary to allow execution of cartograms in a reasonable time. The Fourier series is a trigonometric infinite series used to approximate waveforms. The Fourier transform [11] is an application of this series used to convert measurements in time into measurements in frequency, and the inverse Fourier transform can be used to reverse this operation. A wave of sound over time can be passed through the Fourier transform and a graph of the results would display which frequencies were present in the original wave and at what amplitudes. The Fourier transform is a well-known method to solve partial differential equations (like the diffusion process), quantum mechanics [52] and spectroscopy [21], for example. The transform was first introduced to solve the heat equation given as

$$\frac{\partial^2 y(x,t)}{\partial^2 x} = \frac{\partial y(x,t)}{\partial t}, \tag{2.14}$$

which looks very similar to the one-dimensional case of Fick's second law of diffusion 2.6.

The equation for the Fourier transform can be represented as

$$F(v) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i v t} dt, \tag{2.15}$$

for the forward transform and

$$f(t) = \int_{-\infty}^{\infty} F(v) e^{2\pi i v t} dv, \tag{2.16}$$

for the backtransform.

The full Fourier transform is not needed here, as $F(v)$ is an even series where the Fourier sine coefficients sum to 0. In this case the sine component of the calculation can be dropped, and purely the cosine representation over an even $F(v)$ can be used. Neumann boundary conditions are also applicable to this problem, as the solution to the equation is limited within the bounding box created by the width and height of

the map (L_x and L_y respectively). Considering these factors, the cosine base for the Fourier transform with Neumann conditions is represented as

$$u(r,t) = A_0 + \sum_n c_n cos(\frac{n\pi r}{L}) e^{\frac{-nk\pi}{L}^2 t}, \tag{2.17}$$

where r and t are position and time respectively, k is a physical constant between 0 and N-1, and A_0 is the mean of $f(r)$. The initial conditions of c_0 are defined as

$$c_n = \frac{2}{L} \int_0^L f(r) cos(\frac{n\pi r}{L}) dr, \tag{2.18}$$

for this case.

The base Fourier transform is on its own not a sufficient improvement in speed or algorithmic complexity, taking $O(N^2)$ over discrete time points. FFT is a much faster method to compute the discrete Fourier transform, and outputs a vector of size x_k, just like the standard Fourier transform. FFT has a complexity of $O(NlogN)$ using the widely adopted Cooley-Tukey (CT) approach, used in such packages as libFFTW[25]. The CT algorithm improves on the Fourier transform by splitting up the input matrix between even and odd indices($n = 2m$ and $n = 2m+1$ respectfully). These two matrices are then computed recursively as follows:

$$X_k = \sum_{m=0}^{N/2-1} x_{2_m} e^{\frac{-2\pi i}{N}(2m)k} + \sum_{m=0}^{N/2-1} x_{2_m+1} e^{\frac{-2\pi i}{N}(2m+1)k}. \tag{2.19}$$

By splitting up the original Fourier transform in this way the approach can be computed in a parallel fashion, using the cosine solution to solve for density.

$$p(\mathbf{r},t) = \frac{4}{L_x L_y} \sum_k \tilde{p}(\mathbf{k}) cos(k_x x) cos(k_y y) e^{-k^2 t}, \tag{2.20}$$

where the sum is over all wave vectors $k = (k_x, k_y) = 2\pi(\frac{m}{L_x}, \frac{n}{L_y})$ with m and n as non-negative integers. $\tilde{p}(k)$ is the initial condition discrete cosine transform of $p(\mathbf{r},t) = 0$

$$\tilde{p}(\mathbf{k}) = \frac{1}{4}(\delta_{k_x,0} + 1)(\delta_{k_y,0} + 1) \times \int_0^{L_x} \int_0^{L_y} p(\mathbf{r},0) cos(k_x x) cos(k_y y) dx dy, \tag{2.21}$$

where $\delta_{i,j}$ is the Kronecker symbol.

All of this leads towards the calculation of the velocity field for the diffusive process, which is the end goal towards making this calculation more efficient. This calculation is achieved by combining 2.20 and 2.12.

$$v_x(\mathbf{r},t) = \frac{\sum_k k_x \tilde{p}(\mathbf{k}) sin(k_x x) cos(k_y y) e^{-k^2 t}}{\sum_k \tilde{p}(\mathbf{k}) cos(k_x x) cos(k_y y) e^{-k^2 t}}, \tag{2.22}$$

$$v_y(\mathbf{r},t) = \frac{\sum_k k_y \tilde{p}(\mathbf{k}) sin(k_x x) cos(k_y y) e^{-k^2 t}}{\sum_k \tilde{p}(\mathbf{k}) cos(k_x x) cos(k_y y) e^{-k^2 t}}. \tag{2.23}$$

Run time for 2.20, 2.22 and 2.23 via FFT is $O(L_x L_y log(L_x L_y))$, in keeping with the base analysis of FFT run times. This means that ultimately in the creation of a cartogram the largest limiting factors in terms of run time will the size of the map used, as opposed to the number of locations. An example of the effect of map size can be seen in Table 2.1 where a 375x375 map takes slightly more than half the time to distort compared to a 750x750 map. The process outlined here will continue until all cells of the density matrix contain one uniform value and an equilibrium has been reached.

2.3 Implementation-specific Runtime Improvements

Our original implementation of GCPD-based cartograms took several minutes to process a high-resolution map(4050x2025 px). To increase the efficiency of cartogram creation, we examined the source code in an attempt to introduce acceptable heuristics. Profiler analysis showed that the main expenditures in terms of both function calls and CPU utilization went to the calculation and lookup times for the velocity function for diffusion. This is the process in cartogram creation which manages the force and direction of the diffusion in the density matrix over time. Implementations of this function already relied on caching, and it was the lookup times for values, which occurred eight times per step of the diffusion, that dominated the running time of the process. In order to optimize some of this process the velocity function was stripped of any extraneous calculations. This includes an interpolation which

averaged velocities of a 2x2 area in the density grid. This step was included to verify diffusion did not occur outside of the borders; however, border violations were prevented in a previous step of the diffusion calculation. Therefore this step could be removed without affecting the generation of the final cartogram.

This velocity calculation was also called in two steps when performing the diffusion. The first step performed a more large scale distortion, potentially causing a larger amount of diffusion. After this initial calculation two more modest velocity calculations were performed in order to fine tune the process. In order to reduce the number of velocity calls a threshold of 0.001 was placed on the initial velocity step, so that if only a small amount of diffusion occurred it was not tuned any further, as this tuning would have a very small effect on the overall quality of the distortion.

One of the simplest improvements made to the original Newman cartogram implementation was to remove a serialization step between the creation and distortion of the density matrix and the interpolation step with the original map. This step served two purposes in the original implementation, 1) to save the model so that it could be run against future maps and 2) to transpose the matrix representation. As reading and writing to hard disk is one of the most costly operations a computer can make, this step accounted for a large expenditure in computational run time. In fact for larger maps it could drastically increase the run time. Distortions using the cholera pandemic data on a 4050x2025 px map containing 61 locations required between 7 and 15 minutes using a single core of a 3.10GHz processor. Such lengthy run-times may not be sufficient to support parameter optimization. By integrating interpolation and distortion creation into one set of processes which share objects and memory the run time of large maps is reduced. By removing this read/write step RAM consumption was reduced as redundant objects containing the density matrix could be removed. Increases in performance scale with the size of the map Table 2.1.

Furthermore the use of data structures was not always consistent, with frequent switches from one-dimensional to two-dimensional arrays. These transformations inflated runtime by having to convert from one form to the other. Furthermore there is a cost to memory efficiency to store 4 arrays of thousands of points, as is the case

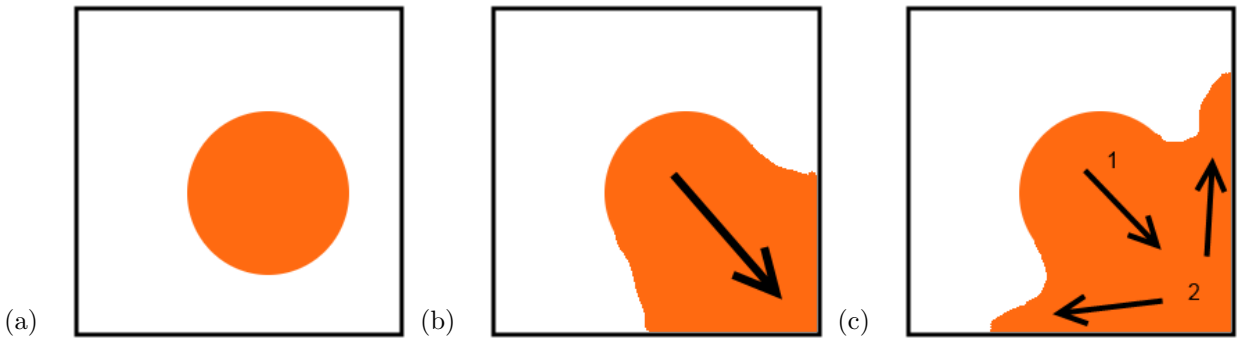


Figure 2.4: Depictions of possible diffusion pathways of isolated sections of the density matrix, where the subsections are not able to access the total area of the map creating artificial borders. (a) represents the starting shape of the object to diffuse, with force pushing towards the bottom right corner. (b) Pushing the path for diffusion into that corner with moderate force, building up pressure on the bottom right corner. (c) Applying a greater amount of force in the diffusion causes the diffusion to push back along the borders of the segment.

with a typical map. As x86 applications such as the GenGIS platform have a memory limit of 4 gigabytes, memory consumption is at a premium. As such these read/write steps were removed and replaced with index calculations which accounted for the expected orientation of the density matrix.

All of these improvements have been applied in the GenGIS implementation of diffusion based cartograms. This method is referred to as "Full GenGIS" in Table 2.1 and accounted for improvements in runtime between this technique and the "Original Gastner-Newman" approach.

One approach to speed-up the runtime that was considered but not applied was divide-and-conquer. Divide-and-conquer is an approach where the computation is split up into several subproblems (in this case, subsets of the map) and solved individually. The individual parts are then reassembled into a solved whole. The specific case of density matrix based diffusion poses a problem for this approach, as there exist no natural divisions of the density matrix which can be hived off and calculated independently. As diffusion relies upon the amount of force each cell in the density matrix exerts upon the surrounding region, splitting regions up would create many

small cartograms with pools of force along the borders. In some cases this force may be coerced to diffuse along the borders, corralling diffusion into areas of the map that would not otherwise have been affected had the original map not been subdivided. Figure 2.4 shows possible examples of just such a scenario, with Figure 2.4c being the worst case, which could cause disjoint sections of the map as well as producing large regions of error.

Such a problem may be alleviated in divide in conquer if at every step of recombination the diffusion process was reintroduced to smooth over such areas of complication. For each recombination step there is no guarantee that any computation would be saved due to the splitting of the density matrix. If high concentrations of density accumulate on cell borders even more movement would be needed to reach a neutral buoyancy. As the height and width of the density matrix are the bounding properties in terms of computational complexity, and divide and conquer would require extra diffusion steps at for recombination to reduce disjoint diffusion, it follows that divide and conquer would not produce a logarithmic complexity in this case. Instead it is much more likely that the runtime would increase to deal with the interpolation of the borders.

2.3.1 Density Graph Reduction

Given that the biggest effect on the run time of a cartogram is the dimensions of the image, an obvious way to speed up the operation is to distort a smaller map. Simply reducing the size of the map however is not adequate, as it compromises the quality of the initial projection. Intuitively a 720 px-wide image is not as clear as a 1480 px-wide image. Therefore rather than operating directly on the initial map the density matrix is the target of reduction. This can be accomplished independently of the map layer interpolation, which abstracts it away from any operation that is exposed to users of the software.

The first step of the process is to choose a reduction percentage (RP) for the density matrix. An initial density matrix is created based on the full map extents, so that no information is lost when populating the initial matrix (Figure 2.5a). A

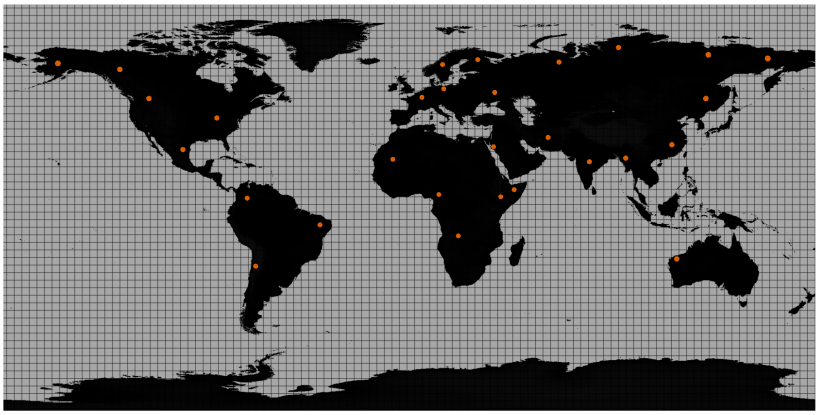
reduced matrix is then created which is RP smaller than this initial matrix. A process outlined in Figure 2.5 is then followed, where every width*RP and height*RP cells of the matrix are collapsed into one cell of the reduced matrix (Figure 2.5b). This means for a 33% reduction value, every 1.5x1.5 section of the original density matrix would be summed together and placed in a cell of the reduced density matrix. Obviously a cell can not be divided in half, so border cases would get values from both sides. During this reduction step the density values for each of the original cells are combined in an additive way, so that their corresponding cell in the reduced matrix is the sum of all parts (Figure 2.5c). The diffusion process is then carried out on the reduced matrix (Figure 2.6d), and as can be seen in Table 2.1, provides a decrease in run time even for small initial matrices. After the diffusion is carried out the matrix reduction step is reversed, replacing cells from the reduced matrix back into their original locations (Figure 2.6e). During the restoration process the values are divided by the RP to account for some of the value increase caused by the additive combination method.

In order to evaluate the affects of the heuristic improvements discussed here against the original implementation of diffusion-based cartograms we compared the run times and RAM consumption on three maps of different sizes (Table 2.1). Density matrix reduction was compared separately in order to assess the performance of the other approaches against the original implementation. RAM consumption was compared in megabytes and excluded any memory consumed by GenGIS prior to starting the cartogram process. In general run time and RAM consumption improves as more of the heuristics discussed above are invoked. The one violation to this trend is in the 375 x 375 map when comparing Pure Gastner-Newman with Full GenGIS, where we see a quicker execution of the former. Due to the small nature of the map and low execution speeds this variation may be due to scheduler interruption. RAM was still observed to decrease between these two runs.

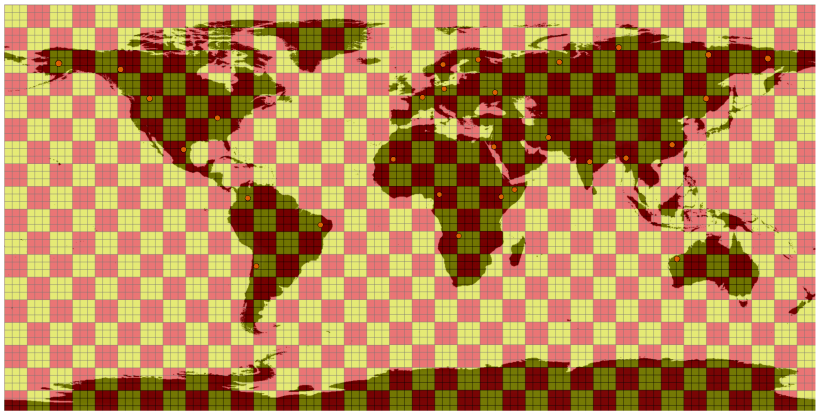
Map	Dim	Technique	Time(s)	RAM(Mb)	# Locations	Min RAM	Max RAM
Grid1	375 x 375	Pure Gastner-Newman	4.6	27.0	20	155.9	182.9
		Full GenGIS	7.1	20.9	20	158.3	179.2
		50% Reduction	4.4	13.0	20	165.3	178.4
Grid2	750 x 750	Pure Gastner-Newman	16.9	87.8	20	177.1	264.9
		Full GenGIS	11.3	81.2	20	179.5	260.7
		50% Reduction	3.2	18.9	20	167.9	186.8
Grid3	1500 x 1500	Pure Gastner-Newman	67.7	417.9	20	228.8	542.3
		Full GenGIS	36.9	312.5	20	229.8	542.3
		50% Reduction	11.5	81.1	20	223.2	304.3

Table 2.1: A comparison of 3 different maps and their run times under different cartogram scenarious. Gastner-Newman refers to the default application of diffusion based cartograms as implemented in the Cart library [54]. Full GenGIS refers to a distortion made with a full density matrix through the GenGIS application, and 50% Reduction refers to a distortion made using a density matrix which has been reduced by half.

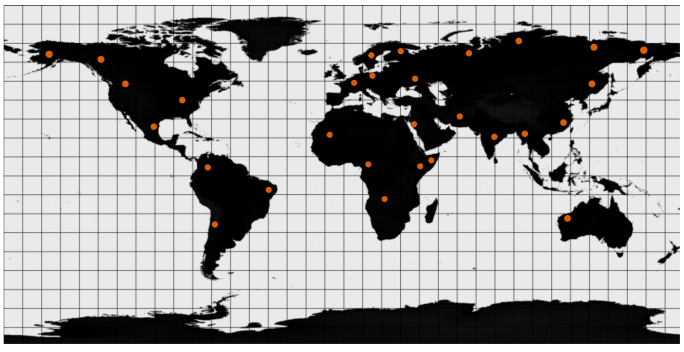
(a)



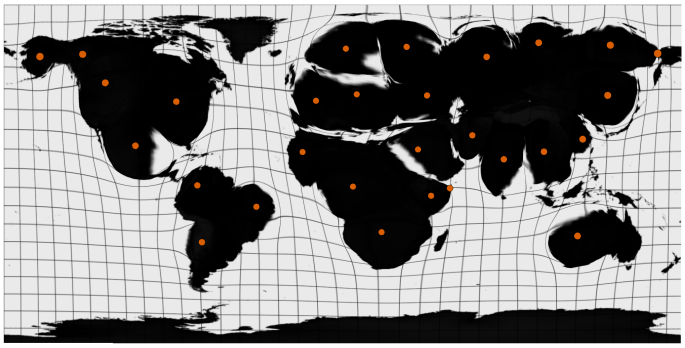
(b)



(c)



(d)



(e)

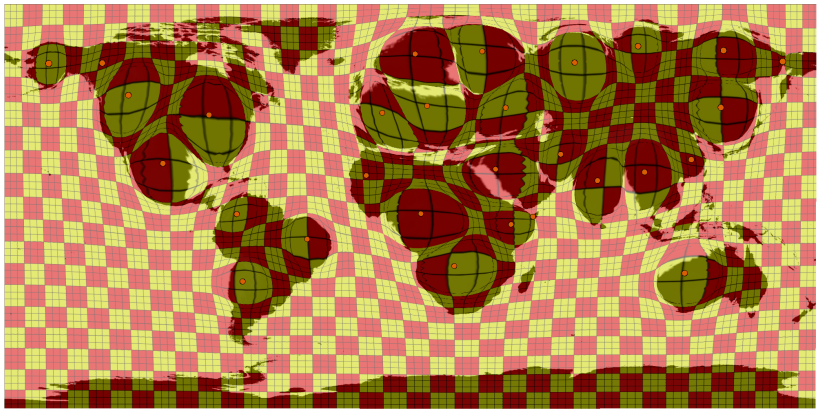


Figure 2.5: An overview of how the density matrix reduction effects the production of a cartogram. (a) creation of the density matrix. (b) reduction of the density matrix, in this case by 33% which results in every adjacent 3x3 section of the matrix being condensed into once cell which is represented here by the yellow and red sections. (c) the resulting map, 1/3 the size of the original. (d) diffusion of the reduced map, creating a scale cartogram of the original map. (e) reversal of the initial reduction, with each reduced cell restored to its initial position in the density matrix.

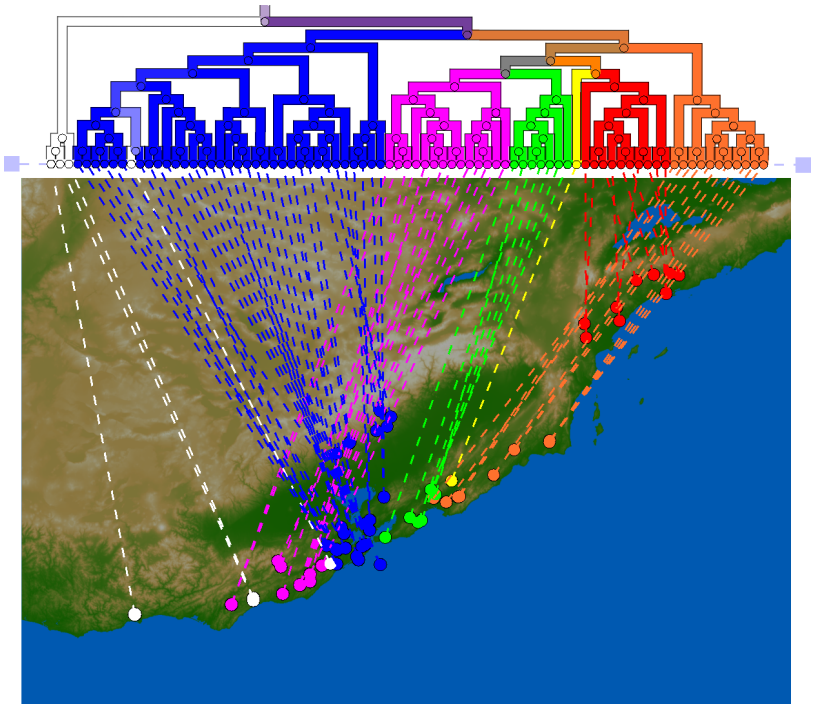
Chapter 3

Results

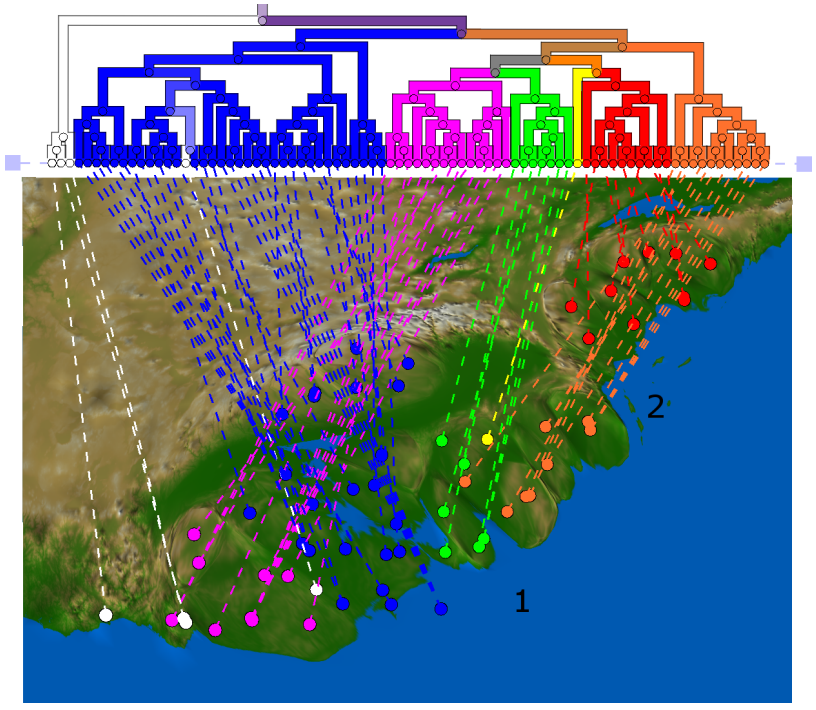
3.1 Visualizing Multi-locus Phylogeography of *Aneides lugubris*

California is one of the world's top 25 hotspots for species diversity [63]. As such it is a test pool for the effects of different ecological and environmental factors on a diverse set of fauna. The California Floristic Province is home to 44 species of salamander alone, 33 of which are endemic to the region according to AmphibiaWeb[45]. A high level of divergence among clades adds to clear phylogenetic distinctiveness with generally low levels of geographic dispersion. Their endemic state and environmental sensitivity may allow them to provide early warning of the effects of climate change. Investigating sources of genetic breaks through geographic and phylogenetic features may provide insight into processes which have shaped diversification of salamanders in the region. Reilly et. al. sequenced mtDNA (DNA from mitochondria, small structures found in eukaryotic cells), commonly used for phylogenetic analysis within species genes *ND*4 and cytochrome b. Sequences were taken from 35 samples of *Aneides lugubris* over 26 locations and combined with 43 more samples from 27 locations retrieved from GenBank for a combined set of 78 samples from 53 locations over 22 counties in California [62]. Samples were divided into 6 mtDNA clades using a Bayesian analysis; the groupings roughly corresponded to Northern, SF Bay/Sierra Nevada, Santa Cruz, Pinnacles, Central Coast and Southern areas of the state.

To create a visualization for this data set GCPD was applied using a global normalization, quantifying how distinct each sample was phylogenetically. Cartograms were created with a location diameter of 15 and a variable multiplier of 10, applying no density matrix reduction. Applying our method provides much greater visual separation, providing greater horizontal displacement. Further separation is also created



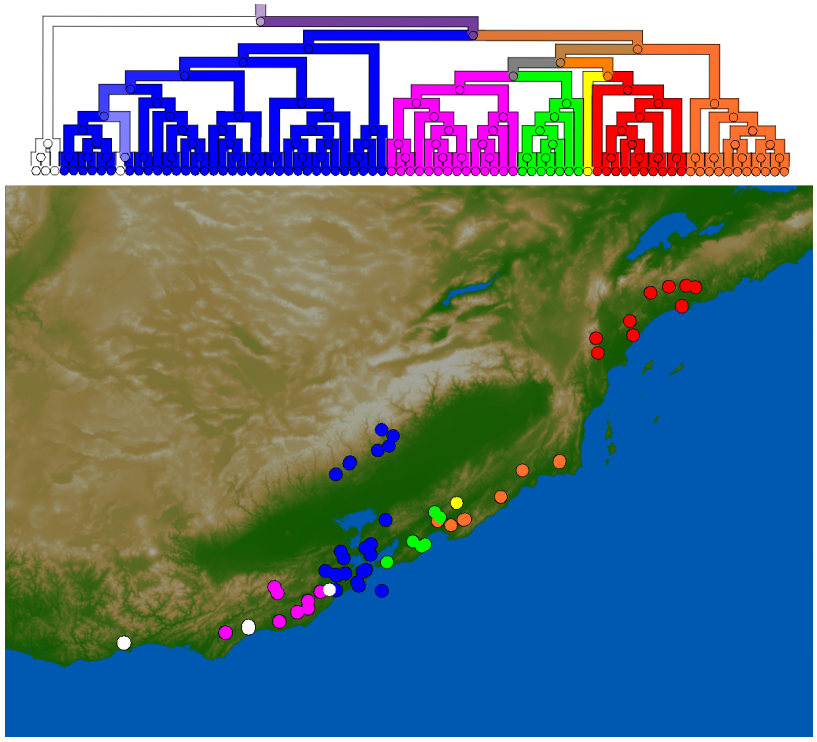
(a)



(b)

Figure 3.1: Projections of *Aneides lugubris* split into 6 mtDNA clades: Northern(Pink), SF Bay/Sierra Nevada(Blue), Santa Cruz(Green), Pinnacles(Yellow), Central Coast(Orange) and Southern(Red). (b) is a cartogram made from the GCPD values with two sites of interest marked: 1)Monterey Bay 2)Transverse Ranges.

(c)



(d)

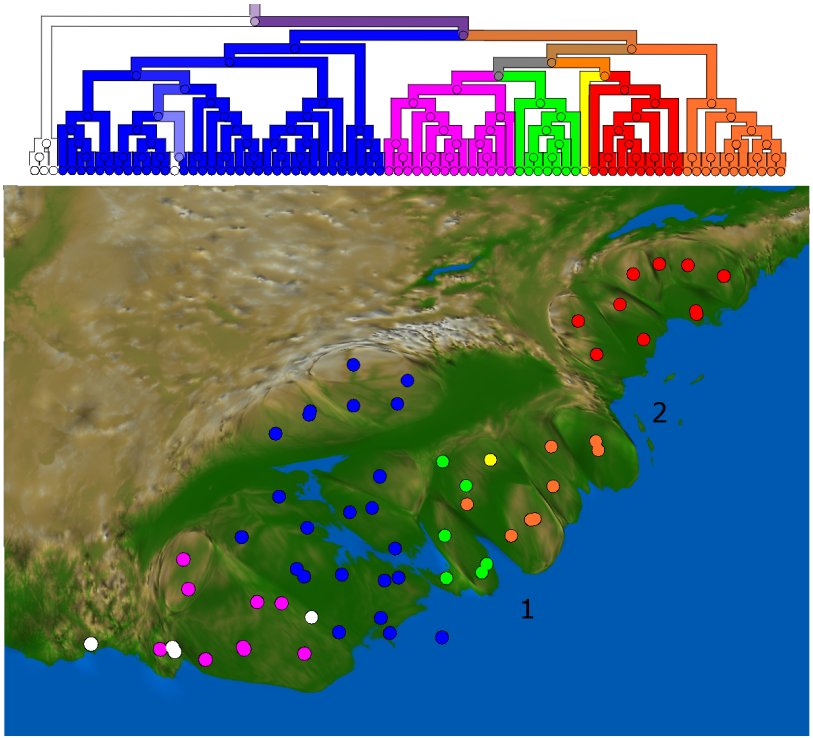


Figure 3.1: Projections of (a) and (b) without phylogenetic drop lines.

between the clades identified by the mtDNA lineage. North-Easterly locations belonging to the SF Bay/Sierra Nevada clade are displaced inland, but maintain their position, being bordered by a mountain range and river. This relative position is interesting given that the diffusion process reduces what was once a significant portion of land taken up by the rivers lowland into a mere sliver. While current conditions of the valley are inhabitable, evidence suggests that this valley was once a lake[34]. The most northerly locations belonging to the salamander outgroup are joined together geographically with other locations from the northern and outgroup clades, which highlights low edge weights in the MST. As this outgroup will have many connections with surrounding locations on the geographic scaffold, its poor GCPD score keeps it from having an effect on the final distortion.

The Monterey Bay area of California had been identified as a North-South break along this region for many species in the area [63], possibly because of an ancient river system which fed the area. The extensive distortion in the Monterey Bay area after a cartogram lends visual evidence to this area as important in the evolution of these salamanders, as the area of the bay and surrounding regions is drastically increased (Figure 3.1b). A phylogenetic break is also present between the Southern and Central clades caused by the east-west Transverse mountain range. Due to the close proximity of these two clades it would be expected that the geographic network used in GCPD would bring them together in diffusion. The effect of this mountain range is evident in the phylogeny and reflected visually by the distinct diffusion of each clade.

3.2 Illustrating local and global diversity patterns in pandemic *Vibrio cholerae*

The GCPD algorithm was applied to the phylogenetic tree created by [41], where the GCPD score of each location was normalized by its number of neighbours in the Delaunay triangulation. Values ranged from 0 to 2.41 with the highest GCPD scores associated with the island of Haiti. A cartogram was then applied for visualization, using a location diameter value of 20 and variable multiplier value of 10 with no

density matrix reduction. To analyse the effect of GCPD the absolute positions of each location were recorded before and after the map was distorted. The most important areas of the map after the distortion were in keeping with the findings from [41] that some samples from Haiti and Nepal share a close phylogenetic relationship.

In order to compare the magnitude of this relationship in the distortion the distance between the location's original position and distorted position were compared against all other locations, shown in Figure 3.2. Furthermore Figure 3.3 shows the relationship between all locations in terms of both how much they moved from each other and the PD of each set. The change in position of many points is not unique, as locations can have several samples associated with them. Each sample in this case will share geographic coordinates, but have varying degrees of phylogenetic relatedness in terms of PD. The difference between how far points moved during diffusion correlates poorly with the GCPD score of those points (R^2=0.02, p value=0.23), which is not an unexpected result. The migration distance of a location on the map depends on its relationship to areas of high diffusion/concentration. An example of this can be seen in Figure 3.3, where the location in the Dominican Republic undergoes the greatest shift (2006km). The source of this shift is the proximity of the Dominican Republic with Haiti, which has a great deal of points which undergo a large magnitude shift. This distortion then pushes the Dominican Republic farther away from its original position. Furthermore it is important to note that not all points in this plot contain a GCPD value, as their phylogenetic information was either unavailable or they were never sampled. These positions were still affected by the diffusion under the same effect which moved the Dominican Republic.

Clear clades were formed in Figure 3.3 between the Haitian and Nepalese samples with very low PDs. Of these location to location relationships, those found in Haiti formed the must uniform cluster over PD and dispersed well over area. Relationships were also visible between the Haitian and Nepalese samples as several of the Nepalese samples were visible in the Haitian cluster. These samples associate with samples from the Bhanke District and Rupandehi district of Nepal, which related most closely with Nepalese samples in the constructed phylogenetic tree.

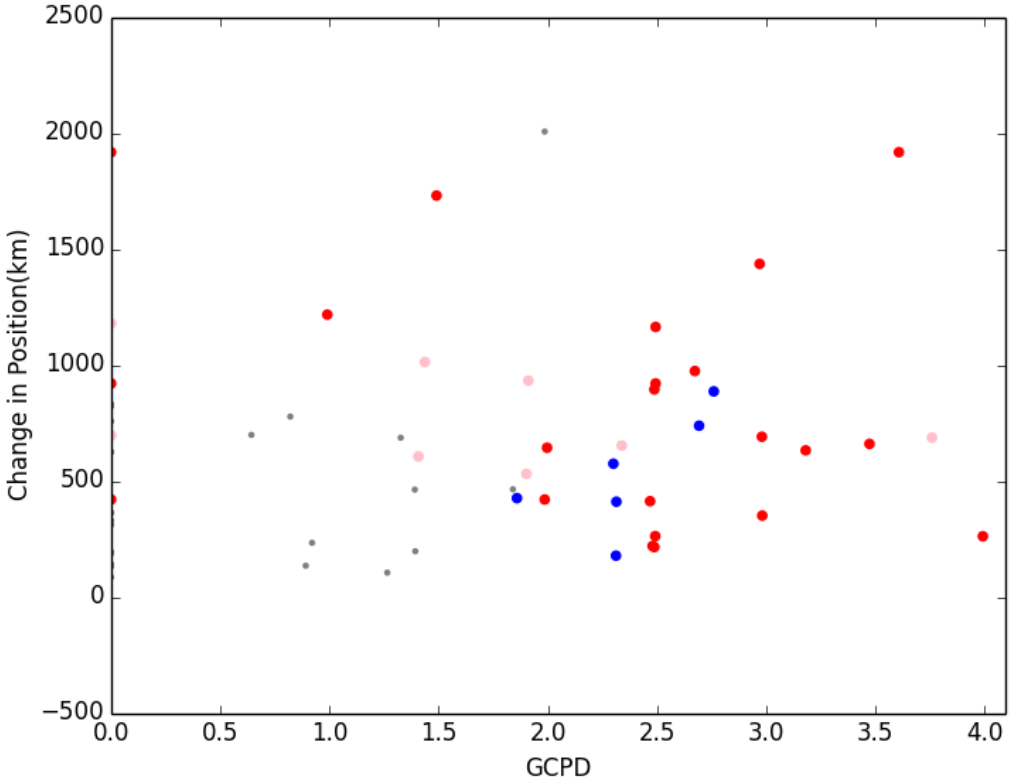


Figure 3.2: Migration distance of each location during cartogram construction. Points highlighted in red and blue belong to Haiti and Nepal respectively, while pink corresponds to Bangladesh and India.

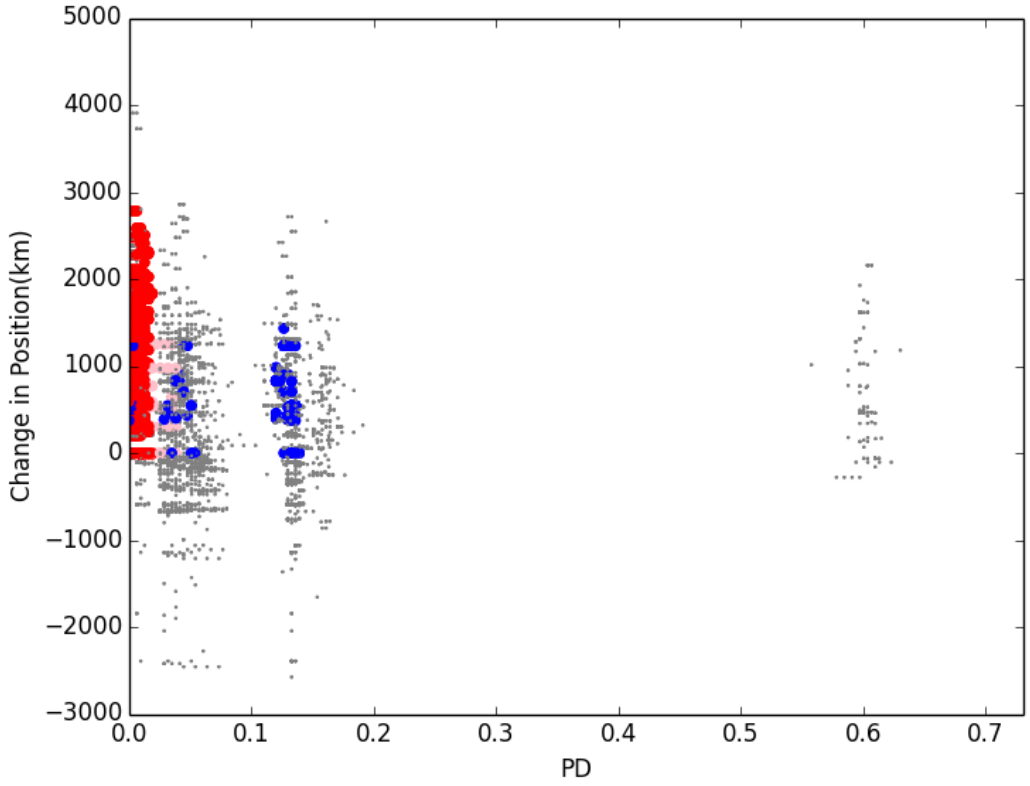


Figure 3.3: Comparison between the phylogenetic distance and the projected coordinates. Here every point is not unique, as the leaves of the phylogenetic tree are used, which may share a many to one relationship with a location. As such multiple points will have the same shift in projection with different PD. Points are coloured based on country association, Haiti(red), South Asia(blue) and other(gray).

(a)

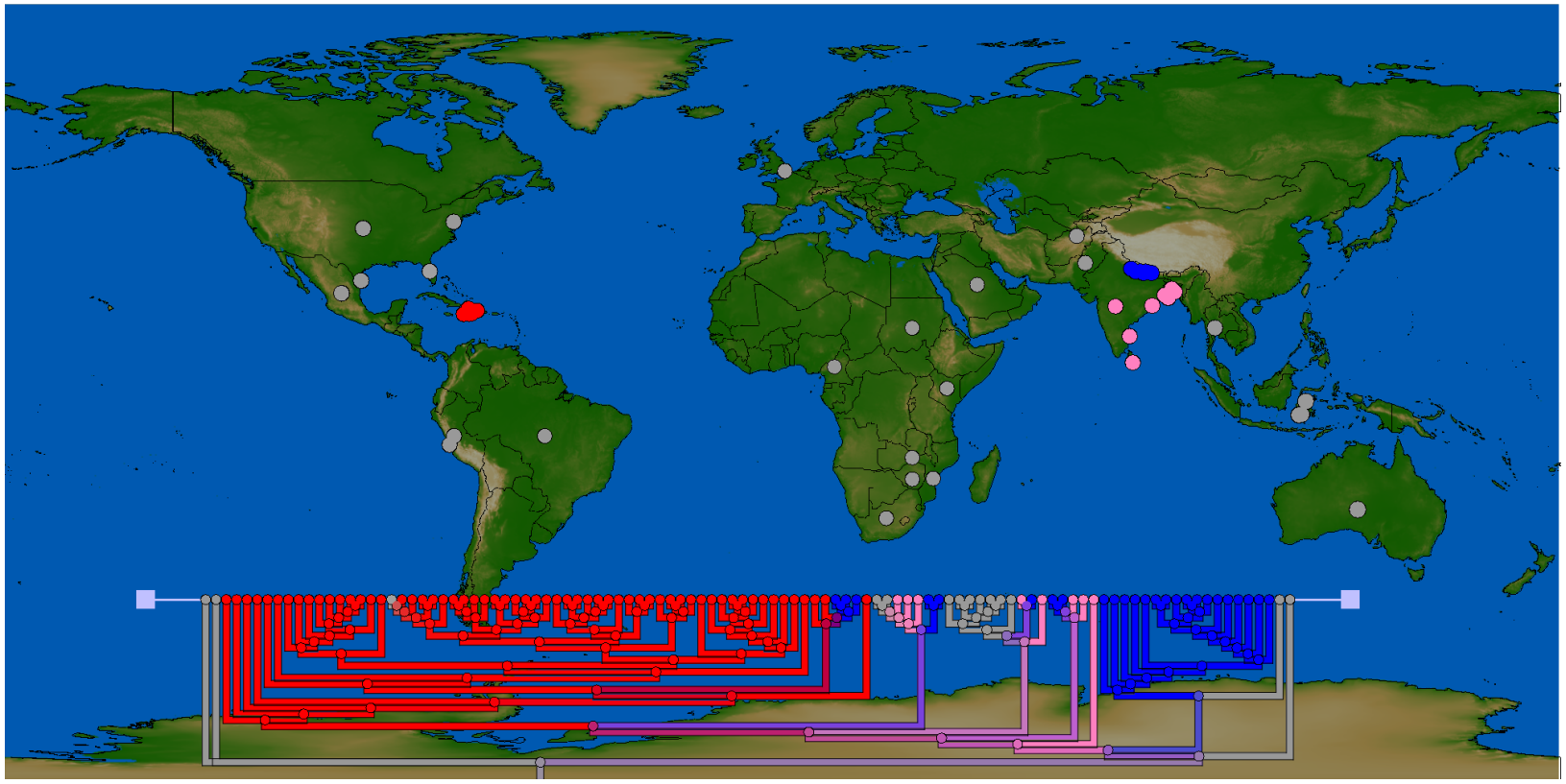


Figure 3.4: (a) An undistorted map of *Vibrio cholerae* data with an overlaid phylogeny showing the phylogenetic relationships amongst strains from Haiti and the Dominican Republic (red), Nepal (blue), South Asia (pink) and other countries (gray).

(b)

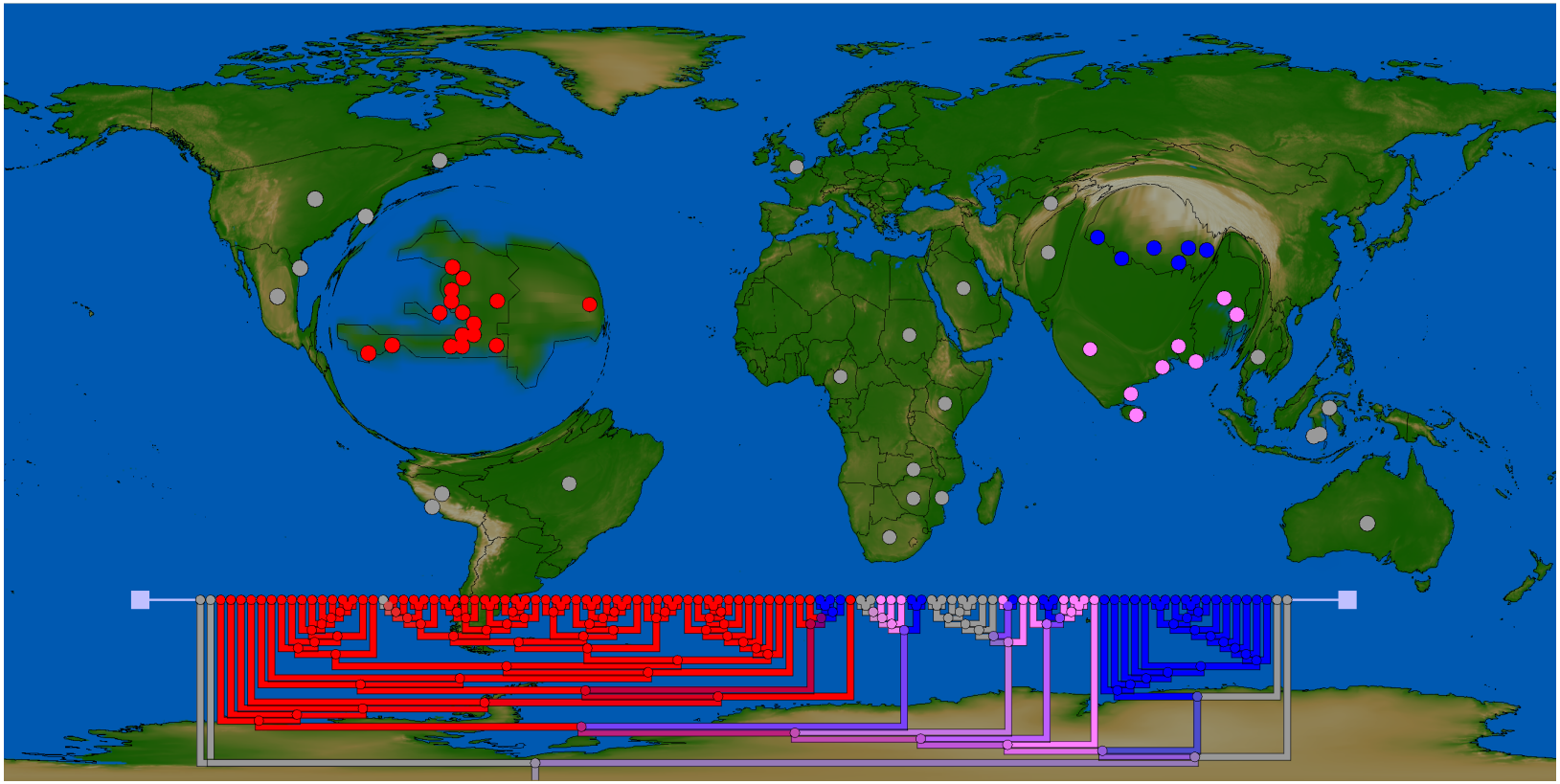


Figure 3.4: (b) A distorted map using GCPD, showing expansions in areas of high phylogenetic diversity.

(c)

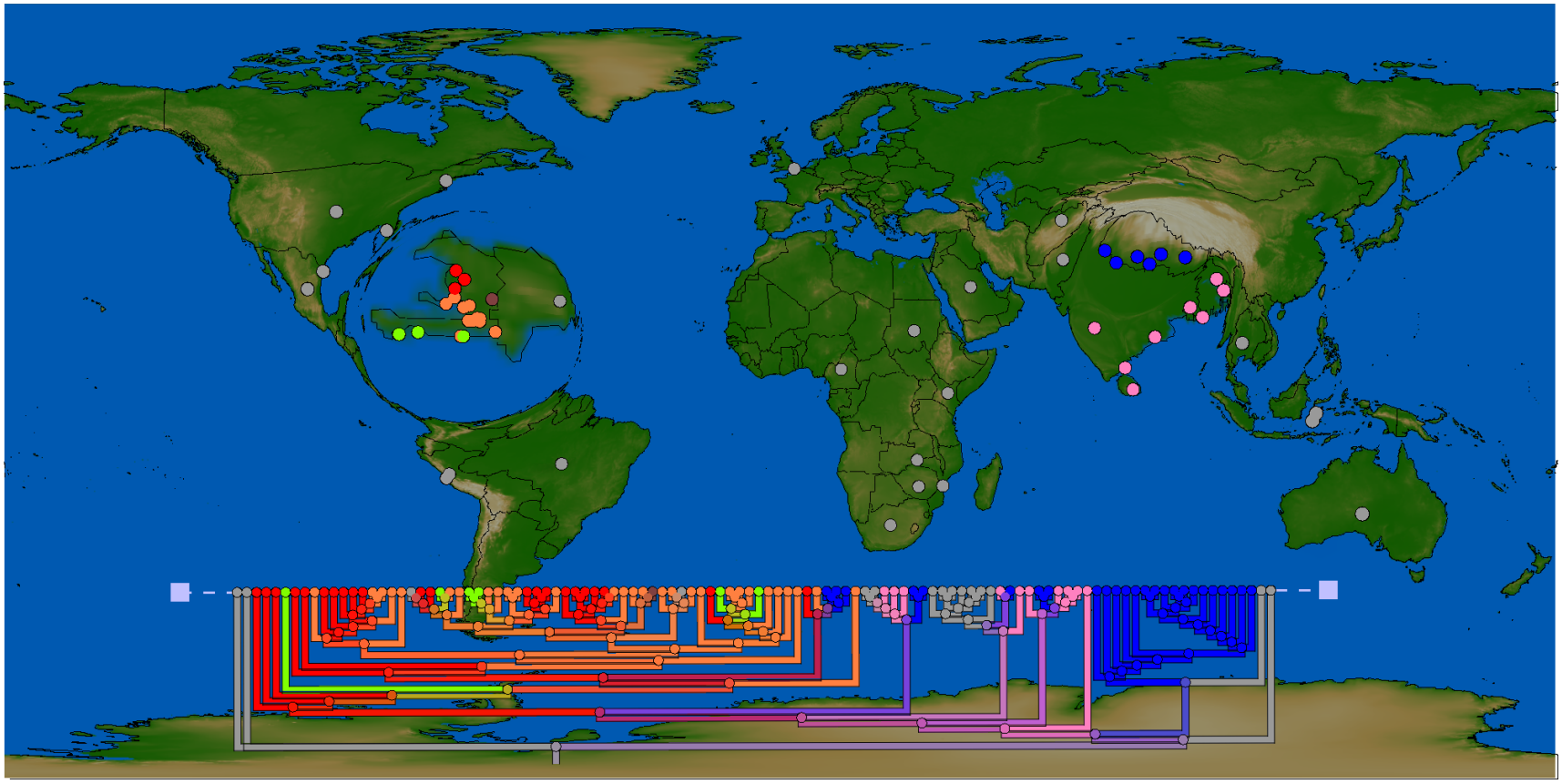


Figure 3.4: (c) A distorted map using GCPD as in Figure 3.5b with the departments coloured orange(Ouest), green(Sud),brown(Centre) and red(Artibonite).

3.3 Parameter Effect and Selection

As location diameter controls the area of the density matrix that is filled with non-neutral values, this parameter controls the distortion on a more global scale allowing locations with large density values to influence surrounding areas. The variable multiplier by contrast accumulates density at individual locations, allowing for very strong local diffusion. Figure 3.5a uses a value of 5 for both parameters, producing a map which is minimally distorted in relation to the original. Figures 3.5b and 3.5c show the effects of changing these parameters, by displaying their effect on the cholera pandemic data set. Larger values for either parameter cause a higher rate of diffusion and in turn more distortion.

When the Location Diameter is increased the density value of Haiti is applied to the Dominican Republic, and a similar effect can be seen between Nepal and India. This sort of distortion can be useful when geographic context is of importance and keeps a balance between the distorted and original projections. In contrast, an increase in the variable multiplier focuses the distortion much more locally. This creates a more fine tuned distortion which focuses on the individual area solely. This approach can be more useful for dense boundary regions where there exist many political borders.

The different parameters controlling density map manipulation do not produce the same maps. The location diameter parameter more strongly preserves the local shape of a region while drawing out its global context. This can be seen when comparing the natural "L" like shape of Haiti between Figure 3.5b and Figure 3.5c. Due to the variable multiplier map relying on the concentration of locations for its distortion, it focuses much more strongly on the upper sections of Haiti around the Artibonite and Ouest regions. This causes a ballooning of the upper region, while not expanding the diameter of the lower portion in proportion to the rest of the map. This problem of proportional swell is solved when location diameter is used, though it produces different features. As location diameter is border agnostic it cannot tell the difference between Haiti and the Dominican Republic in the density matrix and expands the latter country as well. This may not be a desirable feature for maps with subtle or

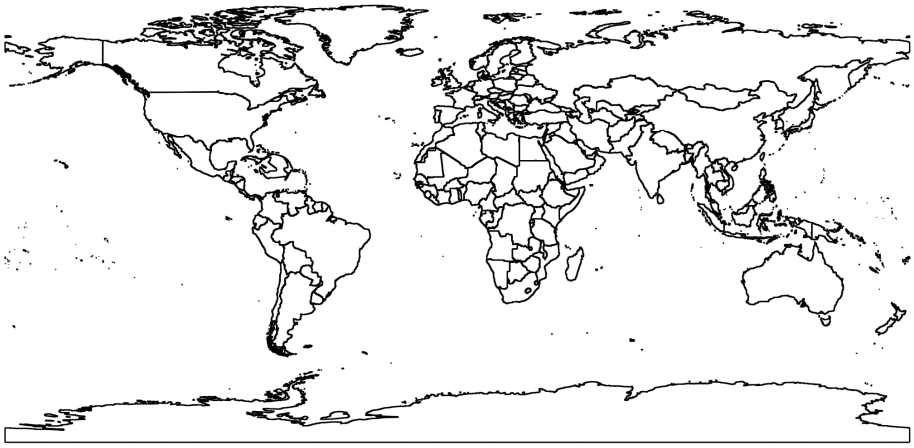
dense features. In such cases the variable multiplier is more appropriate, allowing for growth based solely on these features.

3.4 Effect of Heuristic Density Reduction

While density matrix reduction achieves its goal of reducing the run time of the diffusion calculation, it also creates a loss of information in the transition between matrices. If this loss of information results in a large discrepancy between cartograms its usefulness will be reduced, as distortion parameters that are tuned to a reduced matrix may have different distortion effects on the full matrix.

In order to investigate the effect matrix reduction has on the distortion a world map with the cholera pandemic data was selected as a test case. This map makes an ideal test for this situation due to prominence of Haitian and Nepalese samples generating large amounts of diffusion locally, while the rest of the map remains largely unchanged. In order to record the differences in distortions between matrices a vector file of political borders was distorted in conjunction with the map. The position of the border was recorded before and after distortion, and in each case the displacement of each point recorded. This created an accurate measure of how much each recognizable shape distorts.

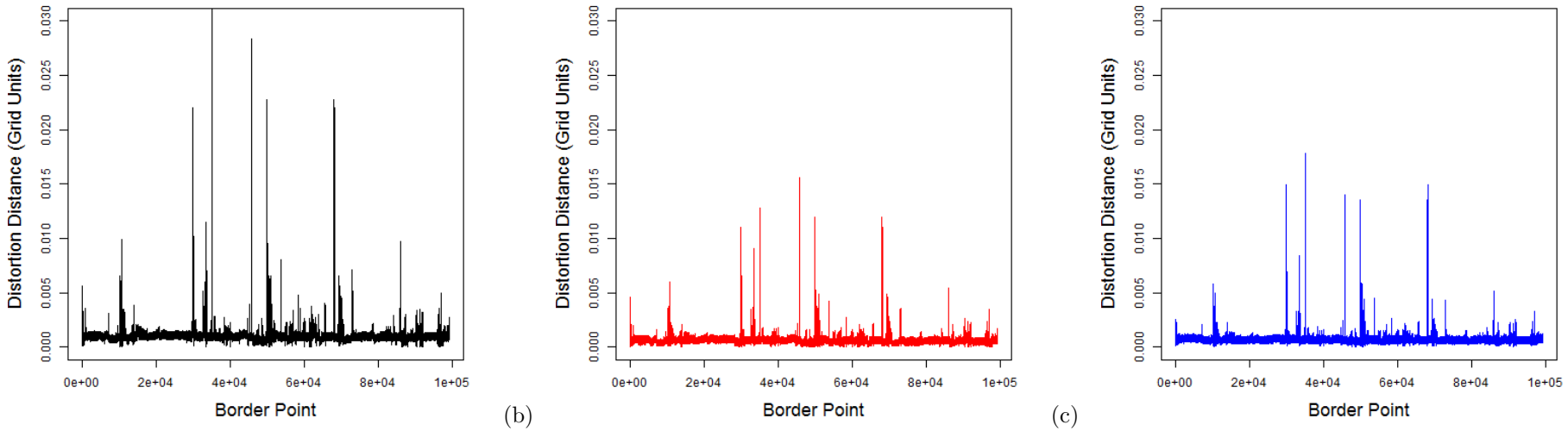
(a)



(b)

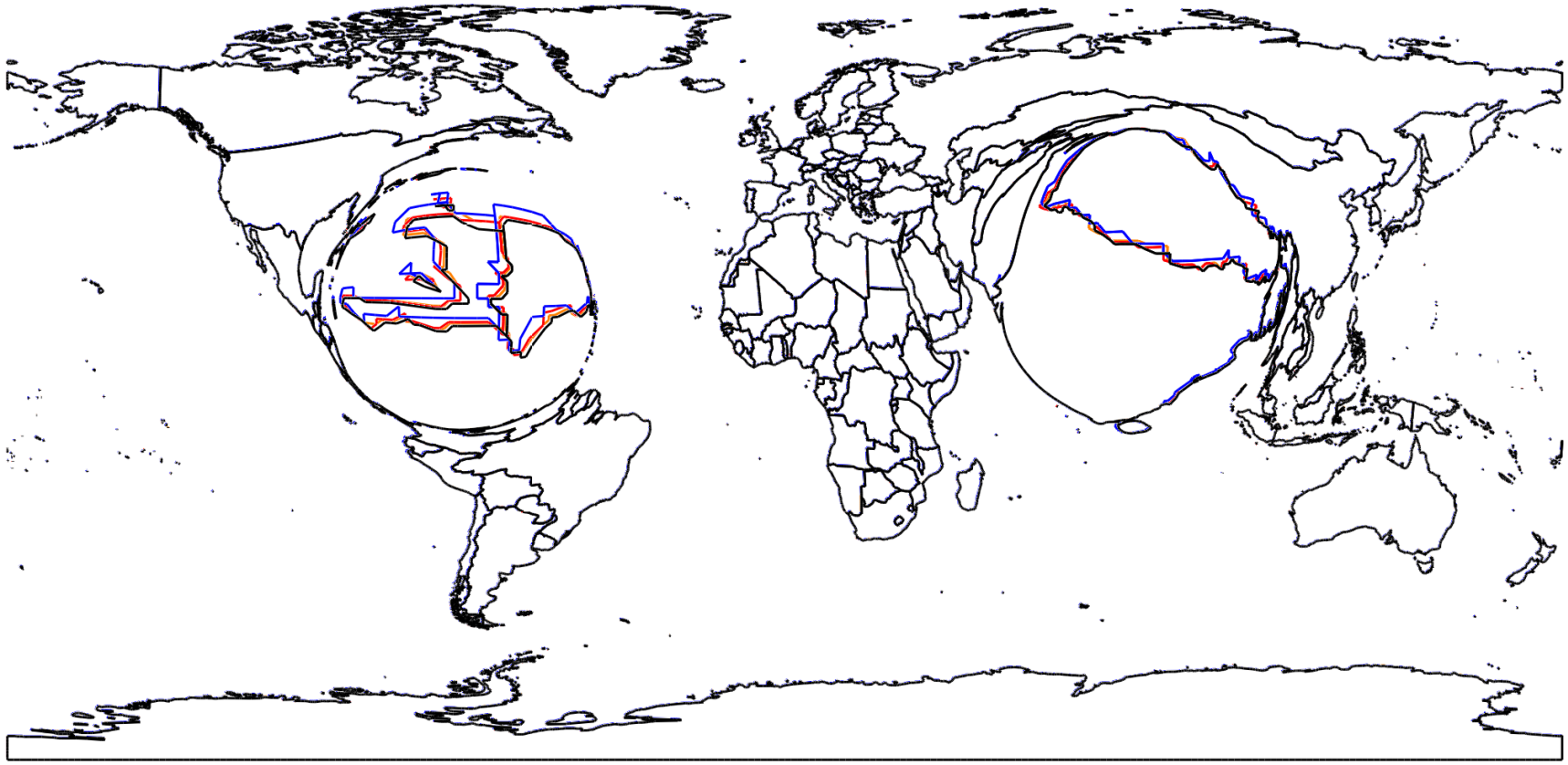
(c)

Figure 3.5: The effects of parameter selection for the *Vibrio cholerae* pandemic using GCPD. (a) A base projection using a location diameter of 5 and a variable multiplier of 5. (b) A projection using a location diameter of 25 and a variable multiplier of 5. (c) A projection using a location diameter of 5 and a variable multiplier of 25.

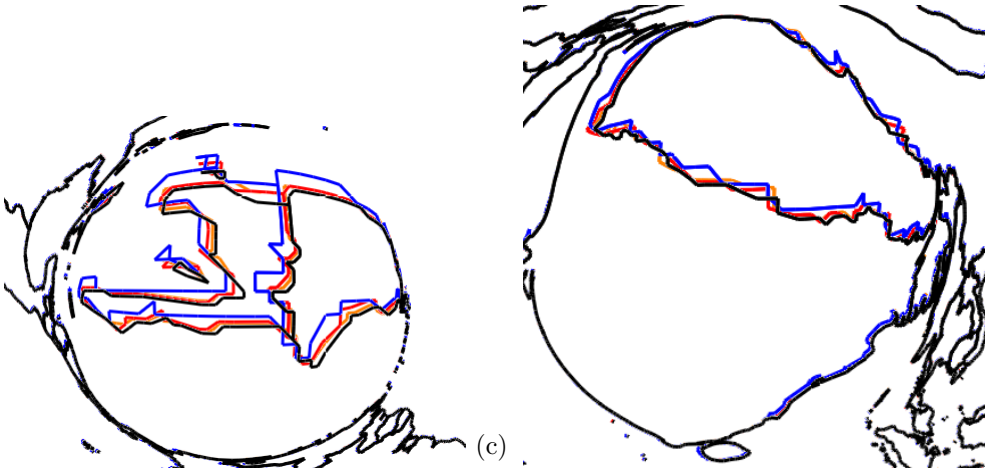


(a) (b) (c)

Figure 3.6: Comparisons in the amount of diffusion between 3 different density matrix reduction values and a fully diffused map. (a) compares the difference in political border placement of a 50% reduced density matrix versus full diffusion. (b) compares the difference between political border placement of a 20% reduced density matrix against full diffusion. (c) compares the difference between political border placement of a 10% reduced density matrix against full diffusion. Distances were compared over the internal grid units of the raster map, which is scaled with a width of 2 and height of 1.2 for this map.



(a)



(b) (c)

Figure 3.7: (a) A comparison of the visual distortion produced by four different values of density matrix reduction. Full density matrix (Black), 50% reduction density matrix (Blue), 20% reduction density matrix (Red) and 10% reduction density matrix (Orange) which is largely obscured by the red border. (b) A cross-section of the full map showing distortion around the Haiti region. (c) A cross-section of the full map showing distortion around the Nepal/India region.

In order to assess the effect of density-grid reduction on cartogram creation three levels of reduction were selected, 10%, 20% and 50% respectively. In each case the diffusion used a location diameter of 20 and a variable multiplier of 10. Run times were not recorded for these comparisons as the time taken to write the differences in each border point to disk has a large influence on the complete runtime. As there is no ground truth in cartogram creation, there exists no "correct" distortion, so the map created using the full density matrix was used to compare the quality of all other distortions. All positions in the shape files of each respective reduced run were compared against their corresponding position in the full density matrix distortion.

As the level of reduction was decreased the level of difference between reduced matrix and full matrix results also decreased. This can be seen in Figure 3.6 when comparing Figure 3.6a with Figure 3.6c as the maximum amount of change has dropped by half. Distances here were compared over the internal grid units of the raster map, with a width of 2 and height of 1.2, which is consistent across a single map. All three graphs share a similar pattern, with core elements of the shape staying very close to

	Min	Max	Avg	StdDev
Full vs 50%	$4.45{\cdot}10^{-6}$	0.03	$1.1{\cdot}10^{-3}$	$1.2{\cdot}10^{-3}$
Full vs 20%	$1.20{\cdot}10^{-6}$	0.015	$6.6{\cdot}10^{-4}$	$5.7{\cdot}10^{-4}$
Full vs 10%	$3.07{\cdot}10^{-5}$	0.018	$6.1{\cdot}10^{-4}$	$5.0{\cdot}10^{-4}$

Table 3.1: A comparison between the Min,Max,Average and Standard Deviation for three separate density matrix reduction values based upon differences in shapefile position. Positions were taken of the internal Grid Units used by the raster maps, consisting of a rectangular regions with a width of 2 and height of 1.2. The results of diffusion using the full density matrix were used as a gold standard for comparison.

their original positions, while similar areas in all three consistently differ the most. This is to be expected as the diffusion from location origins will have very little effect on distant parts of the map. Figure 3.7a shows the differences in political borders between the distortions created by all density matrices, with the majority of the positions agreeing at all levels. The area of highest dissimilarity is the Haiti/Dominican Republic area of the map, which is highlighted in Figure 3.8b. Here we see the highest level of discrepancy between the full density matrix run and the 50% reduced density matrix, which trends generally towards the top left of the map in comparison. As expected the reduction of the density matrix reduced the complexity of distinguishing features on the 50% reduced map, creating a much more blocky outline. As the RP was decreased however, more complex features of the map re-emerged.

From these results we conclude that while there is a difference in projections made using different matrix reduction values, these differences do not seriously compromise the quality of the overall projection. While the variability of different regions of maps will depend on geography and the property driving the distortion, the process of reducing the density matrix provides benefits in both run time and RAM consumption, while still providing a tolerable representation of the final distortion. Its capacity to draft cartograms to evaluate the effect of different parameters and variables in the distortion with minimal drift from the full projection make it a valuable tool.

Chapter 4

Conclusion

We have developed the GCPD, a method to combine phylogenetic and geographic information, by creating a Delaunay triangulation between all geographic points and weighting this graph based on PD. By doing this a measure is created which quantifies the phylogenetic diversity at each location. GCPD addresses some display limitations of current phylogeographic techniques; for example, Daniel et al. used phylogenetic and geographic information to reconstruct and visualize the spread of of H5N1 influenza lineages. Their approach involved overlaying a phylogenetic tree over a Google Earth projection. One problem encountered in this approach was the ability to map multiple terminal nodes of the tree to one location, such as Hong Kong [39]. In order to overcome this problem they adjusted overlapping geographic coordinates from overpopulated sites in an east to west line. Applying GCPD with a diffused visualization would help handle the problem of highly dense sample locations without the need to manually change location coordinates.

The two tunable cartogram parameters, location diameter and variable multiplier, are effective methods to control the amount of distortion created in the map. These parameters are needed to build cartograms from raster rather than vector-based maps. As we are appropriating it for discrete location-based observations a way to distribute highly accurate point based information over a wider area was required. Location diameter propagates the value associated with each location over a larger area, and provides geographic context to areas of import to the final distortion, while the variable multiplier increases the values associated with each location, and provides a highly local context to the final distortion. Depending on what parameter is of interest in a projection, each of these parameters must be tuned in conjunction.

Other approaches to cartogram creation, such as the original Gastner-Newman,

were more primarily concerned with the level of difference between the cartograms they generated and the expected level of distortion [28][37][67]. The methods of evaluation employed in previous studies were concerned primarily with shape errors at a local and global scale. These studies found that in general diffusion-based cartograms performed within a margin of 1-2% of the proposed method in terms of error [37][67]. Sagar even proposed that their morphological approach was best, despite have generally higher levels of error [67], and creating misrepresentations of the target values in some regions in order to preserve global shape without creating disjoint regions.

Here we showed how the reduction in the size of the density matrix used by the diffusion-based cartogram created up to a 2x speed up in run times. This approach also reduced the amount of memory required by the application when compared against other diffusion cartograms. The largest difference between projections made with the full density matrix and a reduced matrix only differed by 1.3% in terms of geographic distance, a relatively small amount that is likely to be acceptable to the majority of users. Our tests of density-map reduction used an original projection map with a width of 4050 px and height of 2025 px. This meant that even fairly drastic reduction in the density matrix resulted in a rich enough map that the discretization process would not significantly alter the final result. While the results of this example are expected to generalize to even moderately sized maps, and projections of similar sizes are widely available from resources such as public domain raster and vector maps at different global scales `http://www.naturalearthdata.com/` or the Oak Ridge National Laboritory Distributed Active Archive Center which provides access to data from NASA's Earth Science Mission `http://webmap.ornl.gov/wcsdown/` it still merits mentioning that significantly smaller projection maps will be more affected by a density matrix reduction. This is not necessarily a drawback as small maps (750 x 750 px) can be distorted in < 30 seconds.

4.1 Summary of Results and Conclusions

Using the GCPD we were able to visualize the spread of *Vibrio cholerae* in a global pandemic. The use of GCPD in conjunction with a diffusion-based cartogram

created a visualization which mirrored results found by previous studies of the pandemic, that Nepalese samples of cholera resembled the Haitian strain most closely in genealogy [36][14]. Further investigation found that Nepalese relief workers were indeed the source of the outbreak, as they had travelled to Haiti without following proper quarantine procedures. In this distortion we found no evidence for a correlation when using a linear regression (r^2=0.02, p value=0.23) between the distance a location moved in the distortion and its GCPD value. This is the expected result, as areas with high GCPD values will cause larger regions of diffusion, which will exert a strong push upon peripheral points. This effect is increased with larger location diameter values.

In this map we also show how the application of cartograms can be used to further analysis. Typical projections of a world map place Haiti as far too small to visualize the dense sampling of cholera within the region. The increase in area afforded Haiti by the density-equalizing map allows for a differentiation of the different departments of the country, while preserving the global context of the outbreak. The distortion clarified the relationships between cholera lineages in specific regions of Haiti and Nepal.

We also visualized the regional dispersion of endemic *Aneides lugubris* through the California region. This region is of special interest in research due to the high level of endemic species and its sensitivity to the ongoing effects of climate change [35]. Here cartograms were combined with the GCPD measure to provide visual confirmation for important geographic regions including Monterey Bay, a region which drew in many species due to connected waterways, via the distribution of closely linked clades. This method was also able to infer the effects of geographic barriers such as mountain ranges and valleys via their effect on phylogenetic relatedness.

4.2 Future Work

Many extensions and refinements of the GCPD can be envisioned. Although the GCPD was developed with raster maps in mind, a variant approach could incorporate vector shape data and thus introduce a density matrix similar to those used by vector based approaches, by using shape information to assign GCPD values to the corresponding regions. This will change the basis of the distortion away from being location driven towards providing uniform distortions for political or geographic boundaries. This modification will remove the local resolution, as observed in Figure 3.5c, while providing stronger signal to studies where boundaries rather than a uniform grid, are the feature of interest.

The effects of the GCPD measure need to be explored further to evaluate their effectiveness in non-epidemiological and visualization roles. Comparisons of GCPD with measures of endemism for conservation study would be one possible avenue of exploration. Comparing the correlation between endemism measures and GCPD could yield high levels of overlap between the two measures, highlighting the potential application of GCPD for conservation studies. This type of comparison would require large scale data sets of species records to be geocoded into location based observation data.

Trees are not the only representation used to show phylogenetic relationships. Networks are becoming an increasingly popular tool for visualizing phylogenetic descent, particularly because of their ability to handle LGT. Due to their popularity and explanatory power extending the GCPD procedure to work with networks would incorporate even more phylogenetic signal into the measure. This would aid in highlighting closely related highly connected regions, and increase the GCPD scores of those locations. As one of the primary representations of a phylogenetic network is a distance matrix, adapting the GCPD calculation to be compatible with arbitrary matrices would also improve the usability of such a measure. This would allow for geographic reconciliation between different distance based scores and the geographic spread of observations.

Bibliography

[1] Achtman, M. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annual Reviews Microbiology 62* (2008), 53–70.

[2] Anderson, R. M., Boily, M., Garnett, G., Rowley, J., and May, R. The spread of HIV-1 in africa: sexual contact patterns and the predicted demographic impact of AIDS. *Nature 352*, 6336 (1991), 581–589.

[3] Andrieu, D., Kaiser, C., and Ourednik, A. Scapetoad. `http://scapetoad.choros.ch/`. Accessed: 2016-03-30.

[4] Barzilay, E. J., Schaad, N., Magloire, R., Mung, K. S., Boncy, J., Dahourou, G. A., Mintz, E. D., Steenland, M. W., Vertefeuille, J. F., and Tappero, J. W. Cholera surveillance during the haiti epidemic-the first 2 years. *New England Journal of Medicine 368*, 7 (2013), 599–609.

[5] Bauer, H.-U., Der, R., and Herrmann, M. Controlling the magnification factor of self-organizing feature maps. *Neural Computation 8*, 4 (1996), 757–771.

[6] Bedford, T., and Neher, R. *Real-time Analysis of Ebola Virus Evolution*, 2015.

[7] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. Genbank. *Nucleic Acids Research 36*, suppl 1 (2008), D25–D30.

[8] Bielejec, F., Rambaut, A., Suchard, M. A., and Lemey, P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics 27*, 20 (2011), 2910–2912.

[9] Blomberg, S. P., Garland, T., and Ives, A. R. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution 57*, 4 (2003), 717–745.

[10] Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. Estimating species richness: the importance of heterogeneity in species detectability. *Ecology 79*, 3 (1998), 1018–1028.

[11] Bracewell, R. The fourier transform and its applications. *New York 5* (1965).

[12] Bunge, J., and Fitzpatrick, M. Estimating the number of species: a review. *Journal of the American Statistical Association 88*, 421 (1993), 364–373.

[13] Chazdon, R. L., Colwell, R. K., Denslow, J. S., and Guariguata, M. R. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern costa rica. Tech. rep., 1999.

[14] Chin, C.-S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., Bullard, J., Webster, D. R., Kasarskis, A., Peluso, P., et al. The origin of the haitian cholera outbreak strain. *New England Journal of Medicine 364*, 1 (2011), 33–42.

[15] Colwell, R. K., and Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 345*, 1311 (1994), 101–118.

[16] Crisp, M. D., Laffan, S., Linder, H. P., and Monro, A. Endemism in the australian flora. *Journal of Biogeography 28*, 2 (2001), 183–198.

[17] de la Torre, J. Organising geo-temporal data with CartoDB, an open source database on the cloud. In *Biodiversity Informatics Horizons 2013* (2013).

[18] Donoghue, M. J., Bell, C. D., and Li, J. Phylogenetic patterns in northern hemisphere plant geography. *International Journal of Plant Sciences 162*, S6 (2001), S41–S52.

[19] Dorling, D. Area cartograms: their use and creation.

[20] Elliott, P., Briggs, D., Morris, S., de Hoogh, C., Hurt, C., Jensen, T. K., Maitland, I., Richardson, S., Wakefield, J., and Jarup, L. Risk of adverse birth outcomes in populations living near landfill sites. *British Medical Journal 323*, 7309 (2001), 363–368.

[21] Ernst, R., and Anderson, W. Application of fourier transform spectroscopy to magnetic resonance. *Review of Scientific Instruments 37*, 1 (1966), 93–102.

[22] Faria, N. R., Suchard, M. A., Rambaut, A., and Lemey, P. Toward a quantitative understanding of viral phylogeography. *Current Opinion in Virology 1*, 5 (2011), 423–429.

[23] Feldmann, H., and Geisbert, T. W. Ebola haemorrhagic fever. *The Lancet 377*, 9768 (2011), 849–862.

[24] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution 17*, 6 (1981), 368–376.

[25] Frigo, M., and Johnson, S. G. The design and implementation of FFTW3. *Proceedings of the Institute of Electrical and Electronic Engineers 93*, 2 (2005), 216–231.

[26] Gardy, J., Loman, N. J., and Rambaut, A. Real-time digital pathogen surveillance — the time is now. *Genome Biology 16*, 1 (2015), 1–3.

[27] Gastner, M. T., and Newman, M. Density-equalizing map projections: Diffusion-based algorithm and applications. *Proceedings of the 8th International Conference on Geocomputation* (2005).

[28] Gastner, M. T., and Newman, M. E. Diffusion-based method for producing density-equalizing maps. *Proceedings of the National Academy of Sciences of the United States of America 101*, 20 (2004), 7499–7504.

[29] Gatherer, D. The 2014 ebola virus disease outbreak in West Africa. *Journal of General Virology 95*, 8 (2014), 1619–1624.

[30] Gire, S. K., Goba, A., Andersen, K. G., Sealfon, R. S., Park, D. J., Kanneh, L., Jalloh, S., Momoh, M., Fullah, M., Dudas, G., et al. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science 345*, 6202 (2014), 1369–1372.

[31] Gotelli, N. J., and Colwell, R. K. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters 4*, 4 (2001), 379–391.

[32] Graham, M., Liang, B., Van Domselaar, G., Bastien, N., Beaudoin, C., Tyler, S., Kaplen, B., Landry, E., Li, Y., Team, N. I. A. G. S., et al. Nationwide molecular surveillance of pandemic h1n1 influenza a virus genomes: Canada, 2009. *PLoS One 6*, 1 (2011), e16087.

[33] Guindon, S., and Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology 52*, 5 (2003), 696–704.

[34] Hall, C. A. *Nearshore marine paleoclimatic regions, increasing zoogeographic provinciality, molluscan extinctions, and paleoshorelines, California: Late Oligocene (27 Ma) to Late Pliocene (2.5 Ma)*, vol. 357. Geological Society of America, 2002.

[35] Hayhoe, K., Cayan, D., Field, C. B., Frumhoff, P. C., Maurer, E. P., Miller, N. L., Moser, S. C., Schneider, S. H., Cahill, K. N., Cleland, E. E., et al. Emissions pathways, climate change, and impacts on california. *Proceedings of the National Academy of Sciences of the United States of America 101*, 34 (2004), 12422–12427.

[36] Hendriksen, R. S., Price, L. B., Schupp, J. M., Gillece, J. D., Kaas, R. S., Engelthaler, D. M., Bortolaia, V., Pearson, T., Waters, A. E., Upadhyay, B. P., et al. Population genetics of vibrio cholerae from nepal in 2010: evidence on the origin of the haitian outbreak. *MBio 2*, 4 (2011), e00157–11.

[37] Henriques, R., Bação, F., and Lobo, V. Carto-SOM: cartogram creation using self-organizing maps. *International Journal of Geographical Information Science 23*, 4 (2009), 483–511.

[38] Illumina. An introduction to next-generation sequencing technology. `http://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf`, 2016.

[39] Janies, D., Hill, A. W., Guralnick, R., Habib, F., Waltari, E., and Wheeler, W. C. Genomic analysis and geographic visualization of the spread of avian influenza (H5N1). *Systematic Biology 56*, 2 (2007), 321–329.

[40] Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P. Global trends in emerging infectious diseases. *Nature 451*, 7181 (2008), 990–993.

[41] Katz, L. S., Petkau, A., Beaulaurier, J., Tyler, S., Antonova, E. S., Turnsek, M. A., Guo, Y., Wang, S., Paxinos, E. E., Orata, F., et al. Evolutionary dynamics of vibrio cholerae O1 following a single-source introduction to haiti. *MBio 4*, 4 (2013), e00398–13.

[42] Kidd, D. M., and Liu, X. GEOPHYLOBUILDER 1.0: an arcgis extension for creating geophylogenies. *Molecular Ecology Resources 8*, 1 (2008), 88–91.

[43] Kistemann, T., Dangendorf, F., and Schweikart, J. New perspectives on the use of geographical information systems (GIS) in environmental health sciences. *International Journal of Jygiene and Environmental Health 205*, 3 (2002), 169–181.

[44] Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological Cybernetics 43*, 1 (1982), 59–69.

[45] Kuchta, S. R., Parks, D. S., and Wake, D. B. Pronounced phylogeographic structure on a small spatial scale: geomorphological evolution and lineage history in the salamander ring species ensatina eschscholtzii in central coastal california. *Molecular Phylogenetics and Evolution 50*, 2 (2009), 240–255.

[46] Layeghifard, M., Peres-Neto, P. R., and Makarenkov, V. Using directed phylogenetic networks to retrace species dispersal history. *Molecular Phylogenetics and Evolution 64*, 1 (2012), 190–197.

[47] Lovett, D. A., Poots, A. J., Clements, J. T., Green, S. A., Samarasundera, E., and Bell, D. Using geographical information systems and cartograms as a health service quality improvement tool. *Spatial and Spatiotemporal Epidemiology 10* (2014), 67–74.

[48] Magurran, A. E. *Ecological diversity and its measurement*. Springer Science & Business Media, 2013.

[49] Matzke, N. J. BioGeoBEARS: biogeography with bayesian (and likelihood) evolutionary analysis in R scripts. *R package, version 0.2 1* (2013).

[50] Morando, M., Avila, L. J., Baker, J., and Sites, J. W. Phylogeny and phylogeography of the liolaemus darwinii complex (squamata: Liolaemidae): evidence for introgression and incomplete lineage sorting. *Evolution 58*, 4 (2004), 842–859.

[51] Munster, V. J., Baas, C., Lexmond, P., Waldenström, J., Wallensten, A., Fransson, T., Rimmelzwaan, G. F., Beyer, W. E., Schutten, M., Olsen, B., et al. Spatial, temporal, and species variation in prevalence of influenza a viruses in wild migratory birds. *PLoS Pathogens 3*, 5 (2007), e61.

[52] Namias, V. The fractional order fourier transform and its application to quantum mechanics. *IMA Journal of Applied Mathematics 25*, 3 (1980), 241–265.

[53] Neher, R. A., Russell, C. A., and Shraiman, B. I. Predicting evolution from the shape of genealogical trees. *Elife 3* (2014), e03568.

[54] Newman, M. Cart: Computer software for making cartograms, 2009.

[55] Newton, A., Allnutt, T., Gillies, A., Lowe, A., and Ennos, R. Molecular phylogeography, intraspecific variation and the conservation of tree species. *Trends in Ecology & Evolution 14*, 4 (1999), 140–145.

[56] Nowbar, A. N., Howard, J. P., Finegold, J. A., Asaria, P., and Francis, D. P. 2014 global geographic analysis of mortality from ischaemic heart disease by country, age and income: Statistics from world health organisation and united nations. *International Journal of Cardiology 174*, 2 (2014), 293–298.

[57] Nuckols, J. R., Ward, M. H., and Jarup, L. Using geographic information systems for exposure assessment in environmental epidemiology studies. *Environmental Health Perspectives* (2004), 1007–1015.

[58] Okabe, A., Boots, B., Sugihara, K., and Chiu, S. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Wiley Series in Probability and Statistics. Wiley, 2009.

[59] Parks, D. H., Mankowski, T., Zangooei, S., Porter, M. S., Armanini, D. G., Baird, D. J., Langille, M. G., and Beiko, R. G. Gengis 2: Geospatial analysis of traditional and genetic biodiversity, with new gradient algorithms and an extensible plugin framework. *PloS One 8*, 7 (2013), e69885.

[60] Petkau, A. Phylogenetic reconstruction and outbreak investigation using IRIDA and SNVPhyl. In *1st ASM Conference on Rapid Next-Generation Sequencing and Bioinformatic Pipelines for Enhanced Molecular Epidemiologic Investigation of Pathogens* (Washington, D.C., USA., September 24-27 2015).

[61] Reif, J. S., Burch, J. B., Nuckols, J. R., Metzger, L., Ellington, D., and Anger, W. K. Neurobehavioral effects of exposure to trichloroethylene through a municipal water supply. *Environmental Research 93*, 3 (2003), 248–258.

[62] Reilly, S. B., Corl, A., and Wake, D. B. An integrative approach to phylogeography: investigating the effects of ancient seaways, climate, and historical geology on multi-locus phylogeographic boundaries of the arboreal salamander (aneides lugubris). *BMC evolutionary biology 15*, 1 (2015), 241.

[63] Rissler, L. J., Hijmans, R. J., Graham, C. H., Moritz, C., and Wake, D. B. Phylogeographic lineages and species comparisons in conservation analyses: a case study of california herpetofauna. *The American Naturalist 167*, 5 (2006), 655–666.

[64] Rode, A. L., and Lieberman, B. S. Using GIS to unlock the interactions between biogeography, environment, and evolution in middle and late devonian brachiopods and bivalves. *Palaeogeography, Palaeoclimatology, Palaeoecology 211*, 3 (2004), 345–359.

[65] Ronquist, F., and Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics 19*, 12 (2003), 1572–1574.

[66] Rosauer, D., Laffan, S. W., Crisp, M. D., Donnellan, S. C., and Cook, L. G. Phylogenetic endemism: a new approach for identifying geographical concentrations of evolutionary history. *Molecular Ecology 18*, 19 (2009), 4061–4072.

[67] Sagar, B. D. Cartograms via mathematical morphology. *Information Visualization* (2013), 1473871613480061.

[68] Skoda, L., and Robertson, J. *Isodemographic map of Canada*, vol. 50. Information Canada, 1972.

[69] Snow, J. *On the mode of communication of cholera*. John Churchill, 1855.

[70] Snyder, L., Peters, J. E., Henkin, T. M., and Champness, W. *Molecular genetics of bacteria*. American Society of Microbiology, 2013.

[71] Stamatakis, A., Ludwig, T., and Meier, H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics 21*, 4 (2005), 456–463.

[72] Stigall, A. L., and Lieberman, B. S. Quantitative palaeobiogeography: GIS, phylogenetic biogeographical analysis, and conservation insights. *Journal of Biogeography 33*, 12 (2006), 2051–2060.

[73] Team, W. E. R., et al. Ebola virus disease in west africathe first 9 months of the epidemic and forward projections. *New England Journal of Medicine 371*, 16 (2014), 1481–95.

[74] Tobler, W. R. A continuous transformation useful for districting. *Annals of the New York Academy of Sciences 219*, 1 (1973), 215–220.

[75] Urwin, R., and Maiden, M. C. Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology 11*, 10 (2003), 479–487.

[76] Wickham, H. *ggplot2: elegant graphics for data analysis*. Springer Science & Business Media, 2009.

[77] Wolf, E. B. Creating contiguous cartograms in ArcGIS 9. *Proceedings of 2005 ESRI International User Conference* (2005).