UNDERSTANDING THE SPATIOTEMPORAL SPREAD OF INFECTIOUS DISEASES USING MATHEMATICAL AND STATISTICAL MODELS AND METHODS OF DATA ANALYTICS

by

Kaitlyn M. Martinez
A thesis submitted to the Faculty and the Board of Trustees of the Colorado School of Mines in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Computational and Applied Mathematics).

Golden, Colorado
Date ______________________

Signed: ______________________
Kaitlyn M. Martinez

Signed: ______________________
Dr. Stephen D. Pankavich
Thesis Advisor

Golden, Colorado
Date ______________________

Signed: ______________________
Dr. Greg Fasshauer
Professor and Head
Department of Applied Mathematics and Statistics
ABSTRACT

An ongoing challenge for the mathematical and statistical study of infectious disease spread is that many standard methods require an assumption of spatial homogeneity, even if the underlying mechanisms of disease transmission are intrinsically spatially-heterogeneous. One of the main goals of this thesis is to relax this assumption for the study of two different infectious diseases, Dengue Fever (DENV) and Ebola Virus Disease (EVD). The spatially-heterogeneous models for Dengue are primarily data driven, relying on proxy data to provide information on the demographic and environmental factors of mosquito borne virus transmission and spread. Modeling with large, diverse data requires effective variable selection and dimension reduction methods as even with “clean” high-dimensional data, the number of variables can quickly outpace the number of observations, leading to overfitting, redundant factors, and difficulties with model interpretation. Thus, prior to building models for Dengue risk, a novel dimension reduction method is proposed and applied. The spatial heterogeneity in the West Africa Ebola epidemic of 2014-2016 is then addressed by incorporating spatial mobility into a stochastic SEIR model by overlaying a directed graph structure over which the population can transition between spatial locations. Distinct spatial mobility structures are examined to explicate the most likely pathways of spatial infection spread. Epidemiological mechanisms are also investigated by estimating distributions for epidemiological parameters, such as the spatially and temporal varying infection/contact rate and the latent period. An empirically adjusted reproductive number is calculated for each spatial location using Bayesian inference methods in order to clarify the spatio-temporal transmission and population heterogeneity that drove the severity of the outbreak at the time.
TABLE OF CONTENTS

ABSTRACT .......................................................... iii

LIST OF FIGURES .............................................. vii

LIST OF TABLES ................................................. xiv

LIST OF SYMBOLS ............................................... xvii

LIST OF ABBREVIATIONS ...................................... xviii

ACKNOWLEDGMENTS ............................................. xix

CHAPTER 1 INTRODUCTION ..................................... 1

CHAPTER 2 A STATISTICAL DIMENSION REDUCTION METHOD WITH APPLICATIONS TO ENVIRONMENTAL AND DEMOGRAPHIC DATA IN BRAZIL .................................................. 4

\(2.1\) Introduction ................................................ 4

\(2.2\) Background and Previous Methods ............................ 5

\(2.3\) Methods .................................................. 9

\(2.3.1\) Preliminaries ...................................... 9

\(2.3.2\) Multi-Objective Feature Selection Method ................. 14

\(2.4\) Results and discussion ..................................... 23

\(2.4.1\) Environmental and Demographic Proxy Data Sets for Dengue in Brazil .................................................. 23

\(2.4.2\) Application of the Unsupervised Feature Selection Method .................................................. 26

\(2.4.3\) Convergence of the Algorithm .......................... 32

\(2.4.4\) Weighting and Selection of Pareto Optimal Point ........ 41
LIST OF FIGURES

Figure 2.1 **Visualization of Linkage Methods** A visual example of how the distance between two clusters is calculated via four linkage methods, single, complete, average, and centroid, used in hierarchical clustering methods ................................................................. 12

Figure 2.2 **Method Schematic** The five steps of the unsupervised multi-objective, clustering feature selection method: (i) Input data set is cleaned and pre-processed to be spatially static and any transformed variables are appended to this set; (ii) Data set is hierarchically clustered using the correlation details of (2.2); (iii) The Pareto front for the two objectives \((N, I)\) is approximated; (iv) The Pareto optimal point is selected using the weighted optimal distance in equation (2.7) to determine a number of cluster representatives and correlation threshold; (v) The dendrogram is cut at a correlation threshold and subset using cluster representatives. ................................................................. 23

Figure 2.3 **Correlation Plots** The magnitude of the correlation for each pair of variables in the Demographic (left) and Environmental data (right) are plotted on a color scale, where dark blue indicates values close to one while white indicates values close to zero. ................................................................. 27

Figure 2.4 **Dendrogram Representation of Hierarchical Clustering** The dendrograms of the demographic (top) and environmental (bottom) data sets are a visual representation of the nested clustering structure produced via single-linkage hierarchical clustering using the distance metric (2.2). Grey boxes and the color scheme represent the clustering structure produced when the tree is cut at a particular height (given by the height of the rectangles), thereby partitioning the data into various clusters. ................................................................. 29

Figure 2.5 **Information Distance** Visualization of the objective function \(I(m, \tau)\), which calculates the information lost due to the subsetting via clustering, for both data sets within the range \(\tau \in (0, 1)\) and \(m \in \{1, ..., 5\}\). Instead of raw information loss, the percentage of maximum information loss is plotted to provide a comparable scale between the two data sets. ................................................................. 31
Figure 2.6  **Evolution of Pareto Fronts** The evolution of a simulated Pareto front with 24 population members is displayed in the \((N, I)\) plane for both the Demographic (top) and Environmental (bottom) data. The first five generations of the algorithm are shown for comparison with the final (750th) generation.  

Figure 2.7  **\(L_1^i\) Difference between Fronts** The \(L_1^i\) difference between the \(i\)th and \((i + 1)\)st generation of a simulated Pareto front is presented with three different population sizes \((24, 96, 768)\) and three generation numbers \((250, 500, 1000)\) for both data sets.  

Figure 2.8  **\(L_2^i\) Difference between Fronts** The \(L_2^i\) maximal difference between the \(i\)th and all subsequent generations of a simulated Pareto front is shown with three different population sizes \((24, 96, 768)\) and three generation numbers \((250, 500, 1000)\) for both data sets.  

Figure 2.9  **Optimal Point Convergence** The location of optimal points selected via the five demonstrated weighting schemes is plotted in the \((N, I)\) plane with 100 generations and three population sizes \((24, 96, 768)\) for both data sets. The optimal point from the final generation is outlined in gray while the previous generations are plotted with transparency that corresponds to their generation number. More transparent points are produced earlier in the algorithm while less transparent points are nearer to the end. Additionally, the grey dashed line represents the final Pareto front of each simulation.  

Figure 2.10  **Standard \(\ell^2\) Error for Optimal Point 1 : 9** The standard Euclidean error of the optimal point selected via the 1 : 9 weighting scheme in generation \(i\) is compared to the optimal point selected in the final generation with three different population sizes \((24, 96, 768)\) and three generation numbers \((250, 500, 1000)\) for both data sets.  

Figure 2.11  **Standard \(\ell^2\) Error for all Optimal Points** The standard Euclidean error of the optimal point in generation \(i\) for each of the five weighting schemes is compared to the optimal point selected by the corresponding weighting schemes in the 1000th generation with three different large population sizes \((192, 384, 768)\) for both data sets.  

Figure 2.12  **Component Error for Optimal Point 1 : 9** The error contributed by the two objectives \((N, I)\) is plotted separately for simulations with 1000 generations and four different population sizes \((12, 48, 192, 768)\) for both data sets. Each component error is scaled to a percent of the maximal value each objective could attain, and as a result, each row is plotted on a different scale.
Figure 2.13 **Selected Pareto Fronts** The most stable Pareto front for both data sets was generated with 768 population members and 1000 generations and is shown on the \((N,I)\) plane with each optimal point highlighted. This illustrative simulation was selected due to the finite convergence of all optimal points except the points selected by the 3 : 1 and 9 : 1 weighting schemes which converged within an acceptable threshold. 43

Figure 2.14 **TF-IDF Ranking of Tokens** The TF-IDF ranking of the top 25 tokens in the four largest clusters in the Demographic data and the top 15 in the six largest clusters in the Environmental data are plotted. 50

Figure 3.1 **Mosquito-Borne Disease Transmission Pathway** The life cycle of the mosquito vector has four stages: (1) a female mosquito lays eggs in water, (2) eggs develop into larve, (3) larve mature into pupae, (4) pupae emerge as adults. Mosquito borne viruses spread when a susceptible vector bites an infectious host, becoming infected. Then when the infectious vector bites a susceptible host the host then can become infected. 59

Figure 3.2 **Brazilian Administrative Regions** Maps that show the political and administrative divisions of Brazil at the Regional (left), State (center), Municipality (right) levels. 65

Figure 3.3 **Training Scatter Plots** The fits of each model, trained either on large or small predictor data, to the training data for the country and regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction. 68

Figure 3.4 **Testing Scatter Plots** The fits of each model, trained either on large or small predictor data, to the testing data for the country and regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction. 69

Figure 3.5 **Maps of Linear Model Results** The observed risk of Dengue is mapped for each municipality to compare to the risk predicted by the regional and country scale models in the two testing years 2015/2016. The color scales are specific to the year and region for which the model was trained. 71
Figure 3.6 **Average State Risk Scatter Plot** The average of the predicted and observed risk is calculated for each of the 27 states for the training and testing data for both the country and regional scale models. The fit of these models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation is equal to the prediction. .................................................. 72

Figure 3.7 **Maps of Aggregated Linear Model Results** The mean observed risk of Dengue is mapped for each state to compare to the risk predicted by the regional and country scale models in the two testing years 2015/2016. The color scales are specific to the year and region for which the model was trained. ................................. 74

Figure 4.1 **Raw Data** Raw WHO data for Sierra Leone, Liberia, and Guinea for the 2014-2016 Ebola outbreak. The left hand panels show the data gleaned from Patient Databases (PD) and the right the Situation Report(SR). The black line in all panels represents the probable cases, the red all confirmed cases, and the green the total probable and confirmed cases. .................................................. 100

Figure 4.2 **Administrative Boundaries** Administratively-defined regions within (a,d) Guinea, (b,e) Sierra Leone, and (c,f) Liberia. ......................... 102

Figure 4.3 **Three Node model** The graph structure for the three node model. ... 105

Figure 4.4 **Accepted Epidemics** The accepted epidemics for the three node model at the SMC threshold of \( \epsilon = 0.20 \) are plotted (grey) with the observed data (black) and the mean simulation (red). ............... 107

Figure 4.5 **Timing of Ebola Outbreak in Guinea** The week in which each prefecture in Guinea recorded its first Ebola case is mapped with the dark colors indicating prefectures that observed cases earlier and the lighter colors representing locations which observed cases later. The range is 1 week to 57 weeks. ................................. 109

Figure 4.6 **Four Node Models** Four node model schema with differing connectivity structures informed by geographical boundaries and timing of epidemic spread. .................................................. 111

Figure 4.7 **Four Node Models with Indicator** Four node model schema with differing connectivity structures and a time indicator function from Guinea 1 to Guinea 2 informed by geographical boundaries and timing of epidemic spread. .................................................. 112
Figure 4.8  **Bayes Factors** Each facet shows the Bayes Factors calculated for a given model by dividing the acceptance rate of the titular model (Model 1) by the modeled indicated by the color scheme (Model 2) for each $\epsilon$ threshold used for SMC-ABC. The Bayes Factors are shown on a log scale and values greater than 1 indicate positive evidence for Model 1 while values less than 1 indicate evidence in favor of Model 2.

Figure 4.9  **Norm Comparison for Schemes 2i, 2, and 3i** The distribution of the norms calculated on all accepted epidemics with $\epsilon = 0.20$ for spatial connectivity schemes 2i, 2, and 3i. Since these values are calculated with accepted epidemics all values are less than 0.20. The top left panel represents the distribution of all 9 norm values $N^2, N^3, n = 1, \ldots, 2$ for the three schemes, the top right only the values of $N^3$, bottom left $N^1$, the bottom right $N^2$ are considered in the distributions.

Figure 4.10  **Accepted Epidemics for Scheme 2i** The accepted epidemics for the Scheme 2i four node model at the SMC thresholds of $\epsilon = 0.75, 0.50, 0.20$ are plotted (grey) with the observed data (black) and the mean simulation (red). While the mean epidemic changes minimally over the change in the threshold the range of the accepted epidemics varies much more at higher (less stringent) thresholds.

Figure 4.11  **Posterior Distribution of the Latent Period Parameter $\alpha$** The posterior distribution approximated when $\epsilon = 0.20$ of the parameter $\alpha$, where $\frac{1}{\alpha}$ represents the latent period, (blue) is compared to the prior (yellow). The mean value of the posterior is indicated by the blue dashed vertical line.

Figure 4.12  **Posterior Distribution of Coefficients $b_i$** The posterior distributions approximated when $\epsilon = 0.20$ of the parameters $(b_{0i}, b_{1i}, b_{2i})$ for $i = 1(\text{Liberia}), 2(\text{SierraLeone}), 3(\text{Guinea}), 4(\text{Guinea})$ are compared to their priors (yellow). The mean value of each of the posteriors is indicated by the blue dashed vertical lines.

Figure 4.13  **Functional Distribution of $\beta_i(t)$** The distribution of the accepted functional values with $\epsilon = 0.20$ of $\beta_i(t)$ are plotted, with the highest density values at each time point denoted by the darkest red.

Figure 4.14  **Prior Functional Distribution** The distribution of prior of $\beta_i(t)$ is plotted as a reference, with the highest density values at each time point denoted by the darkest red.
Figure 4.15 **Marginal and Joint Distributions of Flux Parameters** Panels (A)-(C) are two dimensional joint density plots of pairs of the three fluxes $\rho_{12}, \rho_{13}, \rho_{23}$ with the lightest blue denoting the highest density region and the grey contours indicated equal density values in the plane. Panel (D) is plots the marginal distributions for each of the fluxes, $\rho_{12}$ (blue), $\rho_{13}$ (yellow), $\rho_{23}$ (green) compared to their prior (black). All distributions are generated with $\epsilon = 0.20$.

Figure 4.16 **Distributions of Movement Probabilities** The distribution of the three probabilities generated by the fluxes are plotted for each spatial location besides Guinea 4. Guinea 4 does not have mobility in this scheme. All distributions are generated with $\epsilon = 0.20$.

Figure 4.17 **Distribution of Fluxes for Three Schemes** The distributions of the fluxes $\rho_{ij}$ with $i, j \neq 4$ associated with accepted epidemics with $\epsilon = 0.20$ for Schemes 2, 2i, and 3i demonstrate the mobility differences enforced by the schemes.

Figure 4.18 **Distribution of Probabilities for Three Schemes** The distributions of the probabilities $p_{ij}$ with $i, j \neq 4$ associated with accepted epidemics with $\epsilon = 0.20$ for Schemes 2, 2i, and 3i demonstrate the mobility differences enforced by the schemes.

Figure 4.19 **Basic Reproductive Numbers $R_i(t)$ versus $R_i^{(ea)}(t)$** The mean value of $R_i(t)$ (dashed) and $R_i^{(ea)}(t)$ (solid), calculated from the accepted epidemics and parameter sets when $\epsilon = 0.20$ are plotted for each spatial location. The grey dashed line on all plots highlights the threshold of one.

Figure 4.20 **Functional Distributions of $R_i^{(ea)}(t)$** The distribution of the accepted functional values with $\epsilon = 0.20$ of $R_i^{(ea)}(t)$ are plotted, with the highest density values at each time point denoted by the darkest color. The grey lines represent the 50% (solid) and 90% (dashed) quantiles. The horizontal grey dashed line on all plots highlights the threshold of one.

Figure 4.21 **$R_i^{(ea)}(t)$ in Situational Context** The mean of $R_i^{(ea)}(t)$ for each spatial location is again plotted, however, dotted vertical grey lines annotate each spatial location, indicating epidemic days which correspond to interventions or other events in each region.
Figure A.1  **Environmental Binary Correlation Plot**  The binary correlation between all pairs of tokens derived from the variable names of the Environmental data set are plotted. Dark blue indicates high positive correlation and thus tokens will nearly always occur in the same cluster. Dark red indicates high negative correlation, thus the pair of tokens nearly always occur in different clusters. Colors close to white indicates indicate token pairs that occur in the same or different clusters randomly. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 153

Figure A.2  **Demographic Binary Correlation Plot**  The binary correlation between pairs of tokens derived from the variable descriptions of the Demographic data set are plotted. This plot only includes a subset of the most frequent tokens. Dark blue indicates high positive correlation and thus tokens will nearly always occur in the same cluster. Dark red indicates high negative correlation, thus the pair of tokens nearly always occur in different clusters. Colors close to white indicates indicate token pairs that occur in the same or different clusters randomly. . . . . . . . 154

Figure B.1  **Regional Model Testing Fits**  The fits of each model, broken down by region, trained either on large or small predictor data, to the testing data for the regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 155


Table 2.1 **Summary of Data Set Properties**  Number of variables, represented spatial locations, and temporal observations for both the demographic and environmental data sets at each state of processing prior to applying feature selection. Note that each of the dimension measurements which includes a “*” is a quasi-spatial data set, in the sense that each of the 7 temporal observations are treated as new spatial locations rather than linked to their other temporal observations. .................................................. 26

Table 2.2 **Maximal Values & Scale Parameters** The maximal values and scale parameters used in the evaluation of the method and the selection of Pareto optimal points. ................................................................. 39

Table 2.3 **Convergent Simulations** The results of the optimal point choice using each optimal point weighting scheme for both data sets. Both the output in the objective space and the input space, over which the dimension reduction method is implemented are displayed. The percentages are calculated as percent of maximal for both objectives as provided in Table 2.2. The simulation used had a population size of 768 and 1000 generations. ................................. 45

Table 2.4 **Binary Correlation Calculation Table** For each Token Pair a value of $n_{\star\star}$ is needed to calculate the binary correlation. ........................................ 51

Table 2.5 **Binary Correlation Values for Selected Token Pairs** A small selection of token pairs are selected to demonstrate the utility of a binary correlation measure on the textual analysis of variable names or descriptions that have been clustered. In this case the selected pairs possess some of the greatest absolute binary correlations in the set. Some tokens have been disambiguated from their shorthand version. ................ 53

Table 3.1 **Small & Large Predictor Sets** The sizes of the small and large predictor sets used in this study to build linear models as well as the correlation thresholds and number of cluster representatives which generated the Environmental and Demographic subsets in question. ........ 64

Table 3.2 **Number of Municipalities per Region** The number of municipalities in each of the five regions in Brazil varies between regions and is not related to the geographical size of the regions. ....................... 65
Table 3.3 **Computational time required to train linear models** Time needed to train models on a small ($X^S_i$) and large ($X^L_i$) sets of potential predictor values at both the regional and country modelling scale. Note that the time required for the models at the regional scale represent the time required to train the largest region (in this case the Southeast) as the regional models can be trained in parallel.  

Table 3.4 **Model Sizes** Number of predictor variables in each of the four models. The regional sub-model sizes are indicated and the overall size of the regional models is found by finding the number of distinct variables in all sub-models.  

Table 3.5 **Model Errors** Errors generated by the training and testing data for the regional and country scale models trained on the small and large predictor variables sets. Errors are calculated with unscaled and un-transformed prediction and observed case counts.  

Table 3.6 **Significant Predictor Variables** The top ten significant predictor variables from each of the five regional sub-models as well as the country scale model are described in order. When a variable is a representative of multiple variables, due to the dimension reduction, the cluster description is used. The sign of the coefficient $\beta$ is indicated with parenthesis, which will indicate the overall effect the particular variable has on Dengue risk predictions. The relationship is specific to the precise variable not the cluster as a whole. Some shorthand conventions are used to save space. Spread refers to both Range and SD statistics, DTempR is the Daily Temperature Range, Rel Hum is Relative Humidity, Levels refers to min/max/mean summary statistics, G-NDWI is Green NDWI, and S-NDWI is SWIR NDWI.  

Table 4.1 **Bayes Factor Table** The strength of evidence in support of $m_i$ depends on the the value of $B_{ij}$. The levels of support are indicated with their corresponding ranges for $B_{ij}$.  

Table 4.2 **Standard Priors of Model Parameters** The standard priors used for each parameter in the model are detailed. Additional notes are included to explain non-intuitive aspects of the priors used.  

Table 4.3 **Computational Results:** The number of simulations, the total number of accepted simulations and the resulting acceptance rate is reported for all stages of ABC-SMC for the three node model.  

Table 4.4 **SMC-ABC Results for Six Schemes** The total simulations, acceptances and acceptance rate are reported for each of the six spatial mobility schemes when SMC-ABC uses a threshold of $\epsilon = 0.20$.  

xv
Table 4.5  **Epidemiological Parameter Means, Medians and Credible intervals** The indices $i = 1, 2, 3, 4$ refer to Liberia, Sierra Leone, Guinea 1, and Guinea 2, respectively, and the CI is a 89% equal-tailed credible interval. These values are calculated for the distributions generated by $\epsilon = 0.20$. .......................... 121

Table 4.6  **Flux Parameter Means, Medians and Credible intervals** The indices $i = 1, 2, 3, 4$ refer to Liberia, Sierra Leone, Guinea 1, and Guinea 2, respectively, and the CI is a 89% equal-tailed credible interval. These values are calculated for the distributions generated by $\epsilon = 0.20$. Recall that every flux is a dimensionless difference between two probabilities. . . 124

Table 4.7  **Joint Modal Flux Values** The mode of joint distribution of the three fluxes allows for the calculation of the joint modal values of the three fluxes for $\epsilon = 0.20$. .......................... 129

Table B.1  **Regional Model Errors** Errors broken down by region generated by the training and testing data for the regional scale models trained on the small and large sets of predictor variables. Errors are calculated with unscaled and un-transformed prediction and observed case counts. . . . . 156

Table B.2  **Models Trained on Small Data Set Errors** Errors broken down by region generated by the training and testing data for the country and regional scale models trained on the small set of predictor variables. Errors are calculated with unscaled and un-transformed prediction and observed case counts. .......................... 157
## LIST OF SYMBOLS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E$</td>
<td>Expected value</td>
</tr>
<tr>
<td>$\rho_{ij}$</td>
<td>Entries of correlation matrix</td>
</tr>
<tr>
<td>$d_\rho(x_i, x_j)$</td>
<td>Correlation distance</td>
</tr>
<tr>
<td>$\lfloor x \rfloor$</td>
<td>Floor Function: maps $x$ to greatest integer value less than $x$</td>
</tr>
<tr>
<td>MVN($\mu$, $\Sigma$)</td>
<td>Multivariate Normal with mean vector $\mu$ and covariance matrix $\Sigma$</td>
</tr>
<tr>
<td>N($\mu$, $\sigma^2$)</td>
<td>Normal with mean $\mu$ and variance matrix $\sigma^2$</td>
</tr>
<tr>
<td>$I$</td>
<td>Information loss</td>
</tr>
<tr>
<td>$N$</td>
<td>Subset dimension</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Correlation threshold</td>
</tr>
<tr>
<td>$m$</td>
<td>Number of representative features per cluster</td>
</tr>
<tr>
<td>TF($x$)</td>
<td>Single-cluster frequency</td>
</tr>
<tr>
<td>IDF($x$)</td>
<td>Inverse document frequency</td>
</tr>
<tr>
<td>TF-IDF($x$)</td>
<td>Product of TF and IDF</td>
</tr>
<tr>
<td>$S(t)$</td>
<td>Susceptible Population</td>
</tr>
<tr>
<td>$E(t)$</td>
<td>Exposed Population</td>
</tr>
<tr>
<td>$I(t)$</td>
<td>Infected Population</td>
</tr>
<tr>
<td>$R(t)$</td>
<td>Removed Population</td>
</tr>
<tr>
<td>$R_0$</td>
<td>Basic Reproduction Number</td>
</tr>
<tr>
<td>$R_i^{(ea)}(t)$</td>
<td>Empirically-adjusted Basic Reproduction Number</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EVD</td>
<td>Ebola virus disease</td>
</tr>
<tr>
<td>DENV</td>
<td>Dengue Fever Virus</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
<tr>
<td>NSGA-II</td>
<td>Non-dominated Sorting Genetic Algorithm II</td>
</tr>
<tr>
<td>GSOD</td>
<td>NOAA’s Global Surface Summary of the Day</td>
</tr>
<tr>
<td>ABC</td>
<td>Approximate Bayesian Computation</td>
</tr>
<tr>
<td>SMC</td>
<td>Sequential Monte Carlo</td>
</tr>
<tr>
<td>SR</td>
<td>Situation Report</td>
</tr>
<tr>
<td>PD</td>
<td>Patient Database</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

This endeavour has never been a solitary one, though at times, it can feel that way. To that end, I would like to acknowledge and thank the people and resources that have made this work possible. First and foremost, a never-ending thanks to my family. To my husband who patiently kept me fed and happy, throughout many new adventures and challenges, I am eternally grateful for your support and love that has buttressed me these last five(eight) years. To my parents and sisters, who always work to understand and engage in the complexity that is my work while also reminding me of the bigger picture of what fosters joy and love, I always feel inordinately lucky to be a Martinez. Besides my family, I have many mentors and colleagues to extend gratitude to. To my advisor Steve, I have been honored to have you as my cornerstone at Mines, you have always been there to support and guide me, both academically and personally, throughout the many changes in both our lives in the last five years. To Carrie and Sara, I cannot imagine two better role models to emulate as I mature as a scientist and a woman, I admire your intellectual creativity, grit, and your immensely kind hearts. To my statistical advisors, Grant, thank you for lending your expertise indiscriminately and Aaron, thanks for walking the early days of my graduate career with me always with a smile. To the many research groups and teams I have been a part of, made up of all levels of scientists, particularly the Math Bio Research group at Mines and the Epi team at LANL, thank you for your feedback, ideas, commiseration, and humor on research and life. To my thesis committee, your patience and support means the world to me, particularly now as you are faced with reading this document. Finally, much of this work would not have been possible without the funding from the NSF, the Graduate Research Fellowship Program supporting me for the majority of my degree, giving me flexibility to fully explore what it means to do mathematical research, and the grants, DMS-1614586/DMS-1551229, which opened doors to many travel and internship opportunities.
The spread of infectious diseases has recently captured the attention of the entire world, whose actions now depend crucially upon knowledge of how a particular virus, infection, or condition impacts their health and well-being or that of loved ones and their community. However, in our increasingly connected and digital world, we live with simultaneously more awareness and more risk of being impacted by the spread of one or many infectious diseases. The work in this thesis occurred in perverse harmony with world-wide infectious disease phenomena. It began in 2016 on the heels of the Ebola epidemic originating in West Africa and as Zika spread to the Americas imperiling the 2016 Rio de Janeiro Olympics, and it is being completed as the world is grappling with the reality of the global spread of COVID-19. This symmetry brings into clear focus the critical and timely nature of innovative mathematical, statistical, and computational techniques to better model, describe, understand, and overcome epidemiological challenges.

Over time, scientific inquiry into the mechanisms of disease has frantically tried to employ any tools available keep pace with the ever expanding range of epidemiological questions of interest. The study of mathematical epidemiology is no exception. Evolving from its origins in 1662 and John Graunt’s “Bills of Mortality”, Daniel Bernoulli’s model of the inoculation of smallpox in 1760, and the study of spatial and temporal patterns of Cholera cases by John Snow in 1855, which pinpointed the source of contaminated water in London, mathematical and statistical models of the spread of infectious diseases have become more advanced and nuanced [16]. A critical advancement in the field came not from mathematicians, but public health experts in the early 1900s, when physicians Sir R.A. Ross, W.H. Hamer, A.G. McKendrick, and W.O. Kermack employed mass action ideas to the spread of disease in order to build the now ubiquitous SIR compartmental models of disease dynamics [17]. From
there, the field began to branch out to address various limitations of the existing modeling infrastructure. Stochastic methods emerged to better represent the inherently unpredictable nature of transmission and disease progression, while compartmental models expanded to become more detailed and precise by incorporating additional disease states, age structures, or vector-borne disease spread, among others [17]. Currently, one of the challenges driving the development of epidemiological models is the lack of understanding of spatially-heterogeneous disease spread, especially as the intrinsic disease mechanisms at particular spatial scales are difficult to describe or measure. Thus, the unifying focus of this thesis is to create and examine tools that can aid in clarifying or uncovering the intrinsic dynamics of spatially-heterogeneous disease spread.

In subsequent sections, we examine the spatial and spatio-temporal dynamics of two different types of infectious diseases. Chapters 2 & 3 contain studies of the progression of Dengue fever throughout a particular region of South America. In particular, Chapter 2 investigates and exploits structures in large spatial, demographic, and environmental data to produce lower-dimensional data sets that are more useful in modeling applications. Effective data analytics is critical in mathematical and statistical epidemiology as many mechanisms of infectious diseases are difficult to measure unless at a small scale and many use proxy data to fill the this gap, especially when the scope and scale of the modeling effort is large. To this end, we develop a new, multi-objective feature selection method that uses clustering to generate lower-dimensional data sets while maintaining the intrinsic information within a full data set. In particular, our method is tailored to preserve the original interpretation of variable sets prior to selecting a specific application, and we demonstrate the efficacy and utility of the method on demographic and environmental data that can be used in models of mosquito-borne disease transmission, such as the spread of Dengue fever in Brazil.

Within Chapter 3 we utilize the reduced dimension of the variable space to construct an informative linear statistical model that can be used to quantify the yearly Dengue burden in Brazil. Due to the scale at which they are constructed the linear models are simple; however,
besides providing a case study of the importance of our dimension reduction method, these models provide key insight into the environmental and demographic factors that increase risk of Dengue. We find that models that are more local in nature, rather than universal country-wide models, are more accurate, likely owing to the heterogeneity of intrinsic risk factors, which could be determined by the local nature of the population and environment. In Chapter 4, our analysis will center on developing a novel spatially-heterogeneous extension of a spatially-homogeneous SEIR model. A general infrastructure which allows for mobility of individuals in the population is outlined and then applied to the 2014-2016 Ebola outbreak in Western Africa. The spatial heterogeneity and mobility is incorporated by using graph structures with nodes representing spatial locations and edges serve to control mobility throughout the entire domain. We choose to use an embedded stochastic process so that Bayesian inference methods can be utilized to produce a more thorough understanding of the disease process through the estimation of epidemiological parameters. The impacts of mobility on heterogeneous disease spread are investigated via analysis of various spatial mobility schemes and their feasibility in producing simulations that align with reported data.
CHAPTER 2
A STATISTICAL DIMENSION REDUCTION METHOD WITH APPLICATIONS TO ENVIRONMENTAL AND DEMOGRAPHIC DATA IN BRAZIL

2.1 Introduction

As barriers to the collection, storage, and analysis of data have diminished, researchers have become increasingly inclined to use big data to generate new mathematical and statistical models. While theoretical modeling efforts on their own have provided new insights, the incorporation of measured data allows a model to undertake a larger set of practical applications. For example, large data sets can be used to introduce a variety of environmental, biological, and socioeconomic factors that inform epidemiological models of disease dynamics [20]. By acting as proxies, diverse data streams provide valuable real-world, and often real-time, information about mechanisms and interactions that are difficult or impossible to measure. Despite the massive opportunity that the abundance of accessible data presents, the distinct challenge of utilizing it in an optimal manner remains.

While standard measurement error, sparsity, and bias issues arise within any data stream, often the critical hurdle to overcome when using a particular big data set is how to address its massive dimension. Specifically, it can be difficult to determine which variables or collections of variables are useful, how they should be utilized, and how universal their utility may be. Effective variable selection and dimension reduction methods have become increasingly important in the statistical modeling community, as the use of big data in modeling efforts has expanded [57]. Access to diverse data sources presents the opportunity to incorporate concrete measurements to better inform modeling efforts, but even with “clean” high-dimensional data, the number of variables can quickly outpace the number of observations, leading to overfitting, redundant factors, and difficulties with model interpretation. One typical approach is to apply dimension reduction methods like Principal Component Analy-
sis (PCA) [7], which while effective at extracting information from a data set, can obfuscate the observed processes measured by independent variables. This can prove problematic as many modeling applications are built to support a human decision making process, and interpretation of model components is often just as important as the model output. As an example, epidemiological models are frequently built to support public health efforts, which require them to maintain a basic level of biological interpretation [57]. This additional constraint renders many existing dimension reduction methods, such as PCA, inadequate and necessitates the development of novel tools to select the variables of greatest importance in a given data set.

Because a single set of measurements may be useful to identify and investigate a variety of questions, perhaps even amongst differing fields of research, we aim to construct and perform novel dimension reduction methods without consideration of a specific model or application. In general, this paper is organized as follows. In the next section, we provide some background information concerning previous dimension reduction methods and their associated properties. Next, we outline and discuss a new feature selection method that uses clustering methods to identify, among multiple objectives, a lower-dimensional subspace that preserves the intrinsic information within a given data set. Finally, we demonstrate the utility and efficacy of the method on demographic and environmental data from Brazil.

2.2 Background and Previous Methods

Large data sets present many challenges to researchers. As the number of variables and observations within a data set increases, so does its dimension; however, its intrinsic dimension, i.e. the minimal representation of the data, may not. Additionally, models can be at risk of overfitting when the number of response observations is not much larger than the number of potential variables. Hence, there are numerous advantages to reducing the dimension of a feature space. Basic considerations like storage, computational power, and computational time can be essentially ignored with small data sets, but even performing basic operations on high dimensional data sets adds a layer of complexity to a problem.
The implementation of dimension reduction methods can also aid in the understanding and visualization of a data set, which are critical steps prior to any modeling effort. Finally, dimension reduction has the obvious benefit of improving model prediction performance by uncovering important signals or useful information that can be obfuscated by noise, error, bias, or sparsity of large data sets.

The aim of any dimension reduction method is to identify a lower-dimensional variable subspace embedded within the original high-dimensional space that preserves essential information. Three important considerations when choosing or designing a method are final subspace dimension, information retention/loss, and interpretability [57]. Such methods can be described as employing feature selection or feature extraction and, depending on the procedure, may be unsupervised or supervised.

The majority of dimension reduction methods are supervised, which means that a classifier is linked to each observation in the data set and used in the process of dimension reduction [66]. A classifier could occur in the form of a model output, e.g. the result of a regression model, or in the form of class labels. In the case of a regression model, variable importance is detected through its contribution to the accuracy of the model outputs. Alternatively, dimension reduction methods can be supervised if class labels are included within the data. Class labels are not variable names, but an additional feature chosen to summarize or link other features together. As an example, consider a patient medical file with features that measure age, height, weight, gender, cholesterol level, and various other health measures, which includes an additional class label indicating a physician’s assessment for cardiovascular risk [47]. While in some cases supervision is advantageous, many data sets possess intrinsic information that lend themselves to multiple applications, independent of a classifier. In particular, many data sets require a model output and a dependent variable to define a classifier, severely limiting the scope of their utility by necessitating repeated dimension reduction applications on a single data set used in different models. For this reason an unsupervised method may be preferred.
Unsupervised methods are independent of a classifier, so the search for, or creation of, variables is disjoint from the classification problem. However, the lack of an inherent measure of correctness or accuracy of the method is a major trade-off when choosing a unsupervised method. Instead, success in these methods is evaluated in terms of information retention/loss and, depending on the application, this can be measured by saliency, entropy, smoothness, density or reliability [41]. What is lost in accuracy evaluation is typically gained in scalability to large data sets, computational simplicity, and reduced risk of overfitting [41, 63]. Because the method only needs to be performed once on a data set, a reduction in computational time results and a multi-use reduced data set ready for any application of interest is produced.

Broadly speaking, dimension reduction methods can be characterized according to two distinct approaches: feature extraction and feature selection. Feature extraction methods shrink the data set using statistical or information-based criteria, matrix decompositions, or projection techniques. Perhaps the most celebrated of such methods is Principal Component Analysis (PCA) [7], which generates a lower dimensional subspace embedded within the set of independent variables by creating new features though an eigenvalue decomposition analysis. This and many other extraction methods are naturally free of input from the model or application, and therefore serve as standard unsupervised methods. Often when referring to dimension reduction, studies discuss only feature extraction methods, though there are non-extraction methods that also reduce the dimension of the variable space [66]. While feature extraction methods can be used to identify and construct compact subspaces that capture distinct trends and behaviors in the data, there is a tendency to sacrifice the interpretation of the initial variables in favor of the information retained by the new set of variables. Adjustments can be made to existing methods that attempt to distinguish between groups of variables or to create sparse combinations of features, but these methods become increasingly complex and computationally expensive, while sacrificing the original strength of the associated feature extraction method. Because we view data interpretation to be inherently valuable, these approaches are not robust enough for our goals.
Contrastingly, feature or variable selection methods select a subset of the original variables to serve as the lower-dimensional subspace, rather than creating new variables. This manner of selection process is further categorized into filter, wrapper, and embedded types. Filter methods score and then rank features based on a chosen measure or set of measures. Then, a threshold is set to determine which variables to keep and which to remove from consideration for modeling. The score can be an inherent measure of information by considering the variable independently or incorporate a dependent variable to measure the relevancy of a feature. Wrapper methods use a defined model to search and find combinations of variables that produce the greatest model accuracy. Well-known examples of such methods include simulated annealing, forward selection and backward elimination methods, and genetic algorithms, among others [63]. These methods are criticized as being computationally expensive because they often search through all potential combinations in the variable space, but more efficient search algorithms have been recently developed [41]. Even with efficient search methods, however, wrapper methods often contain a significant risk of over-fitting [63]. Embedded methods create a model in tandem with variable selection through a learned process, often incorporating optimization constraints that can bias the model towards other user defined requirements, like lower model complexity. Classic examples of embedded methods include decision trees, regression type models, and other regularization models, like LASSO or Elastic Net Regression [82]. One crucial difference among these approaches is that filter methods can be unsupervised while wrapper and embedded methods use a classifier or model to learn more about the structure of the data and, therefore, are necessarily supervised. While using the eventual model output may aid in increasing accuracy, it can also be computationally expensive and far less generalizable. Because the ability to extend the results of our dimension reduction method to other models is paramount, we will pursue a filter, rather than a wrapper or embedded, approach in the next section.
2.3 Methods

We have designed and constructed a new, multi-objective clustering method to reduce the dimension of a large data set while maintaining the original interpretation of the measured variables. The method explores the colinearity, covariance, and categorical properties of variables in order to determine the subset of features with greatest utility. Here, utility is defined as a combination of small dimension, variable interpretation, and conservation of the intrinsic information within the full data set. Our method extracts and filters attributes by combining methods of clustering and multi-objective optimization to identify a subset of useful representative factors for the complete data set. The use of correlation-based clustering is a novel way to partition the data before selecting or extracting relevant features. The utility measures we construct are independent of the eventual implementation of the data source; hence, further variable selection can also be performed for a variety of distinct applications.

The main idea of the method is to exploit the inherent colinearity of the given variables, rather than standard measures of distance, in order to select the features that best represent the entire set. In particular, we construct a filter feature selection method, which uses unsupervised hierarchical clustering to group and classify variables and chooses representative variables to retain within each of the discovered classes.

2.3.1 Preliminaries

Prior to describing the new feature selection method, we provide some basic definitions that may be useful throughout the remainder of the chapter. Given a collection of variables \( X = [x_1, \ldots, x_n] \) and an associated data set, we can understand the shape and spread of the variables and data as a whole by their variance, covariance, and correlation structure. The variance of a particular variable vector \( x = x_i, \ i = 1, \ldots, n \) is defined as

\[
\sigma_x^2 = \mathbb{E}[x^2] - (\mathbb{E}[x])^2.
\]
Similarly, if \( x \) possesses \( M \) entries and has mean \( \mu \), the population variance of this sample is given by

\[
\sigma^2 = \frac{1}{M} \sum_{i=1}^{M} (x_i - \mu)^2.
\]

If the mean of the sample is not known \textit{a priori}, we must instead use an approximation, namely the sample mean defined as

\[
\bar{x} = \frac{1}{M} \sum_{i=1}^{M} x_i,
\]

and the sample variance is then defined to be

\[
\hat{\sigma}^2 = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \bar{x})^2.
\]

The definition of the covariance of two variables merely expands on the notion of the univariate variance by measuring the joint variability between any two variables. It is defined by

\[
\sigma_{x,y} = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y].
\]

For a sample of size \( M \) the sample covariance is then

\[
\hat{\sigma}_{x,y} = \frac{1}{M-1} \sum_{i=1}^{M} (x_i - \bar{x})(y_i - \bar{y})
\]

where \( \bar{x} \) and \( \bar{y} \) represent the samples means of the vectors \( x \) and \( y \), respectively. Notice that \( \sigma_{x,x} = \sigma_x^2 \) so that the covariance between a vector and itself is merely its variance. We can further expand our understanding of the relationship between variables by defining the Pearson Correlation Coefficient which quantifies the level of linear relationship (i.e. correlation) between two variables. This quantity is defined by

\[
\rho = \frac{\sigma_{x,y}}{\sqrt{\sigma_x \sigma_y}},
\]

and for a sample of size \( M \) an estimate of the correlation coefficient can be calculated as
\[ \hat{\rho} = \frac{\sum_{i=1}^{M} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{M} (x_i - \bar{x})^2 \sum_{i=1}^{M} (y_i - \bar{y})^2}}. \]

Given a data set containing observations of \( n \) variables, we can further generate \( n \times n \) covariance and correlation matrices by applying the formulas above to each possible pair of variables in the data set so that these matrices are defined entrywise by \( \sigma_{i,j} \) and \( \rho_{i,j} \), respectively.

Next, we detail a variety of clustering methods that will be useful in describing our method in the next section. There are many tools that one can use when clustering data, and these can be broadly categorized into partitional clustering methods and hierarchical clustering methods [40]. Partitional clustering aims to partition a data set into \( k \) clusters, where membership is disjoint or probabilistic, while hierarchical clustering generates a set of nested clustering schemes that elucidate the structure of the clustered data for all numbers of clusters that are possible for the data set. There are other extensions and variations beyond these two basic categories, but most methods can be derived from them.

As we do not assume any inherent knowledge of the underlying cluster structure in a given data set, the use of hierarchical clustering methods is a clear first step in our analysis. Hierarchical clustering algorithms, themselves, fall into two distinct categories: Agglomerative and Divisive [40]. Agglomerative clustering begins with each data element as a single cluster and then at each step, merges the two most similar clusters, until all elements are grouped within a single cluster. Divisive clustering takes the opposite approach, beginning with a single cluster containing all data elements, and then at each step, a cluster is split into two smaller clusters. Implementing the merge algorithm, or split algorithm in the case of divisive hierarchical clustering, is a non-trivial endeavour, and the selected method or algorithm will naturally possess various strengths and weaknesses. For divisive clustering the algorithm must also have a method for deciding which cluster to split, in addition to the optimal method for splitting, making it more complex overall. Thus, we will explore the possibility of using an agglomerative clustering method.
Figure 2.1: **Visualization of Linkage Methods**

A visual example of how the distance between two clusters is calculated via four linkage methods, single, complete, average, and centroid, used in hierarchical clustering methods.

Agglomerative clustering employs various merge schemes that use a defined distance function (or matrix), represented by $d(a, b)$ for $a \in A, b \in B$ and clusters $A$ and $B$, to calculate a dissimilarity score for arbitrary pairs of clusters. The pair of clusters that minimize the dissimilarity score are then merged at that step of the process. To provide the reader with some intuition regarding this process, we include some examples of linkage algorithms, and Figure 2.1 provides a pictorial representation of some of these methods. First, one may consider a maximum or *complete linkage* method, also known as farthest neighbor clustering.
In this case dissimilarity between two clusters $A$ and $B$ is defined by

$$L(A, B) = \max \{d(a, b) : a \in A, b \in B\}.$$  

Contrastingly, the minimum or *single linkage* function, also known as nearest neighbor clustering, uses the minimal value of the distance function. In this case dissimilarity is defined by

$$L(A, B) = \min \{d(a, b) : a \in A, b \in B\}.$$  

Other clustering schemes use average values to determine the clusters that are to be merged. They include the *unweighted average linkage* method, in which dissimilarity is given by

$$L(A, B) = \frac{1}{n_An_B} \sum_{a \in A} \sum_{b \in B} d(a, b),$$

and the *weighted average linkage* method, in which dissimilarity is defined as

$$L(A, B) = \frac{1}{n_An_B} \sum_{a \in A} \sum_{b \in B} w_{ab} d(a, b)$$

where $n_A$ denotes the number of elements in cluster $A$ and $w_{ab}$ represents a chosen weighting scheme. Yet another method uses the centroid of each cluster to determine their similarity. In particular, the *centroid linkage* scheme is defined by using a dissimilarity given as

$$L(A, B) = d(c_A - c_B)$$

where $c_A$ and $c_B$ are the respective centroids of cluster $A$ and $B$. More specifically, the centroid for a cluster $A$ is the group average of all elements of $A$, namely

$$c_A = \frac{1}{n_A} \sum_{a \in A} a.$$  

Other methods exist, such as *Ward’s linkage* method [67, 79], and these typically increase in complexity. In general, the choice of the appropriate linkage method depends greatly on the context in which it is being used. In the next section, we will use a single linkage method based on their correlation structure.
2.3.2 Multi-Objective Feature Selection Method

The method begins with a matrix
\[ \theta = [x_1, x_2, \ldots, x_n] \]
that represents a large, cleaned, and processed data set of measured variables \( x_i \), from which we would like to select a subset of representative features. Throughout, we will use abbreviated notation to represent quantities that are fixed to a particular variable. For instance, the minimum value of a variable taken over all of its \( M \) observations will be denoted
\[ \min(x_i) = \min_{j=1,\ldots,M} (x_i)_j \]
with a similar representation for the maximum value.

To begin the algorithm, selected transforms \( T(\theta) \) are applied to the data in order to augment existing variables \( x_i \) with transformed variables \( x_i^* = f(x_i) \) that exist at a new scale. For the purposes of our later demonstration,
\[ T(\theta) = [f(x_1), f(x_2), \ldots, f(x_n)] \]
is a log transform where
\[ f(x_i) = \log(x_i + |\min(x_i)| + \sigma_i) \]  
(2.1)
and \( \sigma_i \) is the standard deviation of the variable \( x_i \). The augmented data set we consider is then
\[ \Theta = [\theta, T(\theta)] = [x_1, x_2, \ldots, x_n, x_{n+1}, \ldots, x_{2n}] \].
If other transforms are needed, then one may concatenate the transformed variables in the same way as we demonstrate with the log transform. Once all transforms have been applied and we have expanded our feature space as necessary, the dimension reduction process can begin. The data is clustered based on a correlation distance \( d_\rho \), which we define in the following manner[25]. First, denote the correlation coefficients of any two variables \( x_i \) and
where \( \sigma_{ij} \) is the covariance of \( x_i \) and \( x_j \). Then, the correlation distance between \( x_i \) and \( x_j \) is defined as

\[
d_{\rho}(x_i, x_j) = \sqrt{1 - |\rho_{ij}|}.
\] (2.2)

This definition naturally fixes the groups of mutually redundant variables in terms of their colinearity. Next, a single linkage hierarchical clustering scheme is used to generate clusters for all potential correlation thresholds of interest. The method is applied to data sets with unknown clustering structures, making hierarchical clustering preferable to other clustering techniques like \( k \)-means, in which cluster number is imposed \textit{a priori} [38]. Our formulation of the problem motivates the use of a single linkage scheme as opposed to other algorithms because it guarantees that no two elements with a correlation above a certain threshold will be placed within different clusters. In other methods, such as complete linkage or Ward’s linkage, highly correlated variables are consistently grouped in separate clusters due to the averaging or optimization of the clustering algorithm, and this is undesirable for our purposes.

The hierarchical clustering algorithm results in a set of layered clustering schema, represented by a dendrogram \( D_\Theta \), from which to select depending on the choice of a correlation threshold \( \tau \). This threshold refers to the minimum level of positive or negative correlation between two variables required to cluster them. For a particular value of \( \tau > 0 \), \( D_\Theta \) is cut at height \( h = \sqrt{1 - |\tau|} \), a cluster structure \( C_\tau \) with \( C \) clusters is selected, and the cluster membership of each variable \( x_i \) is determined. The variables are then partitioned into \( C \) distinct collections, each of the form \( X_p = \{x_{p1}, \ldots, x_{pn_p}\} \) where \( n_p \) is the number of variables in cluster \( p \) for \( 1 \leq p \leq C \). Once each variable is assigned to a cluster, the goal is to select the variable(s) that best represent the cluster behavior as a whole. An obvious choice of representative variable for cluster \( X_p \) is its medoid, defined by
\[
x_{med}^p = \arg \min_{y \in \{x_1^p, \ldots, x_{np}^p\}} \sum_{i=1}^{np} d_p(y, x_i).
\] (2.3)

The medoid is the variable with minimal distance to every other member of the cluster. For clusters of highly correlated variables, this one representative may suffice, but less correlated clusters may require more than one variable to capture the behavior as the medoid may be biased to one part of a cluster. To address this issue, we consider \(k\) representatives consisting of the \(k\) variables that possess minimal sum distance to the other members of the cluster. The medoid is the first of these \(k\) variables and then the next \(k - 1\) variables with the next lowest sum distances comprise the set of representative factors.

While the choice of the number of representative factors and the correlation threshold can be determined \textit{ad hoc} or via other evaluation metrics on the entire set of clustering schema, our method provides a more rigorous selection criterion by using multi-objective optimization to simultaneously choose the cut height and number of medoids in an objectively optimal way. As the objectives we will consider, namely overall dimension reduction and information loss due to dimension reduction, are often in conflict with each other, a unique solution that perfectly optimizes both is unlikely to exist. Instead, multi-objective or multi-criteria optimization defines a problem space that produces a set of solutions to satisfy multiple objectives at once and illustrates the trade offs between them. To understand how this set is defined, we must first introduce the concept of dominance and a dominance test.

We say that one solution of a multi-objective optimization problem \(\lambda_1\) dominates another solution \(\lambda_2\) if:

1. Solution \(\lambda_1\) performs no worse than \(\lambda_2\) in all objectives
2. Solution \(\lambda_1\) performs strictly better than \(\lambda_2\) in at least one objective.

It follows that \(\lambda_1\) dominates \(\lambda_2\) if and only if \(\lambda_2\) is dominated by \(\lambda_1\) \cite{35}. This definition of dominance aids us in defining the set of solutions which are non-dominated, i.e. those that are as good as or better than all other solutions in all objectives and strictly better than the
solutions they dominate in at least one objective. This non-dominated set in the feasible
decision space is called the Pareto-optimal set, and the mapping of these solutions in the
decision space to the feasible objective space is called the Pareto-optimal front [35].

The two characteristics of the data set that we wish to optimize are (i) the overall di-
mension of the representative subset and (ii) the information loss due to feature selection.
The subset dimension $N(m, \tau)$ can be represented as a function of the number of permissible
representative features per cluster, denoted by $m$, and the correlation threshold $\tau$. It is
determined merely by counting the number of variables selected as representative of their
clusters within a chosen subset. Similarly, the information loss $I(m, \tau)$ within a specific
subset possesses the same parameter dependence. This quantity is determined using a vari-
ation of the Facility Location Distance Problem [22] to measure the distance between the
original data set and the representative subset. More specifically, for a clustering scheme
consisting of $C$ clusters, the $p^{th}$ cluster contains a set of up to $k$ representative variables $x^p_j$,
for $j = j_1, \ldots, j_k$, and the total information loss $I(m, \tau)$ is defined by

$$I = \sum_{p=1}^{C} \sum_{i=1}^{n_p} \left[ \min_{j=j_1, \ldots, j_k} \left| \frac{x^p_i}{||x^p_i||_2} - \frac{x^p_j}{||x^p_j||_\infty} \right| \right]$$

(2.4)

where $x^p_i$ represents the $i^{th}$ redundant variable in the $p^{th}$ cluster, which contains $n_p$ variables.
In the current context, we allow the maximum number of medoids to vary from one to five,
and the correlation threshold to vary within the interval $[\tau_1^*, \tau_2^*]$, where $\tau_1^*$ and $\tau_2^*$ represent
user-defined thresholds for a given application. The range of number of representative vari-
ables was determined ad hoc through early tests. Adding additional representative variables
beyond five, gave increasingly diminishing levels of information gain. The correlation interval
could be the entire range $[0, 1]$ or a narrower set of values chosen to arrive at the best schemes
for higher correlation thresholds separate from those available to lower thresholds. This al-
allows for discrimination between different outcomes of the method, ranging from the goal of
a first pass dimension reduction, which eliminates the low hanging fruit of highly redundant
variables, to a much more aggressive variable reduction objective. Then, for each pair \((m, \tau)\) of cluster representative number and correlation threshold we can calculate the dimension of the representative subset found via the feature selection process and the information loss resulting from the loss of the eliminated variables.

In general, we wish to simultaneously minimize the number of variables and the information loss. Thus, our multi-objective problem is defined as follows:

\[
\min \{ N(m, \tau), I(m, \tau) \} \\
\text{s.t. } (m, \tau) \in \{1, \ldots, 5\} \times [\tau_{1}^{\ast}, \tau_{2}^{\ast}].
\]  

(2.5)

While analytical methods do exist to solve some multi-objective problems and compute parts of or the entire Pareto front, they typically impose various constraints, such as existence of derivatives, and add additional complexity for the user, including appropriate choices of weights to reformulate the problem as a single objective optimization with constraints. For our purposes, both objectives are not continuous functions, and thus are not differentiable. The dimension is a count of discrete elements, while the information loss function experiences discontinuities due to cluster aggregation in the hierarchical algorithm. As a result, it will suffice to employ a numerical approximation of the Pareto front, and this can be achieved using the evolution inspired algorithm NSGA-II [27]. Using the natural evolutionary concepts, such as selection, recombination, and mutation, this algorithm steers a population (set) of individuals (decision vectors) towards optimal or near optimal solutions [35]. There are various algorithms that have been developed since the inception of this field in the 1910s and most fit into three basic categories.

1. Pareto based, which first ranks based on dominance and subsequently ranks by set diversity.

2. Indicator based, which use an indicator to measure the performance of the set

3. Decomposition based, which break the problem into smaller sets of problems and then uses different scalarization techniques to solve each of the sub problems.
The application of this technique was simple enough and lent itself well to Pareto based methods and thus more complicated methods and those that fit into the Indicator based or Decomposition based categories need not be detailed.

Pareto based algorithms dominate the space of multi-objective evolutionary algorithms. Every one of these algorithms combines a Pareto dominance ranking with a secondary ranking procedure to select optimal members of the set. The NSGA-II algorithm is a standard Pareto based algorithm, and while the exact ranking and evolution will differ between algorithms, NSGA-II serves as an exemplar of this universe of algorithms. Additionally, it is the algorithm used in this application so it behooves us to detail it here. The NSGA-II algorithm has two stages. In the first, a population undergoes a generational loop, where the original population is evolved into the next generation of the population. The population $P_i$ is simply the a set of $x \in \mathcal{X}$ where $\mathcal{X}$ is the whole space over which the set of functions $f_i(x)$ are being minimized. So for each $P_i$ consisting of $n$ individuals $x \in \mathcal{X}$ the generational loop, through a method of recombination and mutation, produce a intermediary set $Q_i$. First two elements $x^{(1)}$ and $x^{(2)}$ are selected from the population $P_i$. Then the elements of these two “parents” are recombined using a simulated binary crossover operator [26]. Finally, the resulting vectors are mutated using a polynomial mutation method [52]. Each set of two parents $x^{(1)}$ and $x^{(2)}$ produce two new individuals and so the overall population produces $n$ new individuals to make up a $Q_i$. The second stage of the algorithm is the selection step. The algorithm considers the set $P_i \cup Q_i$ and selects the population $P_{i+1}$ using a multi-objective ranking method. The multiobjective ranking distinguishes the NSGA-II algorithms from other genetic algorithms because it consists of a two level ranking procedure. This allows the algorithm to select based on both the Pareto order and the overall diversity of the set of individuals in the population. The algorithm first sorts the population depending solely on Pareto order using non-dominated sorting [35]. The population is partitioned based on this sorting into layers of individuals that do not dominate each other. The highest layer $\ell$ consists of the elements that are dominated by the rest of the set while the first layer $R_1$
dominates the rest of the set. In this way the population members in the layers $R_1, R_2, \ldots, R_\ell$ can be ranked where the Pareto ranking is given by the subindex $k$ in $R_k$. Clearly, there will be elements that share the same ranking based on Pareto dominance, and so an additional ranking procedure is employed to produce the final ranking. The second ranking procedure evaluates the crowding distance of individuals that reside within the same layer. The crowding distance is calculated for $x \in R_k$ for any $k = 1, \ldots, \ell$ by averaging the contributions $c_i$ over all of the $i = 1, \ldots, m$ objective functions $f_i(x)$, namely

$$c(x) = \frac{1}{m} \sum_{i=1}^{m} c_i(x).$$

(2.6)

For $x \in R_k$ the specific contribution $c_i(x)$ is calculated by finding the nearest neighbors of $x$ and then the maximal space separating $x$ from its nearest neighbors so that

$$c_i(x) = u_i(x) - l_i(x)$$

where

$$l_i(x) = \max \left\{ f_i(y) : y \in \mathcal{X}_l \cup \{-\infty\} \right\}$$

$$u_i(x) = \min \left\{ f_i(y) : y \in \mathcal{X}_u \cup \{\infty\} \right\}$$

and the sets $\mathcal{X}_l$ and $\mathcal{X}_u$ are defined to be

$$\mathcal{X}_l = \{ y \in R_k \setminus \{x\} : f_i(y) \leq f_i(x) \}$$

$$\mathcal{X}_u = \{ y \in R_k \setminus \{x\} : f_i(y) \geq f_i(x) \}.$$

The crowding distance increases as an element gets further from its nearest neighbors, and so elements with the greatest crowding distance will be ranked higher in their layers. With this two level ranking system, the set of $n$ new individuals in a specific population $P_{i+1}$ are selected first by selecting the elements in the highest layers. If the selection must choose between individuals in a particular layer to arrive at exactly $n$ individuals, the crowding
distance is used to determine the final selection. This algorithm generates a computational approximation of the Pareto set for the specified range of cluster representatives and correlation threshold pairs. Correspondingly, the Pareto front consists of ordered pairs for the number of variables in the subset produced by the clustering scheme and the information lost by using these subsets.

Though each point in the Pareto set is an optimal solution, different trade-offs are associated with each choice. For these objective functions $N(m, \tau)$ and $I(m, \tau)$, the global optimal point on the $N, I$ plane is $P = (1, 0)$, as one variable which perfectly represents the entire set without any information loss is ideal. Then, depending on the relative importance of the two functions being minimized, a distance between the ideal point, $P$, and the Pareto front is minimized to select a particular Pareto optimal point.

In most cases, including the application presented within the next section, the objective functions are not on the same scale. Hence, an objective that attains greater values than others can inappropriately skew its importance, as well as, the overall distance between each Pareto point and the optimal point. In order to accurately weigh the tradeoffs between the two objectives, the Pareto front and the global optimal point are standardized to produce the same standard deviation. Let $\sigma_N$ and $\sigma_I$ represent the standard deviation of the sets of information distances and dimensions represented by the Pareto front. Then, the rescaled points are $\left( \frac{N(m, \tau)}{\sigma_N}, \frac{I(m, \tau)}{\sigma_I} \right)$ and $\hat{P} = \left( \frac{1}{\sigma_N}, 0 \right)$, respectively.

Mathematically, we can summarize the optimization problem statement as follows. Denote the feasible set by

$$X = \{1, \ldots, 5\} \times [t_1^*, t_2^*]$$

and the set of all possible objective outcomes as

$$Y = \left\{ y \in (0, \infty)^2 : y = (\hat{N}(m, \tau), \hat{I}(m, \tau)) \text{ for some } (m, \tau) \in X \right\}.$$
Then, we define a domination relation on the set $Y$; in particular, we say $y_1 \in Y$ dominates $y_2 \in Y$, written $y_1 \prec y_2$, if $y_1 = (\hat{N}_1, \hat{I}_1)$ and $y_2 = (\hat{N}_2, \hat{I}_2)$ satisfy

$$\hat{N}_1 \leq \hat{N}_2 \quad \text{and} \quad \hat{I}_1 \leq \hat{I}_2.$$ 

With this, the set of all points that dominate a given output $y \in Y$ is

$$D(y) = \{ u \in Y : u \prec y \text{ and } u \neq y \}. $$

Finally, the Pareto front of the set of outcomes is the collection of all objective values that are not dominated by any others, and this is expressed as

$$P(Y) = \{ y \in Y : D(y) = \emptyset \}. $$

After scaling $N$ and $I$, the user can design the distance over which to minimize. There are many options from which to chose, but for the purposes of this study, the Pareto optimal point is found by identifying the point which minimizes the weighted euclidean distance to the global optimal point. More specifically, let $w_N$ and $w_I$ be weights that form a ratio of relative importance $R_w = \frac{w_N}{w_I}$. The optimal point along the Pareto front is then defined to be

$$\left( \hat{N}^*, \hat{I}^* \right) = \arg \min_{(\hat{N}, \hat{I}) \in P(Y)} \left[ w_N \left( \hat{N} - \frac{1}{\sigma_N} \right)^2 + w_I \left( \hat{I} \right)^2 \right]$$

(2.7)

If $R_w < 1$ then the point chosen will be a pair that prioritizes solutions with relatively less information loss in comparison to the reduction of dimension. The opposite is true if $R_w > 1$, so that the final dimension is more important to the optimization than the information lost. The degree to which each objective is favored depends on the value of $R_w$. Of course, the chosen pair $\left( \hat{N}^*, \hat{I}^* \right)$ is directly associated with a pairing $(m^*, \tau^*)$ that represents the number of cluster representatives and correlation threshold, respectively. This pair determines the
Figure 2.2: **Method Schematic**
The five steps of the unsupervised multi-objective, clustering feature selection method: (i) Input data set is cleaned and pre-processed to be spatially static and any transformed variables are appended to this set; (ii) Data set is hierarchically clustered using the correlation details of (2.2); (iii) The Pareto front for the two objectives \((N, I)\) is approximated; (iv) The Pareto optimal point is selected using the weighted optimal distance in equation (2.7) to determine a number of cluster representatives and correlation threshold; (v) The dendrogram is cut at a correlation threshold and subset using cluster representatives.

cut height \(h = \sqrt{1 - |\tau^*|}\), which produces the clustering scheme to employ and the number of cluster representatives maintained for each cluster in the subset with reduced dimension \(N^*\). A visual overview of the method is provided in Figure 2.2.

### 2.4 Results and discussion

#### 2.4.1 Environmental and Demographic Proxy Data Sets for Dengue in Brazil

Big data provides countless observations of environmental, biological, and socioeconomic factors that can be used to improve epidemiological models of disease dynamics [20, 54]. For instance, the mathematical and statistical modelling of mosquito-borne diseases like Zika, Chikungunya, and Dengue Fever has continued to challenge public health officials for a variety of reasons, including a lack of pharmaceutical interventions (e.g., vaccines), the complex interactions between environmental and socio-demographic factors that exacerbate the spread of an outbreak, and our limited understanding of the critical drivers responsible for emerging and re-emerging epidemics. Such models are frequently built to support public health efforts, and this requires them, and consequently their variables and predictors, to maintain a basic level of biological interpretation. Such a constraint renders many existing dimension reduction methods inadequate, necessitating the use of novel tools to reduce the complexity and size of the input variable space while preserving the biological importance.
of the original variables.

We demonstrate the efficacy of the previously described feature selection method on two data sets that were collected for the primary purpose of improving models of the spread of mosquito-borne disease in Brazil, in particular Dengue fever. Of course, the applications of these data sets are not limited to studying the spread of Dengue or even just epidemiological problems. Therefore, this unsupervised feature selection method can not only reduce the dimension of the set while maintaining the interpretation of the measured variables, but also produce clear and concise data sets that are not limited to specific applications that share the same response variable.

The first data set we consider is the 2010 Brazilian census produced by the Instituto Brasileiro de Geografia e Estatística (IBGE) [18], which conducts a census every ten years in Brazil. The census reports measures of more than 200 different demographic characteristics of the Brazilian population. The 2010 survey included detailed information concerning population sizes, income, inequality, fertility, education, employment, economic activity, health, and household infrastructure within each of the 5564 Brazilian municipalities. This spatial data set has a variable dimension of 232 with one observation per municipality over a 10-year period. As the data is produced by an outside source, it is assumed to be clean, and thus does not require further pre-processing prior to applying of the method.

The second set of observations is an environmental data set created by combining remote sensing satellite data and ground weather station measurements from NOAA’s Global Surface Summary of the Day (GSOD) [56]. This data was collected weekly during a seven year period in Brazil from 2010 to 2016. The remote sensing data was obtained from four satellites (LandSats 5, 7, 8 and Sentinel II) via Descarte Labs computing platform [28], allowing our team to access nearly weekly measurements for Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index for vegetation water content (Green NDWI), Shortwave Infrared Normalized Difference Water Index for water content in bodies (SWIR NDWI), Normalized Burn Ratio (NBR), as well as whether the measurement pixel was
obscured by clouds at a 30 m² resolution. For the time periods during which data was missing, kriging methods were used to interpolate the known values [3]. Upon addressing the voids in temporal resolution, the data was aggregated to obtain the minimum, maximum, and mean values for each of the four satellite indices and the percent cloudy pixels for each of Brazil’s municipalities weekly.

The satellite data set was originally gathered to glean information concerning vegetation health, water content, and burned areas using spectral remote sensing. This particular data is crucial to mosquito population dynamics because mosquitoes require healthy vegetation on which to feed, as blood meals from mammals are only taken by females in the process of reproduction [36, 73]. Additionally, healthy vegetation provides crucial information regarding the water dynamics in the area, and when augmented by the measured water indices, can serve as a proxy for mosquito breeding dynamics. Mosquitoes depend on water to both lay their eggs and, until adulthood, undergo their immature life cycle [36, 73]. Finally, the regions in which neither healthy vegetation nor burned areas exist may indicate urban populations, which not only affect mosquito reproduction, as blood meals are needed for procreation, but also indicate where vector-borne disease will spread.

In addition to the satellite data, NOAA’s GSOD data set provided us with all available daily measurements concerning temperature and humidity from the 613 weather stations in Brazil that provide daily climatological data such as temperature (mean, min, max), precipitation, and humidity. Despite the richness of this set, errors and missing data were common, and with 5564 separate municipalities in Brazil, we were unable to obtain precise measurements for each municipality. Again, kriging methods were employed both spatially and temporally to best approximate the daily mean, minimum, and maximum temperature, as well as the relative humidity. The results were then aggregated temporally to produce weekly temperature and humidity measurements.

Finally, the satellite and weather station data streams were combined into a single spatio-temporal environmental data set, which measured each of the 22 environmental variables
Table 2.1: Summary of Data Set Properties
Number of variables, represented spatial locations, and temporal observations for both the demographic and environmental data sets at each state of processing prior to applying feature selection. Note that each of the dimension measurements which includes a “*” is a quasi-spatial data set, in the sense that each of the 7 temporal observations are treated as new spatial locations rather than linked to their other temporal observations.

<table>
<thead>
<tr>
<th></th>
<th># Variables</th>
<th>Spatial Locations</th>
<th>Temporal Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw Input</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographic</td>
<td>232</td>
<td>5564</td>
<td>1</td>
</tr>
<tr>
<td>Environmental</td>
<td>22</td>
<td>5564</td>
<td>364</td>
</tr>
<tr>
<td><strong>Pre-Processed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographic</td>
<td>235</td>
<td>5564</td>
<td>1</td>
</tr>
<tr>
<td>Environmental</td>
<td>770</td>
<td>38964*</td>
<td>1*</td>
</tr>
<tr>
<td><strong>Transformed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographic</td>
<td>470</td>
<td>5564</td>
<td>1</td>
</tr>
<tr>
<td>Environmental</td>
<td>1540</td>
<td>38964*</td>
<td>1*</td>
</tr>
</tbody>
</table>

weekly for each municipality. In order to apply the dimension reduction method, as it has not yet been extended for spatio-temporal data streams, we generated seven yearly summaries of the environmental data. In particular, a five number summary of each municipality level variable (mean, max, min, range, and standard deviation over all municipalities) was calculated over varying time periods of observations; namely, the full year, the four seasons, and the months of July and January - the two months in the middle of winter and summer, respectively. Thus, for each original spatio-temporal variable we generated 35 summary variables per year. This massively expanded the dimension of our data set resulting in a set with a variable dimension of 770 with seven yearly observations per municipality.

2.4.2 Application of the Unsupervised Feature Selection Method

With both data sets cleaned and pre-processed, feature selection can begin. While the demographic and environmental data sets could be combined, which would synthesize the shared information across the two sources of data, we processed the two sets separately. This choice was made for two reasons. Primarily, these sets possess different inherent dimensions despite the aggregation of the environmental data. The census is a static measure
that contains one temporal measurement, while the environmental data represents seven measurements over time. This difference results in a fundamentally different spatial dimension in our method, as we consider each municipality in each year as a different spatial location, resulting in an environmental data spatial dimension seven times larger than the demographic data. This allows us to evaluate the temporal effects of the data over the seven year period. In theory, prior to the aggregation performed as part of the pre-processing, each weekly measurement could be treated in the same way by generating another quasi-spatial data set. The second justification for maintaining the data sets separately is that the reduced sets produced by our method will still be independent of model supervision, and it is straightforward to envision an application of the census data set for which the environmental data is not utilized and vice versa. Therefore, by processing the sets independently the output is more flexible beyond the current scope of investigation.

The first step of the reduction method entails the application of a log transform (2.1) which doubles the dimension of both environmental and demographic data sets, to 470 and
1540, respectively. A logarithmic transform is a common transform, but other transforms could be applied instead or in addition. A summary of the dimension of each of the data sets as raw input, after pre-processing, and post transformation is displayed within Table 2.1.

At this stage of the method, the dimension of the variable space is at its greatest, and subsequent steps will serve to reduce this quantity through our feature selection method. This process begins by calculating the correlation matrix for each data set. In Figure 2.3 the absolute value of the correlation between pairs of variables is plotted as a heat map, as the magnitude of the correlation, rather than the sign, determines the distance between two variables. With this simple visualization it is clear that the two data sets possess very different correlation structure. The demographic data features multiple large clusters of highly correlated variables, visible in the dark blue triangles in the center of the heat map. These variables are all population counts in one such cluster, such as counts of Urban Population, Female/Male, Over 18, and overall population, which intuitively should be highly correlated and likely contain redundant information. Other than these large groups, the demographic variables are primarily highly-correlated (|ρ| > 0.9) with a small number of other factors and then correlate moderately (0.5 < |ρ| < 0.9) with additional variables. This correlation structure will produce many small clusters for high thresholds and a few looser, but still moderately well-correlated, clusters at lower thresholds. This demonstrates that many of the variables in the demographic data are strongly related and are often measures of similar phenomena observed in different contexts. In contrast, while the features of the environmental data are less correlated overall, they are more highly correlated within smaller cohorts. This is visually apparent in the overall lighter color of the correlation map and the lack of larger distinct clusters. The variegated triangle in the center of the plot, where the correlation (in blue) weakens as the number of correlation values increases, mimics the type of correlation structure that characterizes most of the demographic data. Moderately correlated variables link together in loose clusters signifying some overlapping information,
Figure 2.4: Dendrogram Representation of Hierarchical Clustering

The dendrogram illustrates the hierarchical clustering of the data into various clusters. The height of each node in the tree diagram represents the distance at which two clusters are merged. The color of the boxes and the height at which the clusters are merged are used to represent the distance between clusters. The dendrogram is a visual representation of the hierarchical clustering process, showing the merging of clusters at different levels of dissimilarity.
but not complete redundancy. Besides this loose structure, most of the variables reside
in small classes of highly correlated variables but do not correlate strongly with any other
groups. This indicates that there are distinct categories of features that measure highly
related processes, but most groups measure generally unrelated effects.

These observations implied by the correlation structures are confirmed when both data
sets, using the correlation distance metric (2.2), are hierarchically clustered according to
the single linkage method. The clustering output is a set of nested clustering structures
that begins with each variable as a singleton and progresses by combining clusters until all
variables are grouped within a single cluster. All possible clustering schemes for each data set
are displayed in the dendrograms in Figure 2.4. A particular cluster structure is generated
by a cut height. The height \( h = \sqrt{1-|\rho|} \) at which two clusters are combined is synonymous
with the minimal correlation value for any two members of distinct clusters that are then
combined. The single linkage algorithm guarantees that no two variables with a correlation
value above a chosen threshold will be classified in different clusters when the dendrogram is
cut at the corresponding height. As previously noted, other linkage methods do not preserve
this property and so were not utilized within the algorithm.

At this stage, with 1540 environmental and 470 demographic clustering structures from
which to choose, the clustering results, the variables, and the correlation distance are all used
in the multi-objective optimization to generate a set of optimal correlation thresholds \( \tau \) and
number of variables \( m \) to retain from each cluster produced by the cut height. The two ob-
jective functions to be simultaneously minimized are the reduced subset dimension \( N(m, \tau) \)
and the information loss \( I(m, \tau) \) stemming from this reduction. The quantity \( N(m, \tau) \) is
simple to calculate; it’s monotonically increasing as a function of the correlation threshold
(i.e., it decreases when the threshold decreases) and monotonically increasing in the number
of permitted cluster representatives. On the other hand, \( I(m, \tau) \) is a more complex func-
tional objective. In Figure 2.5 the information distance function is plotted for a range of
correlation values and up to five cluster representatives for each data set. For the census
Figure 2.5: **Information Distance**
Visualization of the objective function $I(m, \tau)$, which calculates the information lost due to the subsetting via clustering, for both data sets within the range $\tau \in (0, 1)$ and $m \in \{1, \ldots, 5\}$. Instead of raw information loss, the percentage of maximum information loss is plotted to provide a comparable scale between the two data sets.
data, we see that information loss increases monotonically as the threshold decreases, with steep jumps in loss at particular correlation values. This monotonicity is not a guaranteed property of the method, as seen in the plot for the environmental data set. There, a spike in the information loss occurs when only one cluster representative is permitted at a correlation threshold around $\tau = 0.8$, which is much greater than any other value in the range of correlation thresholds. The impact of this information loss is mitigated by allowing an additional cluster representative and eliminated entirely once four cluster representatives are permitted. This dynamic motivated the generalization of the method to include multiple cluster representatives. In particular, a single representative structure demonstrated that the medoid can be biased, at times significantly, to a particular part of a cluster whenever clusters are combined, and thus the remaining variables within the cluster are not well represented. By permitting multiple cluster representatives, additional variables can be utilized to preserve the information contained in other regions of a cluster.

### 2.4.3 Convergence of the Algorithm

The Pareto front describing the optimal pairs of each objective can be approximated via the NSGA-II algorithm. Both decision space variables - a continuous range of correlation thresholds and a discrete range of permissible number of cluster representatives - are used to define the multi-objective problem within the algorithm. As this is akin to a genetic algorithm, the population size (in this case, the number of points along the front that we wish to approximate), the number of generations the algorithm will perform, the crossover probability, and the mutation probabilities are all parameters that must be determined to generate output. Foremost in consideration when using this algorithm is the convergence of the approximation to a stable set of points along the Pareto Front. As NSGA II is designed to rank population members by both Pareto dominance and the overall diversity of the population, the convergence of this Multi-Objective Evolutionary Algorithm (MOEA) is not as straightforward as an optimization algorithm that only ranks according to Pareto dominance. The generated set of points will not necessarily converge to a unique set of points.
over the course of the generational algorithm. Instead, from one generation to the next the front will shift to accommodate the diversity constraint, thereby dispersing points to produce the most representative approximation of the front. This dynamic is demonstrated for both data sets in Figure 2.6.

In order to guarantee a stable output of the dimension reduction method (i.e., for a given paradigm under which a point along the Pareto Front Approximation is chosen to be optimal), it is crucial to know which algorithm parameters produce fronts that are stable enough to ensure convergence of the optimal point chosen.

Convergence is explored along a number of experimental dimensions. First, the convergence of a single run is evaluated based on the $\ell^2$ (or sum of squares) difference for points along the front between each generation of the algorithm. We generated runs of various generations $N_{\text{gen}}$ and population sizes $N_{\text{pop}}$, with default mutation and crossover probabilities (0.2 and 0.7 respectively) in order to identify the combinations of generation and population number that produce adequately stable fronts. As greater values of generation and population size increase the computational cost of the method, it is useful to examine the range of values of these quantities that guarantee algorithmic stability. To calculate this set difference, the members are ordered based on the two objectives so that the $n^{th}$ point in one generation is compared to the corresponding $n^{th}$ point in the next. Then, the $\ell^2$ difference between the $i$th and $(i+1)$st generation is defined to be

$$\mathcal{L}_1^i = \frac{1}{N_{\text{pop}}} \sum_{k=1}^{N_{\text{pop}}} \left( |I_k^i - I_k^{i+1}|^2 + |N_k^i - N_k^{i+1}|^2 \right).$$  \hspace{1cm} (2.8)

In Figure 2.7 the difference between subsequent generations of the evolutionary algorithm can be seen for range of total generations and total population members. We note that the $\mathcal{L}_1^i$ difference is a sum of relative errors from the two different objectives. In general, a lesser $\mathcal{L}_1^i$ difference indicates little to no change between consecutive generations. However, merely producing a small $\mathcal{L}_1^i$ difference between two generations, even the last two generations, is not sufficient to demonstrate adequate stability. Instead, we calculate the maximum $\mathcal{L}_1^i$
Figure 2.6: Evolution of Pareto Fronts
The evolution of a simulated Pareto front with 24 population members is displayed in the \((N, I)\) plane for both the Demographic (top) and Environmental (bottom) data. The first five generations of the algorithm are shown for comparison with the final (750th) generation.
Figure 2.7: $L_i^1$ Difference between Fronts

The $L_i^1$ difference between the $i$th and $(i+1)$st generation of a simulated Pareto front is presented with three different population sizes (24, 96, 768) and three generation numbers (250, 500, 1000) for both data sets. The $L_i^1$ difference between a particular generation of the front and all subsequent fronts, namely

$$L_i^2 = \frac{1}{N_{\text{pop}}} \max_{j > i} \sqrt{\sum_{k=1}^{N_{\text{pop}}} (|I_k^i - I_k^j|^2 + |N_k^i - N_k^j|^2)}$$

(2.9)

in order to demonstrate convergence. In Figure 2.8 the maximal differences of every generation of the front are plotted for each execution of the algorithm with various population sizes and generation numbers. In all simulations it is typical to see large values of the $L_i^2$ difference for early generations of each population. However, for some simulations the trend of the maximal differences converges nearly to zero, while others demonstrate an inconsistent and relatively large $L_i^2$ difference across all generations. These latter simulations are considered to be unstable, as they demonstrate that one particular generation of the Pareto front is significantly different than fronts that are generated in subsequent rounds of the algorithm. In contrast, those runs that maintain small differences after an initial burn-in period demonstrate that the subsequent set of Pareto fronts are highly similar. When the
Figure 2.8: $L^2_i$ Difference between Fronts
The $L^2_i$ maximal difference between the $i$th and all subsequent generations of a simulated Pareto front is shown with three different population sizes (24, 96, 768) and three generation numbers (250, 500, 1000) for both data sets.

maximal differences are close to zero, fronts themselves may not be identical, but the differences among the set are minimal, demonstrating stability, if not convergence to a distinct front.

Through this analysis it is clear that population size, rather than the number of generations, is the driving mechanism for stability of the evolutionary algorithm. For a particular application, a minimal difference threshold could be implemented to determine the level of stability required; however, in this application the stability of the front itself is secondary to the convergence of the optimal point, which determines the dimension reduction. A visual representation of how various optimal points converge for a given simulation can be found in Figure 2.9 while the standard $\ell^2$ errors for those points in Figure 2.11. It is clear that for smaller population sizes the optimal point is much more unstable throughout the run whereas as the population increases the optimal point stability increases. To identify runs with sufficient convergence for the optimal points we can repeat the above analysis, but
rather than calculating the difference between two Pareto fronts, we calculate the difference between the two optimal points of different generations. The corresponding results are demonstrated in Figure 2.10 for the Pareto optimal point with greatest dimension reduction of the five weighting schemes considered where \( w_I : w_N = 9 : 1 \) in equation (2.7). This process can be repeated for other weighting schemes as well. Generally, the maximal difference between two optimal points is less than the maximal average difference between two fronts, suggesting that the extremities of a front are the least stable, while the interior points are more stable. At the tails of some of these simulations there are a series of generations which have converged to a unique optimal point at a finite generation number, which we deem “finite convergence”. Typically, we see that the longer the series, the more stable the finite convergence result. This plot also demonstrates that finite convergence requires a large population size - 768 for the census data and the environmental data for all of the optimal points except the 3 : 1 optimal point for the environmental data where finite convergence has not yet been achieved.

If finite convergence is not required or is not produced due to limited computational resources, one can instead evaluate the convergence of the algorithm by establishing thresholds for the tails of each simulation. Figure 2.8 shows that the error between all simulations ranges between 0 and 1, and if we impose an arbitrarily small threshold \( \tau \in (0, 1) \), it will bound the error contributed from each of the components to be less than \( \tau \). To demonstrate the scale of a given threshold in an interpretable way, rather than using the \( L_2 \) difference, we will decompose the total distance into the part contributed by the change in information distance and that of the final dimension change between optimal points. Each component is then converted to a percentage by multiplying by the scale factor used to determine the optimal point and dividing by the maximal value within each data set. These values are given in Table 2.2.

With this rescaling we can plot both components of the maximal generational difference between optimal points on a scale of \((0, 1)\). Figure 2.12 contains a selection of the exper-
Figure 2.9: Optimal Point Convergence
The location of optimal points selected via the five demonstrated weighting schemes is plotted in the \((N, I)\) plane with 100 generations and three population sizes \((24, 96, 768)\) for both data sets. The optimal point from the final generation is outlined in gray while the previous generations are plotted with transparency that corresponds to their generation number. More transparent points are produced earlier in the algorithm while less transparent points are nearer to the end. Additionally, the grey dashed line represents the final Pareto front of each simulation.
Figure 2.10: **Standard $\ell^2$ Error for Optimal Point 1 : 9**
The standard Euclidean error of the optimal point selected via the 1 : 9 weighting scheme in generation $i$ is compared to the optimal point selected in the final generation with three different population sizes (24, 96, 768) and three generation numbers (250, 500, 1000) for both data sets.

Table 2.2: **Maximal Values & Scale Parameters**
The maximal values and scale parameters used in the evaluation of the method and the selection of Pareto optimal points.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Maximal Value</th>
<th>Scale Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$I$</td>
<td>$N$</td>
</tr>
<tr>
<td>Environmental</td>
<td>352.02</td>
<td>1540</td>
</tr>
<tr>
<td>Demographics</td>
<td>298.38</td>
<td>470</td>
</tr>
</tbody>
</table>

Experimental runs for each data set and demonstrates the different behaviors exhibited. This plot shows that the component contributing most to the difference between points is the dimension rather than the information distance. In this way, evaluating convergence primarily along the reduced dimension variable ensures that the information distance has also converged to a value below the required threshold. For cases in which finite convergence occurs, both components of the error tend to 0% rapidly. For cases in which the method does not converge, the percent error continues to oscillate throughout the run on the order of 5% – 40% of the maximal value. For the census data this is on the order of an error of
Figure 2.11: **Standard $\ell^2$ Error for all Optimal Points**

The standard Euclidean error of the optimal point in generation $i$ for each of the five weighting schemes is compared to the optimal point selected by the corresponding weighting schemes in the 1000th generation with three different large population sizes (192, 384, 768) for both data sets.

23 – 188 in the dimension of the resulting variable space, while for the environmental data it constitutes a massive dimensional error on the order of 77 – 616 variables. This result further affirms the previous claim that such simulations can be unstable. Between these two types of behaviors (i.e., convergence and instability), there are simulations that stabilize to a low threshold of error at their tails. While convergence is not finite in these simulations, the percent change between optimal points of different generations after a certain burn-in period can be bounded by a chosen threshold. Of course, the burn-in period can vary, but as the tails are the critical aspect of these runs, the thresholding should be applied only to the final portions of the simulations. In this way, it would be feasible to evaluate the last 10 – 30% of a simulation to determine if these values lie within a given threshold. As the number of tail generations evaluated increases, the likelihood of the optimal point changing within the threshold also increases. Thus, depending on the application of a given data set,
approximate convergence may be acceptable in lieu of finite convergence.

### 2.4.4 Weighting and Selection of Pareto Optimal Point

Once a simulation is determined to have converged adequately, the optimal point can now be used to generate a lower dimensional data set. Here, we have chosen finite convergence of optimal points as the measure of algorithmic success, though a suitable threshold can be utilized, as discussed above. The Pareto front for both data sets which was simulated with 768 population members and 1000 generations is finitely convergent for all weighting schemes except 3 : 1 in both data sets and 9 : 1 in the environmental data. If either of these particular weighting schemes were a primary interest further simulations would need to be run for finite convergence. However in this case we have five optimal points to consider for both data sets and the simulation with 768 members and 1000 generations represents the lowest error for the 3 : 1 and 9 : 1 weighting scheme of all simulations run. Thus we will consider this simulation as the gold standard in further analysis. The Pareto front of the two objectives $N$ and $I$ generated by this simulation for both data sets can be seen in Figure 2.13. Each optimal point is highlighted and the dynamics of the different weighting schemes are clear. Results of finitely convergent simulations are summarized in Table 2.3. They clearly show that for both data sets and the demonstrated weighting schemes, optimal points are produced by moderately high correlation thresholds, e.g. $\tau > 0.65$. Therefore, small changes in correlation thresholds generate vastly different numbers of clusters while maintaining a low level of information loss. Across all optimal point weighting schemes the information loss remains below 12% of the maximal value, and the dimension reduction ranges from 13.5% to 68.9% of the input (maximal) dimension in the environmental data and 10.6% to 42.6% in the demographic data. Most of the weighting schemes use only one cluster representative from each cluster to generate the final subset, demonstrating that in these eight cases of optimal point selection, the option for additional cluster representatives was not necessary to achieve a substantial dimension reduction. Of course, this is not universally guaranteed, as demonstrated by the two extreme examples of weighting schemes. For example, the
Figure 2.12: Component Error for Optimal Point 1 : 9
The error contributed by the two objectives (N, I) is plotted separately for simulations with 1000 generations and four different population sizes (12, 48, 192, 768) for both data sets. Each component error is scaled to a percent of the maximal value each objective could attain, and as a result, each row is plotted on a different scale.
Figure 2.13: Selected Pareto Fronts
The most stable Pareto front for both data sets was generated with 768 population members and 1000 generations and is shown on the \((N, I)\) plane with each optimal point highlighted. This illustrative simulation was selected due to the finite convergence of all optimal points except the points selected by the 3 : 1 and 9 : 1 weighting schemes which converged within an acceptable threshold.

A weighting scheme that satisfies \(w_I : w_N = 1 : 9\) gives rise to demographic data optimal objectives and parameters of \((N, I) = (50, 23.9)\) and \((\tau, m) = (0.791, 1)\), thereby achieving a large degree of dimension reduction with only one representative variable. Contrastingly, for the same weighting scheme the environmental data optimal objectives and parameters are \((N, I) = (208, 40.8)\) and \((\tau, m) = (0.692, 5)\) and produce a considerable dimension reduction using four representative variables. This behavior could be predicted by the information distance objective (Figure 2.5) where additional representative variables significantly reduce the information loss for correlation values in the range \((0.65, 0.80)\).

This method was designed to not only reduce dimension but also preserve variable interpretation; hence, further analysis is needed to evaluate its utility in the latter context. With a chosen optimal point, this can be performed by examining the structure, membership, and
variable interpretation of clusters. We note that each of the weighting schemes explored in this study produces different clustering structures. Across all optimal point selections the method consistently produces many small clusters and a few large clusters, though as the correlation threshold lowers, additional clusters of large size form and the number of smaller clusters decreases. Small clusters in this context tend to consist of 1, 2, or 4 variables. Subsequent analysis demonstrates that, with minor exception, two-variable clusters are composed of an original variable and its log transform. Hence, in such clusters the transform did not contribute additional information to the data set. That being said, even if all variables across a given clustering scheme experience this phenomenon, the log transform can still serve a purpose when the best cluster representative is determined to be a transformed variable rather than a raw version. However, if there is no instance of a transformed variable clustering separately from its source variable, especially if the clustering threshold is high, then one can conclude that the transformation was superfluous. In such a scenario, the analysis could be repeated without it, at a lower computational cost. Note, however, that for most of the optimal point schemes evaluated, this phenomenon does not occur for either the demographic or environmental data set.

At this stage the choice of which point along the Pareto Front will be used to cluster and produce dimension reduction is user determined. For this study we used a minimal weighted distance metric to determine candidate optimal points, and further analysis can be done on any of the five points for each data set. The primary consideration in the selection of the optimal point is naturally how each point satisfies the objective of minimizing dimension and information loss. Recall that the two objectives are inherently in conflict with each other, a loss of dimension will likely increase information loss. However the scale of the trade-off between dimension and information loss may be unbalanced in cases with significant redundancy of variables in the data set. Both data sets considered in this study have this property as demonstrated by the five different optimal points under consideration. In all demonstrated cases the percent of maximal information loss is always lower than the
Table 2.3: Convergent Simulations

The results of the optimal point choice using each optimal point weighting scheme for both data sets. Both the output in the objective space and the input space, over which the dimension reduction method is implemented are displayed. The percentages are calculated as percent of maximal for both objectives as provided in Table 2.2. The simulation used had a population size of 768 and 1000 generations.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>$w_I : w_N$</th>
<th>Objectives</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$N(\tau, m)$</td>
<td>$I(\tau, m)$</td>
</tr>
<tr>
<td>Environmental</td>
<td>9:1</td>
<td>1061 (68.9%)</td>
<td>8.13 (2.31%)</td>
</tr>
<tr>
<td></td>
<td>3:1</td>
<td>696 (45.2%)</td>
<td>15.0 (4.26%)</td>
</tr>
<tr>
<td></td>
<td>1:1</td>
<td>518 (33.6%)</td>
<td>21.8 (6.19%)</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>397 (25.8%)</td>
<td>28.5 (8.10%)</td>
</tr>
<tr>
<td></td>
<td>1:9</td>
<td>208 (13.5%)</td>
<td>40.8 (11.6%)</td>
</tr>
<tr>
<td>Demographic</td>
<td>9:1</td>
<td>200 (42.6%)</td>
<td>8.01 (2.68%)</td>
</tr>
<tr>
<td></td>
<td>3:1</td>
<td>165 (35.1%)</td>
<td>10.7 (3.59%)</td>
</tr>
<tr>
<td></td>
<td>1:1</td>
<td>110 (23.4%)</td>
<td>15.0 (5.03%)</td>
</tr>
<tr>
<td></td>
<td>1:3</td>
<td>78 (16.6%)</td>
<td>19.3 (6.47%)</td>
</tr>
<tr>
<td></td>
<td>1:9</td>
<td>50 (10.6%)</td>
<td>23.9 (8.01%)</td>
</tr>
</tbody>
</table>

percentage of maximal dimension. While information loss increases as dimension decreases, as expected, it increases from such a low level that even in the 1 : 9 weighting scheme both objectives are minimized to the same percent of maximal. Redundancy is implied by this dynamic, since a 10% reduction in dimension does not equate to a 10% increase in information loss. For these two data sets it is clear that we can achieve substantial dimension reduction with information loss that is relatively low cost.

Large clusters of variables present a significant opportunity to evaluate the efficacy and utility of the proposed multi-objective feature selection method. These clusters represent the greatest source of dimension reduction and can reveal the inherent structure of the data. The number of large clusters increases as the correlation threshold is lowered; thus, as the dimension decreases, the information loss increases. However, even with a weighting scheme of 1 : 9 the correlation threshold of 0.69 represents only 11.6% of maximal information information loss for the environmental data set. Similarly the same weighting scheme and a correlation of 0.79 produces 8.01% information loss in the demographic data. In both
cases, the information loss is minimal. Additionally, when the results are calibrated to a typical statistical correlation analysis, two variables with a correlation greater than 0.5 are considered to possess a moderate linear relationship, and a correlation of two variables greater than 0.7 is considered a substantial linear relationship. As such, our method typically removes variables that are highly correlated, and always at least moderately correlated, with another variable in the data set. Therefore, the resulting information loss can be interpreted to coincide with the elimination of the signal contained in a nonlinear deviation between two variables that are moderately to highly correlated. While this information loss is not negligible, it is clearly both quantifiable and manageable.

As a result in subsequent analysis will continue on the optimal point which minimizes the 1 : 9 weighted distance to the theoretical optimal. This point represents the largest dimension reduction conditions of the five weighting schemes considered. This point was selected for further analysis because the primary objective is dimension reduction and reducing the redundancy or co-linearity between variables. Additionally, this choice produced a variety of cluster sizes through the hierarchical clustering process as well as, in the case of the environmental data, employed multiple representative factors, allowing for a robust analysis of this method using one optimal point scheme as an exemplar. Proceeding with this selection reduced the environmental data set from 1540 to 208 variables and reduced the demographic data from 470 to 50 variables.

### 2.4.5 Interpretation of Pareto Optimal Point through Text Analysis

Other than the numerical evaluation of a clustering scheme, the utility of variable and cluster interpretation can be explored via text analysis on variable names or descriptions. The interpretation of the cluster then can be explicitly or implicitly linked to the representative variables of each cluster. An explicit link may occur in the form of a variable description that details the scope of the types of measures in the cluster, while an implicit link can aid in reframing a specific measured variable as a representative measurement of a larger phenomenon. This analysis depends on the information at the analyst’s disposal, often
dictated by the method of data gathering, cleaning, and processing. In the case of the
demographic data, the data is gathered and processed by the source institution, so without
significant effort the analysis is limited to information generated from the raw data. The
associated variable names are largely devoid of meaningful description, as they are named
for brevity. Instead, the data is accompanied by a source of metadata, which augments
each variable name with both a short and long description of the quantity it measures.
These descriptions are in Portuguese, but using the basic translation capabilities of Google
translate, linked with Excel, English descriptions can be evaluated using various text analysis
strategies on an English lexicon. Clearly, translation errors are to be expected, but for
this purpose a general understanding is sufficient. If this was not the case for a specific
application, then additional effort, such as human translation or Portuguese text analysis
could be employed. In contrast, the environmental data was processed primarily by our team,
and as a result, interpretation of the variable names themselves are a more straightforward
description of the signal they measure. Each variable is named for five elements, (1) the
measured data type, (2) the spatial aggregation summary type to generate municipality
level data, (3) the summary time period, (4) the temporal aggregation summary type over
said period, and (5) whether the variable is a log transformation of a raw variable. These five
elements are ordered and evaluated based on a lexicon generated by the range of variables
in the data set.

From either lexicon, a series of basic text analysis techniques can be applied to parse
variable and cluster interpretations. The first step in any textual analysis is to establish a
corpus, or a collection of character strings that are often annotated by additional metadata [6,
62]. In the present context, the meta data generally consists of variable names and cluster
memberships. This is analogous to a corpus generated by a collection of books, where each
line is a character string tagged by the chapter and book in which it resides. The demographic
data corpus is generated by a basic variable definition translated from Portuguese that has
been cleaned for basic translation errors. The corpus for the environmental data is generated
by parsing the five parts of the variable name into a single string for each variable and
applying further cleaning to distinguish among the five elements. After a corpus has been
determined, it is tokenized, meaning each string is divided into single tokens, often words
or numbers, while maintaining the metadata structure linking each token to the variable
and cluster to which it belongs. Tokenization allows for further parsing, particularly in the
demographic data corpus, as a single token determines a number of different properties [62].
The most basic of these properties is whether or not the token is a stop word, that is, a word
that the user knows is not meaningful. This can be a list the user creates or an accepted
list of common English words such as “the”, “and”, and “a”. These tokens can be filtered
and removed from consideration to streamline the textual analysis. After basic filtration
has been performed, the tokenized environmental corpus is ready to be deployed, but the
tokenized demographic corpus requires further cleaning to extract meaningful information.
Because of the complexity of the English language, different tokens can possess the same
or similar meanings but represent different words due to their different inflections. The
simplest example of this is the difference between the singular and plural forms of a noun
like “year”. Currently “year” and “years” are considered to be different tokens. To mitigate
this dynamic we use a technique called lemmatization, which groups together all inflected
forms of a particular word and assigns a single lemma to the group of tokens [62]. This
can be implemented before or after the stop word filtration. An additional step within this
process, which finds common synonyms for groups of tokens with similar meaning, was not
applied in this analysis as many of the words in our corpus are quite technical. Overall,
the demographic corpus contained 247 unique tokens, which was reduced to 228 through
lemmatization, and further decreased to 193 tokens by filtering standard English stop words
and user defined stop words. The environmental corpus produced 28 tokens considering no
stopwords or lemmatization was required.

To begin the text analysis, the frequency of a term must be defined to simultaneously
consider its appearance both within clusters and among them. This is accomplished by
considering both the single-cluster frequency (TF) and inverse document frequency (IDF) of each term [62]. The former is given by

$$\text{TF}(x) = \frac{n(x)}{T}$$

(2.10)

where \( n \) is the number of occurrences of a term in the cluster and \( T \) represents the total number of terms within the cluster. Contrastingly, the IDF is a heuristic that is defined via the conventional wisdom that if a term is common across all specified clusters, then it is not unique to a particular cluster. Conversely, if a term appears in only one cluster, it can distinguish that particular subset from the rest. The IDF for a given term \( x \) is calculated as

$$\text{IDF}(x) = \log \left( \frac{C_T}{C(x)} \right)$$

(2.11)

where \( C_T \) is the total number of clusters, and \( C(x) \) is the number of clusters containing the term \( x \). Finally, a measure of frequency over the entire variable set is defined as the product of term frequency and inverse document frequency, given by

$$\text{TF-IDF}(x) = \text{TF}(x) \cdot \text{IDF}(x).$$

(2.12)

The combined measure TF-IDF quantifies the tokens which are simultaneously commonplace within a cluster but rare outside of it [62]. In this way, each cluster can be characterized by these terms. In Figure 2.14 the tokens with the highest TF-IDF are plotted for a selection of the largest clusters produced after dimension reduction has been performed for both data sets using the optimal Pareto point selected by the weighting scheme 9 : 1.

While the category of tokens that arise in each dataset differs, the analysis of TF-IDF ranking of the tokens in large clusters can elucidate the effects measured by these variables. For instance, the largest cluster is characterized by measurements of income, poverty, and inequality as seen in the top tokens “income”, “capita”, and “poor”. Other tokens which may have little to no meaning on their own can, through familiarity with the dataset, provide further context. Words linked to educational measures, such as “primary” and “complete”, correspond to greater TF-IDF values in this large cluster because educational inequalities
Figure 2.14: TF-IDF Ranking of Tokens
The TF-IDF ranking of the top 25 tokens in the four largest clusters in the Demographic data and the top 15 in the six largest clusters in the Environmental data are plotted in rates of education completion and age delays in starting primary education are linked to inequality measures. This is in contrast to the next largest cluster in which the broad education trends that appear are clustered separately. The final two large clusters are a case study in the impact that the transform can have on a data set. These two clusters are identical, both measuring population counts of various subgroups, with the exception that
Table 2.4: Binary Correlation Calculation Table

For each Token Pair a value of $n_{**}$ is needed to calculate the binary correlation.

<table>
<thead>
<tr>
<th>Has Token Y</th>
<th>Has Token X</th>
<th>No Token X</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has Token Y</td>
<td>$n_{11}$</td>
<td>$n_{01}$</td>
<td>$n_{1*}$</td>
</tr>
<tr>
<td>No Token Y</td>
<td>$n_{10}$</td>
<td>$n_{00}$</td>
<td>$n_{0*}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{1*}$</td>
<td>$n_{0*}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

one is a logarithmic transformation of the other. This is shown at the bottom of the bar chart where the tokens of the two clusters are identical until the last one, in which the token “logarithmic” appears. The relatively small value of TF-IDF for “logarithmic” is a result of its appearance in all clusters except the raw population counts cluster. Of course, if this terms appeared within every cluster then TF-IDF(“Logarithmic”) would be identically zero.

As the environmental dataset is more than four times greater than the demographic dataset, it possesses a greater number of large clusters. The tokens with the greatest TF-IDF values for the six largest clusters are plotted in Figure 2.14. All of these clusters contain more than 50 variables. Notice that tokens with greatest TF-IDF are typically measured data types, like “Temp”, “SWIR-NDWI” and “NBR”, or spatial summarization indicators, such as “Var Mean”, “Var Min”, and “Var Max”. Hence, we can conclude that multiple different temporal summaries of a single spatially summarized variable tend to cluster together. Therefore, most clusters tend to be characterized by a single or pair of spatially summarized variables. Overall, the six large clusters consist of a single temperature cluster, a single NBR cluster, and two clusters measuring different spatial summaries of SWIR-NDWI and the combination of Green NDWI and NDVI. The analysis of large clusters can highlight broad phenomena that are measured by many variables, such as the large income inequality cluster, which consists of 198 variables that can be represented by a single variable, or the many clusters which group GREEN NDWI and NDVI together confirming in the data a relationship between vegetation heath (NDVI) and water content in plants (GREEN-NDWI) that is stronger than other remote sensing measures.

51
Ranking of TF-IDF also facilitates individual cluster analysis by analyzing token frequency within a cluster, while devaluing tokens that are common across multiple clusters. While there is some consideration of the partitioned data set as a whole, the relationships between tokens across the entire data set are still unclear. To address this we calculate the binary correlation for each pair of tokens $X$ and $Y$, which is given by

$$\phi(X, Y) = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{11}n_{01}n_{00}n_{00}}}$$

(2.13)

where $n_{**}$ is the count of the clusters which satisfy the properties displayed within Table 2.4 for token $X$ and $Y$ [6]. Plots of binary correlation matrices for both data sets can be found in Appendix A (Figure A.2, Figure A.1). For clarity, the correlation for a selection of token pairs from each data set is displayed in Table 2.5. While each data set contains many more token pairs than displayed, Table 2.5 represents pairs with large absolute binary correlation values. Token pairs with $\phi = 1$, indicating that they are always found in the same clusters of variables, appear within both data sets. For instance, variables that measure the Relative Humidity and Temperature are always in the same cluster of the environmental data, likely the large temperature cluster revealed by the TF-IDF analysis. The demographic data reveals that “capita” and “poor”, as well as “male” and “female”, are always clustered together. Other than those values identically equal to 1, many token pairs exhibit large, positive correlation values, indicating a propensity to appear within the same cluster. In the environmental data set, the only remote sensing variables with a positive word correlation are those which measure Green NDWI and NDVI. The tokens SWIR NDWI, NBR, and Percent Cloudy Pixels are all negatively correlated with one another, in addition to Green NDWI and NDVI. This indicates that measurements of Green NDWI and NDVI tend to reside in the same clusters, and so capture similar information, while the other remote sensing measures tend not to be clustered together, indicating that they measure predominately independent phenomenon. This could be hypothesized from the individual cluster analysis using TF-IDF, but without considering every cluster, the exact relationship would remain
Table 2.5: **Binary Correlation Values for Selected Token Pairs**

A small selection of token pairs are selected to demonstrate the utility of a binary correlation measure on the textual analysis of variable names or descriptions that have been clustered. In this case, the selected pairs possess some of the greatest absolute binary correlations in the set. Some tokens have been disambiguated from their shorthand version.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Token 1</th>
<th>Token 2</th>
<th>$\phi$(Token1, Token2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Environmental</strong></td>
<td>Relative Humidity</td>
<td>Temperature</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Range (time)</td>
<td>SD (time)</td>
<td>0.962</td>
</tr>
<tr>
<td></td>
<td>Temperature</td>
<td>Daily Range Temperature</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>Relative Humidity</td>
<td>Daily Range Temperature</td>
<td>0.702</td>
</tr>
<tr>
<td></td>
<td>Mean (time)</td>
<td>SD (time)</td>
<td>0.577</td>
</tr>
<tr>
<td></td>
<td>Mean (time)</td>
<td>Range (time)</td>
<td>0.554</td>
</tr>
<tr>
<td></td>
<td>Max (time)</td>
<td>Min (time)</td>
<td>-0.495</td>
</tr>
<tr>
<td></td>
<td>GreenNDWI</td>
<td>NDVI</td>
<td>0.433</td>
</tr>
<tr>
<td></td>
<td>Winter</td>
<td>July</td>
<td>0.424</td>
</tr>
<tr>
<td></td>
<td>Mean (time)</td>
<td>Min (time)</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td>January</td>
<td>July</td>
<td>0.407</td>
</tr>
<tr>
<td></td>
<td>Fall</td>
<td>July</td>
<td>0.402</td>
</tr>
<tr>
<td></td>
<td>Spring</td>
<td>July</td>
<td>0.402</td>
</tr>
<tr>
<td><strong>Demographic</strong></td>
<td>(per) capita</td>
<td>poor</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>female</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>old</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>rate</td>
<td>attendance</td>
<td>0.749</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>logarithmic</td>
<td>-0.700</td>
</tr>
<tr>
<td></td>
<td>female</td>
<td>logarithmic</td>
<td>-0.700</td>
</tr>
<tr>
<td></td>
<td>attend</td>
<td>school</td>
<td>0.672</td>
</tr>
<tr>
<td></td>
<td>primary</td>
<td>complete</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>(per) capita</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>income</td>
<td>poor</td>
<td>0.612</td>
</tr>
<tr>
<td></td>
<td>percentage</td>
<td>rate</td>
<td>-0.594</td>
</tr>
<tr>
<td></td>
<td>household</td>
<td>people</td>
<td>0.585</td>
</tr>
<tr>
<td></td>
<td>population</td>
<td>age</td>
<td>0.560</td>
</tr>
<tr>
<td></td>
<td>education</td>
<td>primary</td>
<td>0.556</td>
</tr>
</tbody>
</table>
unclear. Similarly the TF-IDF analysis alluded to the relationship between “income”, “primary”, “poor” and other words implying economic conditions and inequality. The binary correlations of these tokens confirm this trend as \( \phi \) is positive and usually greater than 0.5 for these pairs of words. Interesting implications also arise for token pairs with strong negative correlations. Intuitively, it makes sense that while the temporal “Range” and “SD” are highly correlated, the temporal “Max” and “Min” would be negatively correlated. The negative correlation of “male” and “female” to “logarithmic” further validates the notion that the transformed measures of the “male” and “female” populations are not uniformly distributed with their raw versions. Textual analysis aids the analyst in understanding the results of this dimension reduction method. It codifies variable relationships revealed by the clustering, can characterize the phenomenon which a particular cluster measures, and bolsters the understanding of the data set as a whole. The extent to which this analysis is needed is left to the user and largely depends on the application intended for the reduced data set. In the anticipated application of this particular data set in a subsequent study, our team applies these two data sets to design Dengue risk models to identify risk factors for Dengue in Brazil. In this way understanding the clustering behavior is critical as an important risk factor in the model will be a representative of a cluster of variables and so risk should be ascribed to the cluster phenomenon not the individual representative variable. For non-diagnostic applications, like forecasting models, the particular meaning of a variable meaning is of minimal importance in comparison to the output of the model, and so textual analysis to understand cluster dynamics has limited use. However, in the case of a particular data stream being disrupted, or halted entirely, the textual analysis can point to candidate variable replacements which measure the same phenomenon to be used in the model. At a minimum the textual analysis can aid in confirming that the application of this dimension reduction is producing a representative data set which is well understood and still measures the scope of the original data set.
2.4.6 Strengths & Weaknesses

As this method is unsupervised, much of the analysis, including the textual analysis, after the selection of the optimal point along the Pareto front will be motivated by the application. The unsupervised nature of this method is one of its primary strengths, as a large amount of dimension reduction is achieved in the absence of a response variable through a filtration rather than extraction method. Filtration methods maintain distinct variables rather than creating new ones, which facilitates the analysis of clusters like the textual analysis. Additionally, the impact of the reduction is quantifiable through the application of the multi-objective optimization, which considers information loss as a counterbalance to the dimension reduction objective. This is analogous to different cost functions used to produce dimension reduction in other unsupervised feature extraction methods, like the calculation of proportion of variance for PCA [7]. The flexibility of this method depending on application also arises from the multi-objective optimization as the trade-off between the objectives can be weighted differently in the choice of the Pareto optimal point. Our method is limited by few factors, namely the correlation metric distance on which the clustering is based. Other metrics could be explored to implement as a clustering distance for dimension reduction that considers measures of redundancy other than the linear redundancy measured by the correlation coefficient. Additionally, this method is mostly limited to a temporally static data set. This limitation has been sidestepped in the case of the environmental data by treating each year of summary variables as a new spatial location. This could clearly be performed for any timescale, up to and including the weekly measurements; however, the size of the data set would be massive, drastically increasing the computational time and storage needed to implement the method. Thus, further exploration of measures of temporal correlation would be required to truly extend this method to spatio-temporal data. Additional adjustments may also be desired during the process of model design, as it may be valuable to withhold the data that would be used for testing purposes within the dimension reduction algorithm in order to measure the robustness of variable relationships across training and testing sets.
2.5 Conclusions

This study introduces a new multi-objective feature selection method to generate lower-dimensional variable subsets of a given data set while maintaining the intrinsic interpretation and significance of the observed variables. In particular, we use correlation-based clustering along with a genetic optimization algorithm and text mining to identify a collection of descriptive representatives for the complete set of observed variables. Furthermore, the method is unsupervised so that no classifier or response variable is necessary to perform the analysis, and the dual objectives presented allow a user to tailor the method to the application of interest while selecting between the competing properties of information retention and dimension reduction. These particular properties cannot be achieved by previous methods (see Sect. 2.2). We demonstrate the utility and efficiency of the method on demographic data from the 2010 Brazilian census and environmental data generated by combining remote sensing satellite observations with weather station measurements in Brazil. In utilizing the proposed method, a significant reduction in the size of the original variable set is achieved with minor information loss. Our work highlights the robust nature of the method and allows for complete elimination of redundant features without requiring information specific to a particular application. A future study will utilize the reduced set of observed variables obtained here to generate a mosquito-borne disease model that will allow epidemiologists to better understand and forecast outbreaks of Dengue fever within Brazil.
3.1 Introduction

In recent years, the steady increase in Dengue Fever (DENV) incidence and other arboviruses such as Zika (ZIKV) and Chikingunya (CHIKV) globally has motivated an elevated focus on quantifying and forecasting risk of these viral diseases. Because the transmission dynamics of mosquito borne diseases are particularly complex and direct measurement of the fundamental mechanisms of their spread (e.g. mosquito population sizes and distributions) is an onerous task, many have turned to proxy data to gain insights into these systems [20]. Fewer limitations in data collection and storage have resulted in an abundance of rich, diverse, and dynamic data sources that measure various aspects of mosquito borne disease spread. As a part of an effort to extend the capabilities of using diverse sources of proxy data at a large spatial scale, this study uses demographic indicators, weather/climate measurements, and remote sensing data simultaneously to quantify the yearly Dengue risk in Brazil for each of the 5564 municipalities and understand how each of the data streams contributes to this risk. Understanding the particular risk factors or indicators is particularly important as effective risk mitigation requires accurate understanding of the drivers of incidence. To this end, though more complex models exist, this study is limited to the construction and testing of linear models to facilitate a clear interpretation of the results and model inputs in order to assess both accuracy and the relative importance of the various data streams. In part, this simplification is also necessary due to the spatial scale at which the subsequent models are constructed, as the shear number of spatial locations simultaneously adds both computational and model complexity. The goal of better understanding the risk factors of Dengue also provides an opportunity to apply the novel dimension reduction method described in Chapter 3 and assess the impact that reduced data sets have on such models. In
addition to providing a substantial foundation for future extensions of this modeling effort, the significant predictors in our models can indicate effective avenues for short-term prevention, as well as strategies for long term public health campaigns to reduce the future burden of these mosquito-borne diseases.

3.2 Background

Dengue Fever (DENV) is a mosquito-borne disease that is nearly ubiquitous in the tropical parts of the world with prevalence in over 100 countries and about 40% of the world population living in areas at risk of infection [71]. The disease is estimated to infect up to 400 million people per year, though this population and region of risk will likely change as the planet warms due to climate change, with nearly a quarter of those infections presenting symptoms. Symptoms include high fever, headache, vomiting, muscle/joint pain, and skin rash and while most cases are not fatal, individuals infected with DENV can develop Dengue Hemorrhagic Fever which is highly fatal [71]. Dengue fever is caused by four related viruses, called serotypes, each of which an individual can contract over their lifetime. The existence of these four different serotypes challenges vaccination and treatment efforts, meaning that interventions must be non-pharmacological in nature, either by control of the vector population or via prevention of contact between the vector and the human host. The virus is spread primarily by the *Aedes Aegypti* mosquito, but also by the *Aedes Albopictus* mosquito [36, 73]. These mosquitoes have evolved to preferentially bite humans over other mammals, and particularly the *Aedes Aegypti* have adapted to live near and amongst humans [36, 73]. Both species lay their eggs in standing water that gathers after rainfall in receptacles like old tires, flower pots, and water dishes [36, 73]. The life cycle of these mosquitoes and the DENV transmission cycle is illustrated in Figure 3.1. Note that in addition to laying eggs in water, all three juvenile stages of mosquito development occur in water, with only the adult mosquitoes living independent of a water source. Patterns of precipitation and humidity thus have significant links to the population growth of the mosquito populations. The complexity of the life cycle dynamics of the vectors, due to the impacts of the environment,
Figure 3.1: **Mosquito-Borne Disease Transmission Pathway** The life cycle of the mosquito vector has four stages: (1) a female mosquito lays eggs in water, (2) eggs develop into larve, (3) larve mature into pupae, (4) pupae emerge as adults. Mosquito borne viruses spread when a susceptible vector bites an infectious host, becoming infected. Then when the infectious vector bites a susceptible host the host then can become infected.

as well as the challenges with observing or measuring the interactions between the vector and the host, have engendered the need to derive relevant information that can capture the complex interactions between the host, vector, pathogen, and environment from *proxy data* (i.e., alternative, indirect data sources). The imperfection of these alternative sources in their relationship to the system in question, is a trade-off when using this data to provide up-to-date local measurements in the absence of other relevant data. Weather station measurements [21, 58, 78] and remote sensing of vegetation health can elucidate the status of the mosquito habitat [44, 59], while demographic indicators can supply information about mosquito interactions with human hosts and their environment [54]. Many of the previous studies using proxy data to model aspects of mosquito borne disease spread and transmission have limited themselves by considering a limited range of data streams and/or have constrained the models to small spatial locations and scales, such as particular cities [11] or small countries [65].

The data available to this particular study is distinct due to its range of data sources, spatio-temporal granularity, and duration, as no other source has approached modeling
DENV at this scale and with these resources. Clinical surveillance data from the Brazilian Ministry of Health provided weekly Dengue case counts in each municipality for the time period of January 3, 2010 to July 17, 2017. The total yearly cases for each municipality was then calculated for each of the seven complete years of data and will serve as the response variable of interest in this study. The exogenous variables considered originate from three separate data streams, on each for demographic, climate/weather, and satellite imagery measurements. Demographic and socioeconomic data were obtained from the Brazil census at municipality and state levels for 2010. They include total 232 variables providing information about population statistics, education, income and poverty levels, and employment status. For indicators of weather and climate, daily temperature and humidity data were collected from weather stations in Brazil from April 1, 2009 to April 1, 2017 using the Global Surface Summary of the Day (GSOD) dataset from the National Oceanic and Atmospheric Administration (NOAA) [56]. Many of the municipalities do not have a single raw measurement from the weather stations as there are only 613 stations in Brazil compared to 5564 municipalities in Brazil. This gap in spatial data was addressed by employing kriging methods to spatially interpolate temperature and humidity values. Finally, land cover and measures of water in the environment were sourced by using multispectral satellite images for each municipality in Brazil on a near-weekly basis from January 2010 to December 2016. These images came from four public-domain global coverage satellites (Landsat 5, Landsat 7, Landsat 8, and Sentinel-2), and were accessed and aggregated using the Descartes Labs’ python-based platform [28]. Within each municipality, we computed the following indicators that provided information about factors related to dengue infections (e.g., water, vegetation): Normalized Difference Vegetation Index (NDVI), Green-based Normalized Difference Water Index (Green NDWI), Shortwave Infrared-based Normalized Difference Water Index (SWIR NDWI), Shortwave Infrared-based Normalized Burn Ratio (SWIR NBR), and the proportion of cloudy pixels. As this study considered models at the yearly temporal scale, yearly weather and satellite-derived time series are obtained at the municipality level by comput-
ing a number of summary statistics on the features (minimum, maximum, mean, standard deviation, and range) by year, season (winter, spring, summer, and fall), and month (only for January and July, which are extreme in terms of dengue infections and weather). This results in a total of 315 weather and 455 satellite features.

3.3 Methods

We first consider a specific region indexed by $i$ and consisting of a collection of $n$ distinct locations. Then, we define $Y_i$ to be the $n \times 1$ vector consisting of the observations of a response variable for each location and $X_i$ to be the $n \times k$ matrix of $k$ predictor variables also with $n$ observations. For each region $i$, a linear model can be constructed so that

$$Y_i = X_i \beta_i + \epsilon_i$$

(3.1)

where $\beta_i$ a $k \times 1$ vector of coefficients assigned to each predictor variable and $\epsilon_i$ is an $n \times 1$ vector of error adjustments for each spatial location [55]. The model can be fit for all $k$ predictor variables, or a subset of $k$ predictor variables by setting the corresponding coefficient to zero for each of the variables out of consideration. This array of models can be vast depending on the number of predictor variables in the data, so it is valuable to use a model selection algorithm in conjunction with model fitting to arrive at the appropriate model for the region $i$. The method used in this study is a step-wise regression method which uses the Akaike information criterion (AIC) as the indicator of model fit [9]. The AIC scores a model based both on its goodness of fit, quantified by the maximum of the likelihood function for the model $\hat{L}$, and number of predictor variables included within the model $k$. The exact formula for this information metric is given by

$$AIC = 2k - 2 \ln(\hat{L}).$$

Using the AIC can mitigate some of the risks of overfitting a model. Step-wise regression builds a regression model from the set of candidate predictor variables by adding and removing predictors in a step-wise manner into the model until there is no longer a statistical difference between the AIC of the models at each step [43].
Once a model $M_i$ has been selected, further analysis on the size of the model and accuracy can be performed. If the region $i$ is one of $m$ subsets of the entirety of observed locations, the model $M_i$ can either be evaluated independently, or as a part of a larger model $M$, which is a collection of the models $M_i$ for all regions $i = 1, \ldots, m$. Often for this study, the collection of models will be analyzed rather than the individual regional models. In addition, as this study is also an application of the dimension reduction method in Chapter 3, models generated by different data sets can be compared to each other using not only error and size metrics, but also measures of computational time. In this way the impact of an unsupervised dimension reduction on a model can be evaluated alongside the overall modeling effort.

The error metrics this study will use are Relative Root Mean Squared Error (RRMSE), Relative Mean Absolute Error (RMAE) and R Error, formulas for which are provided in (3.2). Let $y = Y_i$ be the vector of observed response variables and $\hat{y} = X_i \beta_i$ be the vector of the predicted model values. Then $y_j$ and $\hat{y}_j$ are the values of each for location $j$ and $\bar{y}$ and $\bar{\hat{y}}$ are the mean of each of the two vectors. Then RRMSE, RMAE and R are calculated as

$$
\text{RRMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (y_j - \hat{y}_j)^2} \left/ \max_{j=1, \ldots, n} (y_j) - \min_{j=1, \ldots, n} (y_j) \right.,
$$

$$
R = \frac{\sum_{j=1}^{n} (y_j - \bar{y})(\hat{y}_j - \bar{\hat{y}})}{\sqrt{\sum_{j=1}^{n} (y_j - \bar{y})^2} \sqrt{\sum_{j=1}^{n} (\hat{y}_j - \bar{\hat{y}})^2}}
$$

$$
\text{RMAE} = \frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j| \left/ \max_{j=1, \ldots, n} (y_j) - \min_{j=1, \ldots, n} (y_j) \right.,
$$

The size of the full model for each model generated will be equal to the number of predictor variables in the model selected. Similarly the model time will be equal to the time necessary to train each model on the initial portion of the obtained data. Finally, the predictor variables can be analyzed for their impact on the model. This is important as the model can elucidate the potential mechanisms of risk. There are many disparate ways to rank or
measure impact, however this study will evaluate this on two fronts. First, the $p$-value of each predictor indicates importance it has to the model fit relative to the other variables in the model. This is a ranking comparison, but the other method, if the predictor variables are scaled appropriately, examines the magnitude of the coefficients $\beta_i$ to determine impact and the sign of the coefficients to determine if that impact is positive or negative.

3.4 Results

3.4.1 Data and General Model Set Up

As the primary goal of this study is to build models which quantify the yearly risk of Dengue, as well as reveal potential risk factors for Dengue in Brazil, the response variable in all models described subsequently is a measure of the total number of cases in a municipality over the course of a calendar year. As the accessible data reports weekly case counts, the first step is to calculate the sum over the weeks for each year. Then, through data exploration, it became clear that the logarithmic transform of that sum was better suited to a linear model, so the data had to be prepped to accommodate this transformation. The primary challenge at hand was the existence of a significant number of municipalities each year that recorded zero cases, and thus numerically are not suited to apply a log transform. Typically in these situations there are two general approaches - remove those observations from the historical data or add a small positive number $\epsilon \approx 0$ to the appropriate observations in order to make the transform numerically feasible. Through early tests, it became clear that the performance of the model in both training and testing was improved by employing the second approach, as locations with zero cases still contain meaningful information about the mechanisms of risk. Thus, for every observation which reported zero cases the value $\epsilon = 10^{-4}$ was added prior to applying the log transform.

Recall that a secondary goal of this study is to evaluate the impact of the unsupervised dimension reduction method presented in Chapter 3 on a particular modeling endeavor. To this end, this study will consider two different sets of potential predictor variables which are generated by the dimension reduction method coincide with the data sets generated by the
Table 3.1: **Small & Large Predictor Sets**
The sizes of the small and large predictor sets used in this study to build linear models as well as the correlation thresholds and number of cluster representatives which generated the Environmental and Demographic subsets in question.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Number of Variables</th>
<th>Correlation Threshold</th>
<th>Number of Cluster Reps</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small $X^S_i$</td>
<td>EV</td>
<td>250</td>
<td>0.811</td>
</tr>
<tr>
<td></td>
<td>Census</td>
<td>50</td>
<td>0.790</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>300</td>
<td>-</td>
</tr>
<tr>
<td>Medium $X^L_i$</td>
<td>EV</td>
<td>632</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>Census</td>
<td>110</td>
<td>0.948</td>
</tr>
<tr>
<td></td>
<td>Both</td>
<td>742</td>
<td>-</td>
</tr>
</tbody>
</table>

weighting schemes $w_l : w_N = 1 : 9, 1 : 1$. These data sets will be referred to as the small ($X^S_i$) and large ($X^L_i$) predictor variable data sets. The size of these data sets as well the correlation thresholds and number of cluster representatives which were used to generate them are detailed in Table 3.1. Both the predictor variables and response variables are scaled so that the range of observations for each spatial location have a mean of zero and a standard deviation of one. With every part of the model on a $N(0, 1)$ scale, coefficients and the impact of variables can be directly compared to each other and model outputs.

This study will also consider two different scales for sub-setting the municipality level data. The first will be to build a single model for the whole country meaning every municipality in Brazil is considered to be a part of the same region $i$. The second scale will build five sub-models, one for each administrative region of Brazil and aggregate the results. The five regions are determined by the Brazilian government however, they correspond roughly to the main eco-zones in Brazil: Amazon rain forest (North), dry forest (Northeast), savanna (Center West), Atlantic forest (Southeast), and subtropical grassland (South) as well as Pantanal wet lands that also occur in Center West (Figure 3.2) [4]. These two sub-setting approaches will allow for a comparison of accuracy between the scales, distinguishing between the more local factors which determine risk as opposed to the overall country wide risk factors. The number of municipalities varies from region to region (Table 3.2) and this imbalance is an
Table 3.2: Number of Municipalities per Region
The number of municipalities in each of the five regions in Brazil varies between regions and is not related to the geographical size of the regions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of Municipalities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: North</td>
<td>449</td>
</tr>
<tr>
<td>2: Northeast</td>
<td>1793</td>
</tr>
<tr>
<td>3: Southeast</td>
<td>1668</td>
</tr>
<tr>
<td>4: South</td>
<td>1188</td>
</tr>
<tr>
<td>5: Central</td>
<td>466</td>
</tr>
</tbody>
</table>

additional inducement to minimize the number of predictor variables to consider. If the number of predictors is too large, the linear model can be over determined, and thus over-fit. The available data provides seven years of historical data for the predictor and response variables. With this amount of data, it can be divided into a training and testing set. The model will be trained on the data from 2010-2014. Each of the five yearly observations for each municipality are used simultaneously in the training of the model. The remaining years of data, namely 2015 and 2016, are then used as the testing set.
Table 3.3: **Computational time required to train linear models**

Time needed to train models on a small ($X^S_i$) and large ($X^L_i$) sets of potential predictor values at both the regional and country modelling scale. Note that the time required for the models at the regional scale represent the time required to train the largest region (in this case the Southeast) as the regional models can be trained in parallel.

<table>
<thead>
<tr>
<th>Model Scale</th>
<th>Data Set</th>
<th>Computational Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>Small</td>
<td>20540.067s (~5.71 hours)</td>
</tr>
<tr>
<td>Country</td>
<td>Large</td>
<td>285802.467s (~3.31 days)</td>
</tr>
<tr>
<td>Regional</td>
<td>Small</td>
<td>7784.638s (~2.16 hours)</td>
</tr>
<tr>
<td>Regional</td>
<td>Large</td>
<td>120609.635s (~1.4 days)</td>
</tr>
</tbody>
</table>

### 3.4.2 Comparison of Accuracy, Size and Computational Time

Both data sets $X^S_i$ and $X^L_i$ are used to train the two types of models. All models were trained on the Colorado School of Mines Mio super-computing cluster to utilize a high-memory and high-speed computational platform, as well as, to allow for parallelization of the algorithms when possible. Though the size of the predictor data sets only differ by around a factor of two, the computational time needed to train the models on the larger data set is not proportional to the increase in the size of the data. For instance, in the case of the models formulated at the regional scale, it takes 15 times longer to train the models, while those at the country scale require 14 times the amount of computational time (Table 3.3). The size of models is summarized in Table 3.4, with the regional size breakdowns indicated. Note that the overall size of the two regional models is the same even though the individual regional sub-models differ in size, due to more overlaps in predictor variables across sub-models in the regional model trained on the larger data set. Essentially the regional sub-models trained on the smaller data are more distinct from each other. With a substantial increase in computational time for any model designed with the larger data set, it is important to examine if there is any increase in accuracy due to the larger range of predictor values, and if the improvement is worth the additional training time. The scatter plots of the observations vs the model predictions for training of the models are shown in Figure 3.3 while those for testing are shown in Figure 3.4. It should be noted that the vertical lines of points on the
Table 3.4: Model Sizes
Number of predictor variables in each of the four models. The regional sub-model sizes are indicated and the overall size of the regional models is found by finding the number of distinct variables in all sub-models.

<table>
<thead>
<tr>
<th>Model Scale</th>
<th>Data Set</th>
<th>Overall</th>
<th>Region 1</th>
<th>Region 2</th>
<th>Region 3</th>
<th>Region 4</th>
<th>Region 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>S</td>
<td>166</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Country</td>
<td>L</td>
<td>278</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Regional</td>
<td>S</td>
<td>256</td>
<td>95</td>
<td>110</td>
<td>124</td>
<td>99</td>
<td>87</td>
</tr>
<tr>
<td>Regional</td>
<td>L</td>
<td>256</td>
<td>131</td>
<td>195</td>
<td>205</td>
<td>131</td>
<td>134</td>
</tr>
</tbody>
</table>

left-hand side of all the graphs is due to the inclusion of observations of zero case counts. It is clear that while imperfect all models demonstrate a level of linear fits in both the training and testing data. The testing shows that much of the error in all models comes from the over prediction in municipalities with low or no case counts. All four models tend to over-predict in these municipalities, as baseline conditions often indicate at least a small number of cases. This dynamic is common as zero observed cases can be due to unpredictable stochastic effects as well as a mischaracterization of the model. Additionally, from a public health perspective, a baseline level of preparedness is necessary, and thus some over-prediction is not detrimental.

While it is not clear from the scatter plots which model performs the best on the training data, it is clear that the models trained on the smaller data set perform better on the testing data than those trained on the larger data set. A numerical distinction between the models requires investigation of the error metrics for each of the models (3.2). RRMSE and RMAE are calculated with scaled and transformed predicted and observed case counts Table 3.5. The error metrics are computed and reported for at the scale used for modeling, rather than the scale of the case count data because the process of reversing the scaling and transform is cumbersome, and in the case of the testing data the scale is theoretically unknown. Further modeling and analysis work must be done to either build models at the scale of the data, or to devise appropriate scaling methods to transform risk calculations to general total population
Figure 3.3: **Training Scatter Plots**
The fits of each model, trained either on large or small predictor data, to the training data for the country and regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction.

numbers. As this is a preliminary and exploratory study, this exercise is left to future work. This work will use predicted and observed case counts in reference to these scaled values, or will reference these values as a risk level. The first observation is that the performance of the model depends on whether errors are calculated for its training data set or its testing data set and which errors metrics are being calculated. In training, all models perform similarly according to all three metrics, though the regional models perform slightly better in RRMSE and RMAE, while the country-wide model has the edge in R. When moving to the testing error, performance in RRMSE and RMAE do not change significantly, however there is a clear drop in performance in the R metric, which denotes linearity. This reduction in linearity is particularly stark for the large regional model. Clearly, the strength of the larger regional model, where many variables generated a fine tuned fit to the data, becomes its downfall, as the model is not flexible enough to predict data it has not encountered. Instead
Figure 3.4: **Testing Scatter Plots**
The fits of each model, trained either on large or small predictor data, to the testing data for the country and regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction.

Table 3.5: **Model Errors**
Errors generated by the training and testing data for the regional and country scale models trained on the small and large predictor variables sets. Errors are calculated with unscaled and un-transformed prediction and observed case counts.

<table>
<thead>
<tr>
<th>Model Scale</th>
<th>Data Set</th>
<th>Training/Testing</th>
<th>RRMSE</th>
<th>RMAE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>L</td>
<td>Test</td>
<td>0.189</td>
<td>0.144</td>
<td>0.577</td>
</tr>
<tr>
<td>Country</td>
<td>S</td>
<td>Test</td>
<td>0.178</td>
<td>0.133</td>
<td>0.591</td>
</tr>
<tr>
<td>Regional</td>
<td>L</td>
<td>Test</td>
<td>0.140</td>
<td>0.107</td>
<td>0.281</td>
</tr>
<tr>
<td>Regional</td>
<td>S</td>
<td>Test</td>
<td>0.103</td>
<td>0.0765</td>
<td>0.476</td>
</tr>
<tr>
<td>Country</td>
<td>L</td>
<td>Train</td>
<td>0.165</td>
<td>0.121</td>
<td>0.775</td>
</tr>
<tr>
<td>Country</td>
<td>S</td>
<td>Train</td>
<td>0.176</td>
<td>0.130</td>
<td>0.740</td>
</tr>
<tr>
<td>Regional</td>
<td>L</td>
<td>Train</td>
<td>0.113</td>
<td>0.0812</td>
<td>0.722</td>
</tr>
<tr>
<td>Regional</td>
<td>S</td>
<td>Train</td>
<td>0.120</td>
<td>0.0859</td>
<td>0.679</td>
</tr>
</tbody>
</table>

the smaller models both at the regional and country scale are significantly more accurate in all measures than the larger regional model. They are on par with the large country scale model in both training and testing, however substantial size increase and the prohibitive training time mark the large country scale model as infeasible. These two smaller models
produce errors at the same magnitude, with the country scale model having lower RRMSE and RMAE while the regional model is more linear with a higher R value. Further analysis indicates that the larger regional models perform better on the training set due to better fits for Regions 2, 3, and 4. However, those regions were, with Region 2 being the primary problem, the main sources of error in the testing (Figure B.1, Table B.1). This could indicate a number of things, but as the dimension reduction eliminates redundant variables based on correlation, this could indicate that the larger regional employed the best fitting variable set in training, when another set of more flexible variables may have produced slightly more error in training but significantly less in testing. In other words, less redundancy in the data is able to build models with variables that more accurately indicate risk while more redundancy causes overfitting in training the models.

3.5 Discussion

The extreme computational time and the lack of increase in accuracy signals that any further model analysis should be done on the smaller models. As the models have relatively similar error values in all three metrics, it is valuable to compare which model performs the best in each region. While the country-wide model is not region specific, the metrics can be calculated grouped by region. Table B.2 indicates that neither model performs significantly better in any error metric across any regions. Thus, with the same level of accuracy, the models can serve as compliments to one another. In the case of the country-wide model, the model can aid in identifying overall risk factors of Dengue cases, while the regional models can pinpoint the region specific risk factors, that may or may not differ from the country-wide model. Maps of the observed risk of Dengue are compared to the predicted risk by the Country and the Regional scale models in Figure 3.5. Both models seem to produce predictions that match the overall trend of the observed risk, but seem to struggle with the scale, particularly at the lower levels of observed risk. We know from the scatter plots (Figure 3.4) that the main source of inaccuracy is due to over predictions of low case counts, as well as what appears to be a small number of outlier predictions at the upper end.
Figure 3.5: Maps of Linear Model Results
The observed risk of Dengue is mapped for each municipality to compare to the risk predicted by the regional and country scale models in the two testing years 2015/2016. The color scales are specific to the year and region for which the model was trained.

of the risk scale. One reason for this could be the result of training data on municipality data. As municipalities in Brazil on average have 40,000 people in them, roughly the size of Annapolis, MD, and much smaller than a typical city in the US or even a US county, it begs the question whether a case count reported in a particular municipality actually belongs to one of its neighbors, due to the location of hospitals, doctors offices, or just reporting procedures. It was not feasible to train the models at a larger scale because the census data was produced only at the municipality and state level. Clearly the training on state level data would have caused numerous overfitting issues, and even smaller sub-divisions larger than
Figure 3.6: **Average State Risk Scatter Plot**

The average of the predicted and observed risk is calculated for each of the 27 states for the training and testing data for both the country and regional scale models. The fit of these models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation is equal to the prediction.

Municipalities like micro-regions and meso-regions would have similar issues. Additionally, aggregation techniques for the census from its raw state, is non-trivial, as statistics have to take into account weighted averages, percentages taken on sub-total populations, or formulas which can not be inverted.

Despite these challenges, this theory can be tested crudely by taking the average of the model prediction output as well as the average of the corresponding testing observations at the proper scale over the groupings of municipalities in each of the 27 states of Brazil. Thus we consider the average risk for each state. The scatter plots of the training and testing data Figure 3.6 demonstrate that many of the prediction errors balance, as both models show more linear fits and less spread along the best fit line. It is interesting to note that the country model has a few outliers that significantly over predict on the high end of the risk scale while the regional models see less of this effect. Maps of the mean risk for the country
scale and the regional scale can be seen in Figure 3.7. In all models the state/municipality Federal district is the source of the largest error, with massive over predictions produced. Thus for more distinction in the color scales, it is left blank. This is also true, when looking closely at the non-aggregated results, and is likely due to the particular dynamics of the capital district. Additionally, note that the color scale refers to high and low risk for the region for which the model was trained. In the overall country model, every state’s risk average can be compared directly to every other, but in the regional model risk can only be compared within regions. For the country scale models there appears to be a general trend of over prediction across all states, while the errors in the regional models are not as predictable. The regional model particularly struggles to accurately model the states in the Amazon region, which is not surprising, as the region is more sparsely populated, poorer, and has less health infrastructure. There are costs and benefits to this rudimentary approach, as means can be biased and this average does not take into account the underlying scale of the data. However, this aggregation provides insight that reveals, on average, that the significant errors in prediction at the municipality level are not indicative of a mischaracterization of the model at the state level. This is particularly valuable to assess, as many public health responses operate at both a state and municipality level, just like here in the US. In this respect, recommendations for this model may not be entirely appropriate for the municipal public health, but may be for the state public health response.

Analysis of the variables that make up the models helps to both indicate regional and country specific risk factors, and provide justification for the use of all three proxy data sets in this and other modeling efforts, with demographic data via the census and the environmental data via the weather stations and the remote sensing measurements. Variables are ranked according to the \( p \)-value generated by the model selection algorithm, with the variable with the lowest \( p \)-value denoting the variable which contributes most highly a low AIC value. This indicates that it is important both to the accuracy and parsimony of the model relative to other variables. The model coefficients can be analyzed in a similar fashion;
however, the sign of their coefficients is more informative than their magnitude at this stage of modeling. Table 3.6 reports the top ten variables for each of the regional models as well as the overall country model based on the p-value ranking as well as the sign of the coefficient. If the sign is positive that means as the variable increases risk also increases and vice versa for a negative coefficient. Recall that a variable may be a cluster representative for other measures which correlate highly with it. The more general the variable description, the more variables it represents. For example, the variable “YearMean_MinTemp” is a representative for most of the temperature summaries, temporal or spatial. There are a few exceptions, such as some of the measures of January temperatures. Thus the level of specificity in the
variable description will indicate the specificity of that variable cluster. Despite the use of representative variables, the positive or negative relationship is determined for the precise variable in the model, i.e. though “YearMean_MinTemp” represents many temperature studies the positive relationship indicated in the table means that as the yearly mean minimum temperature increases so does Dengue risk.

Table 3.6 is complicated but there are some patterns that can be extracted. First and foremost, all models require a mix of environmental and demographic predictor variables, as the top ten variables in each list contain a mix of variables from the two predictor data sets. Furthermore both types of environmental variables, weather and remote sensing, are in the top ten of the variable lists as well, indicating that both data sets contain useful information about Dengue risk. Demographically, log population counts are in nearly every model, with positive relationship with dengue risk. In some models the raw population counts is also in the top ten, often serving as an inhibitor of risk, likely correcting for over-inflated risk in high population areas. The number of people living in rural areas, level of income/educational inequality and % Trade sector Workers all have a positive relationship with Dengue risk, while % Vulnerable to Poverty which commutes to work, per captia income, and % Public Sector Workers, have a mitigating impact on Dengue risk according to this table. Environmentally, as the yearly mean minimum temperature increases so does dengue risk, which aligns with what is known about the impact that cold weather has on the mosquito population’s ability to overwinter. The level of healthy vegetation, measured in two different ways via NDVI and Green NDWI, also appears to influence dengue risk in the way the science would suggest, as the level of healthy vegetation increases so does the risk of Dengue. Other environmental variables have varying relationships with Dengue risk, whether the relationship is positive or negative depends on the region the model is trained on.

These conclusions are extracted from only this table. However, for 112 variables out of the 204 variables that occur in more than one model, the relationship the variable has with the risk of Dengue is consistently either strictly positive or strictly negative in all
Table 3.6: Significant Predictor Variables

The top ten significant predictor variables from each of the five regional sub-models as well as the country scale model are described in order. When a variable is a representative of multiple variables, due to the dimension reduction, the cluster description is used. The sign of the coefficient $\beta$ is indicated with parenthesis, which will indicate the overall effect the particular variable has on Dengue risk predictions. The relationship is specific to the precise variable not the cluster as a whole. Some shorthand conventions are used to save space. Spread refers to both Range and SD statistics, DTempR is the Daily Temperature Range, Rel Hum is Relative Humidity, Levels refers to min/max/mean summary statistics, G-NDWI is Green NDWI, and S-NDWI is SWIR NDWI.

<table>
<thead>
<tr>
<th>Region 1</th>
<th>Region 2</th>
<th>Region 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. % Vulnerable to Poverty (commuting): (-)</td>
<td>Rural Pop: (+)</td>
<td>Temperature (Levels &amp; Spread): (+)</td>
</tr>
<tr>
<td>2. Rural Pop: (+)</td>
<td>Year/Summer Healthy Veg levels: (+)</td>
<td>% Trade Sector Workers: (+)</td>
</tr>
<tr>
<td>3. % Trade Sector Workers: (+)</td>
<td>Income/Education Inequality: (+)</td>
<td>Log Total Pop: (+)</td>
</tr>
<tr>
<td>4. S-NDWI-(StandingWater): (-)</td>
<td>% Trade Sector Workers: (+)</td>
<td>% Public Sector Workers: (-)</td>
</tr>
<tr>
<td>5. July DTempR: (-)</td>
<td>Spring Mean NDVI/G-NDWI: (+)</td>
<td>Rural Pop: (+)</td>
</tr>
<tr>
<td>6. July Mean S-NDWI: (+)</td>
<td>Summer Max NDVI/Min G-NDWI: (-)</td>
<td>DTempR Spread: (+)</td>
</tr>
<tr>
<td>7. July Mean S-NDWI spread: (-)</td>
<td>Spring Max G-NDWI: (+)</td>
<td>Summer Min Rel Hum: (+)</td>
</tr>
<tr>
<td>8. Log Total Pop: (+)</td>
<td>% Households with Electricity: (+)</td>
<td>Total Pop: (-)</td>
</tr>
<tr>
<td>9. Per Capita Income: (-)</td>
<td>Jan Avg Max S-NDWI: (+)</td>
<td>Income/Education Inequality: (+)</td>
</tr>
<tr>
<td>10. Jul Min Mean S-NDWI: (-)</td>
<td>Fall SD Rel Hum log: (+)</td>
<td>Year/Spring DTempR: (-)</td>
</tr>
<tr>
<td>Region 4</td>
<td>Region 5</td>
<td>Country</td>
</tr>
<tr>
<td>1. Log Total Pop: (+)</td>
<td>Fall Max Rel Hum: (+)</td>
<td>Temperature (Levels &amp; Spread): (+)</td>
</tr>
<tr>
<td>2. Temperature (Levels &amp; Spread): (+)</td>
<td>Log Total Pop: (+)</td>
<td>Log Total Pop: (+)</td>
</tr>
<tr>
<td>3. Total Pop: (-)</td>
<td>% Public Sector Workers: (-)</td>
<td>Jan Temp (Levels &amp; Spread): (-)</td>
</tr>
<tr>
<td>4. Winter Mean NDVI/G-NDWI: (+)</td>
<td>Jul SD Rel Hum log: (+)</td>
<td>Rural Pop: (+)</td>
</tr>
<tr>
<td>5. Year Max Rel Hum: (-)</td>
<td>Winter Min S-NDWI: (-)</td>
<td>Jan Rel Hum Spread: (+)</td>
</tr>
<tr>
<td>6. Jan Temp (Levels &amp; Spread): (-)</td>
<td>Rural Pop: (+)</td>
<td>% Trade Sector Workers: (+)</td>
</tr>
<tr>
<td>7. Jan Min Temp: (-)</td>
<td>School Attendance Rate: (+)</td>
<td>Year Max Rel Hum: (-)</td>
</tr>
<tr>
<td>8. Summer Max NDVI: (+)</td>
<td>Total Pop: (-)</td>
<td>Fall SD Rel Hum: (+)</td>
</tr>
<tr>
<td>9. Summer SD %CloudyPixels: (-)</td>
<td>Fall Mean NDVI/G-NDWI: (+)</td>
<td>% Public Sector Workers: (-)</td>
</tr>
<tr>
<td>10. % Public Sector Workers: (-)</td>
<td>Fall Min DTempR: (+)</td>
<td>Year Range/Min MaxNBR: (-)</td>
</tr>
</tbody>
</table>
models in which it is present. These variables that have a universal impact on all models could be a strong starting point for interventions. For example, from the demographic data variables the logarithm of the total population, number of people living in rural areas, level of income/educational inequality, % trade sector workers always, % of economically active (working) 10-14 year-olds, increase Dengue risk as they increase while the raw total population, % public sector workers, and % preschool attendees always introduce a mitigating effect in the model. The environmental predictors overlapped much less in the different models, so it is clear that regional variations in climate result in a regional variation in risk factors. The top environmental variables for the Region 1 (Amazon) model indicate that the amount and variation of standing water at various times of the year, particularly in July, plays an outsized role in determining risk. In Region 2 (Northeast), variables measuring the vegetation levels make up a larger portion of the top ten. Regions 4 (South) and 5 (Central) need a mix of all the types of environmental variables, whereas Region 3 (Southeast) and the country wide model rely mainly on the climate and weather indicators of temperature and relative humidity.

3.6 Conclusions

Building and analyzing these models serves two research purposes. First, by training and testing these models on a large and small data set, both generated by the Multi-objective feature selection method, these models demonstrated the importance of effective preemptive data analytics and dimension reduction and provided a case study of the particular usefulness of the method proposed in Chapter 3. The other purpose was to explore the accuracy of linear models for yearly Dengue burden built at a fine spatial scale, and to glean information about particular demographic or environmental variables which serve either as risk or protective factors of Dengue at both the eco-zone/regional and country-wide level. With regards to the first aim, it was clearly demonstrated that larger, highly correlated data, resulted in highly prohibitive computational time without a proportional or even a relatively significant improvement in model accuracy. The large number of variables with
significant linear relationships, causes substantial redundancy and thus a more difficult effort to distinguish between various model designs. Even this simple type of model required substantial dimension reduction, as the model could not be trained on the entire data set in the maximum allotted time of six days on the university’s super-computing system. The design of the feature selection method was geared to a linear model application, so it is not clear what impact its use would have on other types of models. It is clear that at least the reduction in size would be advantageous, but adjustments or extensions of the method, either by changing the clustering metric or by extending capabilities to spatio-temporal data should be considered in future modeling contexts.

The linear models themselves are simple, but despite that simplicity, there are a few key conclusions. Primarily, a mix of proxy data streams, demographic and environmental, employed simultaneously are critical to accurately modeling Dengue risk. The success in these models provides a level of justification for exploring a diverse set of proxy data in any modeling context. Additionally, it was clear that the region in which the model was trained highly impacted the significant predictor variables. This variation can provide insights into the mechanisms of risk in each of the five regions of Brazil as well as the mechanisms that drive Dengue risk across the entire country. While there was a high level of variation in the make up of the models, there were variables, particularly demographic variables, which had a consistent mitigating or protective role in the model. These consistencies can aid as well in developing a better understanding of the risk profile in Brazil. The key weakness in these models was the inability to capture low-level risk, seen in the substantial over prediction in municipalities which observed zero cases. A more nuanced model, where the data first predicts a binary high- or low-level risk and then proceeds to build models in one or both of those categories, could aid be explored as a first step towards improvement. The fundamental model design also engenders some limitations as the model depends on historical data that is fully observed to make its predictions. In other words, to determine the yearly risk level in a municipality one must have the entire year’s worth of data before the model functions.
Adjustments and extensions to this model could explore using climate predictions as input variables to address this in the current model framework, or the model could be retooled to use recent historical data at the beginning of the epidemic year as well as data from the year in question observed during a fixed time period, after which the model could predict the forthcoming risk for the rest of the year.
4.1 Introduction

In many mathematical and data-driven applications, including epidemiology, one often has access to data that features both spatial and temporal variations, but difficulties arise in the incorporation of spatial aspects of the data set within a dynamical model. Therefore, in studying the spread of an epidemic within a population, analysts often make the assumption of spatial homogeneity throughout the region, but this is unreasonable when populations are not “well-mixed” and contacts between individuals depend upon underlying spatial structures within their environment, such as transportation routes [16]. Unfortunately, this assumption is often made despite its shortcomings because the analysis of spatial models that utilize partial differential equations (PDEs) is particularly challenging, especially when stochastic components must be included. Small outbreaks are especially challenging to model and analyze, as similar spatiotemporal conditions can give rise to the progression of an epidemic and alternatively, little to no spread of the disease. Stochastic models aid in such modeling efforts, as this natural uncertainty is encoded directly by a stochastic process [10]. While such processes are often employed in spatially-homogeneous models, little has been done to develop stochastic models that fully describe the spatial aspects of infection dynamics. This is of particular importance as such models allow for the precise fit of parameters from collected field data. In particular, additional information about epidemiological parameters (often, in the form of the their probability distributions), such as the latent period during which exposed individuals develop symptoms and any spatiotemporal variations in the infection and contact rate $\beta$, as well as, knowledge of the potential mechanisms of spatial infection spread, can all be of critical importance to responses during an epidemic and further aid in planning interventions for crises that have yet to emerge.
Stochastic models are especially amenable to analysis via Bayesian inference methods, such as Approximate Bayesian Computation (ABC) [13, 14], where model fits and parameter distributions are estimated simultaneously. Additionally, under the Bayesian paradigm different model proposals can be compared and ranked via Bayes Factors. Finally, though the stochastic framework can complicate standard methods for calculating and analyzing the basic reproduction number, methods for calculating an empirically adjusted reproduction number for embedded stochastic models have been recently developed [19].

Consider the Ebola outbreak which spread around the world between 2014 and 2016 [23]. While this was a large epidemic in comparison to others involving Ebola virus disease, this particular epidemic began, and was maintained, by relatively rare infection and interaction events, both in terms of biological disease processes and the movements of individuals. Thus, in an epidemiological sense it is a “small” outbreak and both the standard well mixed and deterministic assumptions will present fundamental problems. It has been well documented that this outbreak confounded international experts at the time due to its transmission uncertainty [49] and spatial heterogeneity [31, 32]. As a result, this historic epidemic is a rich test case for which a novel spatially heterogeneous stochastic model can be designed, tested, and analyzed.

In general, this chapter is organized as follows. In the next section, we provide some background information concerning Ebola Virus Disease and the particular outbreak centered in West Africa between 2014 and 2016, as well as a survey of previous methods used specifically to model this outbreak and general models that consider spatial heterogeneity or stochastic elements in disease spread. Next, we propose a spatially heterogeneous extension to an embedded stochastic SEIR model, which considers a graph structure of spatial locations connected by edges over which individuals can relocate via random walk as they simultaneously transition through disease states according to a stochastic SEIR process. The various analysis methods that are tailored for use with this proposed class of models, such as Bayesian inference for model fits and parameter estimation, Bayes Factors [45] for model
comparison, and the empirically adjusted reproduction number, are then introduced and described in detail. Finally, we demonstrate the utility and efficacy of the proposed class of models by designing and analyzing collections of models with various spatial mobility structures for the Ebola outbreak in West Africa.

4.2 Background and Previous Methods

4.2.1 Ebola Virus Disease Characteristics

Ebola virus disease (EVD) is a viral hemorrhagic fever effecting humans and primates that is caused by four of the six different species of Ebola virus and is native to tropical regions of sub-Saharan Africa [68, 80]. The Ebola virus first spreads from a natural animal reservoir to a proximal human population. The precise natural reservoir is unknown; however, it has been shown that bats carry the virus without expressing symptoms and have been implicated in earlier outbreaks of EVD [24]. After its initial introduction from the reservoir, the virus spreads within a human population via interaction with infected bodily fluids such as blood, vomit, feces, mucus, urine, and semen, either during direct physical contact with an infectious individual or indirect contact with surfaces and materials that have been contaminated by these fluids. The bodily fluids of an individual who has succumbed to the disease remain infectious even after the death of the infected individual. Symptoms typically present themselves two to 21 days after exposure and usually last for six to 16 days, resulting in either recovery or death [68]. Recovered individuals develop non-permanent antibodies which protect them from the virus for up to ten years. The symptoms of EVD begin as extreme flu-like symptoms and progress to include rash, vomiting, and diarrhea, which causes severe dehydration [68]. In some cases individuals may also begin to experience both internal and external bleeding, further intensifying mortality risk. Overall, EVD has a high mortality rate, ranging from 25% up to 90% depending on the availability of supportive treatment that focuses on rehydration and blood pressure stabilization [69].

Outbreaks of EVD occur intermittently throughout tropical sub-Saharan Africa, but the largest and most deadly outbreak officially began on December 26, 2013 in West Africa caus-
ing 28,646 cases and 11,323 deaths worldwide [23]. The epicenter of this epidemic occurred in the three African countries of Guinea, Sierra Leone, and Liberia and was characterized by the significantly longer period of time that it took to contain, control, and eliminate the spread of the disease, as well as the significant spatial spread in the region and around the world in comparison to previous outbreaks. Many factors, including the extreme poverty, inadequate and dysfunctional healthcare systems, and distrust of government in the region, contributed to the massive spread of the virus [80]. The spread was also worsened by the introduction of the virus to urban developments, whereas previous outbreaks were limited to remote areas, and this factor greatly increased the potential contact rate of an infectious individual. The poor quality of health care systems in the region is considered to be the primary reason for the increased challenge in containing the virus, due to the high risk to those individuals responsible for the care of EVD patients of contracting EVD themselves [80]. Additionally, a new and significant transmission pathway was identified as an instrumental source of spread during the epidemic, particularly throughout Guinea. In particular, EVD was spread via traditional religious burial practices in the region, which included a significant amount of direct contact with the deceased, and hence an increased probability of transmission post-mortem.

4.2.2 Previous Models and Methods

Due to the complexity of transmission pathways and the heterogeneity of the spatial spread, especially within the urban areas of all three of the epicenter countries, modeling efforts for this particular outbreak have been quite diverse. Though the effort to model this epidemic began in real time as a response measure to the crisis, due to the temporary nature of immunity to the particular strain of EVD that ravaged West Africa and the unknown level of vulnerability to other strains, an ongoing modeling effort has continued, even after the afflicted countries were declared disease free, in order to better prepare for future outbreaks in the region. A large number of previous models are compartmental in nature, both deterministic [51, 61] and stochastic [50], ranging from a standard SEIR compartmental
structure to a broad range of disease state subdivisions of the population. The epidemiology of the disease, demonstrating various infection pathways, including a post-mortem infectious period, as well as the high level of uncertainty concerning the duration of latent period, add layers of complexity that many models address by adding compartments to a traditional SEIR structure. These additional disease states include additional infectious classes representing other pathways of disease spread such as hospitalized and deceased individuals, distinctions between reported and unreported cases, and further divisions of the incubation period or the particular type of care or hospitalization reflecting the conditions on the ground [29, 81]. Generally, as the complexity of the compartmental model is increased, additional assumptions regarding the homogeneous spatial mixing of populations are necessary. This assumption was typically addressed not via implementation of spatial heterogeneity, but through meta-population structures that further split the populations into more refined categories of age, risk, or community structures [8, 37, 46]. This particular epidemic spread throughout three separate countries in a distinctly heterogeneous way. Other efforts addressed the spatial heterogeneity by building models at the finest spatial resolution available - districts in Guinea, counties in Liberia, and provinces in Sierra Leone [64]. Despite this high resolution of spatial discretization, models were constructed for an individual location, i.e. a single region or country, ignoring the inherent interactions amongst spatial locations. Most existing models, even highly-complex compartmental ones, struggle to address the spatial heterogeneity inherent within the epidemic, and specifically fail to capture the spread of Ebola within Guinea. This is particularly problematic, as the outbreak began in Guinea and, for the first 12 weeks, spreads only within Guinea’s borders.

Those models that do incorporate interactions between spatial locations in both mathematical and statistical studies are limited in the scope of their applicability because they consider only the mobility of the pathogen and not of individuals themselves, wherein infection can spread across borders, while individuals remain static in their locality. Gravity models are such a model, and have been implemented at a variety spatial scales have been
designed to explore various aspects of the spatial spread of Ebola in 2014-2016 [33, 48, 75].

More generally, in the statistical epidemiology field, the gravity model concept is general-
ized to characterize any spatial auto-correlation structure which could explain the spread of disease using spatial or conditional auto-regressive models (SAR/CAR) [60]. One of the few modeling frameworks that facilitates the movement of individuals as a source of disease spread is that of Agent-based Models (ABMs) [53]. ABMs are in scope due to their high computational cost, constraining modeling efforts to single location studies with relatively small populations.

4.2.3 Spatially Homogeneous Embedded Stochastic SEIR Model

To provide some context for our proposed models, we first consider the standard ordinary
differential equation SEIR model [17] for disease spread, namely

\[
\begin{align*}
\frac{dS}{dt} &= -\beta(t) \frac{SI}{N} \\
\frac{dE}{dt} &= \beta(t) \frac{SI}{N} - \alpha E \\
\frac{dI}{dt} &= \alpha E - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{align*}
\]

(4.1)

with the time invariant total population constraint

\[N = S(t) + E(t) + I(t) + R(t).\]

This model assumes there are four disease states: susceptible, exposed, infectious, and re-
moved (recovered, deceased, and immune) populations denoted by \(S(t), E(t), I(t),\) and \(R(t),\) respectively. The susceptible population is, in this case the entire population of the country (or other relevant spatial region), as no vaccine or direct treatment methods to prevent the spread of the disease were available prior to the onset of the epidemic. Members of the susceptible population may contact an infectious individual with probability \(\mu_1,\) and each interaction then has an associated probability of transmission, given by \(\mu_2.\) These probabil-
ities are then combined to form $\beta = \mu_1 \mu_2$, which represents the characteristic infection rate.

The parameter $\beta$ can be dependent on time or space, and we will consider the situation in which $\beta = \beta(t)$ is a function of time and further dependent on location, i.e. $\beta = \beta(x, t)$, when the model is extended to include spatial heterogeneity. The exposed population represents individuals experiencing the biological latent period (quantified in the model by $\frac{1}{\alpha}$ days) during which they have been infected but present no symptoms and are not infectious. After this time, exposed individuals transition into the infectious compartment, from which they are ultimately removed after an average period of $\frac{1}{\gamma}$ days via either recovery or death. With the standard SEIR model in place, we discretize time and build on its overlying deterministic structure by embedding a stochastic disease process within it. This can be done following existing methods [50], which rely on conditional probabilities and an exponential distribution of waiting times for an individual to transition between population compartments. The resulting discrete-time, spatially-homogeneous stochastic SEIR model [50] is given by

\[
\begin{align*}
S(t + \Delta t) &= S(t) - B(t) \\
E(t + \Delta t) &= E(t) + B(t) - C(t) \\
I(t + \Delta t) &= I(t) + C(t) - D(t) \\
R(t + \Delta t) &= R(t) + D(t)
\end{align*}
\]

where $S, E, I, R$ are defined as in the ODE model and obey the total population constraint as before. Here, $\Delta t$ is a specified time step, $B(t)$ is the population of newly exposed individuals, $C(t)$ is the population of newly infectious individuals, and $D(t)$ is the population of newly removed individuals at time $t$. These transition populations are drawn at each time $t$ from binomial distributions so that

\[
\begin{align*}
B(t) &\sim \text{Bin}(S(t), P(t)) \\
C(t) &\sim \text{Bin}(E(t), p_C) \\
D(t) &\sim \text{Bin}(I(t), p_D)
\end{align*}
\]
where

\[ P(t) = 1 - \exp \left[ \frac{-\beta(t)}{N} I(t) \Delta t \right], \quad p_C = 1 - \exp(-\alpha \Delta t), \quad \text{and} \quad p_D = 1 - \exp(-\gamma \Delta t). \]

This model is useful in describing a single well-mixed population with discrete time steps and a probabilistic infection pathway. However, in order to extend the model to accommodate more than a single spatial location, we must consider how individuals within differing locations may move or interact to spread the disease. As previously mentioned, there are existing methods that rest on assumptions of contact probability and viral strength determined as a function of distance, and these are appropriate for epidemics with clear definitions of distance between locations or many nodes. However, we will propose a novel spatial mixing process in which individuals can move between locations and spread the virus via such interactions within the next section.

4.3 Methods
4.3.1 Spatially Heterogeneous Embedded Stochastic SEIR Model

We consider a graph \( G \) with \( k \) nodes, each representing a spatial location over which the disease process spreads. Population mobility on this graph is implemented via transition probabilities defined on the edges according to a Markov probability matrix

\[
\mathbf{P} = \begin{pmatrix}
p_{11} & p_{12} & \cdots & p_{1k} \\
p_{21} & p_{22} & \cdots & p_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
p_{k1} & p_{k2} & \cdots & p_{kk}
\end{pmatrix}.
\]

Individuals can then move between locations via a random walk on this graph. This structure is a Markov chain [2], as an individual’s movement at the next time step only depends on the current position of that individual. If \( p_{ij} = p_{ji} = 0 \) then no edge connects node \( i \) to node \( j \). While it is feasible to allow \( G \) to be directed graph and enforce unidirectional movement along an edge, we assume that there is some level of continuous movement in both directions along a defined edge. Thus, for every edge either \( p_{ij} = p_{ji} = 0 \) or \( p_{ij}, p_{ji} \neq 0 \).
With the edges and transition probabilities defined, a single flux parameter $\rho_{ij} = p_{ij} - p_{ji}$, which we will refer to as the overall flux probability, is defined for each existing edge. This quantity is necessary as the individual transition probabilities are neither practically nor structurally identifiable [15], and the former has been directly verified via simulation. The overall flux probability is intuitively understood as the net exchange of individuals between two spatial locations and will be practically identifiable for a general Markov random walk on $G$. Additionally, our utilization of a single flux parameter for each edge reduces the dimension of the movement parameter space by a factor of two - from $4k(k-1)$ probabilities to $2k(k-1)$. In this way, we fix the less dominant probability of the pair $(p_{ij}, p_{ji})$ and add the estimated flux probability to uniquely determine the second in the pair when necessary.

This random structure assumes a constant probability of individuals moving between two locations, which is valid for two locations that experience a level of mobility consistently. However, mobility between two locations is not always constant. In the context of the spread of an infectious disease, one type of non-constant movement, known as a “sparking” event, can be extremely important to understanding the spatial spread of the disease. Sparking events are characterized by situations where two locations with typically negligible mobility between them experience a rare transmission event which introduces new cases from a region with infected individuals to a region without. Because of the introduction of the disease to a region without infected cases, mobility that would typically be considered negligible is temporarily vital to the spatial spread of the disease.

To facilitate the incorporation of such sparking events, the proposed model will utilize a matrix $\mathcal{I}$ of spatial and temporal indicator functions. We first define a time interval $T$ centered at a fixed time $t^* > 0$ and possessing a specified duration of $2\tau$, so that $T = [t^* - \tau, t^* + \tau]$. Then, the temporal indicator function on this interval is denoted by

$$I_T(t) = \begin{cases} 
0, & t \notin T \\
1, & t \in T 
\end{cases}$$
and for \( p^*_ij > 0 \) the corresponding spatio-temporal matrix \( \mathcal{I} \) is defined entrywise by

\[
\mathcal{I}_{ij}(t) = \begin{cases} 
0 & (i, j) \notin \mathcal{H} \\
p^*_ij \mathcal{T}(t) & (i, j) \in \mathcal{H}
\end{cases}
\]

where \( \mathcal{H} \subset \mathcal{G} \) is a specified edge subset of the graph. When added to \( \mathcal{P} \), this matrix will implement the short term probability \( p^*_ij \) of population mobility between nodes \( i \) and \( j \) with \( (i, j) \in \mathcal{H} \) that otherwise possess no chance of mobility.

With movement between nodes determined by the probabilities \( \hat{p}_{ij} \) in \( \hat{\mathcal{P}} = \mathcal{P} + \mathcal{I} \) at each time step, we can track individuals that move between nodes from each compartment of the disease process. Consider individuals in a particular disease class \( X \) which is mobile. Then \( x_{ij}(t) \), the number of individuals in compartment \( X \) leaving node \( i \) for node \( j \), at time \( t \) can be calculated using multinomial distributions over all \( k \) locations. Hence, we let

\[
[x_{i1}(t), x_{i2}(t), \ldots, x_{ik}(t)] \sim \text{Multi} (\chi_i(t), [\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{ik}])
\]

where \( \chi_i(t) \) is the number of individuals within compartment \( X \) and location \( i \) that are available to move at time \( t \). To understand the number of mobile individuals at each node and within each disease class requires a simplifying assumption. To help us preserve positivity, we will assume that within any time step, a person first completes the disease process stochastic step, i.e. moving from the susceptible to the exposed state, prior to being considered mobile if appropriate. This imposes a reality in which an individual cannot simultaneously move locations and disease states. The multinomial draws produce \( k \) vectors for the \( i \)th spatial location, generating matrices of mobile \( X \) individuals \( \mathcal{M}_X \)

\[
\mathcal{M}_X = \begin{pmatrix}
x_{11}(t) & x_{12}(t) & \cdots & x_{1k}(t) \\
x_{21}(t) & x_{22}(t) & \cdots & x_{2k}(t) \\
\vdots & \vdots & \ddots & \vdots \\
x_{k1}(t) & x_{k2}(t) & \cdots & x_{kk}(t)
\end{pmatrix}
\]

From these matrices we then calculate the net total individuals in disease class \( X \) entering (or remaining within) spatial location \( i \) from all other locations, denoted \( N^X_i(t) \), by summing the columns of \( \mathcal{M}_X \), namely
\[ N_{i}^{X}(t) = \sum_{\ell=1}^{k} x_{li}(t) \]

With population movement defined in this way, the standard stochastic spatially homogeneous SEIR model is now extended to a spatially heterogeneous model describing the spread of an SEIR disease process over \( k \) nodes, given by

\[
\begin{align*}
S_{i}(t + \Delta t) &= N_{i}^{S}(t) - B_{i}(t) \\
E_{i}(t + \Delta t) &= N_{i}^{E}(t) + B_{i}(t) - C_{i}(t) \\
I_{i}(t + \Delta t) &= N_{i}^{I}(t) + C_{i}(t) - D_{i}(t) \\
R_{i}(t + \Delta t) &= D_{i}(t).
\end{align*}
\]

(4.3)

Here, the newly exposed, infectious, removed, and net mobile individuals at time \( t \) and location \( i \) are defined by

\[
\begin{align*}
B_{i}(t) &\sim \text{Bin}(S_{i}(t), p_{i}(t)) \\
C_{i}(t) &\sim \text{Bin}(E_{i}(t), p_{C}) \\
D_{i}(t) &\sim \text{Bin}(I_{i}(t), p_{D})
\end{align*}
\]

where

\[
P_{i}(t) = 1 - \exp \left[ \frac{-\beta_{i}(t)}{N_{i}(t)} I_{i}(t) \Delta t \right], \quad p_{C} = 1 - \exp(-\alpha \Delta t), \quad p_{D} = 1 - \exp(-\gamma \Delta t)
\]

and

\[
N_{i}^{S}(t) = \sum_{\ell=1}^{k} s_{li}(t), \quad N_{i}^{E}(t) = \sum_{\ell=1}^{k} e_{li}(t), \quad N_{i}^{I}(t) = \sum_{\ell=1}^{k} \iota_{li}(t).
\]

Additionally, the multinomial draws for mobile individuals are defined as follows:

\[
\begin{align*}
[s_{i1}(t), s_{i2}(t), \ldots, s_{ik}(t)] &\sim \text{Multi}(S_{i}(t) - B_{i}(t), [\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{ik}]) \\
[e_{i1}(t), e_{i2}(t), \ldots, e_{ik}(t)] &\sim \text{Multi}(E_{i}(t) + B_{i}(t) - C_{i}(t), [\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{ik}]) \\
[\iota_{i1}(t), \iota_{i2}(t), \ldots, \iota_{ik}(t)] &\sim \text{Multi}(I_{i}(t) + C_{i}(t) - D_{i}(t), [\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{ik}])
\end{align*}
\]

The current structure on mobility assumes that susceptible, exposed and infectious individuals possess the same mobility characteristic probabilities. Depending on the disease this
assumption should be addressed considering some diseases render infectious individuals much less mobile or completely immobile. Of course, the mobility of removed individuals does not impact the system, as the SEIR framework described above assumes immunity (or death), rendering mobility of recovered individuals irrelevant to the disease process; hence, we will not include such a feature in the model.

4.3.2 Bayesian Inference for Model Selection and Parameter Estimation

The above framework presents a rich set of models from which an appropriately calibrated model for a particular disease data set can be identified. There are two primary model design decisions that must be made. First, the number of spatial locations $k$ will likely be informed by the data, however it is valuable to explore a variety of possible spatial divisions, i.e. values of $k$, as inherent spatial divisions may be uncovered that are not apparent in the data. The second decision regards the mobility structure between the $k$ spatial locations. Once $k$ is determined there exists $\frac{k(k-1)}{2}$ of different possible connections between nodes, resulting in, if $k \geq 3$, $\sum_{i=1}^{k} \binom{k}{i} 2^i$ different spatial connectivity structures [12]. Additionally if “sparking” events are suspected the model space expands for each “sparking” event explored. Selection of the appropriate model can be determined primarily by the modeler or via a systematic approach or anywhere on that spectrum. Independent of the choice of model structure, most of the model parameters are unknown and thus must be estimated.

With this constraint and the challenge of selecting a good model structure, it is appropriate to employ Bayesian Inference to explore both the parameter and model space. Parameter estimation via Bayesian Inference has many advantages over other methods because it returns more information about the estimate in the form of distributions rather than a point estimate with error bounds. Additionally prior knowledge about the biological process or the model can be directly encoded to promote accuracy in the form of informative priors. Bayesian Inference rests on the foundation of Bayes’ Theorem [1]:

$$P(\theta|x) \propto P(x|\theta)\pi(\theta)$$
where \( \mathbf{x} \) is the observed data, \( \theta \) is a parameter vector, and \( \pi(\theta) \) represents the prior distribution for the parameters and \( P(A|B) \) is the conditional probability of \( A \) given \( B \). This relationship allows the data to be considered as an observation of an underlying disease process, thus leveraging the information in the data and the known prior information about parameters to uncover distributions for the model parameters.

The most common method for Bayesian Inference is Markov Chain Monte Carlo (MCMC) \([39]\). However, the analytic nature of MCMC, as it needs functional forms of likelihood functions, presents difficulties when models become more complex. In contrast, Approximate Bayesian Computation (ABC) can be used on any estimation process with a defined forward model \([14]\). As a result ABC implemented via a Sequential Monte Carlo (SMC) method is the Bayesian Inference method that will be used to explore this class of proposed models \([14]\).

4.3.3 Approximate Bayesian Computation (ABC) and Sequential Monte Carlo Methods (SMC)

ABC is a method by that one can estimate the true posterior distributions for unknown parameters in the model. ABC approximately conditions on the data \( Y \) to analyze the viability of parameter proposals by comparing observations of sample model output \( Y^* \) generated by a set of proposed parameters \( \theta^* \). The viability of the proposed parameter set is determined using one or more norms that quantify the “closeness” of the sample data observation. The norms that are used depend entirely on the problem set and the desired characteristics of the data that the sample data must emulate. The appropriate choice of norms is an ongoing line of inquiry within the field of ABC methods \([13, 14]\). Over the course of this model selection and parameter estimation procedure, we explored and tested various norms to arrive at the final three that are described below and implemented throughout the ABC simulation procedure.

Often, the type of data one has will determine the most informative norms to use for a given model. In this case, with epidemic data there were various dynamics to consider when selecting norms. First, while Bayesian inference considers the data to be objectively true,
in the sense that it is a particular observation of an underlying model process, it is highly unlikely that the data is error free. Due to the measurement challenges that arise early in clinical surveillance of an ongoing epidemic, it is nearly guaranteed that cases will be missed or recorded incorrectly. Thus, matching the particular data at each time point and each spatial location is of minimal importance. As a result, pointwise norms like a standard sum of squares or $\mathcal{L}_2$ error metric over the entire spatial time series will inflate differences between sample epidemics and the data, which could occur due to errors present in the data. Instead of a pointwise matching process, the chosen norms highlight three particular elements of the data - infection intensity, timing, and duration - which are distinctive and determine the qualitative structure of the epidemic. From a public health perspective, these factors determine the number of cases and impact of a particular epidemic. These characteristics are defined at each location $i = 1, ..., k$ by

1. The number of cases observed during the peak week, given by

   $PC_i = \max_{t \geq 0} W_i(t),$

where

   $W_i(t) = \sum_{s = t_0}^{t_1} C_i(s)$

is the weekly number of cases and

   $t_0 = 7 \left\lfloor \frac{t}{7} \right\rfloor$ and $t_1 = t_0 + 6$

are the first and last days of the week containing day $t$ of the outbreak.

2. The week in which the maximum number of cases are observed (i.e., the peak week)

   $PW_i = \frac{1}{7} \min \{ t \geq 0 : W_i(t) = PC_i \}.$
Additionally, the overall end of the epidemic can be defined invariant of space to be

\[ T = \max\{t \geq 0 : C_i(s) > 0 \text{ for all } s \in [0, t] \text{ for any } i = 1, \ldots, k\}, \]

Using these quantities, two relative norms which are spatial in nature \( \mathcal{N}^n_i, n = 1, 2 \) and one relative non-spatial norm \( \mathcal{N}^3 \), are defined, by which the sample data, denoted by \( * \), can be compared to the established data set, namely

\[
\mathcal{N}^1_i = \frac{|PC_i - PC_i^*|}{PC_i^*}, \quad \mathcal{N}^2_i = \frac{|PW_i - PW_i^*|}{PW_i^*}, \quad \mathcal{N}^3 = \frac{|T - T^*|}{T^*}. \tag{4.4}
\]

In the context of an ABC simulation with a defined “closeness” threshold \( \epsilon \in (0, 1) \), the algorithm progresses as follows in order to acquire \( N \) accepted parameter sets \( \theta^* \):

1. Propose \( \theta^* \) from priors \( \pi(\theta) \)

2. Generate sample data by iterating the model forward using \( \theta^* \)

3. Calculate the \( 2k + 1 \) norm values \( \mathcal{N}^n_i \) for \( i = 1, \ldots, k \) and \( n = 1, 2 \) and \( \mathcal{N}^3 \)

4. If \( \mathcal{N}^n_i < \epsilon \) for every \( i = 1, \ldots, k \) and \( n = 1, 2, 3 \), then accept \( \theta^* \), else reject \( \theta^* \)

5. Repeat until the desired number of acceptances have been obtained

While \( \epsilon \) can be chosen such that \( \epsilon \ll 1 \), ABC is often an iterative process that applies a sequence of more restrictive thresholds to evaluate the convergence of the various posteriors. An extension of ABC, Sequential Monte Carlo Methods (SMC) exploit the convergence of the approximate posteriors to restrict the search space in conjunction with increasingly restrictive \( \epsilon \) thresholds. SMC is especially useful in cases with highly uninformative priors, in which a small subspace of parameter sets \( \theta^* \) are viable. Unfortunately, the embarrassingly parallel nature of ABC simulations must be traded for this advantage of restricting the parameter space in a sequential manner. The algorithm is as follows for a sequence of thresholds \( \{\epsilon_1, \epsilon_2, \ldots, \epsilon_J\} \) and \( N \) large:

1. Iteration \( j = 1 \)
(a) Generate sample epidemics with $\theta_{(1)}^* \sim \pi(\theta)$ and accept/reject using defined norms $N_i^n$, until $N$ acceptances

(b) Calculate weights $\omega_{(1)} = 1/n$ and empirical variances $\tau_{(2)}^2 = 2\hat{\sigma}^2$ of accepted parameter sets $\theta_{(1)}$

2. Iteration $j = 2, \ldots, J$

(a) Generate sample epidemics with adjusted parameter proposal scheme and accept/reject using defined norms $N_i^n$, until $N$ acceptances

i. Draw $\theta^*$ from the previous set $\theta_{(j-1)}$ with weights $\omega_{(j-1)}$.

ii. Perturb $\theta^*$ using a Guassian perturbation kernel to arrival at proposal parameter set $\theta_{(j)}^* \sim K_j(\theta^*, \tau_{(j)}^2 I) = \text{MVN}(\theta^*, \tau_{(j)}^2 I)$.

(b) Calculate normalized weights $\omega_{(j)}$ using an importance sampling calculation and variances $\tau_{(j+1)}^2 = 2\hat{\sigma}^2$ of accepted parameter sets

$$\omega_{(j)} \propto \frac{\pi(\theta_{(j)}^*)}{\sum_{m=1}^N \omega_{(j-1)}^m K_j(\theta_{j-1}^m, \theta_{(j)}^*)}$$

Using ABC SMC, the posterior distributions of the parameters $\theta$ can be approximated more quickly despite the loss of a completely parallelizable process. Each step $j = 1, \ldots, J$ can be performed in parallel, with each simulation being independent of one another. However, the steps themselves must be performed in serial.

To further systematize the process one could extend the parameter space to include a model parameter in ABC methods as in [74]. However, the SMC methods described there do not use SMC to update the priors on the model parameters. Instead, they are simulated from the prior at every stage of the SMC algorithm. Therefore, while this parameterization streamlines all models, it is effectively equivalent to implementing SMC on the models separately due to the uniform priors.
4.3.4 Bayes Factors for Model Selection

In addition to the information produced in the parameter space, ABC SMC facilitates an analysis of model efficacy using Bayes Factors [45]. A Bayes Factor is a likelihood ratio between the posterior probability that observed data $x$ is produced by a model $m_i$ and the same probability for another model $m_j$. In particular, for such models the Bayes Factor $B_{ij}$ is defined as

$$B_{ij} = \frac{P(x|m_i)}{P(x|m_j)}.$$  

Using Bayes theorem, this can be expressed as

$$B_{ij} = \frac{P(x|m_i)}{P(x|m_j)} = \frac{P(x|m_i)}{P(x|m_i)} \frac{P(m_i|x)}{P(m_j|x)} \frac{P(m_j)}{P(m_i)} = \frac{P(m_i|x)}{P(m_j|x)} \frac{P(m_j)}{P(m_i)} = \frac{P(m_i|x)}{P(m_j|x)}.$$  

The ABC SMC methods described above are designed to simulate each model separately, and thus $P(m_i)$ is effectively uniform over the model space. Hence, we find $P(m_i) = P(m_j)$, and the Bayes Factor is merely the ratio of probabilities of each model given the data $x$ or

$$B_{ij} = \frac{P(m_i|x)}{P(m_j|x)}.$$  

As such, these values can be approximated by the acceptance rates for each model. This property of ABC methods allows different models to be directly compared and ranked in relation to one another at any tolerance value $\epsilon$ for which an acceptance rate is defined using established norms for the values of Bayes Factors seen in Table 4.1. With this, not only can various model pairs be ranked, but we can also precisely determine the degree to which one model, relative to another, is more effective at producing acceptable simulations at a particular threshold.

4.3.5 Empirically Adjusted Reproduction Number

With the established compartmental framework it is valuable to calculate and analyze the basic reproduction number and other associated quantitative threshold values. The basic reproduction number is intuitively defined as the expected number of secondary infected
Table 4.1: Bayes Factor Table
The strength of evidence in support of \( m_i \) depends on the the value of \( B_{ij} \). The levels of support are indicated with their corresponding ranges for \( B_{ij} \).

<table>
<thead>
<tr>
<th>( B_{ij} )</th>
<th>Strength of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt; 10^0)</td>
<td>Negative (Supports ( m_j ))</td>
</tr>
<tr>
<td>(10^0) to (10^{1/2})</td>
<td>Barely worth mentioning</td>
</tr>
<tr>
<td>(10^{1/2}) to (10^1)</td>
<td>Substantial</td>
</tr>
<tr>
<td>(10^1) to (10^{3/2})</td>
<td>Strong</td>
</tr>
<tr>
<td>(10^{3/2}) to (10^2)</td>
<td>Very Strong</td>
</tr>
<tr>
<td>(&gt; 10^2)</td>
<td>Decisive</td>
</tr>
</tbody>
</table>

cases directly generated by the introduction of a new primary infected case [17, 76]. For the classical, non-spatial SEIR model (4.1), the associated mathematical quantity is defined to be

\[
R_0(t) = \frac{\beta(t)}{\gamma}.
\]

A simple extension of this value would be to consider a spatial version

\[
R_{0i}(t) = \frac{\beta_i(t)}{\gamma}.
\] (4.5)

This functional value for the basic reproduction number at each spatial location can serve as a baseline for early stage behaviors of the system as the system is perturbed from the Disease Free Equilibrium (DFE) [30, 76]. As a simulation evolves away from the DFE, this value begins to lose its potency as a thresholding parameter that can indicate whether the epidemic is growing or receding. With Other formulations, such as an effective reproduction number, which scales \( R_0(t) \) by \( S(t)/N(t) \) can be used in lieu of the original [30]; however, each new formulation is dependent on the parameterization of the model, not on the underlying dynamics of the infection spread observed or simulated. Thus, we will calculate and analyze an empirically adjusted reproduction number for stochastic compartmental models, denoted by \( R^{(ea)} \) and introduced by Brown, Oleson, and Porter in [19]. The empirically adjusted reproduction number generalizes the original formulation of \( R_0 \) by directly computing the expected number of secondary infections generated by a single infectious individual over the
course of their infectious period and over the lifetime of the epidemic.

To begin, we first calculate the expected number of infections in each location at each time point. Due to the binomial nature of the stochastic transfer between compartments, this is simply $S_i(t)P_i(t)$. To compute the average per infectious individuals, we either divide by $I_i(t)$ if the number of infectious individuals is greater than zero, or define the average number of infections to be zero. As the viral pathogen does not spread across borders in this model, because infection only occurs after the movement of an infectious individual, the resulting matrix of expected infections is diagonal, simplifying the calculation over spatial locations substantially. While this is sufficient to describe the average number of infections at a given time point and spatial location, this is not exactly the expected number of infections caused by a single infectious individual over the course of their infectious period. Thus, the expected average number of infections must be adjusted to account for the entire lifespan of the pathogen. This is done via a weighting term which adjusts the expected number of infections. Therefore, within the proposed model the empirically adjusted reproduction number for each spatial location $i$ and time $t$ is defined to be

$$R_{i}^{(ea)}(t) = \sum_{\tau=t}^{T_i} G_i(\tau)e^{-\gamma(\tau-t)}$$

where

$$G_i(t) = \begin{cases} 
\frac{S_i(t)P_i(t)}{I_i(t)}, & \text{if } I_i(t) > 0 \\
0, & \text{if } I_i(t) = 0.
\end{cases}$$

4.4 Results and Discussion

As an application of the newly defined class of embedded stochastic models and the Bayesian analysis techniques for both parameter estimation and model selection, we explore modeling the Ebola Epidemic in West Africa between 2014 and 2016. The particular spatio-temporal dynamics demonstrated over the course of this epidemic present many modelling
challenges. This particular epidemic possesses some interesting spatial and temporal characteristics that make it particularly difficult to model. The diverse pathways of transmission add the high levels of uncertainty. Additionally, this particular epidemic spread throughout three separate countries in a distinctly heterogeneous way. The flexibility of the embedded stochastic model class defined in the Methods section will serve to address some of these issues by using a simplified compartmental model with a general spatial structure constrained only by the geography of the region.

4.4.1 Data: Ebola in West Africa between 2014 and 2016

The data used for this study was produced by the World Health Organization (WHO), which tracked the Ebola epidemic in the West Africa nations of Sierra Leone, Liberia, and Guinea between 2014 and 2016. The data is publicly available and can be downloaded from the WHO [72]. Raw case count data for Liberia, Sierra Leone, Guinea was obtained that spans the full length of the two year epidemic Figure 4.1. In particular, there are two sources of data - the Situation Report (SR) and the Patient Database (PD). Both sources detail confirmed cases and probable cases on a daily basis, and adding the two provides the total suspected cases over the course of the epidemic. In examining the characteristics of the data, it is strongly advised to use this sum rather than merely the confirmed case count, as many of the early cases that started the epidemic occur in the probable category. Only once the epidemic gained traction do these cases become categorized as confirmed. In addition, the SR does not measure the initial trajectory of the epidemic particularly well, as the epidemic had not gained enough attention, and so was not reported through a weekly situation report. Thus, we will concentrate on using only the data from the PD. Below, we provide the PD and SR data for the country-wide epidemic. One can verify a sharp start to the SR data in all three countries, while the SitRep begins to report data only around 40 weeks. The raw data is discretized to 15 counties in Liberia, 33 prefectures in Guinea, and 16 districts in Sierra Leone. These “state-level” divisions can be further grouped into sixteen distinct regions - five for Liberia, four for Sierra Leone, and seven for Guinea Figure 4.2.
Figure 4.1: **Raw Data**
Raw WHO data for Sierra Leone, Liberia, and Guinea for the 2014-2016 Ebola outbreak. The left hand panels show the data gleaned from Patient Databases (PD) and the right the Situation Report (SR). The black line in all panels represents the probable cases, the red all confirmed cases, and the green the total probable and confirmed cases.
Finally, the raw data can be aggregated into a country-level epidemic time series. For Sierra Leone and Guinea, these divisions are determined by existing administrative divisions in the country, while in Liberia the divisions are more arbitrary, as the country operates primarily at the county level. The availability of weekly, high spatial resolution data for each country facilitates the exploration of a variety of spatial structures to determine an appropriate and computationally feasible representation of the spatial heterogeneity. While the epidemic curves in Liberia and Guinea are typical of epidemic data, the super exponential rise in the Liberia data indicates reporting errors [70]. In general, models will be described and explored in order of increasing spatial complexity, because such complexity has a high computational cost. The ultimate goal is to identify and elucidate the simplest spatial model that achieves a desired level of accuracy and provides reasonable explanations for the spread of the epidemic.

4.4.2 Standard Model Structure, Parameters, Priors and Initial Conditions

For all explored spatial structures there are aspects of the model and Bayesian inference process that remain consistent due to the particular application and computational implementation. Primarily, as an infectious Ebola individual is extremely unlikely to be mobile due to the debilitating nature of the symptoms, our implementation of the model does not consider a mobile infectious class. Thus, for any value of \(i = 1, ..., k\) the model simplifies to

\[
\begin{align*}
S_i(t + \Delta t) &= N_i^S(t) - B_i(t) \\
E_i(t + \Delta t) &= N_i^E(t) + B_i(t) - C_i(t) \\
I_i(t + \Delta t) &= C_i(t) - D_i(t) \\
R_i(t + \Delta t) &= D_i(t)
\end{align*}
\]

with each population defined as in (4.3). Though the data is observed weekly, the model will consider a constant time step as \(\Delta t = 1\) day. An additional simplifying assumption relates to the scope of fluxes considered for a given spatial structure with \(k\) nodes. While it is entirely feasible to define the fully connected graph for a given number of nodes, in this application
our models are limited to defining fluxes between locations that share a physical land border. This is a simplifying assumption which is adopted due to the underdeveloped nature of the region, where frequent air travel is a negligible mode of transfer between locations. In a highly developed country, this simplifying assumption may need to be revisited. The general model structure accommodates a temporally- and/or spatially-dependent $\beta$ by redefining it as a function of $x$ and/or $t$. For this application it is appropriate to consider a temporally and spatially dependent $\beta_i(t)$ to approximate the effects of interventions that reduce the overall infection strength, as well as the inherent spatial heterogeneity of the spread. To simplify the parameter estimation, rather than estimating each daily infection rate we will
estimate a particular functional form of the infection rate, namely

\[
\beta_i(t) = \exp \left( b_0^i + b_1^i t + b_2^i t^2 \right).
\]

A given individual is infectious for 7 – 14 days (6 – 16 days if untreated) and is quantified in the model by \( \frac{1}{\gamma} \). This value is well known and simpler to measure because Ebola symptoms are clear and violent. We will treat this quantity as known in our modeling efforts, and it will be fixed so that \( \gamma = \frac{1}{10} \). In addition, due to the SEIR model structure \( \gamma \) is not structurally identifiable, so it is sufficient to merely estimate \( \beta \). With this model definition and the associated simplifying assumptions, the parameters that require estimation are a single latent period parameter \( \alpha \), the 3\( k \) infection rate functional coefficients \( b_0^i, b_1^i, b_2^i \), and each \( \rho_{ij} \) that is defined for a given spatial connectivity structure.

To ensure that a given set of models for a specific value of \( k \) can be compared to any other, the priors and initial conditions used for ABC SMC will be consistent across every model simulation. Initial conditions will be generated from the data for the exposed and infectious classes, while the susceptible class will be defined as the total population excepting the exposed and infectious individuals in a defined spatial location, as defined by the most current census at the time of the epidemic. Early simulations and tests aided in fixing priors that are centered in the parameter space in an effective manner to ensure a level of computational efficiency, while not impacting the verifiability of the posteriors by being overly informative. The latent time \( 1/\alpha \) has been estimated to be between 2 and 21 days, but a 4 – 10 day time span is typical. That being said, there is some evidence that the latent period for this particular epidemic was significantly longer, ranging up to 30 or 40 days. One of the reasons for this discrepancy is that the recommended quarantine period for individuals exposed to Ebola was 21 days \([34, 42, 77]\), but there were cases in which an individual who had been originally deemed uninfected, later presented symptoms after being quarantined. Test simulations responded poorly to a uniform distribution over the
Table 4.2: Standard Priors of Model Parameters

The standard priors used for each parameter in the model are detailed. Additional notes are included to explain non-intuitive aspects of the priors used.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Prior</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>$\gamma = 1/10$</td>
<td>Fixed for structural identifiability</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$\alpha \sim \Gamma(4, 20)$</td>
<td>Strictly positive Mean: 4 days Standard Deviation: 100 days</td>
</tr>
<tr>
<td>$\rho_{ij}$</td>
<td>$\rho_{ij} \sim \beta(2, 2) - .5$</td>
<td>Re-center Beta distribution to $[-0.5, 0.5]$ to propose positive and negative fluxes</td>
</tr>
<tr>
<td>$p_{ij}$</td>
<td>If $\rho_{ij} &gt; 0$ $p_{ji} \sim U(0, \frac{(1-</td>
<td>\rho_{ij}</td>
</tr>
<tr>
<td></td>
<td>If $\rho_{ij} &lt; 0$ $p_{ij} \sim U(0, \frac{(1-</td>
<td>\rho_{ij}</td>
</tr>
<tr>
<td>$b_{0,1,2}^i$</td>
<td>$(b_{0}^i, b_{1}^i, b_{2}^i) \sim N(\hat{\mu}, 3\hat{\sigma}^2)$</td>
<td>Test Simulations: $\hat{\mu} = (-1.38, -3.07, -6.47)$, $\hat{\sigma}^2 = (0.40, 3.03, 7.80)$</td>
</tr>
</tbody>
</table>

The standard range of latency, as the distinct lack of support for low values of alpha was not present. To address this, while still reflecting the higher likelihood of the standard latency periods, a gamma distribution reflecting a mean latency period of 5 days with a suitably large standard deviation was selected to serve as a prior for $\alpha$. The priors for the parameters in each simulation are defined in Table 4.2.

All simulations were run for a sequence of thresholds

$$\epsilon = (0.75, 0.65, 0.55, 0.50, 0.45, 0.40, 0.35, 0.30, 0.25, 0.20)$$

using the norms $N^n_i$, $n = 1, 2$ and $N^3$ (4.4) where for $n = 1, 2$ the norms are defined for each spatial location $i = 1, \ldots, k$. All norms must be satisfied in all spatial locations simultaneously to generate an acceptance. For the purposes of this data set, $T = 123$ weeks as the final reporting date. Each stage of SMC-ABC was simulated until a sufficient number
of acceptances was achieved, namely $N > 1000$, before calculating the adjusted prior for the next stage and the weights on the sample posterior from which the parameter would be drawn.

4.4.3 Three Location Model

A straightforward starting point to address the spatial heterogeneity of this epidemic is to divide the region of West Africa into three spatial locations mimicking the existing country boundaries. Thus, the original model will be implemented with $k = 3$. Within each spatial location the well mixed assumption holds, and the fluxes $\rho_{ij}$ allow for transfer between the nodes. As each of the three countries share a physical land border with the remaining two, this model considers defined fluxes that mirror the shared land borders and as a result, is fully connected. A schematic of this spatial structure is provided in Figure 4.3. With the
spatial structure defined, the initial conditions of the model are set with

\[ S(L, SL, G) = (4390737, 6232000, 10289445 - 2) \]
\[ I(S, SL, G) = (0, 0, 2) \]
\[ E(L, SL, G) = R(L, SL, G) = (0, 0, 0). \]

The priors are defined as in Table 4.2 for each location and connecting edge which is indicated by the graph structure shown in Figure 4.3. Results for each \( \epsilon \) threshold are summarized in Table 4.3. As expected the acceptance rate drops as the threshold becomes more restrictive but this model can generate \( N > 1000 \) accepted epidemics for each threshold. Figure 4.4

Table 4.3: **Computational Results:**
The number of simulations, the total number of accepted simulations and the resulting acceptance rate is reported for all stages of ABC-SMC for the three node model.

<table>
<thead>
<tr>
<th>Epsilon</th>
<th>Total Simulations</th>
<th>Total Acceptances</th>
<th>Acceptance Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>1000000</td>
<td>3060</td>
<td>3.06e-03</td>
</tr>
<tr>
<td>0.65</td>
<td>1000000</td>
<td>4739</td>
<td>4.74e-03</td>
</tr>
<tr>
<td>0.55</td>
<td>2000000</td>
<td>1694</td>
<td>8.47e-04</td>
</tr>
<tr>
<td>0.50</td>
<td>7400000</td>
<td>2559</td>
<td>3.46e-04</td>
</tr>
<tr>
<td>0.45</td>
<td>10000000</td>
<td>1660</td>
<td>1.66e-04</td>
</tr>
<tr>
<td>0.40</td>
<td>13550000</td>
<td>1159</td>
<td>8.55e-05</td>
</tr>
<tr>
<td>0.35</td>
<td>40700000</td>
<td>1186</td>
<td>2.91e-05</td>
</tr>
<tr>
<td>0.30</td>
<td>103950000</td>
<td>1359</td>
<td>1.31e-05</td>
</tr>
<tr>
<td>0.25</td>
<td>393680000</td>
<td>1013</td>
<td>2.57e-06</td>
</tr>
</tbody>
</table>

shows the accepted epidemics for \( \epsilon = 0.20 \) plotted against the observed data and the pointwise mean of the accepted simulations. It is clear that this model is able to produce acceptable epidemics under this norm paradigm; however, there are a few clear patterns of deviations that the model and norms are not able to resolve. Across all three spatial locations, the early epidemic behavior of the simulated epidemics is highly variable, often presenting double or triple of the observed cases in the first 20-30 weeks of the epidemic. This dynamic is particularly problematic in the node representing Guinea as there are simulated epidemics that present nearly the observed peak cases in the first 20 weeks. If this initial growth
Figure 4.4: **Accepted Epidemics**
The accepted epidemics for the three node model at the SMC threshold of $\epsilon = 0.20$ are plotted (grey) with the observed data (black) and the mean simulation (red).
in the simulations for all three nodes was limited to only a slight over estimation of case counts this discrepancy could be attributed almost in its entirety to surveillance errors in data collection. This dynamic often occurs early in epidemics as the focus and resource allocation for disease surveillance takes time to ramp up. Despite this potential bias in the data, it is unlikely the under count would amount to nearly 100 cases per week in each location over the course of nearly 20 weeks. Instead the rapid spread in all three locations might be attributed to an mis-characterization of the model which promotes case transfer via mobility at inflated rates early in the epidemic. Due to the binomial and multi-nominal nature of transfer between disease compartments and spatial location respectively, both the mobility and infection dynamics will be particularly sensitive to population sizes. This aligns with the striking deviations in the simulated epidemics for Guinea, as the total population of Guinea was 10 million as opposed to the 4 million and 6 million in Liberia and Sierra Leone. Building on this hypothesis, rather than adding additional norms to constrain the early disease dynamics we will explore a different spatial structure over which the model can be applied.

4.4.4 Four Location Model

As demonstrated by the three location model, the dynamics of spread in Sierra Leone and Liberia are adequately described when considered as a single spatial location despite the spatial heterogeneity of the epidemic itself, though there are systematic errors in the simulated epidemics early in the epidemic across all three countries. In particular, the three location model is inadequate to describe the spread of Ebola in Guinea using the well-mixed assumption for the entire country. One reason this difference could occur is that the overall outbreak in Guinea is smaller than those in Liberia and Sierra Leone, despite being the source of all cases in those two countries. It is well known that the well-mixed assumption provides a worse approximation of the disease dynamics of smaller epidemics [10, 16]. Another potential explanation that would account for the inadequacy of the three location model is the particular spatiotemporal dynamics of infection early in the course of
Figure 4.5: **Timing of Ebola Outbreak in Guinea**

The week in which each prefecture in Guinea recorded its first Ebola case is mapped with the dark colors indicating prefectures that observed cases earlier and the lighter colors representing locations which observed cases later. The range is 1 week to 57 weeks.

The epidemic. For this reason, it is valuable to consider the distinct order in which spatial locations within the country present new cases. Figure 4.5 illustrates the week in which each county observed its first Ebola cases. The epidemic begins in the southeastern section of Guinea within the Guéckédou prefecture that borders both Sierra Leone and Liberia. It rapidly spreads north reaching the northern border in the first 20 weeks. In the Western part of Guinea, the first case is observed in the coastal urban capital prefecture of Conakry in week ten. This instance of infection could have only stemmed from the southeastern part of the nation, as that region was the only one with observed cases in any of the three prefectures during that time. The exact mode of infection, which managed to span the entire breadth of the country, is unknown. However, we conjecture that after this initial transfer, the western epidemic spreads primarily independent of the outbreak in the eastern region. Indeed this concurrent spread is observed in other modeling efforts [48]. The first observed
cases are limited to the coastal prefectures during the first 20 – 25 weeks, and the only prefecture which shares a land border with the eastern region with cases is Tougué and it observes its first case in week 57, over a year after the epidemic started. There is effectively a physical separation between the two regions aligned with the least populous areas of the country which could serve to severely dampen population mixing between the two regions. Therefore, a reasonable model adjustment is to divide Guinea into two distinct regions, which we will denote by Guinea 1 and Guinea 2, representing the eastern and western collection of counties, which are each internally well-mixed. With this division, all four spatial locations have more similar population sizes, mitigating the impact that a disproportionately sized spatial location would have on disease spread and mobility. As the mobility between the these two regions and the other two countries is unknown, different assumptions can be explored via a set of spatial connectivity schemes.

We define three different graph structures, in which each subsequent scheme deletes an edge from the previous one, as seen in Figure 4.6. We begin with the fully connected graph, informed by the shared borders between the locations. If a region doesn’t share a physical border with another, we assume that the possibility of transfer between the two regions does not exist. The additional mobility limitations in the form of deleted edges are motivated by a similar argument that was used to split Guinea into two regions. Cases were first observed in Guinea 1, then Guinea 2, then Liberia and finally Sierra Leone. It is clear that the first cases in Libera and Sierra Leone are a result of spread from Guinea 1, as the outbreak in Guinea 2 remains isolated to the urban capital. Then we can additionally define three parallel models wherein nodes 3 and 4 (representing Guinea 1 and Guinea 2) are connected via a temporary edge, $\rho_{34}I_T(t)$. This only allows a positive flux $\rho_{34}$ to act on the nodes for a set period of time $[t^* - \tau, t^* + \tau]$. We implement a sparking event to simulate a sparking event which passes the infection from Guinea 1 to Guinea 2 that is informed by the observed geographical distinctness of the two regions of Guinea, especially the hubs of infection for the two regions that appear to be separated from each other. In all the following simulations
Figure 4.6: **Four Node Models**
Four node model schema with differing connectivity structures informed by geographical boundaries and timing of epidemic spread.

\[ t^* = 70 \text{ and } \tau = 7 \] limiting the mobility to a week before and a week after the first case is observed in Guinea 2. By setting these values, our simulations will more closely mirror the data, however, this is an assumption which could be relaxed by exploring a range of \( t^* \) and \( \tau \). As \( \rho_{34} > 0 \) and the primary purpose is to allow for a sparking event to occur through the mobility of infected individuals from Guinea 1 to Guinea 2, we will assume this mobility is unidirectional with \( p_{43} = \rho_{34} \) and \( p_{43} = 0 \). The schematics for each of these spatial connectivity schemes are illustrated in Figure 4.7. While these are not the only spatial connectivity schemes possible in the model space, by using intuition gained through analysis of the data, we are able to limit the scope of the models explored. This study serves as a demonstration of the application of the model class and the Bayesian inference methods for model selection and parameter estimation, not as a comprehensive exploration of the entire model space.

All six of these spatial connectivity schemes have consistent initial conditions defined by

\[
S(L, SL, G_1, G_2) = (4390737, 6232000, 4253117 - 2, 6036328) \\
I(S, SL, G_1, G_2) = (0, 0, 2, 0) \\
E(L, SL, G, G_2) = R(L, SL, G, G_2) = (0, 0, 0, 0).
\]
Figure 4.7: Four Node Models with Indicator
Four node model schema with differing connectivity structures and a time indicator function from Guinea 1 to Guinea 2 informed by geographical boundaries and timing of epidemic spread.

The priors are defined as in Table 4.2 for each location and connecting edge which is indicated by each of the graph structures shown in Figure 4.6 and Figure 4.7. SMC-ABC was run for all six spatial schemes beginning at an error threshold of $\epsilon = 0.75$ and ending at $\epsilon = 0.20$. Due to the obscure nature of the acceptable parameter space, embedded in the large proposed parameter space, the adjusted posterior variance was calculated to be $\tau_{(j+1)}^2 = 0.5 \hat{\sigma}^2$ to facilitate computational efficiency. This adjustment, carries the risk of a degenerate posterior distribution, however this risk is alleviated by the large number of acceptances required at each stage of SMC-ABC. Additionally tests of SMC-ABC with the original variance calculation showed only negligible differences between posteriors using the two different variances at early stages of the algorithm.

All six models proposed were able to generate acceptances at the final threshold, though clear differences in the suitability of the spatial structures become clear. Table 4.4 summarizes the results for the final stage of SMC-ABC with $\epsilon = 0.20$ for all six proposed model structures. Note that due to increasingly low acceptance rates for the less suitable models, not all schemes found $N = 1000$ acceptances in the final stage. An interesting dynamic that can be observed from the entirety of the the SMC results is that, the “best” model depends
Table 4.4: SMC-ABC Results for Six Schemes

The total simulations, acceptances and acceptance rate are reported for each of the six spatial mobility schemes when SMC-ABC uses a threshold of $\epsilon = 0.20$.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Epsilon</th>
<th>Total Simulations</th>
<th>Total Acceptances</th>
<th>Acceptance Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheme2i</td>
<td>0.20</td>
<td>9346996000.00</td>
<td>1011</td>
<td>1.08e-07</td>
</tr>
<tr>
<td>Scheme2</td>
<td>0.20</td>
<td>5645614000.00</td>
<td>448</td>
<td>7.94e-08</td>
</tr>
<tr>
<td>Scheme3i</td>
<td>0.20</td>
<td>9781627000.00</td>
<td>739</td>
<td>7.55e-08</td>
</tr>
<tr>
<td>Scheme3</td>
<td>0.20</td>
<td>2738693000.00</td>
<td>82</td>
<td>2.99e-08</td>
</tr>
<tr>
<td>Scheme1</td>
<td>0.20</td>
<td>3168408000.00</td>
<td>22</td>
<td>6.94e-09</td>
</tr>
<tr>
<td>Scheme1i</td>
<td>0.20</td>
<td>2089687000.00</td>
<td>12</td>
<td>5.74e-09</td>
</tr>
</tbody>
</table>

heavily on the error the modeler is willing to tolerate. At less stringent thresholds it is not clear which model(s) will generate the most acceptances at the later restrictive thresholds. Thus selecting a model prior to reaching a desired error tolerance is inadvisable. This is abundantly clear when one looks at the evolution of the Bayes factors for any given model shown in Figure 4.8. The stark contrast of the evolution of the Bayes Factors for Schemes 1 and 2i are demonstrative of this uncertainty early in the SMC-ABC Algorithm. At higher $\epsilon$ thresholds, Scheme 1 first displays Bayes Factors greater than one when compared to most models, but once $\epsilon \leq 0.35$ the values shift rapidly to indicate evidence for other models. The opposite dynamic occurs for Scheme 2i, where its comparatively low acceptance rate generates Bayes Factors that are significantly less than one until the same threshold is reached and the evidence in favor of Scheme 2i increases. While thresholds with $\epsilon > 0.35$ may be adequate for some modeling endeavours, this study aims to explore the models which are supported at low levels of error tolerance. Thus any ranking, or evidence in favor of a model will be analyzed using the Bayes Factors for when $\epsilon = 0.20$. The resulting matrix $\mathbf{B}$ is generated by dividing the acceptance rate for the model indicated by the column by the acceptance rate for the model indicated by the row, and is given by
Figure 4.8: Bayes Factors
Each facet shows the Bayes Factors calculated for a given model by dividing the acceptance rate of the titular model (Model 1) by the modeled indicated by the color scheme (Model 2) for each $\epsilon$ threshold used for SMC-ABC. The Bayes Factors are shown on a log scale and values greater than 1 indicate positive evidence for Model 1 while values less than 1 indicate evidence in favor of Model 2.

$$B = \begin{pmatrix}
1 & 1i & 2 & 2i & 3 & 3i \\
1 & - & 0.83 & 10.71 & 14.50 & 3.77 & 10.45 \\
1i & 1.21 & - & 12.97 & 17.55 & 4.57 & 12.65 \\
2 & 0.09 & 0.08 & - & 1.35 & 0.35 & 0.98 \\
2i & 0.07 & 0.06 & 0.74 & - & 0.26 & 0.72 \\
3 & 0.27 & 0.22 & 2.84 & 3.84 & - & 2.77 \\
3i & 0.10 & 0.08 & 1.03 & 1.39 & 0.36 & - \\
\end{pmatrix}. \quad (4.7)$$

Bayes factors generate pair-wise comparisons, not an overall group ranking, though a ranking can be implied from the values for the pair-wise comparisons. The first thing to note is that
the only proposed spatial structure $i$ with $B_{ij} > 1$ for all $j$ is Scheme 2i. Thus, by raw ranking of acceptance rates, Scheme 2i is the model which has the most evidence in the data in favor of its structure. This evidence is substantial in the case of Scheme 2i compared to Scheme 3, strong when compared to either Scheme 1 or 1i, but only slight when compared to Scheme 2 and Scheme 3i - see Table 4.1. In general, Schemes 2, 2i, and 3i stand apart from the other three models as displaying evidence in the data to support them over the others. Though the evidence in which ranks Scheme 2i over Scheme 3i over Scheme 2 is slight, the evidence for less connected spatial structures across the board is substantial as the most highly connected schemes, Scheme 1 and 1i, show strong evidence against them no matter which of the other models they are compared to. Additionally all spatial structures that included the sparking event between Guinea 1 and Guinea 2, in the form of the indicator function all have Bayes Factors greater than 1 when compared to their non-indicator parallel model. This is particularly striking, as the indicator function presents a large source of rejections independent of any parameter choice. It is extremely common that, due to the indicator function, the epidemic never spreads to Guinea 2, causing an otherwise favorable parameter scheme to be rejected. This is a built in limitation of the model structure, which should intuitively favor non-indicator schemes all else being equal. However all the spatial structures with indicator functions built in, overcame this limitation and performed on par or better than their counterparts.

Additional work on the three best schemes - Schemes 2i, 3i and 2 - in the form of additional runs at lower epsilon values and/or analysis on the quality of the acceptances from each scheme using the defined norms or subsequent metrics can be done to elucidate the differences between the three schemes. For example, Figure 4.9 shows the distribution of the values which the norms take for the three different schemes accepted epidemics. The top left panel, shows the distribution of all norms, and it is clear that Scheme 2 has a bias to accepted epidemics with norm values closer to 0.20. The remaining three panels separate the three norm components, and while the norm on the peak number of cases $N_i^1$ shows
little difference between the three schemes, the peak week norm $\mathcal{N}_i^2$ indicates that Scheme 2 encounters difficulties in this context, while the final week norm $\mathcal{N}_i^3$ demonstrates the advantage Scheme 2i has over the other two Schemes. This analysis does not serve as a replacement for further stages of SMC, however it can indicate the likelihood of acceptances overall at a lower $\epsilon$-threshold under each norm.

4.4.5 Scheme 2i Results

Accepted Epidemics
Figure 4.10: **Accepted Epidemics for Scheme 2i**

The accepted epidemics for the Scheme 2i four node model at the SMC thresholds of $\epsilon = 0.75, 0.50, 0.20$ are plotted (grey) with the observed data (black) and the mean simulation (red). While the mean epidemic changes minimally over the change in the threshold the range of the accepted epidemics varies much more at higher (less stringent) thresholds.
For the purposes of demonstration of the results we will limit further analysis to only those results from Scheme 2i. Figure 4.10 shows the evolution of accepted epidemics for \( \epsilon = 0.75 \), \( \epsilon = 0.50 \), and \( \epsilon = 0.20 \). As the threshold becomes more restrictive the mean epidemic gets shorter and the peak moves to the right (later in the epidemic) and increases in value across all four locations. We can see clearly that the main problematic feature from the three location model, where the early weeks of the epidemic were poorly modeled, particularly in Guinea, is resolved by the division of the country into two separate regions and the reduction of edges connecting Guinea 2 to the rest of the graph structure. In fact all three of the best model structures resolve this problem, further solidifying the hypothesis of disconnected concurrent spread in Guinea as well as the possibility of a sparking event as the mode of spread to the urban region of Guinea. None of these models seem to address the mismatch of peak timing for the epidemic in Liberia. The nature of the super exponential growth and subsequent at the entire country level is difficult to match with a SEIR structure, even with spatio-temporal infection and contact rates \( \beta_i(t) \). While further spatial divisions and other model structures may address this, it is likely that the data itself has not captured the true situation on the ground. This is particularly suspect in Liberia, were the super exponential rise of the data is indicative of reporting errors [70]. This may be due to a number of factors, such as the healthcare system being overwhelmed by the explosion of the epidemic or the governmental structures which produce the data, as unlike the other two countries, there is no unifying federal system which runs the public health response particularly the collection of data from each of the subregions of the country. Both situations are bound to introduce intrinsic errors in the data. Thus, as the dynamics fit the other three regions so well, our model could be revealing a more accurate picture of the truth of the epidemic in Liberia.

**Epidemiological Parameter Estimation**

The epidemiological parameters in the model are \( \alpha \) the inverse of the latent period and the coefficients \( b_{0,1,2} \) which define \( \beta_i(t) = \exp (b_0 + b_1 t + b_2 t^2) \). The SMC-ABC algorithm produces approximate posterior distributions for these parameters. Figure 4.11 shows the
Figure 4.11: **Posterior Distribution of the Latent Period Parameter** $\alpha$

The posterior distribution approximated when $\epsilon = 0.20$ of the parameter $\alpha$, where $\frac{1}{\alpha}$ represents the latent period, (blue) is compared to the prior (yellow). The mean value of the posterior is indicated by the blue dashed vertical line.

approximated posterior distribution for $\alpha$ as compared to its prior. Similar plots for each of the coefficients for the beta function are shown in Figure 4.12. A numerical summary for each of the epidemiological parameters can be found in Table 4.5. It is useful to more deeply examine the numerical values of $\alpha$ as it represents the time that an individual spends after being infected before becoming infectious. This category is particularly important in our model, as it facilitates the spread of infection through the mobility of the exposed class of individuals. The posterior in Figure 4.11 indicates a mean latent period of approximately 31 days but ranging from 11 days at minimum to 71 days at maximum. While the bulk of the distribution is centered at values that are higher than the biology suggests, as the range of the latency period for Ebola is supposed to range from 2-21 days while 4-10 days
Figure 4.12: Posterior Distribution of Coefficients $b_{i,1,2}^j$

The posterior distributions approximated when $\epsilon = 0.20$ of the parameters ($b_{0}^i, b_{1}^i, b_{2}^i$) for $i = 1$(Liberia), 2(Sierra Leone), 3(Guinea1), 4(Guinea2)) (blue) are compared to their priors (yellow). The mean value of each of the posteriors is indicated by the blue dashed vertical lines.

is typical, this dynamic is to be expected due to the simplified nature of our underlying compartmental model and the mode of infection transfer between multiple locations. To accommodate the absence of other infection pathways which generate spread such as hospital
The indices $i = 1, 2, 3, 4$ refer to Liberia, Sierra Leone, Guinea 1, and Guinea 2, respectively, and the CI is a 89% equal-tailed credible interval. These values are calculated for the distributions generated by $\epsilon = 0.20$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Median</th>
<th>ETI CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>0.0317</td>
<td>0.0288</td>
<td>[0.0175, 0.0545]</td>
</tr>
<tr>
<td>$b_0^3$ (G1)</td>
<td>-0.7258</td>
<td>-0.6834</td>
<td>[-1.89, 0.33]</td>
</tr>
<tr>
<td>$b_1^3$ (G1)</td>
<td>-4.3997</td>
<td>-3.8127</td>
<td>[-18.22, 7.69]</td>
</tr>
<tr>
<td>$b_2^3$ (G1)</td>
<td>-19.0920</td>
<td>-18.6516</td>
<td>[-42.59, 2.61]</td>
</tr>
<tr>
<td>$b_0^4$ (G2)</td>
<td>-1.7691</td>
<td>-1.7621</td>
<td>[-2.96, -0.57]</td>
</tr>
<tr>
<td>$b_1^4$ (G2)</td>
<td>3.6917</td>
<td>3.5251</td>
<td>[-4.05, 11.86]</td>
</tr>
<tr>
<td>$b_2^4$ (G2)</td>
<td>-13.7259</td>
<td>-12.9623</td>
<td>[-29.28, -0.41]</td>
</tr>
<tr>
<td>$b_0^1$ (L)</td>
<td>-1.2870</td>
<td>-1.1514</td>
<td>[-2.87, -0.01]</td>
</tr>
<tr>
<td>$b_1^1$ (L)</td>
<td>-2.3085</td>
<td>-1.3326</td>
<td>[-18.08, 9.90]</td>
</tr>
<tr>
<td>$b_2^1$ (L)</td>
<td>-18.8828</td>
<td>-17.8319</td>
<td>[-43.40, 3.55]</td>
</tr>
<tr>
<td>$b_0^2$ (SL)</td>
<td>-1.3620</td>
<td>-1.2574</td>
<td>[-2.89, -0.068]</td>
</tr>
<tr>
<td>$b_1^2$ (SL)</td>
<td>-3.1685</td>
<td>-2.3765</td>
<td>[-18.39, 9.92]</td>
</tr>
<tr>
<td>$b_2^2$ (SL)</td>
<td>-18.0510</td>
<td>-16.7445</td>
<td>[-47.95, 5.90]</td>
</tr>
</tbody>
</table>

Infections or infections caused by a deceased individual, the latency period becomes inflated to match the data and generate the observed spread. Particularly, as the only mode in this model framework of infection transfer between two spatial locations is that due to the movement of an exposed individual, there is an additional bias towards longer latency periods. Besides the model structure there has been some postmortem research after the West Africa Epidemic, showing that the standard quarantine of 21 days was insufficient as there were a non-negligible number of people with longer latency periods, a dynamic that was newly observed during this particular outbreak of EVD. While this research does not fully align with the latency periods this distribution suggests, it at least lends credibility to an over estimation of the mean latent period even without considering the model’s built in limitations.

Similar analysis of the numerical meaning of the coefficients $b_{0,1,2}$ is meaningless, as the three coefficients do not independently impact the dynamics of the simulated epidemics. Instead, it is necessary to examine the functional form of $\beta_i(t)$ for each set of accepted co-
Figure 4.13: **Functional Distribution of $\beta_i(t)$**

The distribution of the accepted functional values with $\epsilon = 0.20$ of $\beta_i(t)$ are plotted, with the highest density values at each time point denoted by the darkest red.

Efficients. To do so, the pointwise, temporal quantiles of $\beta_i(t)$ are calculated and plotted in Figure 4.13 where the color ranges from the darkest point representing the highest density value for $\beta_i(t)$ to the lightest representing the lowest density value. For reference, Figure 4.14 is the same plot generated by a random sample of the coefficients from the priors. It is clear that in all four locations there is a convergence to an estimate of the functional form for $\beta_i(t)$ for all $t > 0$. This is not due to a convergence of the coefficients themselves as Figure 4.12 demonstrates widely varying behavior of each individual distribution. Further-
Figure 4.14: **Prior Functional Distribution**
The distribution of prior of $\beta_i(t)$ is plotted as a reference, with the highest density values at each time point denoted by the darkest red.

more, examination of the evolution of the posteriors demonstrates that the approximations of individual posteriors do not follow a clear convergence pattern across all twelve coefficients. The level of convergence to a point estimate depends both on the spatial location and the time. Overall, Guinea 1 demonstrates the largest spread and variability across the epidemic, particularly at the beginning. This is understandable as the epidemic begins in Guinea 1 with a small number of cases, and any further propagation of the epidemic depends more heavily on chance rather than the precise value of an infection and contact rate. The variability later in the time series in all four locations can be mostly disregarded as much of that variability is counteracted by zero cases during that time period. Careful examination of the infectious probability

$$P_i(t) = 1 - \exp \left[ -\frac{\beta_i(t)}{N_i(t) I_i(t) \Delta t} \right]$$
Table 4.6: **Flux Parameter Means, Medians and Credible intervals**

The indices \( i = 1, 2, 3, 4 \) refer to Liberia, Sierra Leone, Guinea 1, and Guinea 2, respectively, and the CI is a 89% equal-tailed credible interval. These values are calculated for the distributions generated by \( \epsilon = 0.20 \). Recall that every flux is a dimensionless difference between two probabilities.

<table>
<thead>
<tr>
<th>Flux ((\rho_{ij}))</th>
<th>Mean</th>
<th>Median</th>
<th>ETI CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho_{12} )</td>
<td>-0.0579</td>
<td>-0.0610</td>
<td>[-0.1578, 0.0529]</td>
</tr>
<tr>
<td>( \rho_{13} )</td>
<td>0.2683</td>
<td>0.2736</td>
<td>[0.0672, 0.4578]</td>
</tr>
<tr>
<td>( \rho_{23} )</td>
<td>0.2686</td>
<td>0.2754</td>
<td>[0.0579, 0.4686]</td>
</tr>
</tbody>
</table>

confirms this intuition. Similar quantile plots for \( P_i(t) \) can be found in the appendix clearly demonstrate less variability when case counts in a spatial location are near or at zero.

**Mobility Parameters**

The fluxes \( \rho_{ij} \) between two spatial locations \( i \) and \( j \) and their underlying directional movement probabilities \( p_{ij}, p_{ji} \) are critical to the ultimate spread of the disease. These parameters are critical due to the design of the model, as infection can only spread across a border between two spatial locations via the movement of individuals. While the mobility of susceptible individuals impacts the overall number of potential contacts which could spread infection, the mobility of the exposed class is indistinguishable from their susceptible counterparts and is the primary mode of transmission across space. Thus the model can uncover the potential pathways of disease spread. A positive flux \( \rho_{ij} \) indicates an overall flow towards spatial location \( i \) from \( j \). Alternatively, a negative flux favors movement to location \( j \) from \( i \). Table 4.6 catalogues the mean, median, and 89% equal tailed credible interval for each of the three constant fluxes in this model while Figure 4.15 demonstrates the final posterior for each compared to the prior as well as joint posterior densities for the pairs of fluxes. The joint posterior densities aid in extending the understanding of the individual fluxes because even though the fluxes are treated as independent random variables in the parameter estimation, in the model, they do impact each other as the mobility of one location does not occur in a vacuum.
Figure 4.15: **Marginal and Joint Distributions of Flux Parameters**

Panels (A)-(C) are two dimensional joint density plots of pairs of the three fluxes $\rho_{12}, \rho_{13}, \rho_{23}$ with the lightest blue denoting the highest density region and the grey contours indicated equal density values in the plane. Panel (D) is plots the marginal distributions for each of the fluxes, $\rho_{12}$ (blue), $\rho_{13}$ (yellow), $\rho_{23}$ (green) compared to their prior (black). All distributions are generated with $\epsilon = 0.20$. 
These results signal a few notable dynamics. First, the flux between locations 1 (Liberia) and 2 (Sierra Leone) is near zero, with the weight of the distribution slightly favoring movement towards Sierra Leone. With a near zero flux, it is worth examining the individual probabilities that result from said flux. If the mobility between each location is high in both directions, this could be compelling evidence for the merger of the two locations due to the high level of mixing present. The high level of mobility is not present in these results however. Figure 4.16 demonstrates in the panels for Liberia and Sierra Leone that the dominating probability for each location is the probability that an individual, exposed or susceptible, does not change location. The probabilities \( p_{11} \) and \( p_{22} \) are greater than 0.50 and have means close to 0.75 while the probabilities \( p_{12} \) and \( p_{21} \) are less than 0.25 and means close to 0.10. Clearly, the system considers these two spatial locations as distinct, with only small levels of movement being necessary and sufficient for the spread of the infection between the two locations.

In contrast to the dynamics between Sierra Leone and Liberia, the mobility of individuals in Guinea 1 is drastically different. Both \( \rho_{13} \) and \( \rho_{23} \) indicate an overall flow of individuals away from Guinea 1. The mean for both is 0.26 meaning that the probability of leaving Guinea 1 for one of the other locations is on average 0.26 higher than the probability of entering Guinea 1. In fact credible intervals indicate that that 89% of all realizations of both fluxes are positive, meaning that a different mobility dynamic favoring entering Guinea 1 is very rare. A high level of flow away from Guinea 1 matches intuition both about the model and the situation on the ground observed during the epidemic. With regards to the model, the epidemic is started on Day 0 in Guinea 1, and thus there must be movement from Guinea 1 to at least one of the other spatial locations to perpetuate the spatial spread. Additionally as the data suggests, the case counts in Guinea 1 remain relatively low throughout the epidemic, while the bulk of the infection occurs in Sierra Leone and Liberia. There is evidence that the case counts maintain low levels in Guinea 1 for longer periods of time due to rare re-seeding of epidemics from Liberia and Sierra Leone[33]. This matches a mobility dynamic.
which while favoring leaving Guinea 1, does maintain a low, but consistent probability of entry. From what we know on the ground is that the borders at the nexus of these three regions are extremely porous, and though the region is technically divided between three countries for administrative and political purposes, it was difficult to implement closed borders to slow the spread of the virus from Guinea 1 [46]. Despite the clarity of the overall flux away from Guinea 1, the distribution of these positive fluxes have wider spreads, indicating a larger range of acceptable movement dynamics within that limitation. This is especially clear when observing the individual probabilities for Guinea 1 as seen in the final panel of Figure 4.16. The three distributions of the probabilities $p_{33}, p_{31}, p_{32}$ have wide ranges of high density regions ranging anywhere from 0 to nearly 0.75 for all three probabilities. Clearly, the three probabilities must sum to one, so all three cannot be simultaneously either high or low, instead they must balance under the constraints of the fluxes. Up to this point in the analysis, each parameter has been mostly treated as independent, with its individual posterior distributions and summary statistics being compared to other parameters. Analyzing various joint probability density functions of combinations of the parameters can aid in elucidating the shared dynamics. Looking back at Figure 4.15, the joint densities of pairs of the fluxes are visualized using contour and color density plots. The regions of lightest blue are the regions with the highest joint density. In this case the modal regions align well with the modes of the individual posteriors. From this it can be deduced that the average behavior of the fluxes overall is that there is a near zero, though slightly negative flux between Liberia and Sierra Leone, the fluxes involving Guinea 1 are likely to be large and equally positive, favoring movement away from Guinea 1. Additionally, as $\rho_{12}$ decreases, $\rho_{23}$ decreases slightly and $\rho_{13}$ increases slightly to compensate. Similarly the relationship between $\rho_{23}$ and $\rho_{13}$ is definitively negative with $\rho_{13}$ increasing to compensate for a reduction in $\rho_{23}$.

Furthermore, it is possible to calculate the 3D joint kernel density estimate to find the three fluxes values that occur simultaneously with the highest density. This can also be done for the three probabilities that result for each node. The modal values for each of
Figure 4.16: Distributions of Movement Probabilities
The distribution of the three probabilities generated by the fluxes are plotted for each spatial location besides Guinea 4. Guinea 4 does not have mobility in this scheme. All distributions are generated with $\epsilon = 0.20$.

These joint distributions are detailed in Table 4.7. Across the board the modal values align with high density regions of the individual distributions, but the interactions between the three different variables may shift the values slightly from the distinct means or medians. For example the means and medians of $\rho_{13}$ and $\rho_{23}$ are nearly identical, but the joint modal values indicate slightly more movement from Guinea 1 to Sierra Leone than to Liberia. This dynamic is only revealed by exploring the joint posteriors.

Recall that the two schemes that produce the most acceptances after Scheme 2i are Schemes 2 and 3i. Necessarily Scheme 3i has $\rho_{23} = p_{23} = p_{32} = 0$, however both the overall flux and the corresponding probabilities in Scheme 2i and 2 are not all equal to zero. Thus to produce acceptable epidemics without this pathway, Scheme 3i must compensate in with adjustments to the remaining probabilities and fluxes in its mobility structure. Figure 4.17
Table 4.7: Joint Modal Flux Values

The mode of joint distribution of the three fluxes allows for the calculation of the joint modal values of the three fluxes for $\epsilon = 0.20$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\rho_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Modal Value</td>
<td>-0.060</td>
</tr>
</tbody>
</table>

and Figure 4.18 compare the distributions of the fluxes and probabilities connecting nodes 1 (Liberia), 2 (Sierra Leone), and 3 (Guinea 1) for the three best schemes. For the purposes of this analysis, the fluxes and probabilities relating to the fourth location (Guinea 2) are not considered as the probability of staying in Guinea 2 is consistently greater than 0.75, so the mobility involving this location is minimal for all three schemes. Clearly the differences between both the fluxes and the associated probabilities for Schemes 2 and 2i are minimal as the displayed distributions for each of the shared parameters are near mirrors of each other. As such the remaining difference between the two mobility structures is the constant mobility allowed between Guinea 1 and 2 in Scheme 2. On the other hand, there are distinguishable differences between the distributions for fluxes and probabilities when comparing Scheme 3i to Schemes 2 and 2i. The mean of $\rho_{12}$ and $\rho_{13}$ does not change significantly, however the spread of the distribution narrows, indicating a more narrow range of feasible flux values.

Though the range of $\rho_{12}$ is smaller, there is negligible change in the distributions of $p_{12}$ and $p_{21}$ for Schemes 2i and 2 as opposed to Scheme 3i indicating that the probability pairs are more constrained to each other as compared to limited in individual values. Thus the balance of the movement probabilities needs to be more precise for the epidemic to progress in an acceptable manner with the mobility structure defined for Scheme 3i. In contrast, there is an obvious change for $p_{13}$ which contributes to less standard deviation in $\rho_{13}$. Though the probability of moving from Liberia to Guinea 1 is still low, it is more consistently near to the mean probability of 0.10. As a result, for acceptable epidemics to be simulated using
Figure 4.17: Distribution of Fluxes for Three Schemes

The distributions of the fluxes $\rho_{ij}$ with $i, j \neq 4$ associated with accepted epidemics with $\epsilon = 0.20$ for Schemes 2, 2i, and 3i demonstrate the mobility differences enforced by the schemes.
Figure 4.18: Distribution of Probabilities for Three Schemes
The distributions of the probabilities $p_{ij}$ with $i, j \neq 4$ associated with accepted epidemics with $\epsilon = 0.20$ for Schemes 2, 2i, and 3i demonstrate the mobility differences enforced by the schemes.
Scheme 3i, a low consistent level of movement from Liberia to Guinea 1 is required to balance a moderate probability in the other direction.

In addition to the changes in the fluxes there are changes to the probability of staying in Guinea 1 and Sierra Leone when employing Scheme 3i. As movement is no longer permitted between those locations, an additional likelihood of staying in either of those locations is predictable. There is a slight rightward shift towards higher probabilities of staying in Sierra Leone, and a significant shift towards a probability of one for staying in Guinea 1. Thus one of the primary accommodations for zero flux and zero probability of movement are a combination of low, consistent movement from Liberia to Guinea 1, more precise balance of probabilities $p_{12}, p_{21}$ and $p_{13}, p_{31}$, and finally higher probabilities of staying in Sierra Leone and Guinea 1. This implies that though a mobility structure with movement among Guinea 1, Liberia, and Sierra Leone is favored, as demonstrated by Scheme 2i, there is an alternative mobility structure that is more static that can produce acceptable epidemics.

**Basic Reproduction Number**

As detailed in the Methods section, there are a variety of tools that modelers use to quantify the epidemiological definition of the basic reproduction number. Recall that from an epidemiological perspective the basic reproduction number is the average number of secondary infections caused by a single infectious individual over their infectious period. Often this is used in the disease free case to analyze under certain parameter schemes whether an epidemic is likely to persist or die out. In the midst of an epidemic these $R$ values are generally tracked over time and space and used to indicate further growth $R > 1$ or decay $R < 1$ of the epidemic. The underlying compartmental model suggests one such quantification of the basic reproduction number (4.5). It is sufficient to use this definition to explore the types of reproduction numbers suggested by the underlying compartmental structure as the infection and contact rate $\beta$ varies in space and time and scaling by the proportion of susceptibles in the population is negligible because $\frac{S_i(t)}{N_i(t)} > 0.99$ for all $i = 1, 2, 3, 4$ and $t \geq 0$. Alternatively, the empirically adjusted reproduction number (4.6) returns to
the epidemiological definition of the basic reproduction number and uses the parameters to quantify the expected number of secondary infections resulting from one additional infectious individual at time $t$ and spatial location $i$ and adjusts this average to account for the infectious period by weighting by a pathogen lifespan term that monotonically approaches zero. Then, the sum of the adjusted average number of secondary infections occurring after time $t$ is calculated for each time point to get the basic reproduction number at each time and spatial location. While both the parameterized reproduction number $R_{i0}(t)$ and the empirically adjusted reproduction number $R_{i}^{(ea)}(t)$ represent the same epidemiological value in theory, in practice, depending on the model, can become entirely meaningless. Figure 4.19 demonstrates how the calculating $R_{i0}(t)$ and $R_{i}^{(ea)}(t)$ with the accepted sample parameter set and epidemic curves can communicate very different meanings with regards to the growth or decay of the epidemic. Notably, the parameterized reproduction number $R_{i0}(t)$ suggests that the epidemic is decaying (i.e., values less than one) in three of the four locations in the spatial locations at a time where the epidemics are clearly growing in size. Namely we know that the cases are rising until around week 40 (day 280) in all of the regions. The exponential rise at the end of the epidemic of $R_{i0}(t)$ is an artifact of $\beta_i(t)$, that could be accounted for through additional adjustments, however it is clear that this version of the reproduction number is basically meaningless in the context of this model. In contrast the average dynamics of $R_{i}^{(ea)}(t)$ appear to match the general behavior of both the observed data and the accepted simulated epidemics. Figure 4.20 shows more detailed plots with the quantiles of $R_{i}^{(ea)}(t)$ calculated from the accepted sample from SMC represented by the range of color darkness, with the darkest region representing the median value of $R_{i}^{(ea)}(t)$ at each time point. While the accepted sample of parameters and epidemics do produce variability in the empirically adjusted reproduction number, there is a clear trend that the time series for each spatial location follows. The peak and the decay occur in the same places, with only the magnitude presenting the uncertainty. The level of uncertainty depends on the spatial location. Clearly $R_{4}^{(ea)}(t)$ (Guinea 2) has low levels of uncertainty across the board,
Figure 4.19: **Basic Reproductive Numbers** \( R_i^0(t) \) **versus** \( R_i^{(ea)}(t) \)

The mean value of \( R_i^0(t) \) (dashed) and \( R_i^{(ea)}(t) \) (solid), calculated from the accepted epidemics and parameter sets when \( \epsilon = 0.20 \) are plotted for each spatial location. The grey dashed line on all plots highlights the threshold of one.

indicating the dynamics of that location are very distinct and must follow certain trends to produce acceptable epidemics. The increasing uncertainty in the other three regions with Guinea 1 demonstrating the most uncertainty, is intuitive, due to the connections between the spatial locations, and the reality that a secondary infection (a new exposed individual) may not become infectious and cause its own secondary infections in the location in which it was generated.
Figure 4.20: **Functional Distributions of $R_{ia}(t)$**

The distribution of the accepted functional values with $\epsilon = 0.20$ of $R_{ia}(t)$ are plotted, with the highest density values at each time point denoted by the darkest color. The grey lines represent the 50% (solid) and 90% (dashed) quantiles. The horizontal grey dashed line on all plots highlights the threshold of one.

Finally, even though this model formulation avoided implementing or encoding any interventions that would reduce transmission within regions, or between regions, the calculated reproduction number may uncover links between interventions on the ground and the average number of secondary infections. Additionally, it can be useful in formulating hypotheses concerning the efficacy of different intervention strategies or the accuracy of real time analysis. Figure 4.21 shows the mean of $R_{ia}(t)$ for each spatial location, while highlighting various timestamps of events that occurred in the location [5]:

- Day 90: WHO publishes its first report on the crisis, coinciding with the first reported cases in Liberia and Sierra Leone
Figure 4.21: $R_i^{(ea)}(t)$ in Situational Context
The mean of $R_i^{(ea)}(t)$ for each spatial location is again plotted, however, dotted vertical grey lines annotate each spatial location, indicating epidemic days which correspond to interventions or other events in each region.

- Day 170: Sierra Leone closes all schools and closes borders to Liberia and Guinea
- Day 216: Liberia closes borders and brings in military to enforce quarantines
- Day 220: Sierra Leone brings in military to enforce quarantines
- Day 270: Aid workers and journalists attacked and killed in Nzérékoré (Guinea 1). Reflects distrust of international response.
- Day 283: Festivals for the holiday season prohibited in Guinea
• Day 310: Liberia reports slowed transmission rates due to improved cultural mortuary practices

• Day 394: Guinea, Sierra Leone, and Liberia report lowest transmission rates since August 2014.

The timestamp on day 90 aligns with the peak value of the reproduction number in Guinea 1, Sierra Leone and Liberia, signaling that once attention was paid to the epidemic, in the form of WHO awareness and tracking, the number of secondary cases starts to decay in at least in those three regions. This does not mean the epidemic is not propagating, as the reproduction values stay above 1, instead it signals that the rate of spread is not accelerating. An attempt at closing the borders is made in Sierra Leone on Day 170 and in Liberia on Day 216, a change that the model does not impose, since the fluxes are constant. It is not clear how effective this intervention is because while $R_{(ea)}^i(t)$ continues to decrease from its peak, it is still well above 1 after Day 170/216, and only gets below 1 on average after Sierra Leone enlists the help of the military to enforce quarantines on Day 220 and after the implementation of improved mortuary practices in Liberia that were reported to have been effective on Day 310, but occurred prior. Epidemiologists have noted that the porous borders in the region greatly contributed to the spread of the disease, supporting the hypothesis that closing the borders was not entirely successful at halting the spatial spread. In Guinea, distrust of international aid was high, which mitigated the impacts of interventions to slow the spread. When aid workers and journalists are killed in a prefecture in Guinea 1 on Day 270, this challenge of successful interventions is highlighted by $R_{(ea)}^i(t) \approx 2$ in both Guinea 1 and Guinea 2 representing still rapid spread in those two regions. It is only after subsequent efforts by the Guinean government, like cancellation of festivities on Day 283, that the reproduction number experiences decay across all three regions. Finally, on Day 394 all three countries reported their lowest transmission rates since August 2014, which matches the dynamics of the reproduction number this model suggests as this is the first time in which all four locations have a mean $R_{(ea)}^i(t) \leq 1$ signalling the eventual extinction
of the epidemic.

### 4.5 Conclusions

This chapter introduces a spatially-heterogeneous extension of an embedded stochastic SEIR model that considers the spread of an infectious disease across a graph structure with \( k \) different spatial locations and mobility along the edges of the graph by the various disease compartment populations. As demonstrated by the application of the model to the Ebola Outbreak in West Africa between 2014 and 2016, the proposed class of models can be adapted to fit a variety of spatial structures and mobility schemes to capture standard infectious disease spread, as well as rare behaviors such as sparking events. Additionally, the SEIR structure allows for estimation of important epidemiological quantities and analysis of the basic reproductive number, which can be compared and validated against other methods and models. The design of the class of models allows for estimation of parameter distributions and model fits via Bayesian inference, as well as model comparison using Bayes Factors. The use of Approximate Bayesian Computation methods, particularly Sequential Monte Carlo ABC methods, allows for greater flexibility and increased computational efficiency. This additional efficiency is significant, as the process of parameter estimation and model fits of the stochastic model can incur a high computational cost depending on the rarity of the events captured in the data and the size and complexity of the model. The range of complexity available within the class of models can be scaled up or down depending on the application, but in addition to a possible increase in computational complexity, ABC methods require norms that gauge “closeness” of proposed simulations, which must be tailored to the model proposed. The summary norms detailed in the Methods section provide more flexibility and a greater rate of acceptances even though the number of norms increases with the number of spatial locations. The norms themselves are also subject to additional scrutiny, as various adjustments to the existing norm structure may be appropriate. For example, rather than a single norm evaluating the end of epidemic, in this case it could have been valuable to introduce \( k \) separate norms that enforce an appropriate duration of the epidemic at each
spatial location. Overall, the flexibility of the model and associated analytic methods is both an advantage to exploit and a challenge that must be addressed depending on the application.

An ongoing challenge for future applications is to address the present procedure for determining the spatial mobility structure. The number of spatial locations and the mobility structures in this study are determined *ad hoc*, but in theory the space of potential spatial mobility structures is infinite. This study limited the exploration space by starting with small numbers of spatial locations and subsequently imposing assumptions on the mobility that are based on the observed data. However, it is possible to relax some of these constraints and build a system by which the model space can be explored more holistically. For example, the four location model is a step in the direction of further systematization, as a similar method could be applied to generate a number of mobility schemes for a system with $k$ spatial locations. Additional complexity in the application of the operator that introduces “sparking events”, by varying the timing, duration, and location of such events, could also be considered when constructing a desired model space to explore. An algorithmic method for determining spatial structures would be particularly valuable in the case of an ongoing crisis, as the course of the epidemic and the particulars of future spatial spread would necessarily be unknown, and thus well suited to the adaptable nature of the class of models proposed herein.
The work in this thesis canvassed statistical and mathematical tools and data analytic strategies to provide insight into the spatio-temporal spread of infectious diseases around the world. We examined a recent Dengue epidemic in Brazil and a massive outbreak of Ebola in West Africa as particular case studies for which data - historical clinical surveillance and proxy data alike - and models can be used to enhance the study of infectious disease in highly spatially heterogeneous regions.

Novel data analysis methods and simple linear models were employed along with large spatial proxy data to model the risk of Dengue in Brazil at a fine spatial resolution in Chapters 2 and 3. The environmental and demographic data sets gathered to serve as a proxy for the difficult to observe mechanisms of mosquito borne disease were large and contained a significant amount of redundancy among the measured variables. To address this issue prior to building models, we proposed a novel multi-objective feature selection method which exploited the correlation structures in the data to cluster and extract representative variables that facilitate the construction of lower dimensional data sets. The method considers the optimization of two distinct objectives, namely the simultaneous minimization of information loss and final dimension of the subset. As a result, the feature selection method is unsupervised, i.e. independent of a model or response variable, similar to the standard unsupervised feature extraction method PCA. Unlike PCA, however, the proposed method possesses the highly-desirable property that all original variables are maintained throughout the dimension-reduction process.

Chapter 3 uses reduced data sets generated by the feature selection method to build linear models for yearly Dengue burden. While simple conceptually, the linear models furthered two distinct research goals. Initially, by training these models on two data sets of different sizes,
the analytic and computational benefits of our novel dimension-reduction method, utilized prior to designing and building a model, are demonstrated. From the presented analysis, it is clear that larger data sets required significantly more time investment in the build stage, but did not generate significantly more accurate results than their smaller counterparts. The linear models also served to elucidate potential mechanisms of Dengue risk, either region-specific or more general. Both the regional and country scale models utilized a mixture of demographic and environmental predictor variables to build parsimonious and accurate models. Thus, besides the implications of the models themselves on the understanding of spatially-heterogeneous Dengue risk profiles in Brazil, this effort also provides the impetus to explore the utility of diverse sets of proxy data in any epidemiological model for the spread of Dengue.

The design and implementation of a spatially-heterogeneous extension of a embedded stochastic SEIR model to the 2014-2016 Ebola outbreak is detailed in Chapter 4. This study, compared to the Dengue study, focused much more on model design and analysis rather than methods of data analysis. A spatial mobility graph structure was overlaid on a stochastic SEIR disease process to allow for the movement of individuals from one location to another, which serves as the only mode of spatial infection spread. The design of the model supported the use of ABC, a Bayesian inference method, to simultaneously perform parameter estimation and model selection by using clinical surveillance data to compare model simulations to the observed data. In this way various spatial mobility structures could be investigated by varying, within a set number of spatial locations, the connectivity of the graph structure. In addition, distributions for epidemiological parameters, such as the location and time specific infection/contact rate and the universal latent period, were generated thereby allowing for the analysis of an empirically adjusted reproduction number $R_i^{(ea)}(t)$ for each spatial location $i$.

The contrasts of the two approaches to studying Dengue and Ebola in this thesis provide the direction and motivation for further developments in the mathematical and statistical
study of infectious disease. The study of Dengue was focused on the use of data and modeling at a fine spatial resolution, which precipitated limitations in model complexity, while the approach to Ebola stands in contrast with a relatively coarse spatial resolution but a considerable increase in model complexity. Future applications in this direction will continue to bring these two juxtaposed approaches closer together by applying increasingly nuanced models to explore the mechanisms of spatial spread of infectious disease throughout the world.
REFERENCES CITED


Appendix A includes figures which supplement the results in Chapter 3.

Figure A.1: **Environmental Binary Correlation Plot**
The binary correlation between all pairs of tokens derived from the variable names of the Environmental data set are plotted. Dark blue indicates high positive correlation and thus tokens will nearly always occur in the same cluster. Dark red indicates high negative correlation, thus the pair of tokens nearly always occur in different clusters. Colors close to white indicates indicate token pairs that occur in the same or different clusters randomly.
Figure A.2: Demographic Binary Correlation Plot

The binary correlation between pairs of tokens derived from the variable descriptions of the Demographic data set are plotted. This plot only includes a subset of the most frequent tokens. Dark blue indicates high positive correlation and thus tokens will nearly always occur in the same cluster. Dark red indicates high negative correlation, thus the pair of tokens nearly always occur in different clusters. Colors close to white indicates indicate token pairs that occur in the same or different clusters randomly.
Appendix B includes figures and tables which supplement the results in Chapter 4.

Figure B.1: Regional Model Testing Fits
The fits of each model, broken down by region, trained either on large or small predictor data, to the testing data for the regional scaled models is demonstrated via a scatter plot of the scaled observations versus the scaled predictions. The dashed black line represents a perfect linear fit, or where the observation equals the prediction.
Table B.1: **Regional Model Errors**

Errors broken down by region generated by the training and testing data for the regional scale models trained on the small and large sets of predictor variables. Errors are calculated with unscaled and un-transformed prediction and observed case counts.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Region</th>
<th>Train/Test</th>
<th>RRMSE</th>
<th>RMAE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>L 1</td>
<td>Test</td>
<td>0.23</td>
<td>0.17</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>S 1</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>L 2</td>
<td>Test</td>
<td>0.31</td>
<td>0.25</td>
<td>0.14</td>
<td></td>
</tr>
<tr>
<td>S 2</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>L 3</td>
<td>Test</td>
<td>0.17</td>
<td>0.13</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>S 3</td>
<td>Test</td>
<td>0.12</td>
<td>0.09</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>L 4</td>
<td>Test</td>
<td>0.27</td>
<td>0.23</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>S 4</td>
<td>Test</td>
<td>0.24</td>
<td>0.20</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>L 5</td>
<td>Test</td>
<td>0.13</td>
<td>0.09</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>S 5</td>
<td>Test</td>
<td>0.11</td>
<td>0.08</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>L 1</td>
<td>Train</td>
<td>0.14</td>
<td>0.10</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>S 1</td>
<td>Train</td>
<td>0.15</td>
<td>0.10</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>L 2</td>
<td>Train</td>
<td>0.17</td>
<td>0.12</td>
<td>0.60</td>
<td></td>
</tr>
<tr>
<td>S 2</td>
<td>Train</td>
<td>0.18</td>
<td>0.12</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>L 3</td>
<td>Train</td>
<td>0.15</td>
<td>0.11</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>S 3</td>
<td>Train</td>
<td>0.16</td>
<td>0.12</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>L 4</td>
<td>Train</td>
<td>0.18</td>
<td>0.13</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>S 4</td>
<td>Train</td>
<td>0.19</td>
<td>0.14</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>L 5</td>
<td>Train</td>
<td>0.12</td>
<td>0.08</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>S 5</td>
<td>Train</td>
<td>0.13</td>
<td>0.08</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>
Table B.2: **Models Trained on Small Data Set Errors**
Errors broken down by region generated by the training and testing data for the country and regional scale models trained on the small set of predictor variables. Errors are calculated with unscaled and un-transformed prediction and observed case counts.

<table>
<thead>
<tr>
<th>Model Scale</th>
<th>Region</th>
<th>Train/Test</th>
<th>RRMSE</th>
<th>RMAE</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country 1</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Regional 1</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Country 2</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>Regional 2</td>
<td>Test</td>
<td>0.21</td>
<td>0.15</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Country 3</td>
<td>Test</td>
<td>0.15</td>
<td>0.12</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>Regional 3</td>
<td>Test</td>
<td>0.12</td>
<td>0.09</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Country 4</td>
<td>Test</td>
<td>0.24</td>
<td>0.20</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>Regional 4</td>
<td>Test</td>
<td>0.24</td>
<td>0.20</td>
<td>0.63</td>
<td></td>
</tr>
<tr>
<td>Country 5</td>
<td>Test</td>
<td>0.14</td>
<td>0.09</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Regional 5</td>
<td>Test</td>
<td>0.11</td>
<td>0.08</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Country 1</td>
<td>Train</td>
<td>0.18</td>
<td>0.13</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Regional 1</td>
<td>Train</td>
<td>0.15</td>
<td>0.10</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>Country 2</td>
<td>Train</td>
<td>0.19</td>
<td>0.14</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Regional 2</td>
<td>Train</td>
<td>0.18</td>
<td>0.12</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Country 3</td>
<td>Train</td>
<td>0.18</td>
<td>0.14</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>Regional 3</td>
<td>Train</td>
<td>0.16</td>
<td>0.12</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Country 4</td>
<td>Train</td>
<td>0.20</td>
<td>0.16</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Regional 4</td>
<td>Train</td>
<td>0.19</td>
<td>0.14</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>Country 5</td>
<td>Train</td>
<td>0.15</td>
<td>0.10</td>
<td>0.53</td>
<td></td>
</tr>
<tr>
<td>Regional 5</td>
<td>Train</td>
<td>0.13</td>
<td>0.08</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>