Spatiotemporal Patterns and Models of Ebola Virus Disease Epidemic in West Africa

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Sciences

in the

College of Graduate Studies

University of Idaho

By

Vasile A. Suchar

Major Professor: Michelle M. Wiest, Ph.D.

Committee Members: Christopher Williams, Ph.D.; Erkan O. Buzbas, Ph.D.

Department Administrator: Christopher Williams, Ph.D.

May 2018

Authorization to Submit Thesis

This thesis of Vasile A Suchar, submitted for the degree of Master of Science with a Major in Statistical Science and titled "Spatiotemporal Patterns and Models of Ebola Virus Disease Epidemic in West Africa," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: _____ Date: _____

Michelle M. Wiest, Ph.D.

Committee Members: _____ Date: _____

Christopher Williams, Ph.D.

_____ Date: _____

Erkan O. Buzbas, Ph.D.

Department

Administrator: _____ Date: _____

Christopher Williams, Ph.D.

Abstract

The purpose of this study was to investigate the utility of exploratory analytical techniques using publically available data in informing interventions in case of outbreaks infectious diseases. More exactly, spatiotemporal and multivariate methods were used to characterize the dynamics of the Ebola Virus Disease (EVD) epidemic in West Africa, and propose plausible relationships with demographic/social risk factors. The analysis showed that there was significant spatial, temporal, and spatiotemporal dependence in the evolution of the disease. For the first part of the epidemic, the cases were highly clustered in a few administrative units, in the proximity of the point of origin of the outbreak, possibly offering the opportunity to stop the spread of the disease. Later in the epidemic, high clusters were observed, but only in Liberia and Sierra Leone. The spatial-temporal models suggested that Montserrado area was a source of new cases for the neighboring areas, while NW Sierra Leone region was a sink for new cases from the neighboring areas. Also, there were region-specific population responses to the outbreak. Social attributes effects were significant - although small – pointing towards the importance of hypothesized social attributes of the population in the outbreak dynamics. Overall, the analysis suggests that infrastructure, access to and use of health services, and connectivity possibly accelerated and magnified the spread of EVD. The spatial, temporal, and spatiotemporal patterns of epidemic can be clearly shown – with evident application in the early stages of management of epidemics. Moreover, spatial-temporal models with fairly high predictive power could have been proposed even during the peak of the EVD epidemic – given that the data were available. But these efforts were not possible due to limited access or absence of quality data. While many of these deficiencies might have been unavoidable due to the severity of the epidemic and the limited resources of both affected countries and intervening agencies, open access to quality data, integration of local knowledge and customs, and continuous monitoring at the lowest spatiotemporal resolution possible during outbreaks might be useful, save time, resources, and allow for more effective decision support tools to be created in real time in the future. There are six supplemental information (SI) files: SI 1 covers the methodology for the case counts calculations. SI 2-4 covers the local spatial autocorrelation analysis for the three distance matrices considered. SI 5 presents the prospective spatiotemporal clustering analysis, and SI6 presents the complete spatiotemporal modeling results.

Acknowledgments

I would like to thank my committee for all their help and support.

Special thanks go to Steven Radil for his suggestions.

This research was supported by the EPSCoR Program, National Science Foundation #IIA-1301792, the Mountain West Clinical and Translational Research - Infrastructure Network, NIH, National Institute of General Medical Sciences (NIGMS) #1U54GM104944-01A1, and NIGMS #P20GM104420

Dedicated to my wife Livia

A constant source of love, support, and inspiration.

Table of Contents

Authorization to Submit ……..…… ii
Abstract ……..……...…. iii
Acknowledgements ……..……..…. iv
Dedication ……..…….…...…... v
Table of Contents ……..…….…..… vi
List of Figures …….…..... viii
List of Tables ……... ix
Chapter 1: An Exploration of The Spatiotemporal Patterns Of Ebola Virus Disease Epidemic in West Africa ……...…… 1
1.1. Introduction …… 1
1.2. Methods …….. 2
1.2.1. Data processing and description …….. 2
1.2.2. Spatial analysis …… 3
1.2.3. Temporal analysis ……..... 4
1.2.4. Spatial-temporal analysis ……..… 4
1.3. Results …… 5
1.3.1 Spatial analysis ……. 5
1.3.2 Temporal analysis …….…… 7
1.3.3. Spatiotemporal analysis …….…….. 8
1.4. Discussion ……..…… 11
1.4.1 Spatial analysis …….……... 11
1.4.2. Temporal analysis …….…… 13
1.4.3. Spatiotemporal analysis …….…… 13
1.4.4. Methodology limitations …….…... 14

1.4.5. Data sources limitations ……..…… 15

1.5. Recommendations ……..… 15

1.6. Acknowledgments …… 16

Chapter 2: Spatiotemporal Modeling of Ebola Virus Disease Epidemic in West Africa Using Open Access Data Sources ……..…. 17

2.1. Introduction ……..…….. 17

2.2. Data processing and description ……..…….. 19

2.3. Modeling approach ……..…….. 20

2.4. Results ……..….. 22

2.4.1. Models A and B detailed results ……..…... 25

2.5. Discussion ……..…… 28

2.6. EVD models comparison and concluding remarks ……..…..... 31

2.7. Acknowledgments …… 33

References ……... 34

List of Figures

Figure 1.1: Weekly case counts ……...… 2

Figure 1.2: Global Moran's I ……...… 5

Figure 1.3: Local Moran's I ……..... 6

Figure 1.4: Retrospective space-time clusters …….. 8

Figure 2.1: Relative contributions of models' components …… 25

Figure 2.2: Estimated random intercepts …… 26

Figure 2.3: Figure 2.4: Observed vs. predicted case counts ……... 27

Figure 2.4: Observed vs. one-week-ahead predicted case counts ……... 28

List of Tables

Table 1.1: Spatiotemporal clusters …… 9

Table 1.2: Counties in high observed/expected clusters ……. 10

Table 2.1: Summary of EVD models …….. 18

Table 2.2: dss and rmse for the NB models …… 23

CHAPTER 1

AN EXPLORATION OF THE SPATIOTEMPORAL PATTERNS OF EBOLA VIRUS DISEASE EPIDEMIC IN WEST AFRICA

1.1. Introduction

The West African Ebola epidemic of 2014 - the largest in history - arose in a much different cultural setting than previous outbreaks. Previous outbreaks had occurred in isolated villages, whose people had experience with Ebola and were unlikely to travel great distances to seek medical care. In contrast, continuous movement of people from their villages (even while very sick), across borders from Guinea to either Sierra Leone or Liberia, and into urban centers, drove the rapid spread of Ebola to neighboring West African countries, into cities, in a matter of days (WHO, 2016a). Over three years, 28,616 confirmed, probable, and suspected cases have been reported in West Africa, resulting in 11,310 deaths (WHO, 2016a). The magnitude of this epidemic and the difficulty containing it suggests the need for better understanding of dynamics of the Ebola Virus Disease (EVD).

While it is well recognized that interventions such as isolation of patients and safe and sanitary funerals and burials played a vital role in controlling the epidemic, as did the people's own adaptation (Chowell and Nishiura, 2014; Richards, 2015; Rivers et al., 2014), the purpose of this study was to investigate if exploratory analytical techniques using publically available data can provide insights into epidemic dynamics. The analysis of spatiotemporal-distributed disease data can be used to identify the presence or absence of areas with significant differences in risk (Sherman et al., 2014), identify possible periodical patterns in the behavior of the disease (Marek et al., 2015), and propose effective responses to outbreaks (Martins-Melo et al., 2012). Thus, spatial, temporal, and spatiotemporal analysis were used to assess the patterns of the EVD epidemic, identify the areas and time intervals of high risk, and identify the associated risk factors which can influence the risk of infection.

The paper is structured as follows: Section 1.2 describes the data processing and the statistical methodology. Section 1.3 presents the results of my analysis. Section 1.4 discuss the results, while section 1.5 contains the recommendations and conclusions.

1.2. Methods

1.2.1. Data processing and description

To calculate the daily and weekly number of cases in each administrative district, a dataset provided by OCHA ROWCA on the Humanitarian Data Exchange (HDX, 2015), that compiled the number of cases released by various sources including the WHO, national health ministries, and other sources was used. The dataset recorded daily cumulative total, confirmed, probable and suspected cases, as well as new cases and the number of deaths from March 24, 2014, up to March 28, 2015. The records cover six countries in West Africa: Guinea, Liberia, Mali, Nigeria, Senegal, and Sierra Leone, at various administrative units' levels.

For the study, the most severely affected countries by the EVD epidemic were selected: Guinea, Liberia, and Sierra Leone, with a total of 63 administrative units. Additional data was collected from published reports (Fink and Sheri, 2014; HumanitarianResponse, 2016; WHO, 2016b). As a result, the coverage of the case counts was extended, and some of the missing entries (June 2014 -August 2014), and errors (end of 2014 – beginning of 2015) were corrected. Figure 1.1 shows the weekly case counts based on the original and appended datasets. Supplementary material file S1 describes briefly the method used in calculating the case counts. The final datasets had daily and weekly Ebola virus counts and rates for the three

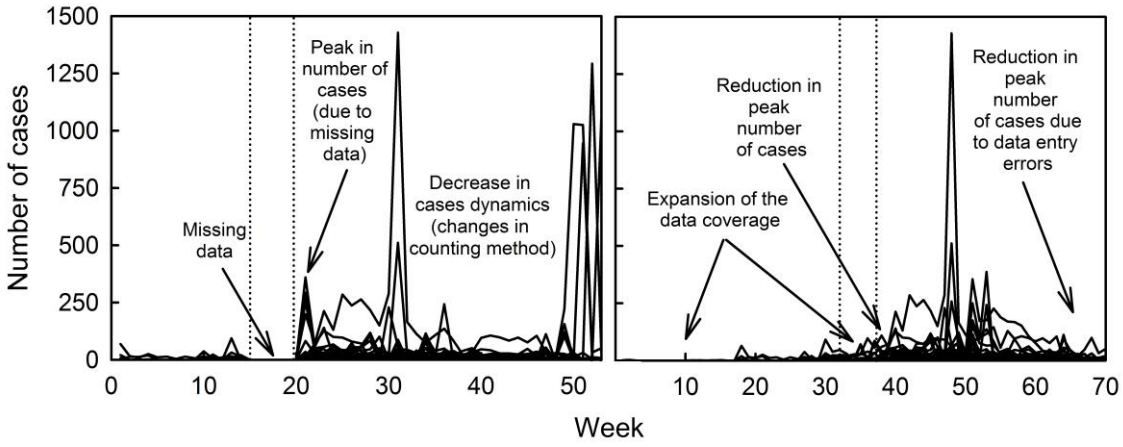

Figure 1.1: Weekly case counts calculated from the original dataset (left), and calculated from the appended dataset (right).

West African countries (aggregated over 63 administrative units) from December 06, 2013 to March 28, 2015.

1.2.2. Spatial analysis

Spatial analysis methods were used to evaluate the geographical distribution of the weekly Ebola infection rates. The administrative units from which the case counts were recorded were considered the units for the analysis.

The presence of spatial dependence was assessed using Global and Local Moran's I indexes for each of the 70 weeks of the epidemic considered in our dataset. The rates of infection were used instead of case counts since, generally, the number of cases is correlated to the underlying population size, and sometimes spatial autocorrelation may be detected only as an artifact of the spatial distribution of the population (Bivand et al., 2008). For the local Moran's I analysis, Holm p-value adjustments were used to assess the significance of each test (Brunsdon and Comber, 2015).

In the analysis, three distance measures were considered: first two are commonly used in spatial analysis: a *contiguity based neighbors matrix*, and a *centroid based distance matrix* (Bivand et al., 2008). Since it was suggested that the dispersal of Ebola virus is supported by the proximity of infected people to main roads (Hui-Jun et al., 2015), a *population-weighted road distance matrix* (Mitze, 2012) was also considered as the third distance matrix. To calculate the population-weighted road distance matrix, a list of major cities with their complementary population sizes for the three states considered was compiled from various internet sources (Brinkhoff, 2015; Wikipedia, 2015), but not limited to them. The web-based information available for these countries is scarce, inconsistent, and the town names were often different from source to source. In the end, a list of 83 cities from across all administrative units was compiled. The road-based distances (in km) between all them was calculated using the `ggmap` function in R (Kahle and Wickham, 2013) and Google Maps for the city pairs unrecognized by the R package. The population-weighted distances between administrative units were calculated as described by Mitze (Mitze, 2012).

The contiguity based weight matrix was row-standardized, and inverse-distance weight matrices were generated for the centroid and road distances. The analysis was

conducted in R-language (R Core Team, 2016), using the R packages `PBS mapping`, `spdep`, and `ape` (Bivand and Piras, 2015; Paradis et al., 2004; Schnute et al., 2015).

1.2.3. Temporal analysis

To test for temporal dependence, a multivariate ARMAX model was considered (Shumway and Stoffer, 2011). The multivariate ARMAX model expressed the counts of new Ebola cases, in a given administrative unit, as a linear combination of the trend and past counts of Ebola cases in all the other administrative units.

The case counts at time t were expressed as:

$$y_{t,i} = \propto_i + \beta_i t + \sum_{k=1}^{K}\left(\phi_{i,i} y_{t-k,i} + \sum_{j=1}^{N} \phi_{i,j} y_{t-k,j}\right) + w_{t,i} \quad (1)$$

for each of the $i = 1, 2, \ldots, N$ administrative units.

Where: $k = 1,2, \ldots, K$ is the k-order time lag. $j = 1,2, \ldots, N$ indicates the administrative units $j \neq i$. $y_{t,i}$ represents the case counts at time *t* and location i. $y_{t-k,i}$ and $y_{t-k,j}$ are the case counts at time *t-k* and locations *i* and *j*, respectively. And $w_{t,i}$ term represents correlations between the residuals over the locations *i*. Residual are assumed to be independent over time.

The analysis was conducted in R-language (R Core Team, 2016), using the R package `vars` (Pfaff, 2008a, b). Model residuals were checked to see if they fit the model assumptions: tests for the absence of serial correlation (Portmanteau test), heteroscedasticity (multivariate ARCH test), and normality (Jarque-Bera test). Non-normality and conditional heteroscedasticity are not often a concern for the validity of the models, especially in this case where the model was not considered final, but may help better the model deficiencies and the underlying properties of the data (Luetkepohl, 2011).

1.2.4. Spatial-temporal analysis

A third exploratory analytical approach looks at the retrospective spatiotemporal cluster analysis for the high and low incidence of the weekly Poisson-distributed count cases at each location. An analysis was conducted using the `SaTScan` software for the spatial and space-time scan statistics (Kulldorff, 2009).

For each location and time step, the scan analysis expects, under the null hypothesis, that the number of cases is proportional to the administrative unit population size. The alternate hypothesis is that there is an elevated risk within the scanning window as compared to outside (Kulldorff, 1997, 2009; Kulldorff et al., 1998). A maximum likelihood ratio test statistic and a p-value are calculated using Monte Carlo integration (for this study set at 9999). Identified clusters are ordered based on their likelihood ratio test values (Kulldorff, 1997, 2009). The program scans for clusters of geographical size between zero and some user-defined upper limit, called population at risk. The authors recommend to use values of 50% for the upper limit of the population at risk especially when in doubt. It should be noted that population at risk is not referring to susceptible as defined in SIR models, but rather as a geographical susceptibility. In the current research, several upper limits of the percent of population at risk (10 to 50%) were tested, and the results were compared.

Further, the usefulness of the spatial-temporal analysis for real-time prioritization of interventions was evaluated. To do so, prospective spatial-temporal analyses were conducted at monthly intervals, starting with week six of the epidemic. The prospective analysis identifies spatial-temporal clusters that are current, i.e. existent (or "live") at the end date of the dataset analyzed.

1.3. Results

1.3.1 Spatial analysis

Global Moran's I: Figure 1.2 presents a summary of the global Moran's I values over the 70 week period. Each pair of plots for the three weight matrices indicate the changes in the global Moran's I values, the p-value and the 0.05 level of significance line over time. For all distance matrices considered, we observe a pattern of alternating significant positive autocorrelation with non-significant autocorrelation. Moran's I values tend to increase over time, indicating an increase in the spatial autocorrelation as the disease

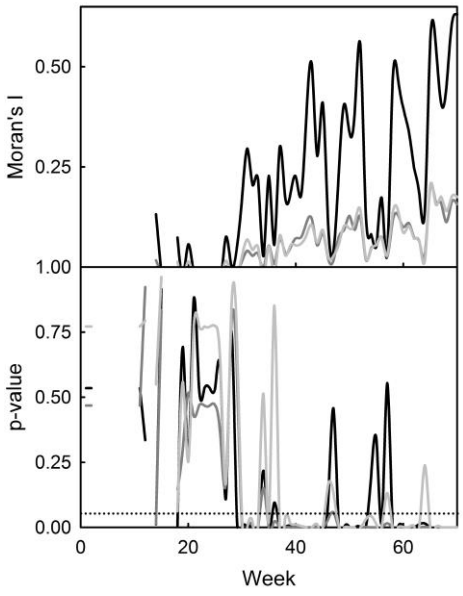

Figure 1.2: Global Moran's I values (top) and p-values (bottom) for contiguity (black), centroid (dark grey) and road (light grey) distance matrices.

evolved, and intervention measures take place. Moran's I values ranged from -0.03 to about 0.2 for the centroid and road distance matrices, and from -0.04 to 0.63 for the contiguity weight matrix. Week 30 in the epidemic period seem to be the first week with significant positive autocorrelation for all weight matrices.

Local Moran's I: The complete local Moran's I analysis can be found in Supplementary material files S2 to S4. Due to the low number of administrative units with non-zero cases,

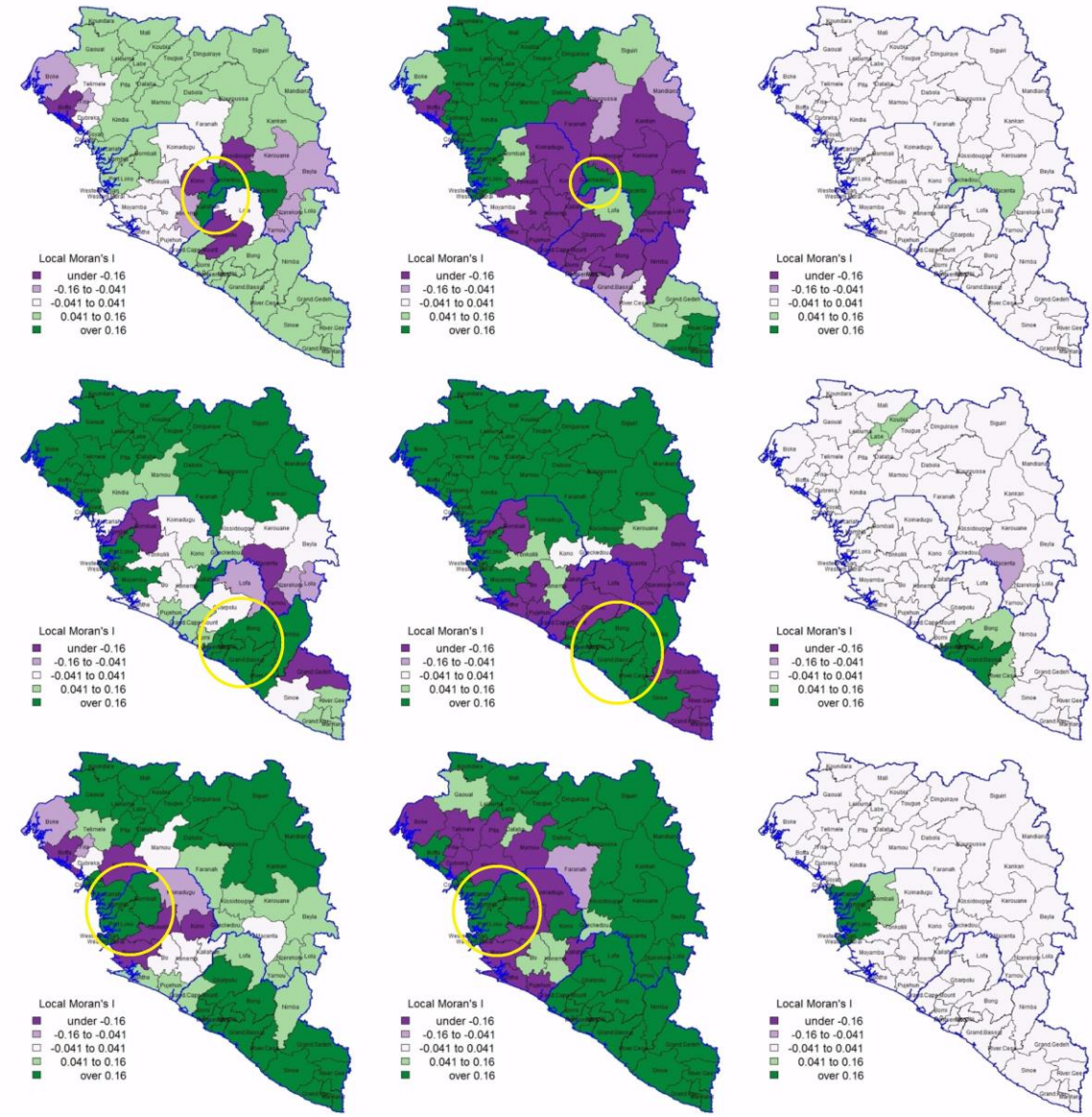

Figure 1.3: Local Moran's I values for the contiguity, centroid and road distances (left to right columns), for weeks 27, 43, and 66 (top to bottom rows). Yellow circles indicate districts with Local Moran's I p-values (Holm's method) <0.05.

until week 27 there were only a few time periods when local Moran's I values could be calculated. Overall, the results for all weight matrices were similar, but the clusters for the population-weighted road distance were less significant. Figure 1.3 shows the local Moran's I plots (p values are not shown) for three representative time periods. Initially, there is a cluster of the disease around three administrative units, Gueckedou and Macenta (Guinea), then in Lofa (Liberia). From week 27 to week 32-33 a hot spot of the epidemic can be identified at the border area of the three countries Kailahun (Sierra Leone), Lofa (Liberia) and Gueckedou (Guinea). The disease seems still localized in that area (surrounding counties have very dissimilar values). Up to week 40, the epidemic continues to be mainly clustered in the tri-state area, but significant values can be observed in Liberia, in the Montserrado area which becomes the center of a cluster of the Liberia outbreak, on and off until week 63. From week 51 until the end of the covered epidemic period, another cluster of significant autocorrelation can be observed in the NW region of Sierra Leone (Port Loko, Bombali, and Kambia).

Overall, the local spatial analysis highlights the initial cluster of the Ebola epidemic in the tri-state area, followed by a second cluster in Liberia, and a third in the NW Guinea. The results indicate that for several weeks, the outbreak was fairly localized, but later as it spread in West Africa, affected more heavily the highly populated areas, and their neighbors. A cluster of similar low values can be seen in the NE of Guinea almost for the entire duration of the epidemic.

1.3.2 Temporal analysis

The analysis was conducted for the daily cases data with maximum time lag of five days. Bigger time lags could not be tested due to overfitting. The analysis for weekly cases data could not be fit due to overfitting, while the model for the weekly infection rates leads to computational errors.

Most models for the daily cases data with time lags of five days had multiple R-squared values above 85%, and only one administrative unit showed nonsignificant temporal dependence (Dinguiraye). The models with higher time lags performed better than the ones with lower lags. The complete temporal analysis is available by request. The hypothesis of no serial correlation was rejected, suggesting that the model does not fully capture the temporal

dependence component. Also, the hypothesis of normality was rejected, while the heteroscedasticity test could not be performed due to overfitting.

1.3.3. Spatiotemporal analysis

The results of the 20% and 50% population at risk spatiotemporal clustering analysis are shown in Figure 1.4 and for 10% to 50% population at risk in Table 1.1. The choice of different percentages of the population at risk yielded a wide range of number of clusters, from eleven to one cluster(s) for 10% and 50% population at risk, respectively. However, regardless of the values considered, we see clusters of significantly higher than expected case counts centered on Liberia and Sierra Leone from week 35 to 64, and clusters of significantly lower than expected case counts in the first 35 weeks centered in Guinea, and along the border of the affected area. The same pattern was observed in the local spatial analysis, with the outbreak moving from the point of origin (in the tri-state area) directionally towards parts of Liberia and Sierra Leone. Looking at the results of the 20% population at risk clustering analysis, two stages of the epidemic can be distinguished: First, two clusters of significantly lower case counts for weeks one to 36 (clusters 3 and 4), followed by two high case counts clusters in Liberia for weeks 36 to 53 (clusters 2 and 5), and by one high case counts cluster in Sierra Leone for weeks 43 to 65 (cluster 1). The same pattern was observed during the spatial analysis, with an initial period when the outbreak was highly localized, followed by an

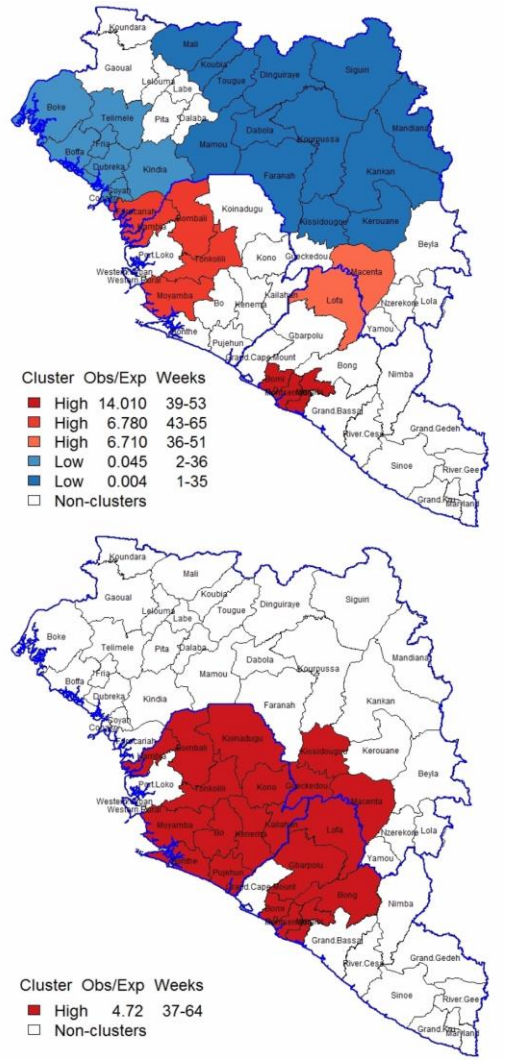

Figure 1.4: Retrospective space-time clusters for 20% (top) and – 50% (bottom) population at risk.

Table 1.1: Spatiotemporal clusters

Cluster	Administrative units	Weeks	Observed /expected cases ratio	Log Likelihood ratio	p-value
10% population at risk					
1.	Montserrado, Bomi, Margibi	39 to 53	14.01	9688.80	<0.001
2.	Western Rural, Western Urban, Port Loko	41 to 68	8.83	8750.47	<0.001
3.	Tonkolili, Bombali, Bo, Moyamba	43 to 56	5.71	2016.15	<0.001
4.	Dalaba, Pita, Labe, Mamou, Lelouma, Tougue, Koubia, Kindia	1 to 35	0	1210.20	<0.001
5.	Lofa, Macenta	36 to 51	6.71	1187.39	<0.001
6.	Kankan, Kerouane, Mandiana, Kouroussa, Kissidougou, Beyla	1 to 35	0.0059	1182.84	<0.001
7.	Nimba, Yamou, Bong, Grand Gedeh, River Cess, Nzerekore, Grand Bassa	1 to 34	0.0073	1066.95	<0.001
8.	Boke, Boffa, Fria, Telimele, Gaoual, Dubreka	1 to 35	0.041	828.11	<0.001
9.	Conakry	1 to 35	0.059	824.36	<0.001
10.	Kono, Gueckedou, Kailahun, Kenema	30 to 64	2.21	533.95	<0.001
11.	Forecariah, Kambia, Coyah	1 to 35	0.0020	488.15	<0.001
20% population at risk					
1.	Port Loko, Kambia, Western Rural, Western Urban, Moyamba, Forecariah, Bombali, Tonkolili	43 to 65	6.78	9921.75	<0.001
2.	Montserrado, Bomi, Margibi	39 to 53	14.01	9688.80	<0.001
3.	Siguiri, Dinguiraye, Mandiana, Kouroussa, Kankan, Dabola, Tougue, Faranah, Koubia, Kerouane, Kissidougou, Mamou, Mali	1 to 35	0.0042	2426.32	<0.001
4.	Boffa, Fria, Dubreka, Boke, Telimele, Conakry, Coyah, Kindia	2 to 36	0.045	2027.60	<0.001
5.	Macenta, Lofa	36 to 51	6.71	1187.39	<0.001
30% population at risk					
1.	Pujehun, Grand Cape Mount, Bonthe, Bo, Kenema, Bomi, Moyamba, Kailahun, Montserrado, Gbarpolu, Tonkolili, Margibi, Kono, Western Rural, Port Loko, Lofa	35 to 64	4.93	14407.70	<0.001
2.	Dinguiraye, Tougue, Dabola, Koubia, Kouroussa, Siguiri, Mali, Mamou, Dalaba, Labe, Faranah, Lelouma, Pita, Kankan, Mandiana, Koinadugu, Kissidougou, Kindia, Koundara, Gaoual	1 to 35	0.0029	3647.86	<0.001
40% population at risk					
1.	Pujehun, Grand Cape Mount, Bonthe, Bo, Kenema, Bomi, Moyamba, Kailahun, Montserrado, Gbarpolu, Tonkolili, Margibi, Kono, Western Rural, Port Loko, Lofa, Gueckedou, Bong, Western Urban, Grand Bassa, Bombali	35 to 64	4.78	21808.09	<0.001
50% population at risk					
1.	Kenema, Bo, Kailahun, Pujehun, Kono, Grand Cape Mount, Tonkolili, Gbarpolu, Bonthe, Gueckedou, Moyamba, Bomi, Lofa, Koinadugu, Montserrado, Bombali, Margibi, Port Loko, Kissidougou, Bong, Western Rural, Macenta, Kambia, Western Urban	37 to 64	4.72	22969.37	<0.001

risk results, we see one high case counts cluster over the Liberia and Sierra Leone.

The prospective clustering analysis at various time periods during the epidemic is presented in Supplementary material file S5. Table 1.2 summarizes the counties in high clusters in the prospective spatial-temporal analysis. Five of the six initial counties in high clusters are still present in high clusters at the end of the study period. Also, at least 50% of counties in a "live" high cluster at the end of the month are still in a high cluster a month later.

Table 1.2: Counties in high observed/expected clusters (black) at the end of the time period

County/week	14	18	22	26	30	34	38	42	46	50	54	58	62
Beyla													
BoCounty													
Boffa													
Bombali													
Bomi													
Bong													
Bonthe													
Conakry													
Coyah													
Dubreka													
Fria													
Gbarpolu													
Grand_Bassa													
GrandCape													
GrandGedeh													
Gueckedou													
Kailahun													
Kambia													
Kankan													
Kenema													
Kerouane													
Koinadugu													
Kono													
Lofa													
Lola													
Macenta													
Margibi													
Montserrado													
Moyamba													
Nimba													
Nzerekore													
Port_Loko													
RiverCess													
Telimele													
Tonkolili													
Western_area_rural													
Western_area_urban													
Yomou													
Percent*	**0.50**	**0.75**	**1**	**0.5**	**0.92**	**0.61**	**0.63**	**0.5**	**0.87**	**0.96**	**0.96**	**0.68**	**0.82**

*Percent of counties that remains in a high cluster at the end of the next prospective analysis

1.4. Discussion

1.4.1 Spatial analysis

Global Moran's I: The selection of the "best" weight matrix was data-driven (Dray et al., 2006). The contiguity matrix lead to the best model performance or best Moran's I (Getis and Aldstadt, 2004). The contiguity matrix also satisfied other proposed recommendations, such as: (1) under-specified matrices (fewer neighbors) should be preferred instead of over-specified weight matrices (extra neighbors), and (2) variables showing a good deal of local spatial heterogeneity should probably be modeled by fewer links in weight matrix (Getis and Aldstadt, 2004). The selection of the contiguity matrix as the "best" should be interpreted with caution, since Moran's I test can incorrectly suggest the presence of spatial autocorrelation in the presence of other effects (Bivand et al., 2008; Viton, 2010)(Viton, 2010).

Assuming that the contiguity matrix is capturing the spatial autocorrelation most accurately, the global Moran's I analysis indicates high spatial heterogeneity (since each county is more alike its neighbors – with little influence from more distant districts). This may be an indicator of movement primarily occurring within neighboring counties, which lead to a spatial aggregation of cases. The pattern of alternating significant positive autocorrelation, where there was an EVD emergence shortly followed by a EVD diffusion (non-significant autocorrelation), may suggest several things: (1) people were leaving the high-risk areas, as soon as the disease re-emerged, (2) localized medical response targeted preferentially, and effectively, the re-emergence areas or (3) it is an artifact due to timing of case reporting. In time, there is an overall positive trend in the global Moran's I values. This suggests that, as efforts were increasingly directed towards the treatment and prevention of the disease, the new outbreaks became more localized – a possible indicator that the intervention efforts were effective.

Since it was suggested that the dispersal of Ebola virus was supported by the proximity of infected people to main roads (Hui-Jun et al., 2015), the population-weighted road distance weight matrix was expected to yield the highest Moran's I values (i.e. had the best explanatory power). Instead, the contiguity matrix lead to higher Moran's I values. The global spatial analysis did not support or negate this hypothesis, but it suggested that it is

more plausible that there were other risk factors than main road proximity facilitating the spread of the EVD. For example, patterns of marriages and family ties may not adhere to the infrastructure in the region, and hence the transmission patterns in rural areas could have followed these family connections through rough and sometime treacherous routes (Richards, 2015).

Local Moran's I: Overall, the results for all weight matrices were similar. But the clusters for the population-weighted road distance were less significant. It suggests the for this particular weight matrix, the underlying process was more stable (homogeneous) within the data, and the local values had about the same contribution to the global statistic. This suggests that the proximity to main roads probably was a factor of risk, but, as suggested by the global spatial analysis, was not the only factor (i.e., population size played also a role, as expected).

The local spatial analysis highlights the initial cluster of the Ebola epidemic in the tri-state area, followed by a second cluster in Liberia, and a third in the NW Guinea. The results indicate that for several weeks, the outbreak was fairly localized, but later as it spread in West Africa, affected more heavily the highly populated areas, and their neighbors. A cluster of similar low values can be seen in the NE of Guinea almost for the entire duration of the epidemic.

A significant result proposed by the local spatial analysis was that for several weeks the disease was fairly localized in the tri-state area. This can be interpreted in two ways: (1) it is possible that there was an opportunity to contain the outbreak for a fairly large period of time, or (2) despite the sustained efforts to contain the initial outbreak, the EVD broke and spread in all West Africa. Regardless of the interpretation, the results reiterate the need for strong, sustained containment efforts right from the beginning of any outbreak. Epidemic models usually propose exponential growth in the number of cases. And this was true, at least in Liberia and Sierra Leone. Before that exponential growth, however, an initial built-up period of several months (resembling an Allee-like effect) seemed to have been present in the EVD population.

1.4.2. Temporal analysis

The important results of the time series analysis were: (1) there was a strong temporal dependence in the changes in Ebola cases, and (2) the time series were not fully capturing the dynamics of the disease. A limitation of the method used consists in the ARMA assumption that the error term is white noise, approximately normally distributed with mean zero. But, by using count data, we violate this assumption since negative observations cannot occur. Moreover, the ARMA model approach ignores the fact that the data is discrete instead of continuous. Another limitation of the approach is that the administrative units with zero counts were eliminated from the analysis, and the missing counts had to be converted to zeros – which might alter the temporal dynamics of the epidemic. Therefore, in this context, the method can be used solely as a relative evaluation of the temporal dependence.

Overall the results discussed so far suggests that in the case of EVD epidemic there was both strong spatial and temporal dependence. Thus, it illustrates the need for spatiotemporal analysis.

1.4.3. Spatiotemporal analysis

The spatiotemporal clustering analysis indicated there was significant clustering of cases in time and space. There were significantly higher than expected case counts centered on Liberia and Sierra Leone from week 35 to 64.There were clusters of significantly lower than expected case counts in the first 35 weeks centered in Guinea and along the border of the affected area. The same pattern was observed during the spatial analysis, with an initial period when the outbreak was highly localized – followed by an explosion of cases in Liberia, and then in Sierra Leone. If we examine the 50% population at risk results, we see one high case counts cluster over the Liberia and Sierra Leone from week 37 to 64. A cluster of large size is indicative of areas of exceptionally low rates outside of the circle (Kulldorff, 2009). This confirms again the directionality of the spread of the Ebola virus disease: although it started in Guinea, it spread towards Liberia and Sierra Leone, with less than expected case counts in Guinea. While this result is hardly surprising in the context of the previous discussion, it still highlights the fact that Guinea had less than expected case counts for the entire epidemic.

The choice of percent population at risk value had a significant effect on the results. In this type of analysis, population at risk is not equivalent to the percent of population susceptible to the disease as defined in SIR models, but rather it refers to the spatial population at risk. Since the transmission chain was spatially limited (Ajelli et al., 2015; Faye et al., 2015; Lau et al., 2017), it is plausible that the results using smaller population at risk values are more reliable.

An important result came from the prospective spatial-temporal analysis (Supplementary file S5 and Table 1.2). The analysis was intended to assess the potential of the method to identify "live" spatial-temporal clusters of the EVD and their evolution throughout the extent of the epidemic. Five out of six counties present in the early high clusters were still present at the end of the study period. An important question was: can the current "live" high clusters be used to predict the "live" clusters a month from now? The proportion of counties present in "live" clusters for two consecutive months ranged from 50 to 100%. Therefore, it is a strong possibility that concentration of resources in a current "live" cluster may be the *necessary* strategy to reduce the severity of an epidemic, but by no means should it be expected to be the *sufficient* strategy. The method was already used in identifying outbreaks of shigellosis in Chicago (Jones et al., 2006), and is currently used in the daily automated spatiotemporal analysis in New York for early outbreak detection of 35 reportable diseases (Greene et al., 2016). The current research indicated the tool can potentially be useful in effective early response to other diseases, such as Ebola, as long as the data can be collected, recorded, geocoded and analyzed in real-time.

1.4.4. Methodology limitations

There are a few limitations inherent to the methods themselves. For example, the spatial autocorrelation analysis is sensitive to spatial scale effects, the different polygons' shapes and sizes, and border effects. Further, the spatial autocorrelation test can incorrectly suggest the presence of spatial autocorrelation in the presence of other effects, the temporal analysis is clearly suggesting other effects beside the autoregressive terms, and the correlation analysis may be spurious. Even under the assumption that the previously mentioned limitations are not an issue in this case, the range of conclusions based on this analysis is rather limited. Spatial, temporal, and spatiotemporal patterns can be shown - with clear

application in the management of the epidemic – but the drivers of the epidemics and the effects of interventions cannot be accounted for with the data we used. Finally, our dataset did not contain data on individual cases, and is based on aggregate data at the county level. Therefore, we recognize that some associations we are observing could suffer from ecological fallacy.

1.4.5. Data sources limitations

It is well-recognized that reliable and accurate information is essential to evaluate and improve the delivery of health services. In the case of epidemics of emerging infectious diseases and crises, data collection is often difficult if not impossible. Therefore, considerable efforts were placed to collect standardized, high quality data in West Africa prior and during the EVD disease. Still, it is still possible that the quality of the dataset may raise questions about the reliability of the analytic results:

First, the original dataset of case counts had gaps, incorrect counts, and unexplained drops in the cumulative case counts in several administrative units. Second, the early symptoms of Ebola are, for the most part, indistinguishable from malaria, and posed a major challenge when identifying probable cases. While measures were taken to amend these problems, there is a certain degree of uncertainty about the calculated daily/weekly new cases. Third, it has been documented that there was under-reporting of the number of Ebola cases (Westcott, 2014; Zavis and Healy, 2016).

1.5. Recommendations

This research investigated if exploratory analytical techniques using publically available data can inform interventions in case of infectious diseases outbreaks. More specifically, the methods were used to evaluate the dynamics and causes of the EVD epidemic in West Africa. The results showed that there was significant spatial, temporal, and spatiotemporal dependence in the evolution of the disease. For the first part of the epidemic, the cases were highly clustered in a few administrative units, in the proximity of the point of origin of the epidemic, offering the opportunity to stop the spread of the disease, pending a robust, directional intervention. Later in the epidemic, high clusters were observed only in

Liberia and Sierra Leone. The spatial-temporal analytic tool `SaTScan` may be used effectively during the evolution of an epidemic.

Based on these results, we believe that these exploratory techniques can be useful for monitoring purposes, as tools for early detection of potential outbreaks. Early in an outbreak, data is usually sparse, the potential risk factors are largely unknown, and resources are limited. The presented methods have the advantage of being fairly straight forward, require rather low resources, while the results are quite reliable. The presented analysis can indicate if and where disease clusters are. Later in the epidemic, as interventions and behavioral changes are shaping the dynamics of the outbreak, they have to be taken in consideration in the analytic efforts. This requires more sophisticated modeling approaches (which has been extensively addressed in the literature), better data, integration of local knowledge and customs, and substantially higher resources.

1.6. Acknowledgments

This publication was supported by the EPSCoR Program, National Science Foundation #IIA-1301792, the Mountain West Clinical and Translational Research - Infrastructure Network, NIH, National Institute of General Medical Sciences (NIGMS) #1U54GM104944-01A1, and NIGMS #P20GM104420. Special thanks go to Steven Radil for his suggestions, and the anonymous referees, whose comments have greatly improved the paper.

CHAPTER 2

SPATIOTEMPORAL MODELING OF EBOLA VIRUS DISEASE EPIDEMIC IN WEST AFRICA USING OPEN ACCESS DATA SOURCES

2.1. Introduction

Over the past several years, scientific models have been developed to better understand the drivers of the spread and eventual containment the 2014-2016 West African Ebola Virus Disease (EVD) epidemic. The heavily affected countries were poor, recently war-torn states with highly mobile populations that had no previous experience with Ebola. Transmission rates varied widely between regions. Thus, it was difficult to predict where resources are going to be needed as the epidemic slows, as it was to predict when the epidemic began. While in many places most transmission events occurred in the community and between family members (Ajelli et al., 2015; Faye et al., 2015), the continuous movement of people from their villages and crossing borders from Guinea to either Sierra Leone or Liberia was the driver behind the rapid spread of Ebola to neighboring countries (WHO, 2015). The isolation of patients and safe and sanitary funerals and burials played a vital role in controlling the epidemic (Chowell and Nishiura, 2014; Rivers et al., 2014).

Models of the EVD epidemic have taken many different forms. Table 2.1 summarizes some of these results. First, models confirmed that initially there was little control of the EVD outbreak, and that additional resources were needed in order to contain the spread of the disease. Later, models showed that were differences in the EVD outbreak dynamics between and within the affected West African countries. Interventions and behavior changes were found responsible for significant reductions in transmission pathways. EVD dispersed mostly between geographically closed regions, but the dispersion was seemingly directional, with some West African regions significantly more affected then others.

To explain the observed differences in the EVD outbreak dynamics, a wide variety of variables were considered in the models. From those, population size, time to travel to large population centers, proximity to roads and hospitals, precipitation, and temperature seasonality were found to have a significant effect on the case counts. In the current research,

Table 2.1

Model	Conclusions	Suggestions	Reference
SEIR	The number of secondary infections dropped by May – July 2014 in Guinea and Sierra Leone, but not in Liberia	Increased intervention efforts in Liberia are needed.	Althaus (2014)
SEIR	Resources needed in Montserrado to control the EVD outbreak, by November 2014, exceeded those committed by aid groups.	Increased intervention efforts are needed.	Lewnard et al. (2014)
Incidence decay with exponential adjustment (IDEA)	There was weak evidence for the occurrence of epidemic control in West Africa as a whole, and essentially no evidence for control in Liberia by August 2014	Improved control measures are needed, especially in Liberia, if catastrophe is to be averted	Fisman et al. (2014)
Math models of the effective reproduction number	Effective control of the epidemic can be achieved if the number of secondary transmissions per infected individual is cut in half.	Increased intervention efforts are needed, as well as prevention of cross-border transmission.	Nishiura and Chowell (2014)
spatial agent-based models	Up to August 2014, most infections were acquired in hospitals and in households: interventions significantly reduced these transmission pathways.		Merler et al. (2015)
Exponential and polynomial models	There were significant differences in the growth patterns of EVD cases at different spatial scales in West Africa.	Behavior changes, differences in intervention control, or disease-specific features might be responsible for the observed dynamics.	Chowell et al. (2015)
survival, hazard, Poisson and chain-binomial transmission models	Case isolation and safe burials significantly reduced the transmissibility at chiefdom level in Sierra Leone.	Population density, proximity to Ebola treatment centers, cropland coverage & temperature were associated with EVD transmission.	Fang et al. (2016)
Spatiotemporal growth model & SEIR	There were significant differences in the evolution of the outbreak in the different regions in West Africa	For effective interventions, continuous monitoring at the district level is needed.	Santermans et al. (2016)
Phylogenetic and Bayesian generalized linear models	EVD tended to disperse mostly between geographically closer regions. The spread was more prevalent within country borders.	Economic output, population density, traveling times to large settlements & climatic factors were not associated with the EVD dispersal.	Dudas et al. (2017)

we tested if behavioral and cultural attributes of West African population could improve the model of EVD outbreak, and, thus, better explain the observed differences in the disease dynamics. The work presented here evaluated the methodologies by which these population characteristics can be incorporated, and measured if their inclusion improves the model fit of the Ebola epidemic. We compiled a comprehensive dataset of Ebola cases and demographic

characteristics to map and analyze the spatiotemporal transmission patterns at the administrative unit level in Guinea, Liberia and Sierra Leone.

The rest of the paper is structured as follow: Section 2.2 examines the data processing and description. Section 2.3 presents the modeling methodologies considered. Sections 2.4 and 2.5 present the results and discussion of the models considered, and details several "best" models. Section 2.6 compares our results with other EVD modeling papers and concludes the paper.

2.2. Data processing and description

To calculate the daily and the weekly number of cases, we used a dataset provided by OCHA ROWCA on the Humanitarian Data Exchange (HDX, 2015), that compiled the number of cases released by various sources from March 24, 2014, up to March 28, 2015. For this study, I selected only the three countries most severely affected by the Ebola outbreak: Guinea, Liberia and Sierra Leone, with a total of 63 administrative units. Additional data was collected from published reports (Fink and Sheri, 2014; HumanitarianResponse, 2016; WHO, 2016b). As a result, I extended the coverage of the case counts, I imputed missing data and corrected some errors (Suchar et al., 2018). The final datasets had daily and weekly EVD outbreak counts and rates for three countries: Guinea, Liberia and Sierra Leone (63 administrative units) from December 06, 2013 to March 28, 2015. Details on the methodology can be found in the Supplemental Information (SI) file.

For demographic and health information about West Africa, datasets provided by USAID on the Demographic and Health Surveys (DHS) Program website (USAID, 2016) were used. DHS collects, analyzes, and disseminates population, health, HIV, and nutrition data in over 90 countries. More details about the data collected in West Africa can be found in the DHS reports specific to each country (INS, 2013; LISGIS, 2014; SSL, 2014). For this research, the focus was on surrogate variables for factors hypothesized to be possible risk factors in the transmission of diseases: percent having access to bicycles, motorcycles or cars, percent of women that had a hospital delivery, had a doctor or medical professional present at delivery, education level, literacy, and percent reading newspapers, listening to radio or watching TV, percent of the population sharing a toilet with other households, number of children living at home, and the mean number of STD and sexual partners. These factors

represent access to information, transportation, healthcare, and behavior that might modify the risk of exposure. The data were aggregated to the administrative unit.

To account for the spatial autocorrelation, three distance measures were considered: a *contiguity based neighbors matrix*, a *centroid based distance matrix* (Bivand et al., 2008), and a *population-weighted road* distance (Mitze, 2012). Details on the methodology used to calculate the population-weighted road distance can be found in Mitze (2012) and Suchar et al. (2017).

2.3. Modeling approach

To evaluate the EVD outbreak in West Africa, we used a negative binomial (NB) spatial-temporal mixed effects model with fixed social explanatory variables, and autoregressive temporal and spatial-temporal random effects. To fit the model, an endemic-epidemic multivariate time-series method was used, with the fixed social explanatory variables accommodated by the endemic component, and the spatial-temporal interactions accommodated by the epidemic component (Hohle et al., 2016; Meyer et al., 2016; Paul and Held, 2011). The endemic-epidemic multivariate time-series model for the spread of infectious diseases used and its applications were described in detail in a series of papers and its R package `surveillance` vignette (Hohle et al., 2016; Meyer et al., 2016; Paul and Held, 2011). The response variable is case counts Y_{ij} from county *i* and time step *j* with $i = 1,2,\dots I$ and $j = 1,2,\dots J$. In the endemic-epidemic formulation, the response variable has a negative binomial (NB) distribution conditional on the past observations:

$$\begin{gathered} Y_{ij} \sim NB\left(\mu_{ij}, \psi_i\right) \\ \mu_{ij} = e_{ij}\nu_{ij} + \lambda_{ij}Y_{ij-1} + \phi_{ij}\sum_{k\neq i} w_{kj-1}Y_{kj-1} \end{gathered} \quad (1)$$

Where μ_{ij} is the mean of the NB distribution, and ψ_i is the overdispersion parameter ($\psi_i = 0$ corresponds to Poisson (POI) distribution). The mean is decomposed in endemic and epidemic components. The endemic (end) component $e_{ij}\nu_{ij}$ explains a baseline rate of infection in the population. The epidemic component has an autoregressive (**AR**) effect $\lambda_{ij}Y_{ij-1}$ accounting for the spread of the disease within county *i*, and a neighborhood (**NE**)

effect $\phi_{ij} \sum_{k \neq i} w_{kj-1} Y_{kj-1}$ accounting for effects of other counties to county *i*. w_{kj} are spatial weights, while e_{ij} correspond to the population size offset for county *i*.

Both fixed and random effects can be used in the model. In the case of fixed effects, standard likelihood inference is performed, while, in the case of random effects (Gaussian independent (NI), correlated (NC), or conditionally autoregressive (CAR)), the inference is based on penalized quasi-likelihood method (Hohle et al., 2016; Meyer et al., 2016; Paul and Held, 2011).

In the case of EVD there is no endemic component (baseline rate of infection in the population) as proposed by the authors of the endemic-epidemic model. Instead, we used the endemic component to accommodate the social covariates considered. Since the model could not accommodate a very large number of covariates, two strategies were considered: (1) principal components (PC) analysis was used to reduce the dimensionality of the social variables, and the first four PCs were used as covariates in the models; (2) of the 24 factors considered to represent access to information, transportation, healthcare, and behavior that might modify risk of exposure, five were included in the final model based on the exploratory data analysis: ethnic homogeneity, % reading newspapers, % hospital delivery of the last child, % health professional present at the birth of last child, and mean children at home.

For each weight matrix, four models were fit using this modeling approach: MODEL 1: fixed AR and NE effects without population offset for the social fixed effects; MODEL 2: fixed AR and NE effects with population offset for the social fixed effects; MODEL 3: random AR and NE effects without population offset for the social fixed effects; and MODEL 4: random AR and NE effects with population offset for the social fixed effects.

To compare the goodness-of-fit of these models we used square error score (ses), logarithmic score (logs), ranked probability score (rps), and Dawid–Sebastiani score (dss). The dss has been argued to be more appropriate for model performance assessment than the traditional statistics, since they may account for the uncertainty associated to point predictions (Czado et al., 2009; Paul and Held, 2011). These scores are readily available in the `surveillance` package (Hohle et al., 2016). Since these methods are not widely used and in order to provide a familiar baseline to our readers, the root mean squared error (rmse) was

also calculated for the models. The model performance measures logs, rps, and dss are strictly proper scoring rules, meaning that they are unique, minimized penalties based on the predictive distribution proposed by the model and the observed quantities (Czado et al., 2009; Paul and Held, 2011). They take in account the uncertainty associated with the predictions, which is not the case for the root mean squared error (rmse) which accounts only for the differences between the observed and predicted values. Square error score (ses) depends on the mean of the predictive distribution, while Dawid–Sebastiani score (dss) accounts for both the mean and the variance of the predictive distribution.

A probability integral transform (PIT) histogram for count data was used to qualitatively check the predictive distributions. PIT histograms verifiy if the observed values have the predicted distribution. If so, the PIT histogram for a well calibrated model should be uniform. Biases cause skewness, while under- and over-dispersion causes U-shape and inverse U-shape PIT histograms (Czado et al., 2009; Hohle et al., 2016).

Finally, for validation purposes, one-week-ahead model predictions were compared with the observed case counts.

2.4. Results

The complete goodness-of-fit scores for the models considered are presented Supplementary material file S6: Tables S1 and S2. A subset of these results is presented in Table 2.2.

The goodness-of-fit measures indicated that:

1. In general, model performance measures agreed that equivalent models using social PC as covariates are better than models using the selected social variables as covariates.

2. Random effects models have higher rmse than the fixed effects models.

3. Model performance measures did not agree which model is the best. Bold values in Supplementary material file S6: Tables S1 and S2 indicate the best models relative to each goodness of fit metric.

4. All PIT histograms (Supplementary material file S6: Figures S1 and S2) exhibit a decay on the right side – indicative of biases over predictions.

Table 2.2: Dawid–Sebastiani score (dss) and root mean squared error (rmse) for the Negative Binomial (NB) models using as covariates: (1) ethnic homogeneity, read newspaper, hospital delivery, health professional present at delivery, and mean number of children at home and (2) social principal components 1-4. Note: overdispersion parameter was significantly greater than 0 for all models.

			Selected covariates		**PC covariates**	
Weights	**Pop. offset**	**ar+ne**	**dss**	**rmse**	**dss**	**rmse**
Contiguity	No	Fixed	6.6	11.4	8.9	9.3
	No	Random	5.8	14.0	10.7	12.8
	Yes	Fixed	6.0	10.6	5.4	8.7
	Yes	Random	6.6	13.3	8.9	11.8
Centroid distance	No	Fixed	5.2	4.2	4.8	3.8
	No	Random	12.7	10.1	1.1	8.7
	Yes	Fixed	5.1	4.1	4.3	2.2
	Yes	Random	5.2	10.0	4.8	8.8
Population-weighted road distance	No	Fixed	5.5	4.7	5.0	4.3
	No	Random	10.8	9.2	1.1	7.9
	Yes	Fixed	5.3	4.6	4.3	2.3
	Yes	Random	5.5	9.2	5.0	8.0

Regarding the EVD outbreak dynamics:

1. All models agree that there were high autoregressive (AR) in areas near the origin of the EVD outbreak (Lofa, Macenta, Kailahun and Kenema), in NW Sierra Leone (e.g., Port Loko and Western Urban), and very strong AR contribution in Montserrado and surrounding area (Liberia). (Supplementary material file S6: Figures S3-S10).

2. Models disagree about the neighbor effect (NE) contributions to the case counts. In general, high NE contributions were observed for counties neighboring Montserrado. In Sierra Leone, counties with high AR contributions had also high NE values. In the case of the fixed effects models, this was observed for the contiguity distance matrix only. For the random effects models, the distance matrices had an influence on the amplitude of the NE effect.

3. For the covariates contributions, only Gueckedou region had values higher than the other districts, and in only few of the fixed effects models.

4. There are substantial differences between districts in the relative AR and NE contributions to the case counts. The estimated random intercept for the AR and NE components of the random effects models (Supplementary material file S6: Figures S11 and S12) showed large differences in the AR coefficients between Montserrado and Lofa, and the most of the remaining counties. In contrast, NW and SE Sierra Leone had relatively low AR intercepts. For the NE effect, we see almost the opposite: high heterogeneity across the three countries, highest intercept values for the NW and SE Sierra Leone, and average values for Montserrado. The AR component was significant for all models except for the fixed effects

for centroid, and road distances. For the random effects models, the district level AR random coefficients were not significant.

5. The covariate effects were significant, although their contribution to the total case counts was small. Principal component 1 (PC1) was significant in models 1, 3, and 4, PC2 in models 1 and 3, while PC 3 and 4 were significant in all models. The direction of the relationship was positive for all principal components, except PC2.

Model selection: The model performance measures did not agree on which model fit the data the best. This can be explained by the properties of the statistics considered. The logarithmic score (logs) is sensitive to extreme case counts, while the probability score (rps) score is less sensitive to outliers. Also, probability score (rps) and squared error score (ses) are highly dependent on the size of the counts, which makes high counts dominant of the mean score (Czado et al., 2009). Dawid–Sebastiani score (dss) accounts for both the mean and the variance of the predictive distribution and less influenced by the size of the counts. The root mean squared error (rmse) is also sensitive to extreme counts, and weight more the larger errors than the smaller ones.

Taking into consideration all these, the model with no population offset, principal components as covariates, population-weighted road distances weights and random AR and NE components might be the best. It has the smallest dss, rps, and logs scores, and mid-values for ses and rmse. For comparison purposes, the model with population offset, principal components as covariates, centroid distance weights and fixed AR and NE components was included in the following discussion. It had the lowest rmse, comparable logs, rps, and ses values, but a much higher dss score than the "best" model. The following discussion refers to the two selected models, which will be referred as **Model A** (no population offset, principal components as covariates, population-weighted road distances weights and random AR and NE components) and **Model B** (population offset, principal components as covariates, centroid distance weights and fixed AR and NE components).

2.4.1. Models A and B detailed results

The PIT histograms for the selected models exhibited skewness and similar decay in the right side, suggesting a central tendency bias, and predisposition towards over-prediction. This might be related to the high number of counties with extreme counts (many with zero counts, and some with extremely high counts). When the relative contributions of the components are examined (Figure 2.1), both models agreed on the very high AR contributions in Montserrado, and moderate AR contributions near the point of origin on the EVD outbreak. Model A indicated moderate AR contributions in NW Sierra Leone, while Model B proposed fairly high AR contributions for the area. The big differences between models were in the NE component: Model A suggested high NE contributions in Kailahun and Kenema (SE Sierra Leone), and moderate NE contributions in Margibi (Liberia) and NW Sierra Leone, while Model B suggested no/low NE contributions to the case counts. Both models agree that the

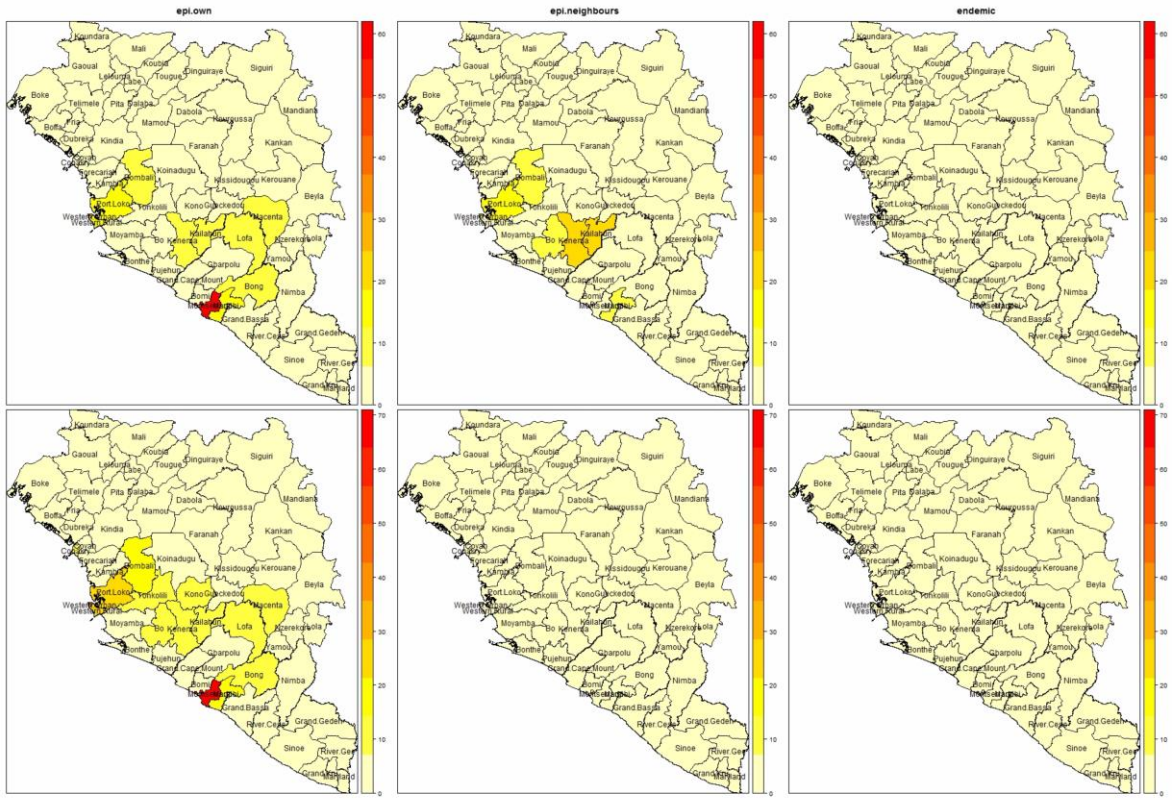

Figure 2.1: Model A (top) and Model B (bottom) relative contributions of the AR, NE, and covariates components (from left to right). The color indicates the relative contribution of the model component to the case counts in each district (e.g., pale yellow – low contribution, red – very high contribution).

covariates had similar and low relative contributions to the total case counts, as indicated by the uniform low heat color in Figure 2.1, right column.

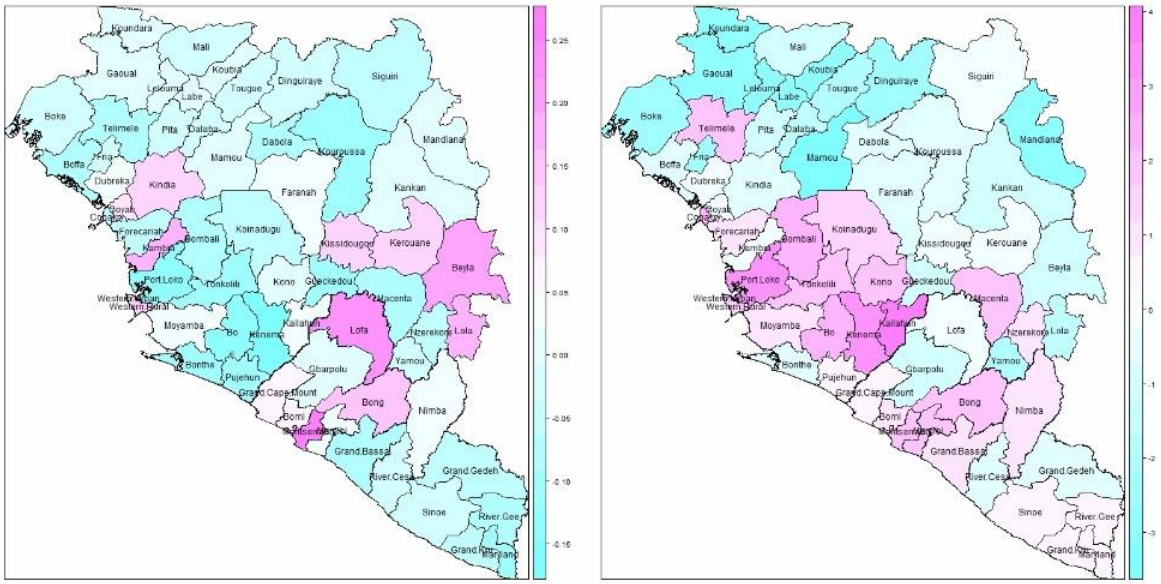

Figure 2.2: Model A estimated random intercepts for AR (left) and NE (right). The colors indicate the random coefficients estimates for each district from low (green) to high (violet) values.

The random intercepts values for the Model A (Figure 2.2) indicated higher values for the Lofa - Montserrado corridor for the AR components and relatively high values for the NE component in Sierra Leone. Most of the administrative units in Guinea had small values for both the AR and NE components, with the exception of Coyah and Kindia (higher AR values), and Conacry and Telimele (higher NE values). Looking at the EVD fitted components in each administrative units and at the two selected models predictions (Figure 2.3, Supplementary material file S6: Figure S13), we can see that Model B was consistently over-predicting the NE components in districts with low case counts. The AR components were comparable in both models. In most cases, the total predictions were similar for both models considered, especially for high counts districts. Model A had a tendency to be less

precise for the medium case counts districts (Figure 2.3, Supplementary material file S6: Figure S13).

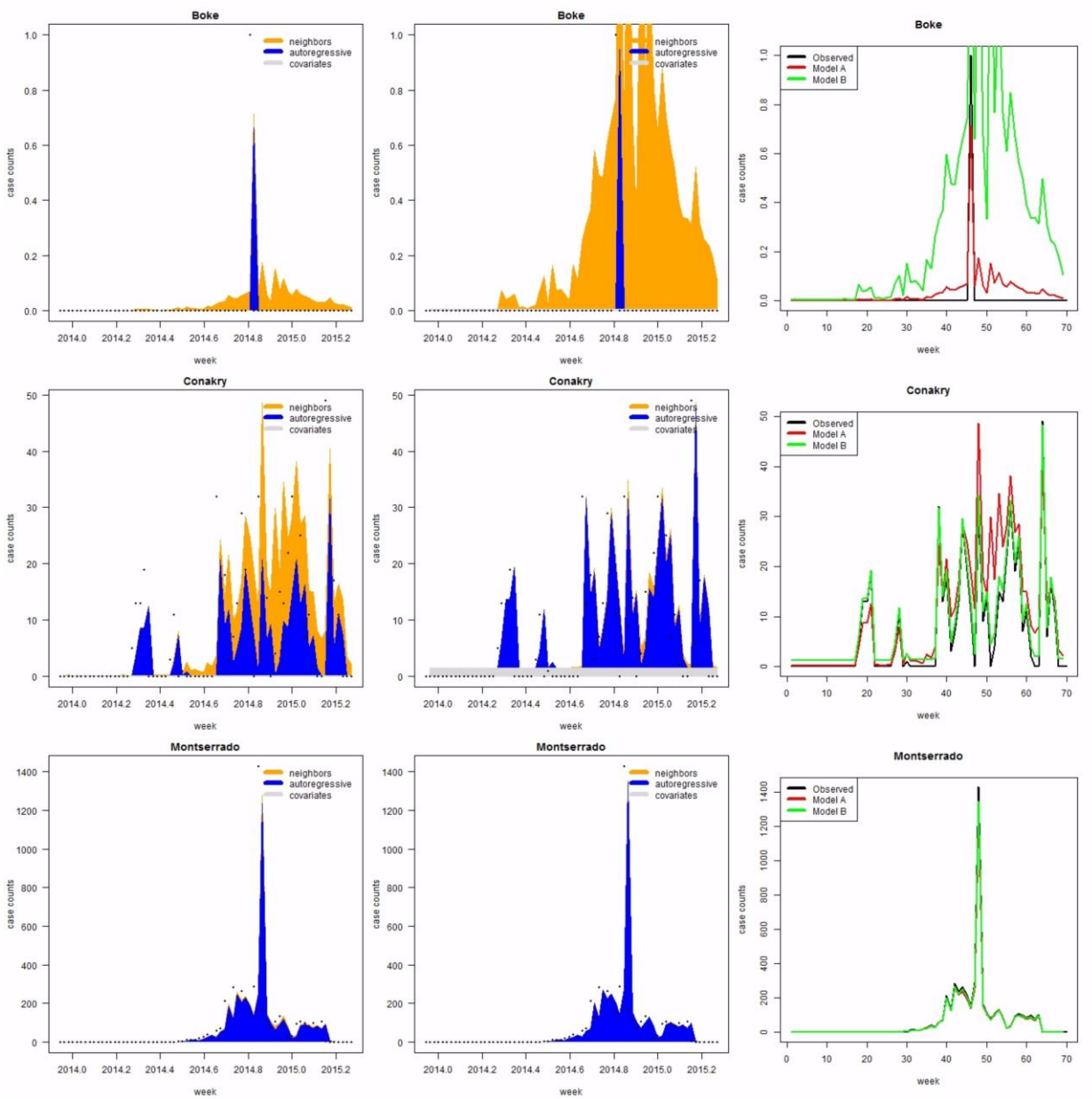

Figure 2.3: Fitted components and model predictions for selected counties (Model A –left, Model B – middle, observed vs. predicted - Right). In the left and middle plots, yellow, blue, and gray areas indicate the NE contribution, AR, and covariates contribution, respectively, to the total case counts, while the dots indicate the actual observed values. In the right plots, black lines indicates the observed values, while the red and green lines indicate the MODELS A and B fitted values.

Overall, as the model selection scores indicated, Model A (no population offset, principal components as covariates, population-weighted road distances weights and random AR and NE components) seem to have captured fairly well the dynamics of EVD outbreak in West Africa.

When it comes to the models' predictive power, Model B over-predicted the case counts in administrative units with low values, while Model A over-predicted cases in regions with medium values. Both models predictive power was fairly similar in administrative units with high case counts **(**Figure 2.4, Supplementary material file S6: Figure S14**).**

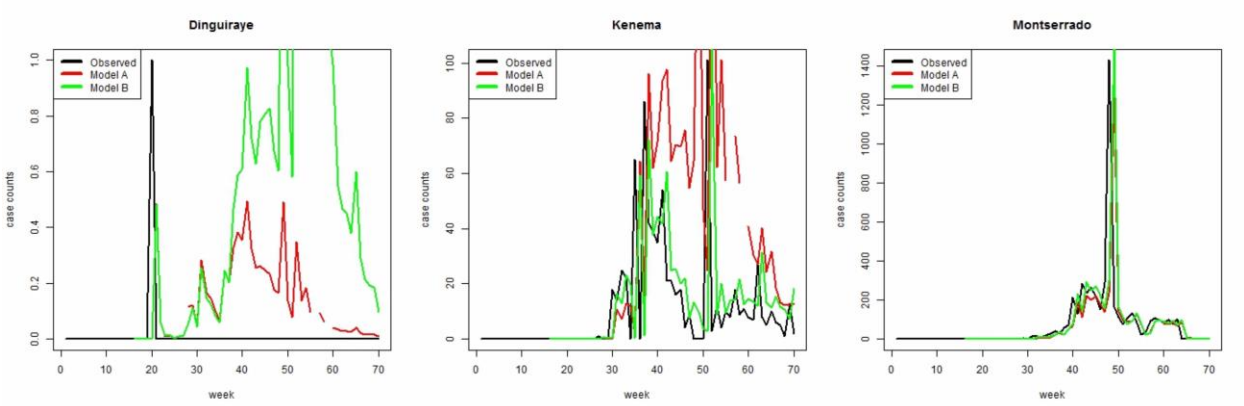

Figure 2.4: Observed vs. one-week-ahead predicted case counts. Black lines indicates the observed values, while the red and green lines indicate the MODELS A and B predicted values.

2.5. Discussion

Local conditions might have favored the persistence of the EVD in the Gueckedou and surrounding areas in the beginning of the outbreak. In general, the covariates contribution to the case counts was low, except for Gueckedou in some of the fixed effects models. But the models suggested larger AR contributions around the point of origin of the disease (Figure 2.1). The local spatial analysis (Suchar et al., 2017) showed that for several weeks the disease was fairly localized in the tri-state area, and an initial built-up period of several months might have been present in the EVD population. The spatial analysis also suggested that tri-state area as a place where the disease stayed localized for quite some time before it exploded in the rest of the West Africa. In combination with the current result, it seems that some specific

conditions may have facilitated the longevity of the transmission chain in this region, beyond what the demographic and economic can capture.

Montserrado area was a source of new cases for the neighboring areas, while NW Sierra Leone region was a sink for new cases from the neighboring areas. The relative contributions of autoregressive and neighboring effects were highly variable – suggesting differences in the processes driving the outbreak in different regions in West Africa. The estimated random intercept for the AR and NE components of random effects models (Supplementary material file S6: Figures S11 and S12) showed large differences in the AR coefficients between Montserrado and Lofa, and the most of the remaining counties. In contrast, NW and SE Sierra Leone had relatively low AR intercepts. For the NE effect, we saw almost the opposite: high heterogeneity across the three countries, highest intercept values for the NW and SE Sierra Leone, and average values for Montserrado. The spatial analysis of the data has also indicated that there are cycles altering significant positive clustering with non-significant autocorrelation, probably diffusion of the disease, as well as local clustering around the Montserrado and Lofa and NW Sierra Leone, respectively. The current results complement the spatial analysis, and suggests that the outbreak (at least) in the areas in vicinity of Montserrado was sustained by infected people living the city, while in NW Sierra Leone, the outbreak was sustained by people coming from other parts of the country.

There were region-specific population responses to the outbreak. The self-sustained AR infection rates were fairly similar in all districts, while the NE effect was more variable (Figure 2.1 and Supplementary material file S6: Summary S1 and S2). Higher NE contributions were especially observed in Sierra Leone relative to that in the other two countries. This suggested that the within-district evolution of the disease was generally similar regardless of the location, but there was some heterogeneity associated with behavior and mobility.

Social attributes may have influenced the outbreak dynamics. The covariate effects were significant - although small – pointing towards the importance of hypothesized social attributes of the population in the outbreak dynamics. Increases in percent reading newspapers were associated to lower case counts. Higher literacy and access to information are potential indicators of higher economic status that may decrease the vulnerability to infectious diseases

of different groups, but also may point towards the role of information in the decision-making process.

Increases in percent of health professional present at birth resulted in increases in case counts, while percent of births at a hospital decreased the case counts. These results are seemingly contradictory, but it may just reflect the ambivalent role the health system played in transmission facilitation (as proposed to be the case earlier during the outbreak), and epidemic control (later).

Higher number of children at home and higher ethnic homogeneity were associated with decreases in case counts. These were more specific to rural Guinea – less affected by the EVD outbreak. Thus, they should not be interpreted ad litteram, but rather in the context of the community location, social structure and behavior.

As discussed in Suchar et al. (2017), Principal component 1 (PC1) was significant in models 1, 3, and 4, PC2 in models 1 and 3, while PC 3 and 4 were significant in all models. The direction of the relationship were, positive for all principal components, except PC2. The interpretation of the relationship between the PC and the case counts is less evident. PC1 was highlighting the differences between urban centers and suburban areas, especially in education and literacy rates, PC2 seemed to capture the sources of variation in socioeconomic status, PC3 weighted access to transportation methods heavily, while PC4 had positive loadings on hospital delivery and doctors present at birth, and negative weights on prevalence of STDs and number of sex partners. The results were less intuitive, since factors that we usually associated with diminished exposure and infection risk (such as higher education, literacy, and access to healthcare), seemed in this analysis to have been associated with both high and low case counts. It is possible that the results are reflecting the differences between countries in social interactions, cultural believes and behavior, differences that were not captured by our model.

Circulation corridors and population size facilitated the spread of EVD. The distance matrices that led to the "best" models were population-weighted and centroid matrices. This confirms that the continuous movement of people from their villages and crossing borders from Guinea to either Sierra Leone or Liberia was the driver behind the rapid spread of Ebola to neighboring countries (WHO, 2015). Population size was accounted for in both the best

models: in Model A within the population-weighted road distance, while in model B in the population offset.

The models captured the outbreak dynamics fairly accurately. Looking at the EVD fitted components in each administrative units and at the two selected models predictions (Figure 2.3, Supplementary material file S6: Figure S13), we can see that Model B was consistently over-predicting the NE components in districts with low case counts, while Model A had a tendency to be less precise for the medium case counts. In the case of Model B, this is probably a result of using a fixed effects (one coefficient-fits-all) for the NE component. Although Model B systematically over-predicted the lower case counts, the loss in predictability of Model A for medium case counts was weighted more in the rmse calculations – which probably this explains the higher rmse score for Model A. Regardless the level of over-prediction is fairly low relative to the heterogeneity in the data.

Models can potentially be useful in effective response to other diseases. Both models predicted fairly well the case counts one-week-ahead in the regions most affected by the outbreak. While their performance diminished in the regions with low and medium case counts, the overall predicted case counts were fairly close to the actual values, as opposed to some earlier models that predicted outlandishly high case counts. Our model predictions went only one week ahead, but the uncertainty associated with more distant predictions would have made them entirely useless. The overall results confirm that the modeling approach can potentially be useful in effective responses to diseases. Still, both models over-predicted the case counts to some degree, but we consider this to be less of an issue than under-prediction, as a decision support tool during active interventions.

2.6. EVD models comparison and concluding remarks

This models are similar in form to those of Dudas et al. (2017); Fang et al. (2016); Santermans et al. (2016): it used a Negative-Binomial function for the within-location case counts for all three countries - as Dudas et al. (2017) and Santermans et al. (2016), has spatial and temporal components – as Fang et al. (2016) and Santermans et al. (2016), and some covariates – as Fang et al. (2016) and Dudas et al. (2017).

Not surprisingly, our results had a lot in common with their conclusions, for example: (1) there was a strong heterogeneity of the spatial and temporal dynamics of the epidemic; (2) access to transportation corridors increased the risk of EVD spread in Sierra Leone. Also, it showed that infection risk was positively associated with proximity to health care workers on the entire West Africa, confirming the results proposed for Sierra Leone (Fang et al., 2016). Our hypothesized region-specific drivers and social responses to the outbreak are in agreement with Dudas et al. (2017) results, that the EVD epidemic in West Africa can be better categorized as an assembly of sub-epidemics, with different transmission characteristics and spatial connectivity, than a single epidemic. There were also differences in the relative magnitude of neighboring effects, but that might be a result of the different spatial scale (districts in our analysis vs. chiefdoms in (Fang et al., 2016)).

The choice of covariates used in the EVD models was wider. For example, Dudas et al. (2017) found that most geographic, administrative, cultural, and climatic variables were not significantly associated with the virus dispersal, except for the population size, and distance from the nearest settlement with more than 50,000 inhabitants, while Fang et al. (2016) found that transmission risk was associated with high population density, proximity to Ebola treatment centers (ETC), and high coverage of cropland. In our models, all covariates used were significant (ethnic homogeneity, % reading newspapers, % hospital delivery of the last child and % health professional present at the birth of last child, and mean children at home), but their contribution to the case counts was fairly low. The models using principal components evaluated over the 24 socio-demographic variables considered fared better, but this was expected since they lump together the effects of all the covariates considered, rather than of selected few. But they still have rather low explanatory power, relative to the autoregressive and neighboring effects. It is not that they were not influential, but they are static – i.e. they are not capturing the behavioral changes during the outbreak, and the effect of interventions. Which stresses out the need to characterize the population behavior during epidemics, in order to correctly inform the health decision-making process. This is not a new result, but just reinforces the conclusions reached earlier during the outbreak, such as Chowell and Nishiura (2015) and Santermans et al. (2016).

This research has shown that models with fairly high predictive power could have been proposed even during the peak of the EVD epidemic. But these efforts were not possible due to limited access or absence of quality data. While many of these deficiencies were unavoidable due to the severity of the epidemic and the limited resources of both affected countries and intervening agencies, we believe that real-time, open access to the quality data, and continuous monitoring at the lowest spatiotemporal resolution possible during outbreaks might be useful, save time, resources, and allow for more effective decision support tools to be created in real time in the future.

2.7. Acknowledgments

This publication was supported by the EPSCoR Program, National Science Foundation #IIA-1301792, the Mountain West Clinical and Translational Research - Infrastructure Network, NIH, National Institute of General Medical Sciences (NIGMS) #1U54GM104944-01A1, and NIGMS #P20GM104420.

References

Ajelli, M., Parlamento, S., Bome, D., Kebbi, A., Atzori, A., Frasson, C., Putoto, G., Carraro, D., Merler, S., 2015. The 2014 Ebola virus disease outbreak in Pujehun, Sierra Leone: epidemiology and impact of interventions. Bmc Med 13, 281.

Althaus, C.L., 2014. Estimating the Reproduction Number of Ebola Virus (EBOV) During the 2014 Outbreak in West Africa. PLoS currents 6.

Bivand, R., Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. Journal of Statistical Software 63, 1-36.

Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., 2008. Applied spatial data analysis with R. Springer, New York, NY 374 pp.

Brinkhoff, T., 2015. City Population. http://www.citypopulation.de/.

Brunsdon, C., Comber, L., 2015. An introduction to R for spatial analysis ans mapping. Sage Publications, UK.

Chowell, G., Nishiura, H., 2014. Transmission dynamics and control of Ebola virus disease (EVD): a review. Bmc Med 12.

Chowell, G., Nishiura, H., 2015. Characterizing the Transmission Dynamics and Control of Ebola Virus Disease. Plos Biol 13, e1002057.

Chowell, G., Viboud, C., Hyman, J.M., Simonsen, L., 2015. The Western Africa ebola virus disease epidemic exhibits both global exponential and local polynomial growth rates. PLoS currents 7.

Czado, C., Gneiting, T., Held, L., 2009. Predictive Model Assessment for Count Data. Biometrics 65, 1254-1261.

Dray, S., Legendre, P., Peres-Neto, P.R., 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Model 196, 483-493.

Dudas, G., Carvalho, L.M., Bedford, T., Tatem, A.J., Baele, G., Faria, N.R., Park, D.J., Ladner, J.T., Arias, A., Asogun, D., Bielejec, F., Caddy, S.L., Cotten, M., D'Ambrozio, J., Dellicour, S., Caro, A.D., Diclaro, J.W., Duraffour, S., Elmore, M.J., Fakoli, L.S., Faye, O., Gilbert, M.L., Gevao, S.M., Gire, S., Gladden-Young, A., Gnirke, A., Goba, A., Grant, D.S., Haagmans, B.L., Hiscox, J.A., Jah, U., Kugelman, J.R., Liu, D., Lu, J., Malboeuf, C.M., Mate, S., Matthews, D.A., Matranga, C.B., Meredith, L.W., Qu, J., Quick, J., Pas, S.D., Phan, M.V.T., Pollakis, G., Reusken, C.B., Sanchez-Lockhart, M., Schaffner, S.F., Schieffelin, J.S., Sealfon, R.S., Simon-Loriere, E., Smits, S.L., Stoecker, K., Thorne, L., Tobin, E.A., Vandi, M.A., Watson, S.J., West, K., Whitmer, S., Wiley, M.R., Winnicki, S.M., Wohl, S., Wölfel,

R., Yozwiak, N.L., Andersen, K.G., Blyden, S.O., Bolay, F., Carroll, M.W., Dahn, B., Diallo, B., Formenty, P., Fraser, C., Gao, G.F., Garry, R.F., Goodfellow, I., Günther, S., Happi, C.T., Holmes, E.C., Kargbo, B., Keïta, S., Kellam, P., Koopmans, M.P.G., Kuhn, J.H., Loman, N.J., Magassouba, N.F., Naidoo, D., Nichol, S.T., Nyenswah, T., Palacios, G., Pybus, O.G., Sabeti, P.C., Sall, A., Ströher, U., Wurie, I., Suchard, M.A., Lemey, P., Rambaut, A., 2017. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature advance online publication.

Fang, L.Q., Yang, Y., Jiang, J.F., Yao, H.W., Kargbo, D., Li, X.L., Jiang, B.G., Kargbo, B., Tong, Y.G., Wang, Y.W., Liu, K., Kamara, A., Dafae, F., Kanu, A., Jiang, R.R., Sun, Y., Sun, R.X., Chen, W.J., Ma, M.J., Dean, N.E., Thomas, H., Longini, I.M., Halloran, M.E., Cao, W.C., 2016. Transmission dynamics of Ebola virus disease and intervention effectiveness in Sierra Leone. P Natl Acad Sci USA 113, 4488-4493.

Faye, O., Boëlle, P.-Y., Heleze, E., Faye, O., Loucoubar, C., Magassouba, N.F., Soropogui, B., Keita, S., Gakou, T., Bah, E.H.I., Koivogui, L., Sall, A.A., Cauchemez, S., 2015. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. The Lancet Infectious Diseases 15, 320-326.

Fink, D.G., Sheri, 2014. Tracing Ebola's Breakout to an African 2-Year-Old.
Fisman, D., Khoo, E., Tuite, A., 2014. Early epidemic dynamics of the west african 2014 ebola outbreak: estimates derived with a simple two-parameter model. PLoS currents 6.
Getis, A., Aldstadt, J., 2004. Constructing the spatial weights matrix using a local statistic. Geogr Anal 36, 90-104.

Greene, S.K., Peterson, E.R., Kapell, D., Fine, A.D., Kulldorff, M., 2016. Daily Reportable Disease Spatiotemporal Cluster Detection, New York City, New York, USA, 2014–2015. Emerging Infectious Diseases 22, 1808-1812.

HDX, 2015. Sub-national time series data on Ebola cases and deaths in Guinea, Liberia, Sierra Leone, Nigeria, Senegal and Mali since March 2014.Humanitarian Data Exchange. Retrieved November 7, 2015 from https://data.hdx.rwlabs.org/dataset/rowca-ebola-cases

Hohle, M., Meyer, S., Paul, M., 2016. surveillance: temporal and spatio-temporal modeling and monitoring of epidemic phenomena, R package version 1.12.1 ed.

Hui-Jun, L., Jun, Q., David, K., Xiao-Guang, Z., Fan, Y., Yi, H., Yang, S., Yu-Xi, C., Yong-Qiang, D., Hao-Xiang, S., Foday, D., Yu, S., Cheng-Yu, W., Wei-Min, N., Chang-Qing, B., Zhi-Ping, X., Kun, L., Brima, K., George, F.G., Jia-Fu, J., 2015. Ebola Virus Outbreak Investigation, Sierra Leone, September 28–November 11, 2014. Emerging Infectious Disease journal 21.

HumanitarianResponse, 2016. Sierra Leone Situation Reports. Government of Sierra Leone. https://www.humanitarianresponse.info/.

INS, 2013. Enquete demographique et de sante et a indicateurs multiples

(EDS-MICS 2012). Institut National de la Statistique (INS) and ICF International, Conakry, Guinée and Calverton, Maryland, USA.

Jones, R.C., Liberatore, M., Fernandez, J.R., Gerber, S.I., 2006. Use of a Prospective Space-Time Scan Statistic to Prioritize Shigellosis Case Investigations in an Urban Jurisdiction. Public Health Reports 121, 133-139.

Kahle, D., Wickham, H., 2013. ggmap: Spatial Visualization with ggplot2. The R Journal 5, 144-161.

Kulldorff, M., 1997. A spatial scan statistic. Communications in Statistics: Theory and Methods 26, 1481-1496.

Kulldorff, M., 2009. SaTScanTM v9.0: Software for the spatial and space-time scan statistics. http://www.satscan.org. http://www.satscan.org.

Kulldorff, M., Athas, W.F., Feuer, E.J., Miller, B.A., Key, C.R., 1998. Evaluating cluster alarms: a spece-time scan statistic and brain cancer in Los Alamos, New Mexico. American Journal of Public Health 88, 1377-1380.

Lau, M.S.Y., Dalziel, B.D., Funk, S., McClelland, A., Tiffany, A., Riley, S., Metcalf, C.J.E., Grenfell, B.T., 2017. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. Proceedings of the National Academy of Sciences 114, 2337-2342.

Lewnard, J.A., Mbah, M.L.N., Alfaro-Murillo, J.A., Altice, F.L., Bawo, L., Nyenswah, T.G., Galvani, A.P., 2014. Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis. Lancet Infectious Diseases 14, 1189-1195.

LISGIS, 2014. Liberia Demographic and Health Survey 2013. Liberia Institute of Statistics and Geo-Information Services (LISGIS), Ministry of Health and Social Welfare, National AIDS Control Program, and ICF International, Monrovia, Liberia.

Luetkepohl, H., 2011. Vector autoregressive models. EUI Working Papers, 33.
Marek, L., Tuček, P., Pászto, V., 2015. Using geovisual analytics in Google Earth to understand disease distribution: a case study of campylobacteriosis in the Czech Republic (2008–2012). International Journal of Health Geographics 14, 1-13.

Martins-Melo, F.R., Ramos, A.N., Alencar, C.H., Lange, W., Heukelbach, J., 2012. Mortality of Chagas' disease in Brazil: spatial patterns and definition of high-risk areas. Trop Med Int Health 17, 1066-1075.

Merler, S., Ajelli, M., Fumanelli, L., Gomes, M.F., Piontti, A.P., Rossi, L., Chao, D.L., Longini, I.M., Jr., Halloran, M.E., Vespignani, A., 2015. Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical

interventions: a computational modelling analysis. The Lancet. Infectious diseases 15, 204-211.

Meyer, S., Held, L., Hohle, M., 2016. Spatio-temporal analysis of epidemic phenomena using the R package surveillance. Journal of Statistical Software In press.

Mitze, T., 2012. Empirical modeling in regional science: towards a global time-space-structural analysis. Springer.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289-290.

Paul, M., Held, L., 2011. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. Stat Med 30, 1118-1136.

Pfaff, B., 2008a. Analysis of Integrated and Cointegrated Time Series with R, 2nd ed. Springer, New York, NY.

Pfaff, B., 2008b. VAR, SVAR and SVEC Models: Implementation Within R Package vars. Journal of Statistical Software 27, 1-32.

R Core Team, 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Richards, P., 2015. How a people's science helped end an epidemic. Zed Books 300 pp.

Rivers, C.M., Lofgren, E.T., Marathe, M., Eubank, S., Lewis, B.L., 2014. Modeling the impact of interventions on an epidemic of ebola in sierra leone and liberia. PLoS currents 6. Santermans, E., Robesyn, E., Ganyani, T., Sudre, B., Faes, C., Quinten, C., Van Bortel, W., Haber, T., Kovac, T., Van Reeth, F., Testa, M., Hens, N., Plachouras, D., 2016.

Spatiotemporal Evolution of Ebola Virus Disease at Sub-National Level during the 2014 West Africa Epidemic: Model Scrutiny and Data Meagreness. Plos One 11.

Schnute, J.T., Boers, N., Haigh, R., 2015. PBSmapping: Mapping Fisheries Data and Spatial Analysis Tools. http://CRAN.R-project.org/package=PBSmapping.

Sherman, R.L., Henry, K.A., Tannenbaum, S.L., Feaster, D.J., Kobetz, E., Lee, D.J., 2014. Applying Spatial Analysis Tools in Public Health: An Example Using SaTScan to Detect Geographic Targets for Colorectal Cancer Screening Interventions. Preventing Chronic Disease 11, E41.

Shumway, R.H., Stoffer, D.S., 2011. Time series analysis and its applications. With R examples, 3rd ed, New York, NY 595 pp.

SSL, 2014. Sierra Leone Demographic and Health Survey 2013. Statistics Sierra Leone (SSL) and ICF International., Freetown, Sierra Leone and Rockville, Maryland, USA.

Suchar, V.A., Aziz, N., Bowe, A., Burke, A., Wiest, M.M., 2018. An exploration of the spatiotemporal and demographic patterns of Ebola Virus Disease epidemic in West Africa using open access data sources. Applied Geography 90, 272-281.

Suchar, V.A., Aziz, N., Bowe, A., Wiest, M., 2017. An Exploration of the Spatiotemporal and Demographic Patterns of the Ebola Virus Disease Epidemic in West Africa Using Open Access Data Sources. Journal of Applied Geography under review.

USAID, 2016. Demographic and Health Surveys (DHS) Program. http://dhsprogram.com/.
Viton, P., 2010. Notes on spatial econometric models, Proceedings of the City and Regional Planning Conference.

Westcott, L., 2014. Sierra Leone Grapples With Spike in Ebola Numbers. @newsweek.
WHO, 2016a. Situation Report. Ebola virus disease. 10 June 2016. World Heath Organization. http://www.who.int/csr/disease/ebola/en/.

WHO, 2016b. WHO | Disease Outbreak News (DONs). http://www.who.int/csr/don/en/. WHO.

Wikipedia, 2015. Wikipedia: The Free Encyclopedia. Wikimedia Foundation Inc. http://www.wikipedia.org.

Zavis, A., Healy, M., 2016. Ebola cases in West Africa may be vastly underreported, WHO says. @latimes.