Estimating the Impact of Cultural Variation on Epidemic Behavior

A Thesis
Presented in Partial Fulfillment of the Requirements for the
Degree of Master of Science
with a
Major in Statistical Science
in the
College of Graduate Studies
University of Idaho
by
Noha Aziz

Major Professor: Michelle Wiest, Ph.D.
Committee Members: Ben Ridenhour, Ph.D., Timothy R. Johnson, Ph.D.
Department Administrator: Christopher Williams, Ph.D.

November 2015

Authorization to Submit Thesis

The thesis of Noha Aziz, submitted for the degree of Master of Science with a major in Statistical Science and titled, " ESTIMATING THE IMPACT OF CULTURAL VARIATION ON EPIDEMIC BEHAVIOR," has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ____________________________________ Date: _______________
Michelle Wiest, Ph.D.

Committee Members: ____________________________________ Date: _______________
Ben Ridenhour, Ph.D.

____________________________________ Date: _______________
Timothy R. Johnson, Ph.D.

Department
Administrator: ____________________________________ Date: _______________
Christopher Williams, Ph.D.

Abstract

Ebola was discovered in 1976 (WHO, 2015). It has killed many people throughout history, but the 2014 West Africa epidemic was an apocalypse-sized outbreak. This epidemic would have harvested lives five times more than the lives that all known Ebola epidemic has harvested combined (BBC, 2015).

At a basic level, the success of interventions depends on behavioral responses to disease conditions, as well as risk perception and cultural norms that might conflict with interventions. Such sociological conditions can be very fluid and influenced by circumstances on the ground as well as the evolution of opinions within the community.

Sierra Leone, Liberia and Guinea have a unique combination of geographic, sociocultural and political factors (Walker & Whitty, 2015). Those factors combined together in one region formed an ideal environment for the epidemic to explode. The absence of accurate critical data hindered understanding the dynamics of the epidemic and created a hazy and unclear picture of the situation and the conditions that helped the Ebola outbreak to evolve. Unfortunately, most models of emerging epidemics do not account for cultural variation which impacts the case incidence within and across affected countries.

This thesis begins to the fill the gap in understanding how sociocultural characteristics can affect the dynamics of an emerging infectious disease. To accomplish this, we focused on Liberia as a case study. We evaluated the association of social and cultural variables and actual case counts by using principal component analysis (PCA), autoregressive integrated moving average (ARIMA) models and cross correlation between the counties to explore the relationship between sociological and cultural factors and the epidemic.

Evaluating the dataset using the previously mentioned methods showed that there is evidence that some behavioral responses might have played a role in spreading the Ebola epidemic. For instance, continues movement of the people between the counties was associated with spread, whereas there were not enough evidence for the role of other demographic attributes such as people’s level of education.

Acknowledgements

It is with great gratitude I thank Dr. Michelle Wiest for her guidance and help in developing and completing this research. Her curiosity and interests toward the dynamics that helped the spread of EVD epidemic was the first spark that initiated the research and work behind this thesis. The continuous guidance and directing from Dr. Michelle Wiest while working on finishing the thesis was the torch that lighted the way for me toward the right path of the field of research.

Thank you to committee members Dr. Ben Ridenhour, Dr. Timothy R. Johnson who helped and enlightened the work on this thesis.

I would like to thank the Social Ebola group at the Center for Modeling Complex Interactions at the University of Idaho for their help and support. Many other friends and family members that were not mentioned here.

Dedication

To my family:

My parents: Mohammmedamin Aziz

Lutifa Khuja

My husband: Hani Alturkostani

My children: Yazan and Yamen

Thank you for your valuable and unlimited support

Table of Contents

List of Figures

List of Tables

Chapter 1: Literature Review

Ebola Virus

Ebola virus disease was first isolated from Myriam Louise Ecran, a 42-year-old Belgian nursing sister who was working at the Yambuku Mission Hospital in Congo. She died from the Ebola Virus Disease (EVD) after she was infected by the patients she was taking care of (CDC, 2015).

The word “Ebola” is a corruption of the Ngbandi (language spoken in Demographic Republic of Congo) name Legbala, meaning “white water” or “pure water” (CDC, 2015). The reason behind choosing the word Ebola to describe this virus was because it is the name of the Ebola River which is located near a village in Yambuku, Demographic Republic of Congo, one of the two first neighboring locations where the Ebola virus appeared in 1976. The other one was in Nzara, Sudan (WHO, 2015).

According to Kirch (2008), “Ebola hemorrhagic fever is an acute viral illness caused by the Ebola virus which belongs to the Filovirus group.” The CDC (2014) defined Ebola as “a rare and deadly viral hemorrhagic fever.”

Ebola symptoms are fever greater than 40°C and one or more of the following clinical findings: severe headache, muscle pain, rash on the trunk, vomiting, diarrhea, inflammation of the in the back of the throat, abdominal pain, bleeding not related to injury, retrosternal chest pain, the presence of abnormal quantities of protein in the urine, deficiency of platelets in the blood which causes bleeding into the tissues, bruising, and slow blood clotting after injury. (CDC, 2011). Kirch (2008), listed the symptoms as “acute fever, diarrhea (often bloody), nausea, vomiting, and headache and at a later stage, nosebleeds, conjunctival infection, dysphagia and affection of central nervous system.”

Those symptoms are similar to symptoms of other infectious disease which make it difficult to distinguish between Ebola and malaria, typhoid fever and meningitis (WHO, 2015). In order to confirm the infection with Ebola virus the following tests needs to be run:

- antibody-capture enzyme-linked immunosorbent assay (ELISA)
- antigen-capture detection tests
- serum neutralization test
- reverse transcriptase polymerase chain reaction (RT-PCR) assay
- electron microscopy

- Virus isolation by cell culture

Maximum precautions must be taken in collecting and dealing with samples as they considered an extreme biohazard risk (WHO, 2015), as transmission from person to person happens by "direct contact or by drops through mucous membranes or indirectly by infected blood, secretions, semen and vomit (Kirch, 2008) ." The human-to-human transmission according to the WHO (2015) happens "via direct contact (through broken skin or mucous membranes) with the blood, secretions, organs or other bodily fluids of infected people, and with surfaces and materials (e.g. bedding, clothing) contaminated with these fluids."

Although the Ebola virus first appeared in two neighboring locations in Congo and Sudan, the two epidemics were caused by two different strains of Ebola virus: Sudan Ebola virus and Zaire Ebola virus (Feldmann, H., & Geisbert, T., 2011). Those two viruses are considered the source of almost all human cases. In regions of Gabon, Republic of the Congo, and DRC, Zaire Ebola virus is responsible for the emergence of the epidemic. Whereas in Sudan and Uganda, it was Sudan Ebola virus (Feldmann, H., & Geisbert, T., 2011).

The reservoir species - natural host - or the source of the Ebola virus disease is unknown, which makes it difficult to predict the occurrence of outbreaks (Sanchez, 2003). It is thought that rodents and bats are the source of Ebola virus but there is not any firm link between these two species and the virus. As a result of not knowing the actual source of the Ebola virus, there is no proven treatment available for the EVD. The only treatment that medical care facilities could offer to Ebola virus patients is "supportive care-rehydration with oral or intravenous fluids- and treatment of specific symptoms" (WHO, 2015) which improves survival chances for Ebola patients.

West African Epidemic

The 2014 Ebola epidemic started on 26 December 2013 with a 2-year old boy who was living in Meliandou a remote Guinean village that is located in a forested area between the borders of Guinea, Liberia and Sierra Leone. The boy became sick by a "mysterious illness characterized by fever, black stools, and vomiting" (WHO, 2015), which was followed by his death two days later. After the boy's death, the Ebola virus spread to his family, which could have happened because of their direct contact with him. Most of the early cases that were detected spread by "caring for a sick relative, preparing a body for burial, or delivering a baby (WHO, 2015).” The continuous movement of people from their villages and crossing borders from Guinea to either Sierra Leone or Liberia was the main cause behind the quick spread of Ebola to neighboring countries (WHO, 2015). Since March 2014, 28,575 cases were reported with 11,313 deaths in 5 countries in West Africa: Guinea, Sierra Leone, Liberia, Senegal, and Nigeria (WHO, 2015) (Figure 1).

Figure 1 : West Africa Map
(source :Ensheng Dong by ESRI ARCMap)

The first two cases of the epidemic in Liberia were detected in 30 March 2014 at Lofa County near the Guinea border (Figure 2). On 7 April 2014, 4 more cases were found in Lofa and one in Monrovia. During the rest of April and May there were no more cases in Liberia, however in June, the epidemic reappeared in Lofa with 41 cases. By mid-June Ebola cases started to appear in Monrovia the capital of Liberia. As a result of the high population density and the circumstances that surround the Ebola mechanism of infection, the number of Ebola cases escalated very quickly in Monrovia. The huge numbers of Ebola cases led the Liberian president to announce a three month state of emergency with strict measures in hopes to control the spread of Ebola virus (WHO, 2015).

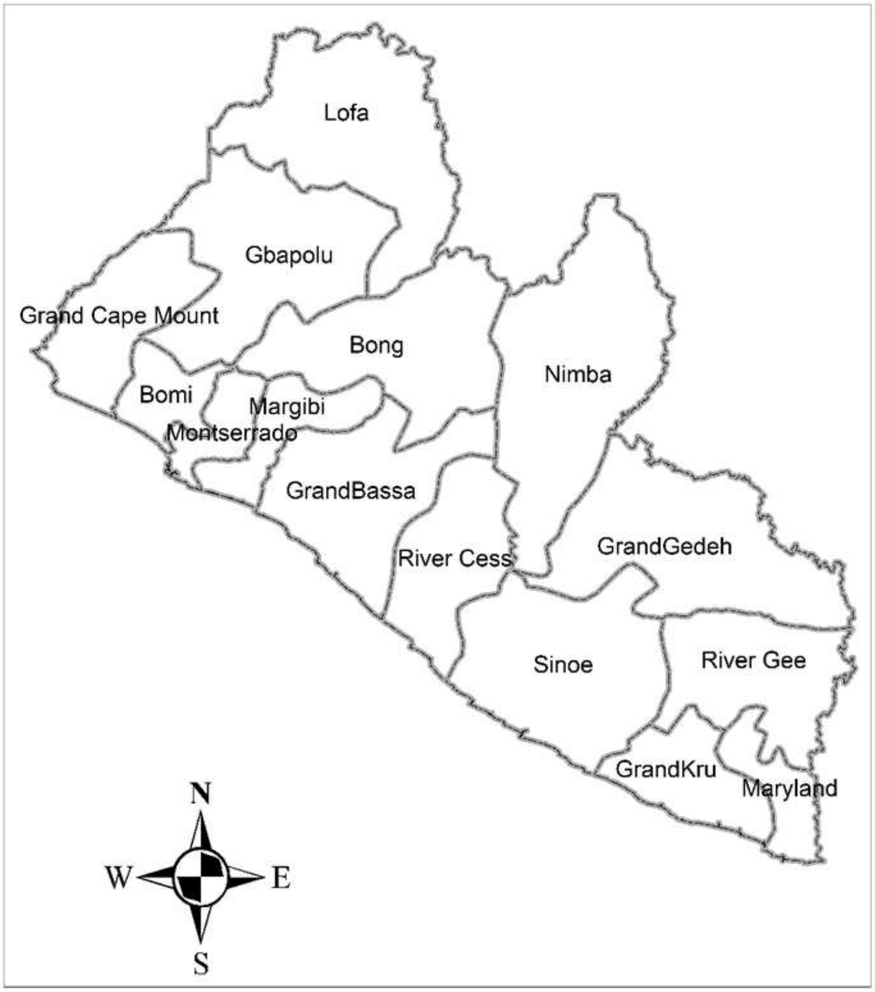


Figure 2 : Liberia Map

source :Ensheng Dong by ESRI ARCMap

The huge spread of the Ebola virus in West Africa, which made it one of the major epidemics to hit West Africa, was fueled by several unique factors that distinguish the West Africa area from any other region had the virus in the past. Sierra Leone, Liberia and Guinea have a unique combination of geographic, sociocultural and political factors (Walker & Whitty, 2015). Those factors, combined together in one region, formed an ideal environment for the epidemic to explode. This means that any factor by itself cannot be blamed for the wide spread of Ebola virus nor the delayed care and treatment of sick individuals (Walker &

Whitty, 2015). Long years of war in West Africa left the countries in the region with weak civil infrastructure. Hospitals, health care workers and medical supplies were a fragile triangle that were expected to treat an extreme number of very ill patients (Suffredini, 2015). Most of the people in those countries had no direct access to hospitals, and they would go to a community-based care with limited capacity for treatment (Suffredini, 2015). Even when hospitals were available, basic medical care supplies such as: laboratory monitoring, diagnostic imaging, IV therapy, supplemental oxygen, mechanical ventilation, or dialysis were scarce (Suffredini, 2015). The limited number of hospitals were suffering from a shortage in trained medical staff and left to face too many challenges by themselves.

The health workers in West Africa had little to no experience in handling and treating epidemics and viruses. According to (Walker & Whitty, 2015), "many healthcare workers lost their homes and were ostracized from their communities due to stigma. Some healthcare workers were forced to work in unsafe environments due to financial hardship and others were obligated to attend to sick friends and relatives in their community, increasingly as patients feared attending health facilities." Those healthcare workers were working in a very difficult circumstances. They needed to wear personal protective equipment (PPE) to protect themselves from being infected. The heat and humidity of West Africa and the number of patients they needed to take care of, led the healthcare workers to shorten the time of helping each case into minutes. (Walker & Whitty, 2015). In later stages of the epidemic, trained international health workers arrived in West Africa to help, but they faced communication challenges with the local community (Fowler, 2014).

The sociocultural role in spreading the Ebola virus consisted of the Ebola patient's family role, continuous movement of people, burial practices, local healers and midwives. At the beginning of the epidemic, most of the patients that were admitted to hospitals would not survive from the virus, which caused the relatives of sick individuals to avoid their admission to hospitals or any medical facility to seek help and treatment (Fowler, 2014).

Traditional West African burial practices mostly consist of washing and touching the deceased body during funeral ceremonies (Walker & Whitty, 2015). Ebola virus is transmitted through bodily fluids and secretions, and even after death the deceased person remains infectious. Eighty-five Ebola cases in Guinea were traced down to a single source which was a funeral of an Ebola virus victim (Walker & Whitty, 2015).

Midwives played a huge role in spreading the Ebola virus. For instance, the family of the index case and the nurse and village midwife who attended the child also died from Ebola. In all, 14 people, including funeral attendees and a health worker from the local hospital died of Ebola. The family of the health worker was subsequently infected (Ki, 2014).

The role of political infrastructure in spreading the Ebola virus was multifaceted. One of the reasons was the people's mistrust of the government. The governments of West African countries responded differently in each country and took advantage of the public fear of the epidemic. For example, the Liberian government saw that huge spread of Ebola was the perfect opportunity to seek more international help (Walker & Whitty, 2015). On the other end in Sierra Leone, the government reported only the confirmed cases and ignored the suspected ones in fear of public panic (Walker & Whitty, 2015).

Illustrative Attempts to Predict the Epidemic Using Models

Since the first Ebola case at December 2013 (WHO, 2015), there have been many attempts to understand the dynamics of this epidemic and the factors that promoted its wide spread. Here, we review three such analyses as examples of modeling approaches that have been used thus far.

Rainisch *et al.*, (2015) attempted to predict the regional spread of the 2014 West Africa Ebola virus. They used case counts, population data and distances between affected and non-affected districts to predict the spread of the virus. The general linear regression model which uses Sum of Inverse Distance (SID) as distance matrix served as a good predictor for the timing of the Ebola virus to reach the non-affected counties. Sometimes the model would show early prediction in some counties before the actual recorded cases of that counties. This discrepancy in prediction might be due to underestimated case numbers, one of the problems that faced this paper and many others. While this paper did account for the distances of counties from each other, it did not consider people's travel patterns between them. Travel patterns on can have huge influence disease spread. Considering that this model is easy to adapt, if this information became available the model used in this paper could be adjusted to consider the travel patterns as a fixed variable in the model.

Another paper that detailed a compartmental mathematical model for the epidemic was Lewnard *et al.*, (2014). This paper tried to predict the control of Ebola virus in

Montserrado County through three different ways: EVD units, case ascertainment and protective kits. This paper attempted to avoid the problem of underestimated or unreported number of cases by calculating the “worst case scenario” for the number of cases that could have been occurring. The authors note many limitations that faced this approach. Assuming equal rates of transmissions between dead and alive people was one of the major limitations, as the number of transmissions between live and dead cases varied widely. Considering that the paper only studied the cases in Montserrado, it did not take into account the continuous movement of people going in and out of the county. The model assumed that sanitary burial for EVD deceased cases would not produce any more transmission to people, which is not scientifically proven. This reflects one of the major challenges that faced the several attempts to model the sociocultural effects on EVD spread. Knowing the exact mechanism of sociocultural impacts on the spread of EVD, which makes it difficult to produce an accurate mathematical model to predict the epidemic.

Merler *et al.*, (2015) modeled the epidemic in Liberia using the most recent data available, which allowed this paper avoid the inaccurate results Lewnard *et al.,* (2014) paper. Merler *et al*. used an agent based model to model the movements of people who are not infected with EVD and could be attending health care facilities or taking care of an infected person. These parameters are used to investigate the effectiveness of EVD treatment units, safe burials and protective kits. There were no data available for hospital admission nor reported case rates numbers, therefore, the authors estimated the numbers from previous outbreaks. The authors note that they assumed that distributing the protective kit would mean that the individuals will use it, which is not necessarily true that the kits will be used. They did not consider the mobility due to commuting patterns which might affect the Ebola virus spread model as mentioned before. A strength of this paper is that they created an extendable model in case of more data or pharmaceutical interventions became available.

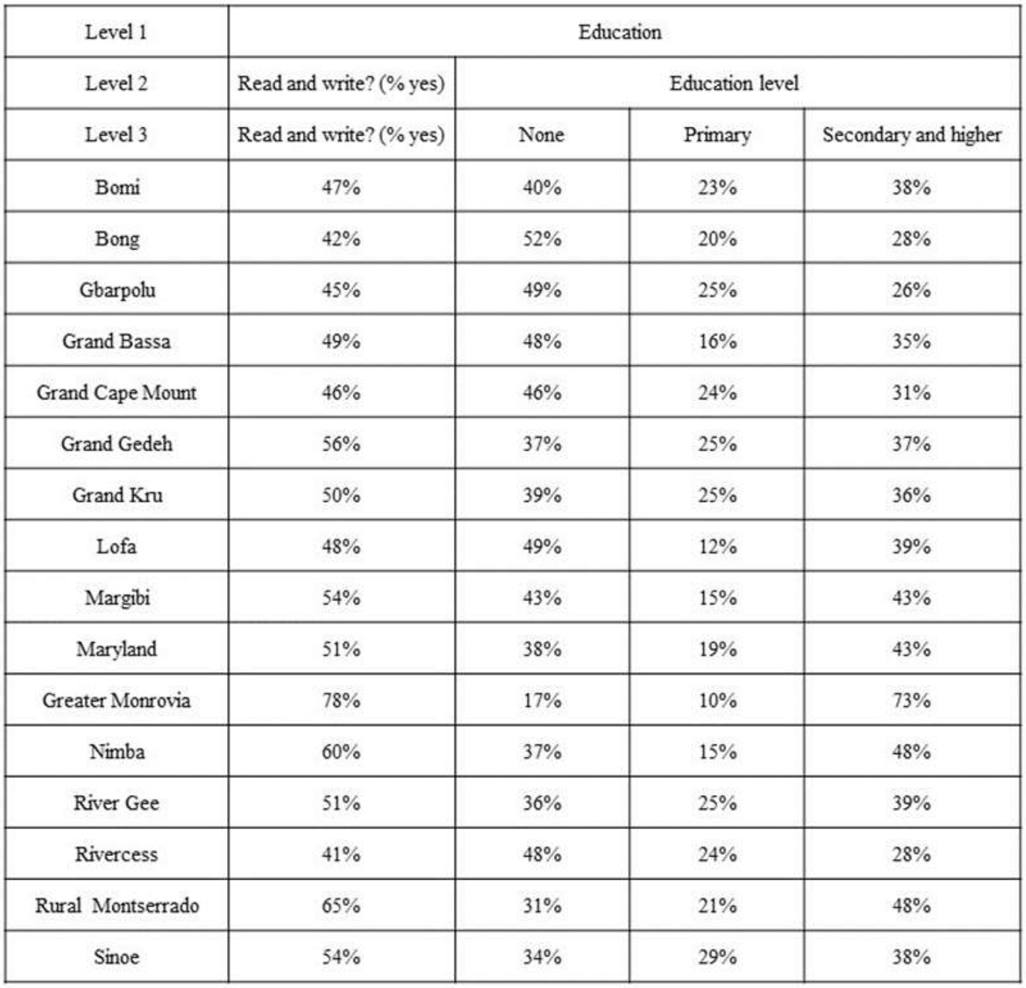
The previous studies Rainisch *et al.,* (2015), Lewnard *et al.,* (2014) and Merler *et al.,* (2015) focused only on EVD case counts, population size, distances between the counties and the effectiveness of EVD treatment units, safe burials and protective kits. There was no use of any cultural or demographic data that might affected or helped the spread of the EVD epidemic on West Africa.

In this thesis, this gap is being filled by understanding the relationship between the cultural behaviors of people in Liberia and the emergence of Ebola virus. Data were compiled from three different sources to form one combined dataset that contained cultural and demographic data and case counts of all 15 counties in Liberia. Then PCA was applied to explore the similarities between the 15 counties and categorize them. After performing PCA on the data, ARIMA model and then CCF were applied between the counties that shared the same category. The CCF was applied to explore any links between the counties that might have helped the spread of EVD. This work helps explain how the epidemic was spreading one region more than another. I hope this work can inform future attempts of modeling emerging infectious disease.

Chapter 2: Data Collecting and Cleaning

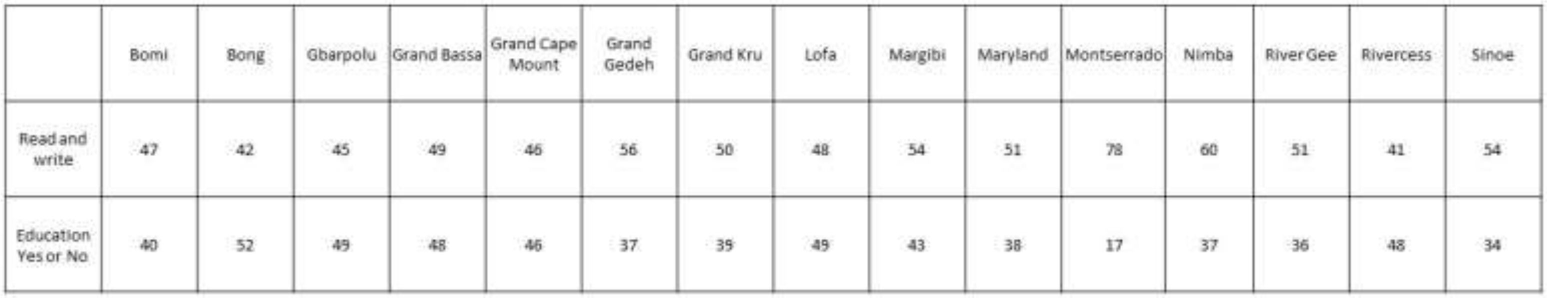
In order to investigate the associations between the cases and social and demographic characteristics of Liberian counties, available data needed to be gathered from different sources and merged into one dataset. This chapter explains the process of gathering the data from three different sources, which made a complete combined dataset that contained demographic, social characteristics, case counts and population of each county. The data were compiled from three different sources: Vinck *et al.,* (2011), these data contained different demographic and ethnicity variables such as level of literacy, ethnic distribution, etc.; (Humanitarian Data Exchange, 2015), these data contained the Ebola case counts for different counties of different countries from different sources; (Geohive, 2008) data which contained the total population for each county.

The first dataset, which was obtained from Vinck *et al.,* (2011), is a study implemented in Liberia in November and December 2010. The study was based on a nationwide survey in cooperation with local organizations. In order to work with these data properly, several steps had to be performed to adjust the dataset and put it in the same format as the other sources (Table 1), which allowed for merging all three data sets into one complete data set. First, the data were transposed to have the counties as rows and the different variables as columns. Then, the original dataset variables were divided into three levels: general title, a question and the different answers to the question. First level (the general title) and second level (the question) were removed and the third level (the answers of the questions) was kept. Then, the data include two categories for Montserrado County: Greater Monrovia which contained the capital and its surroundings and Rural Monserrado which contained the rural area in Montserrado. A decision had to be made to choose one of the two categories to represent Montserrado in the data. As merging the two categories were not an option. The numbers in the dataset were in percentage that reflected the total weight of the observations according to the counties populations. After observing the numbers of both categories, it was found that rural Montserrado had low percentage of surveyed individuals compared to all other fifteen counties, where Greater Monrovia had the larger percentage of surveyed individuals. This observation led to the exclusion of Rural Montserrado and considering Greater Monrovia to represent Montserrao County in the data, especially given most cases in Montserrado were from Monrovia. Finally, the category name was changed from Greater Monrovia to Montserrado (Table2).



Level 1	Education			
Level 2	Read and write? (% yes)	Education level		
Level 3	Read and write? (% yes)	None	Primary	Secondary and higher
Bomi	47%	40%	23%	38%
Bong	42%	52%	20%	28%
Gbarpolu	45%	49%	25%	26%
Grand Bassa	49%	48%	16%	35%
Grand Cape Mount	46%	46%	24%	31%
Grand Gedeh	56%	37%	25%	37%
Grand Kru	50%	39%	25%	36%
Lofa	48%	49%	12%	39%
Margibi	54%	43%	15%	43%
Maryland	51%	38%	19%	43%
Greater Monrovia	78%	17%	10%	73%
Nimba	60%	37%	15%	48%
River Gee	51%	36%	25%	39%
Rivercess	41%	48%	24%	28%
Rural Montserrado	65%	31%	21%	48%
Sinoe	54%	34%	29%	38%

Table 1: Example of the Data Before Adjusting



	Bomi	Bong	Gbarpolu	Grand Bassa	Grand Cape Mount	Grand Gedeh	Grand Kru	Lofa	Margibi	Maryland	Montserrado	Nimba	River Gee	Rivercess	Sinoe
Read and write	47	42	45	49	46	56	50	48	54	51	78	60	51	41	54
Education Yes or No	40	52	49	48	46	37	39	49	43	38	17	37	36	48	34

Table 2: Example of the Data After Adjusting

The second dataset was obtained from (Humanitarian Data Exchange, 2015). These data came from United Nations Office for the Coordination of Humanitarian Affairs Regional

Office for West Africa and Central Africa (OCHA ROWCA). In order to prepare the data, Liberia dataset that was reported from the government source were extracted. The data from the government were the most complete, reliable and consistent data. These data contained case counts for the 15 counties for the longest period of time (7 months). The United Nations Children's Fund (UNICEF) source contained case counts for only 6 counties plus the national counts for about two and a half months. The World Health Organization (WHO) source contained case counts for only 4 counties plus the national counts for about four months. The county level data had 16 categories, the 15 counties plus a category named “national”. National counts had to be removed, as it did not represent any specific county.

These dataset had 6 categories for case counts (probable, suspected, cases, confirmed, new cases and death). New cases counts was the only category that were extracted from this dataset. First, Death counts were removed. Next the number of case counts for the different categories were compared to have a better understanding of the differences between the different case counts categories and determine which category represents the right case counts for all new cases.

Suspected case according to the WHO report is

- “any person, alive or dead, suffering or having suffered from a sudden onset of high fever and having had contact with: - a suspected, probable or confirmed Ebola case; - a dead or sick animal”
- “ any person with sudden onset of high fever and at least three of the following symptoms: headaches, vomiting, anorexia / loss of appetite, diarrhea, lethargy, stomach pain, aching muscles or joints, difficulty swallowing, breathing difficulties, hiccup”
- “ any person with inexplicable bleeding”
- ” any sudden, inexplicable death.”

A probable case according to the WHO report is

- “any suspected case evaluated by a clinician”
- “any deceased suspected case (where it has not been possible to collect specimens for laboratory confirmation) having an epidemiological link with a confirmed case.”

These definitions meant that suspected and probable cases are linked together and the numbers of cases for both categories overlapped and that both the suspected and probable cases categories are not precise. This conclusion led to the exclusion of both categories from the case counts. Confirmed cases according to the WHO are cases that has "Laboratory-confirmed diagnostic evidence of Ebola virus infection." This definition means that this category had the cumulative cases from both suspected and probable. This category was not consistent and reliable. The cases category combined all the counts (suspected, probable and confirmed cases), which resulted in overestimated numbers. Furthermore, "cases" was also cumulative. The previous observations on the different case counts for all the cases categories meant the new cases category was the most consistent and reliable case counts.

The epidemic in these data set lasted about 59 weeks. Started at 24 August 2014 and ended at 9 February 2015. These data set started the case counts at 24 August 2014 with counts for Lofa, Montserrado, Magribi and Nimba. The case counts occurred for only those counties because those counties had the epidemic before the others. The rest of the counties had no case counts at this date. This pattern occurred several times throughout the dataset where some counties lacked some count records at some weeks. It was assumed that absence of reports meant there were no new cases.

In order to have proper case counts, case records for each county in every date were needed. To solve this problem, a function was created to check if a date existed for each county in the dataset, starting from the first recorded date. If it did not exist, then the function will create a date and fill it with zeros as an indicators that there was no records at this date. After completing this process a dataset for the case counts of each of the 15 counties for every day for the whole 59 epidemic weeks were created.

Next, the created data set was merged with the total population for each county (Table 3), which was obtained from (Geohive, 2008). Finally, the dataset was merged with the Vinck *et al.*, (2011) dataset to have a final dataset that contained the total population, different demographic and ethnicity variables and the case counts for every day for the whole 59-epidemic week.

County	Total Population
Bomi	84,119
Bong	333,481
Gbarpolu	83,388
Grand Bassa	221,693
Grand Cape Mount	127,076

Grand Gedeh	125,258
Grand Kru	57,913
Lofa	276,863
Margibi	209,923
Maryland	135,938
Montserrado	1,118,241
Nimba	462,026
River Gee	66,789
River Cess	71,509
Sinoe	102,391

Table 3: County Populations

The Vinck *et al.*, (2011) dataset was important to understand the different patterns of the epidemic across the 15 counties of Liberia. But the problem with this dataset was it has 160 variables, the interpretation of so many variables at once would be intractable. Several steps were applied to reduce the 160 variables. First, some of the variables were eliminated as it was not related to the thesis subject such as questions regarding the Charles Taylor trial. After selecting out variables that were not in place of interest for the modeling, 48 variables remained. Abbreviations were made to the remaining variables because it would be easier to distinguish between them in figures. Second, the ethnicity variable were separated which contained 16 variables, then the religion variable were separated which contained 4 variables. After that, the rest of the data, which we refer to as the social and political dataset, contained 28 variables. The available data were searched to look for the major factors that could affect the Ebola case counts and then to explored the relationship between those factors and the Ebola epidemic. Those estimates were compared to the county-level case totals that were reported on February 2015.

Name	Description
Educated.1	Read and write
Educated.2	Not read and write
Income	Income less than 1.25 USD a day
Info. Source	Information from family and friends
Media habits.1	Listen to radio news

Media habits.2	Never read news paper
Media habits.3	Watch TV at least occasionally
Health.1	Priority health
Health.2	Access to health care
Family	Have relationship with family
Groups	Member of religious group
Technology.1	Cellphone ownership
Technology.2	Radio ownership
Technology.3	TV ownership
War effect.1	Lost a friend due to war
War effect.2	Witnessed rape or sexual abuse during war
War effect.3	Took part in war
Safty.1	Area not safe
Safty.2	Area safe
Police.1	Improve police capacity
Police.2	Know location of police station
Ethnicity.1	Have relations with other ethnic groups
Ethnicity.2	Have interactions with other ethnic groups
Ethnicity.3	Some tribes are favored more than others
Ethnicity.4	Have problems with other ethnic groups
Cause of war	Cause of war are greed and corruption

Table4: List of the Abbreviation Used in Social and Political PCA

Chapter 3: Modeling the association between social and cultural components and epidemic data, analysis and results

After collecting, cleaning and rearranging the data, a complete combined dataset that contained demographic, social characteristics, case counts and population of each county was created. Analysis had to be made about this dataset using the three different categories of the counties characteristics: religion, ethnicity and social and political and counties infection rates. The 15 counties were categorized onto three different infection rate levels according to the case rate per 10,000. Principal component analysis (PCA) were performed on each of the three categories: religion, ethnicity and social and political using prcomp function in R. After that, ARIMA was applied on each county and cross correlations between the counties to evaluate the patterns within each county as well as the associations between counties.

Principal Component Analysis:

The main purpose of PCA is finding the maximum variance of a linear combination of the variables. This allows one to discover the directions where the data is most spread out.

Suppose we have a sample of n observation vectors $x_1, x_2, \ldots . x_n$ that form a group of points in the q-dimensional space. To maximize the variance, a natural axes of the variables $x_1, x_2, \ldots . x_q$ must be found with an origin at $x_1, x_2, \ldots . x_n$ mean vector which will be called $\bar{x}$. To rotate the axes each x_i will be multiplied by V which is an orthogonal matrix.

$$r_i = Vx_i$$

Since $V'V = I$

$$r_i' r_i = (Vx_i)'(Vx_i) = x_i' V' V x_i = x_i' x_i$$

This operation rotated the axes of $x_i$and transformed x_i to the point $r_i$which has the same distance from the origin.

Most of the time, the first two principal components will contain the linear combination of maximum variance. Therefore the first component will be $r_1 = v_{11} x_1 + v_{12} x_2 + \ldots + v_{1q} x_q$. (Rencher and Christensen, 2012).

The first two components of the religion PCA reported 84.54% of the variation, the first two components of the ethnicity PCA reported 41.23% of the variation and first two components of the political and social PCA reported 56.78% of the variation. The PCA helped identify the similarities in counties within each rate level.

ARIMA Modelling:

ARIMA is the abbreviation for autoregressive integrated moving average. This model is applied on time series data to aid in forecasting or prediction. In order to apply the ARIMA model to any time series, the time series needs to be ”stationary” which means that the mean and variance must not to change over time and do not follow any trends (Shumway & Stoffer , 2006). There are many methods two make a time series stationary, the most common method is “differencing“ which is computing the difference between consecutive observations in the data. The differencing could be applied once and mostly it will not exceed twice (Shumway & Stoffer, 2006).

Suppose y_t is a time series that needed to be forecasted, the ARIMA full model for this time series will be written as

$$y'_t = c + \emptyset_1 y'_{t-1} + \ldots + \emptyset_p y'_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t \qquad (1)$$

The left hand side of this equation is y'_twhich is the differenced series is of y_t. The right hand side includes lagged values and lagged errors of y_t.Those two terms are called “predictors”.

The ARIMA model can be denoted by ARIMA (p, d, q) model, which requires three different entries to calculate the model (Shumway& Stoffer, 2006). p is the autoregressive model order which is the lags of the stationarized series in the forecasting equation, d is the differencing time and q is the moving average model order which is lags of the forecast errors. Once the three components are combined, equation (1) transforms into

$$\left(1 - \emptyset_1 B - \cdots - \emptyset_q B^p\right) \quad (1 - B)^d y_t \;=\; c + \left(1 + \theta_1 B - \cdots - \theta_q B^p\right) e_t \qquad (2)$$

Where the first term from the left hand side represents the autoregressive part (p), the second part represents the difference (d) and the term on the right hand side represents the moving average.

ARIMA models can model seasonal data through ARIMA seasonal model. The ARIMA seasonal model includes additional terms to the original model to predict the seasonality. The ARIMA seasonal model is denoted by

$$ARIMA\ (p, d, q)(P, D, Q)_m$$

Where (p, d, q) represents the non-seasonal part of the model and $(P, D, Q)_m$ represents the seasonal part of the model (otexts.com, 2015). Equation (2) in seasonal ARIMA model with $ARIMA\ (1,1,1)(1,1,1)_r$ will transform into

$$(1-\emptyset_1 B)\ (1-\Phi_1 B^r)(1-B)\ (1-B^r)\ y_t\ =\ (1+\theta_1 B)(1+\Theta_1 B^r)e_t \qquad .$$

PCA and rates per county analysis:

After performing the PCA analysis and observing the case counts and rates per county per day, the situation was illustrated in all 15 counties with the help of the case rates side by side with the counties PCA analysis. The idea was to categorize the counties according to the rates per county into three levels. The first group included the high rate counts, or counties that had rate counts larger than 20 cases per 10,000. The second group was comprised of the medium rate counts which included the counties with rate counts between 19 and 10 cases per 10,000. The third group included the low rate counts which contained the counties that had the rate counts less than 9 cases per 10,000. Then, the three different PCAs of each category (social and political PCA, religion PCA and ethnicity PCA) were examined to identify any common characteristics that the counties which belonged to the same case counts.

High rate counts interpretation of demographic & cultural similarities:

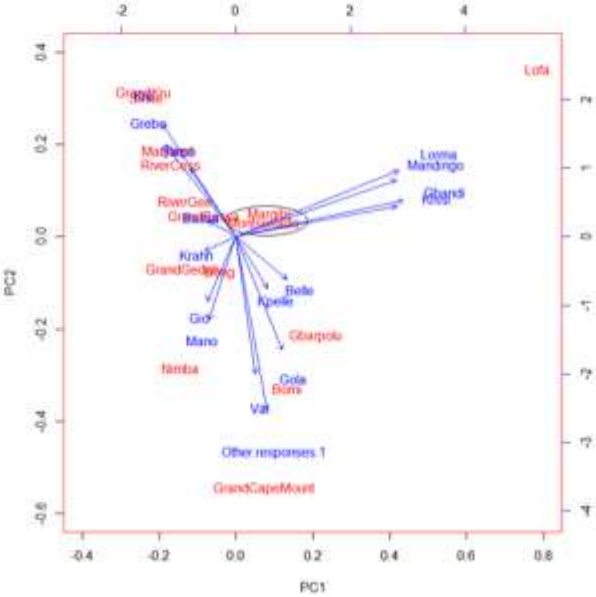


Figure 3: Ethnicity PCA for High Rate Counties

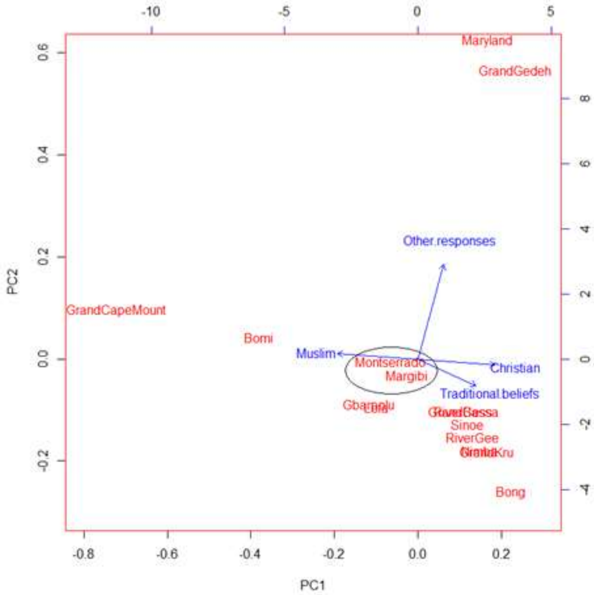


Figure 4 : Religion PCA for High Rate Counties

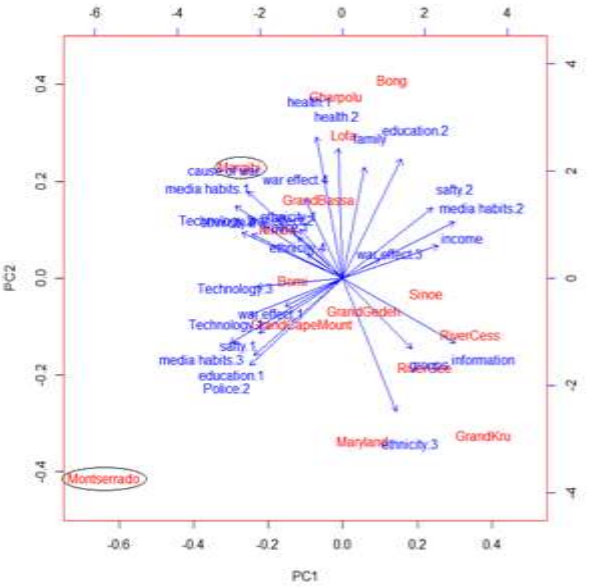


Figure 5: Social and Political PCA for High Rate Counties

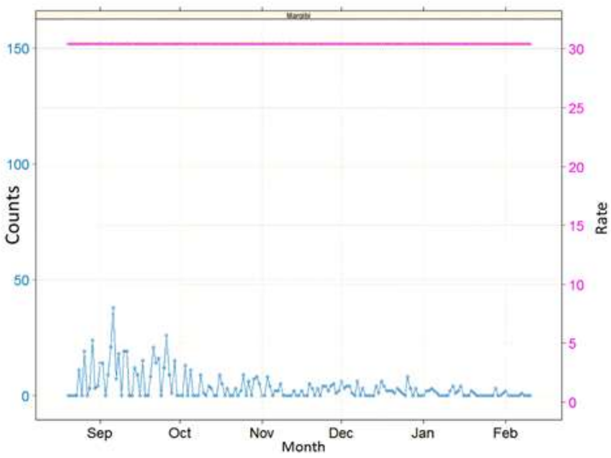


Figure 6 : Magribi Case and Rate Counts

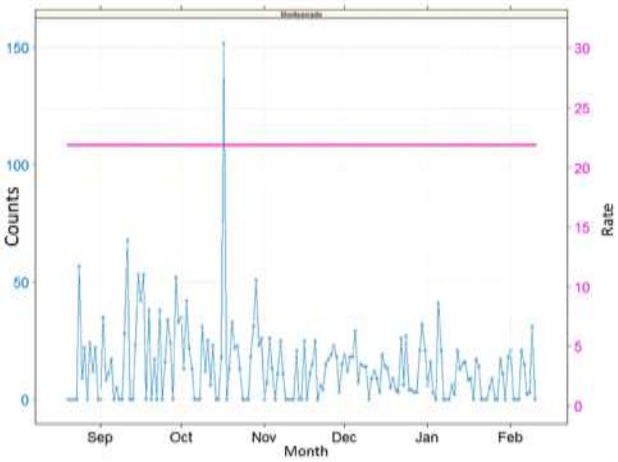


Figure 7 : Montserrado Case and Rate Counts

It was found that the counties with high rate counts were mostly the urban areas with high population density, which are Montserrado, Margibi (Figures 6, 7).

According to the Magribi and Montserrado location in the middle of both the Religion and Ethnicity PCA plots (Figures 3, 4), they shared the same characteristic which is a mixture of all religions and ethnicities. This characteristic, which would be explained by the geographic location of Magribi on the south of Montserrado, along with high population density, seems the most unique attributes which could explain the high rate counts in those two counties.

The highest rate counts were in Montserrado (Figure 7), which is the county that contains the Capital Monrovia. Monrovia contains Liberia's largest slum "West Point" slum. This slum was a huge hub for Ebola virus disease. It has 75,000 people living there without proper living condition such as: running water and good sewage system Lewnard *et al*., (2014). The PCA of the social and political characteristics of Montserrado which is in the far bottom left corner of the plot (Figure 5) shows that, in this county, most of the people read and write which means most of the population are educated, know the location of the nearest police station, own devices such cellphones, televisions, and watch television at least occasionally. Many think that Montserrado is not a safe area, and many were displaced during the war.

The political and social analysis of Margibi (Figure 5) was different than Montserrado. Magribi is located in the upper left area of the political and social PCA plot. The characteristics that are located in this area of the plot are: problems with ethnic groups which was mainly existed as part of the inherited culture of hatred and violence between the different ethnic groups, and the unfair treatment between from the ethnic groups to each other Vinck *et al.,* (2011). People feel the need to improve police capacity to keep everybody safe, they have relations with other ethnic groups and interact with them at least once weekly. They own radios and use it mostly to listen to the news, they took part in the war and they think that the cause of war was greed and corruption.

Medium rate counts interpretation of demographic & cultural similarities:

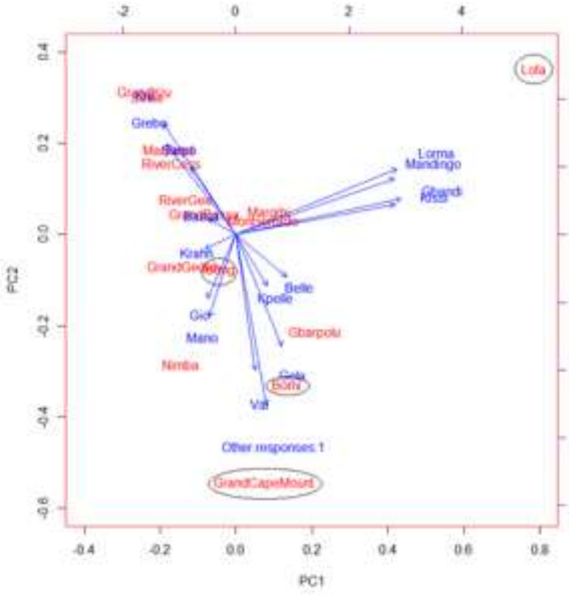


Figure 8 : Ethnicity PCA for Medium Rate Counties

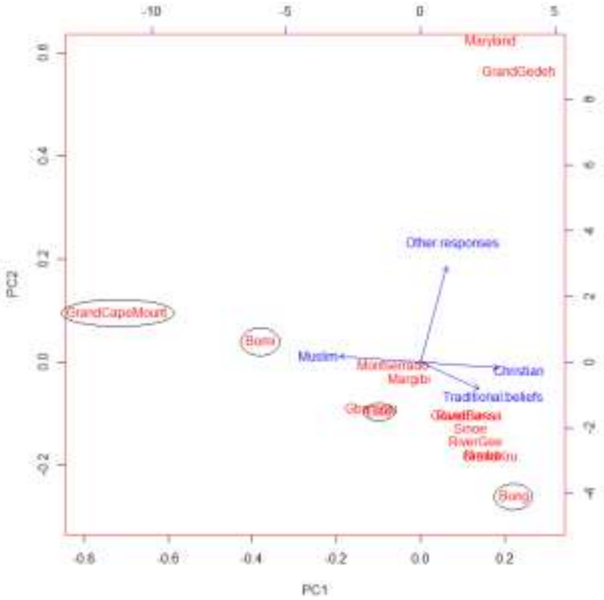


Figure 9 : Religion PCA for Medium Rate Counties

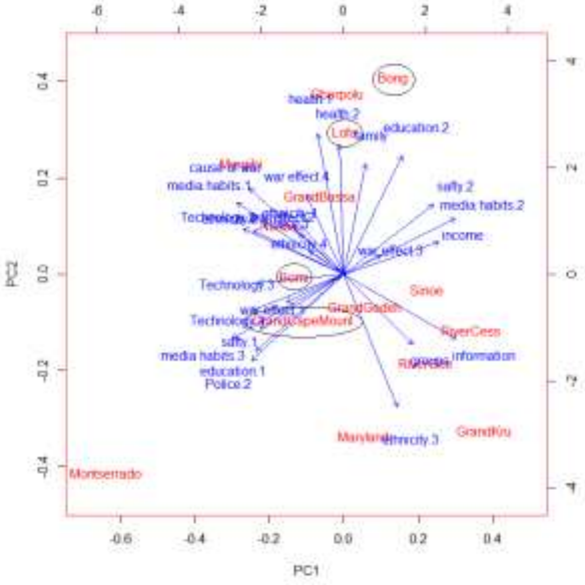


Figure 10 : Social and Political PCA for Medium Rate Counties

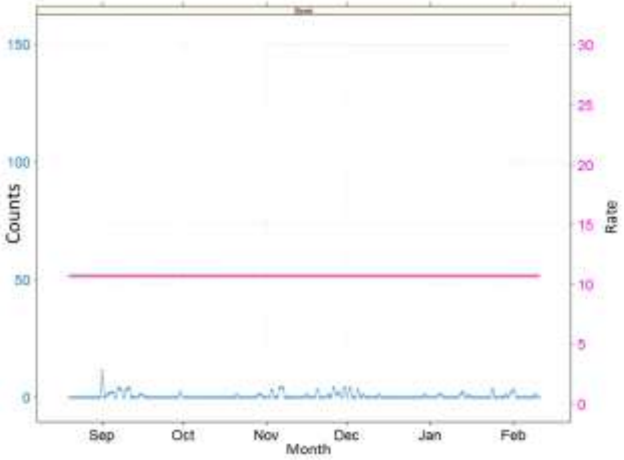


Figure 11 : Bomi Case and Rate Counts

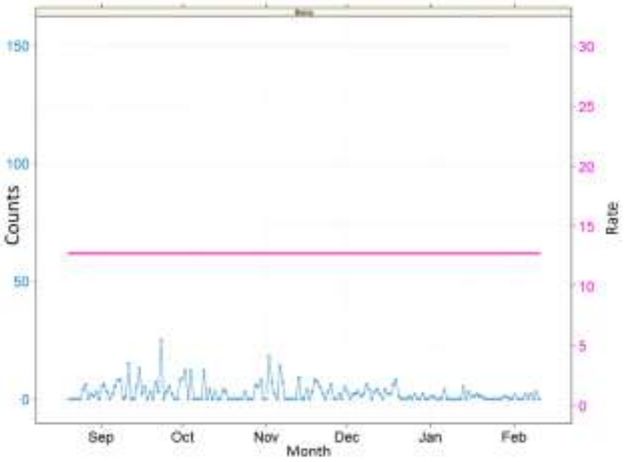


Figure 12 : Bong Case and Rate Counts

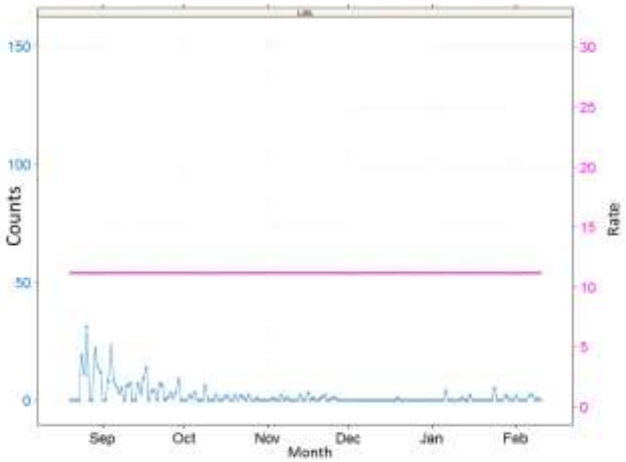


Figure 13 : Lofa Case and Rate Counts

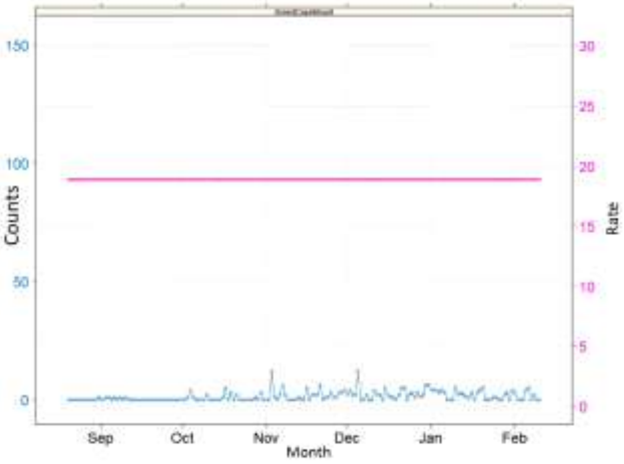


Figure 14 : Grand Cape Mount Case and Rate Counts

This category contained Grand Cape Mount (Figure 14), Bomi (Figure 11), Bong (Figure 12) and Lofa (Figure 13). Bomi and Grand Cape Mount share the same characteristics on the social and political PCA as Montserrado (Figure 10) which could be explained that they are geographically located next to each other and on the north side of Montserrado. Those characteristics were that most of the people read and write, which indicates most of the population are educated. They tend to know the location of the nearest police station, own devices such as cellphones and televisions, watch television at least occasionally, and they think that it is not a safe area. The ethnicity and religion PCA (Figures 8, 9) confirmed the similarity between Bomi and Grand Cape Mount. They are located on the far south area of the ethnicity plot which means they have the same tribes which is Gola and Vai. They are located on the far left area of the religion PCA which means that they share the same religion as

Muslims. Those characteristics combined together might help explain the similar case counts after considering the cross correlation function. The two counties having the same religion and ethnicities means that the two populations would likely have similar burial and funeral practices which would provide perfect environment for infections between dead bodies and people attending the service.

On the other hand, Lofa and Bong were located on the upper part of the social and political PCA (Figure 5), which means they shared illiteracy which led to the other factor that they do not read newspapers. They have good relationships with their family and have access to healthcare. The surveyed individuals think that they live in a safe area, have incomes less than 1.25 dollars per day, and witnessed a rape or sexual abuse during the war. The religion PCA (Figure 9) showed that they are located on the bottom area of the plot which means that they shared traditional beliefs as religion with Lofa tending to the Muslim religion and Bong toward Christian religion. This tendency could be explained by the difference in ethnicity that can be clearly visible on the ethnicity PCA. Bong is located on the left area of the ethnicity PCA which means that it is consisted of the Mano and Gio ethnicity, where Lofa is located on the far right corner which means it is consisted of the Loma, Mandingo, Gbandi and Kissi ethnicities.

Rare rate counts interpretation of demographic & cultural similarities:

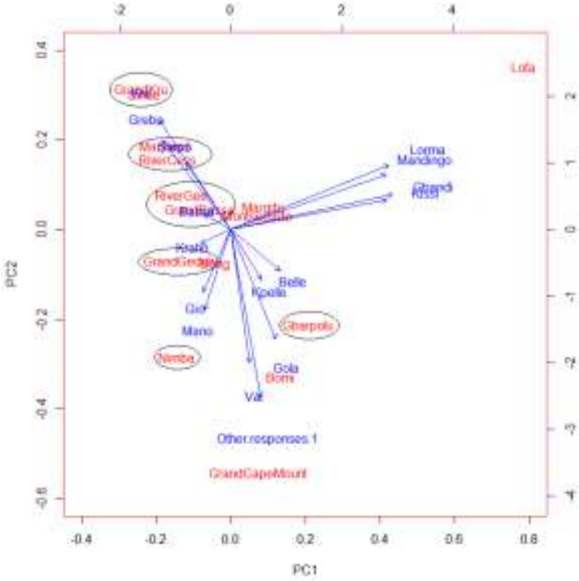


Figure 15 : Ethnicity PCA for Rare Rate Counties

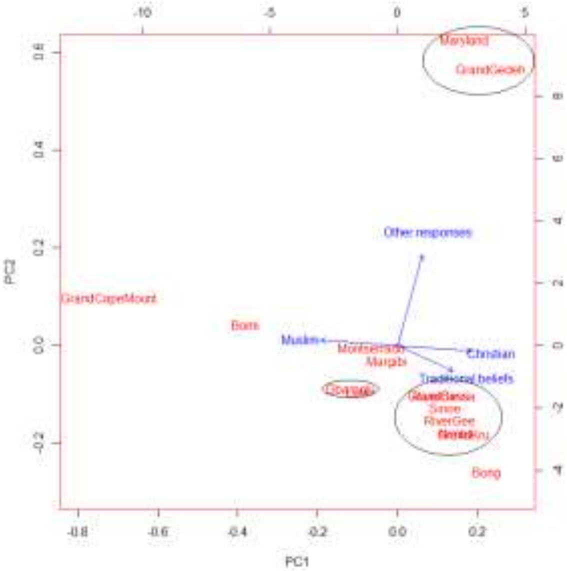


Figure 16 : Religion PCA for Rare Rate Counties

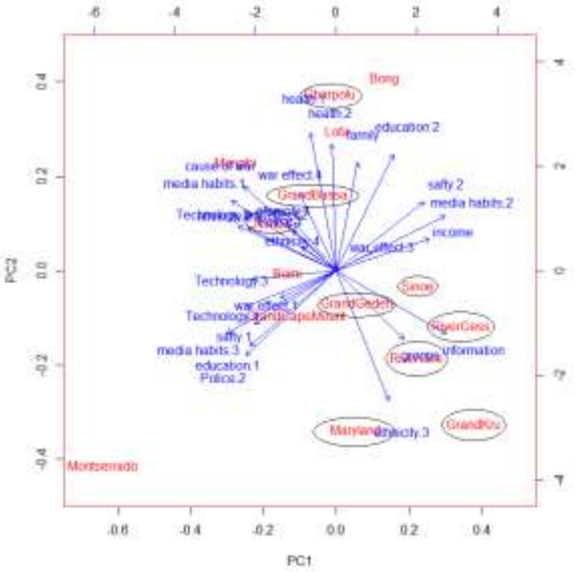


Figure 17 : Social and Political PCA for Rare Rate Counties

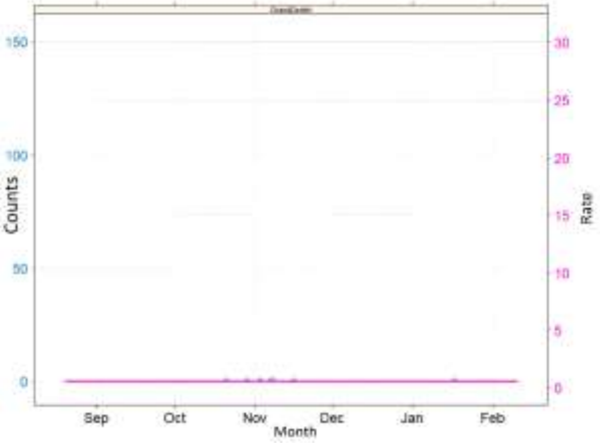


Figure 18 : Grand Gedeh Case and Rate Counts

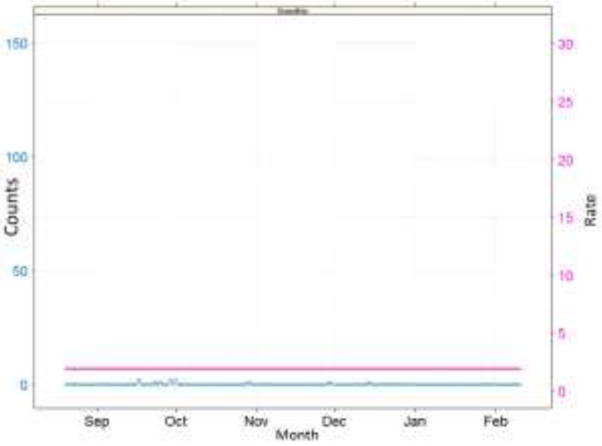


Figure 19 : Grand Kru Case and Rate Counts

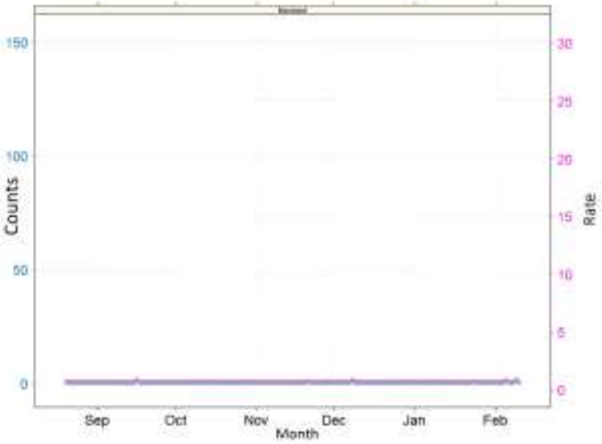


Figure 20 : Maryland Case and Rate Counts

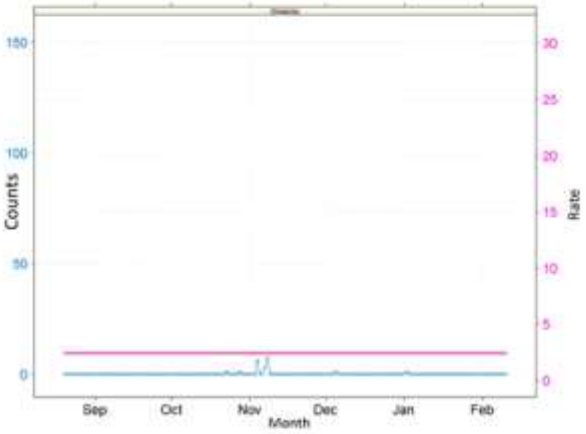


Figure 21 : Gbarpolu Case and Rate Counts

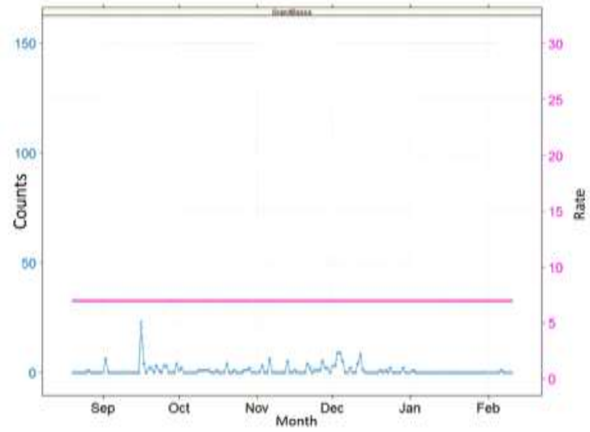


Figure 22 : Grand Bassa Case and Rate Counts

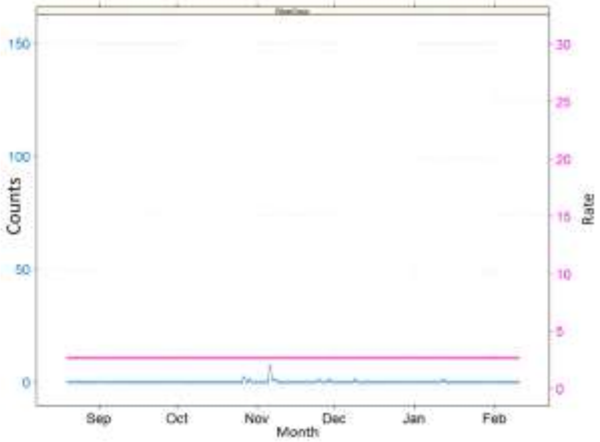


Figure 23 : River Cess Case and Rate Counts

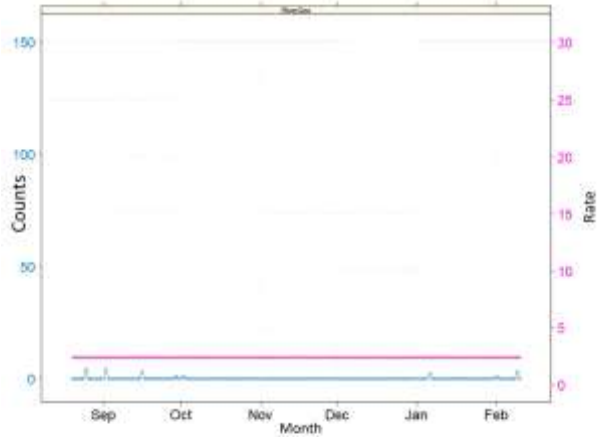


Figure 24 : River Gee Case and Rate Counts

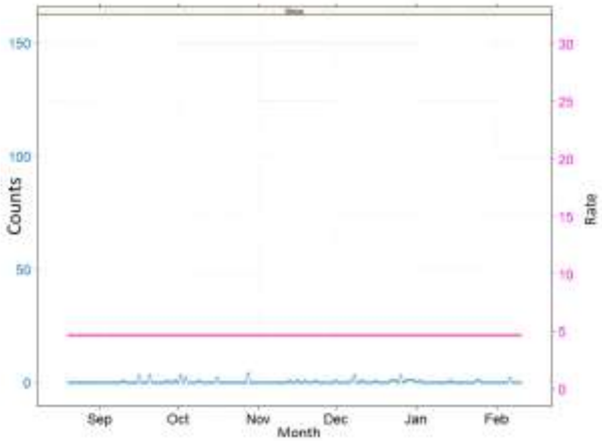


Figure 25 : Sinoe Case and Rate Counts

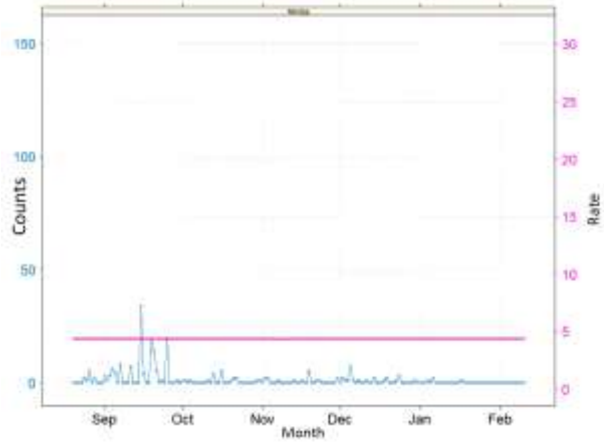


Figure 26 : Nimba Case and Rate Counts

This category contained Gbarpolu (Figure 21), Grand Bassa (Figure 22), Grand Gedeh (Figure 18), Grand Kru (Figure 19), Maryland (Figure 20), Nimba (Figure 26), River Cess

(Figure 23), River Gee (Figure 24) and Sinoe (Figure 25). There are no unifying characteristics all of these counties, however, some are similar to each other.

Grand Bassa and Nimba share the same characteristics in the religion and the political and social PCA (Figures 16, 17) analysis. The political and social PCA (Figure 17) analysis showed that they are located in the middle upper left area of the political and social PCA plot, which means that they are sharing the same characteristics. Those characteristics are problems with ethnic groups which was mainly existed as part of the inherited culture of hatred and violence between the different ethnic groups, and the unfair treatment between from the ethnic groups to each other Vinck *et al.,* (2011). Improve police capacity to keep everybody safe. They took part in the war, and they think that the cause of war was greed and corruption. They have relations with other ethnic groups and interact with them at least once weekly. They own radios and use it mostly to listen to the news. Interacting at least weekly with other ethnic groups and listening to the news on the radio might been factors on spreading the awareness upon this county which helped to reduce the rate counts.

The religion PCA (Figure 16) showed Grand Bassa and Nimba on the bottom right area of the plot which means that they have mainly traditional beliefs as religion with tendency toward Christian religions. The ethnicity PCA showed different result where Grand Bassa were on the middle left with Bassa ethnicity and Nimba on the bottom left with Gio and Mano ethnicity.

Grand Gedeh, Sinoe, River Gee, River Cess, Grand Kru and Maryland have the same characteristics they were all located on the bottom right area of the political and social PCA plot (Figure 17). The characteristics of this area are people obtain their information through friends and family, individuals tend to be members of religious groups and they think that the government favors some tribes more than others. Having their information through family and friends might had helped spread the awareness of the epidemic that helped them to lower the rate counts.

After taking a look at the ethnicity PCA (Figure 15) Grand Kru, Sinoe, Maryland, River Gee and River Cess have the same ethnicity which is Sarpo, Grebo and Kru whereas Grand Gedeh has a more weight on the Krahn ethnicity.

The religion PCA (Figure 16) revealed that River Cess, Sinoe, River Gee and Grand Kru have traditional beliefs as religion whereas Maryland and Grand Gedeh shared the same response for religion as none of the offered choices in the survey would represent their religion.

According to the ethnicity PCA, Belle and Kpelle was the dominant ethnicity at the Gbarpolu region. The PCA of religion reflected the uniqueness of this region in religion as it has traditional beliefs and Islam as religion.

Understanding the Relationship between the Case Counts and within the Counties Themselves:

In order to understand the development of the case counts between and among the counties themselves, two steps process were applied to the different case counts of the 15 different counties. The steps were: 1- ARIMA modeling, 2- Cross-Correlation function.

After examining the available dataset, it was decided to apply seasonal ARIMA models on this dataset. The observation of the case counts showed that there are some seasonality toward the weekends as it seemed that in some counties the case counts were collected and recorded as a whole on weekends. Using the ARIMA models with very low (one or two) frequency will not result in any seasonal terms, which left me with either choosing 7 as weekly seasonal term or 30 as monthly seasonal term. ARIMA model with daily seasonality with frequency =30 was used because the AICc results between the 7 days and 30 days seasonality was lower with this frequency (Table 4). The auto.arima () function in R were used to obtain the three different numbers in ARIMA model. This function in R uses a variation of the Hyndman and Khandakar algorithms to search through models to return the best ARIMA model according to either AIC, AICc or BIC value.

	AICc	BIC	AIC
7 Days Frequency	955.3	961.55	955.23
30 Days Frequency	942.29	954.69	942.05

Table 5 : Bong County Different Frequencies

After Choosing the 30 days frequency for seasonal ARIMA model, 30 days of the data were dropped (since no future datasets are available) and predicted by ARIMA model. This process was performed to examine the validity of the model in predicting the future counts of Ebola cases in each county. The produced forecast plot lays on a small blue shaded area which represents 80% prediction interval and a larger grey shaded area which represents 95% prediction interval.

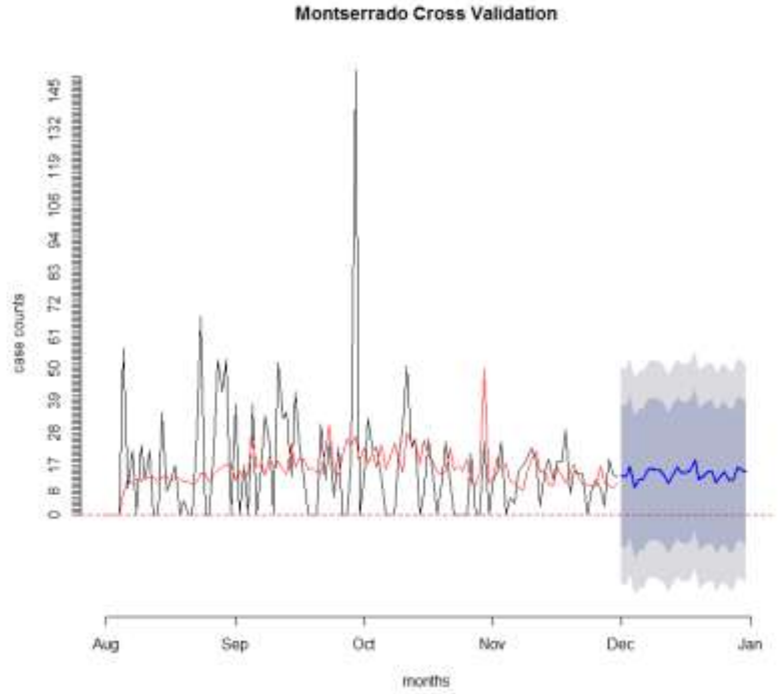


Figure 27:Predicted Montserrado Case Counts

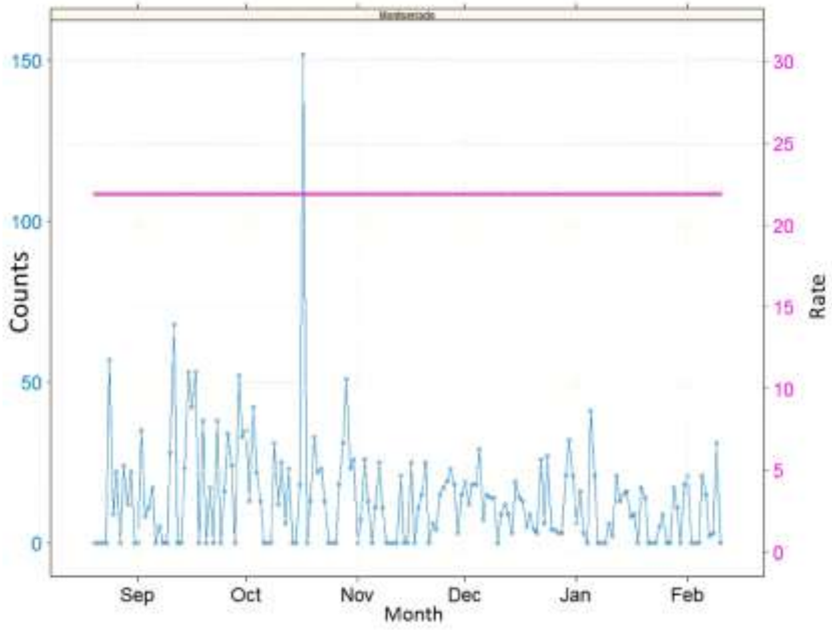


Figure 28 : Montserrado Actual Case Counts

The resulting ARIMA model plot showed relatively close numbers between the predicted model (Figure 27) and the actual case counts (Figure 28) of Montserrado county.

In the resulting figures, the Y axis at the ARIMA model represents the daily case counts in the county. The X axis represents the months, because the dates were divided by 30.

Why Cross-Correlation:

The cross-correlation function is used to explore the relationship between two series x (i) and y (i). This method measures the degree of similarity or correlation between the two series. CCF was applied between each of the counties and Montserrado County as it had the most case counts, and was in the “middle” in terms of demographics and cultural measurements. As for the CCF between the rests of the counties, different PCA plots were examined and the CCF was applied on the counties that shared the same characteristics according to the social and political PCA plot, ethnicity PCA plot and the religion PCA plot. The Y axis reflects the value of the autocorrelation between the two components of the cross correlation. The higher the ACF value the stronger the correlation between the components, the lower the value the weaker the correlation. The X axis reflects the lag between the two components of the cross correlation. The negative value on the X axis represents a correlation between the first component and the second component at where the first leads the second that time point. The positive value on the X axis represents a correlation between the first component and the second component where the first lags the second at that time point.

Cross-correlation was applied on the rates and case counts between the counties in two different processes, the two results showed that the cross correlation between the counties for rates and case counts are exactly similar. The following (Figure 29, 30) are an example of the cross correlation between Bomi and Montserrado counties for the rates and case counts.

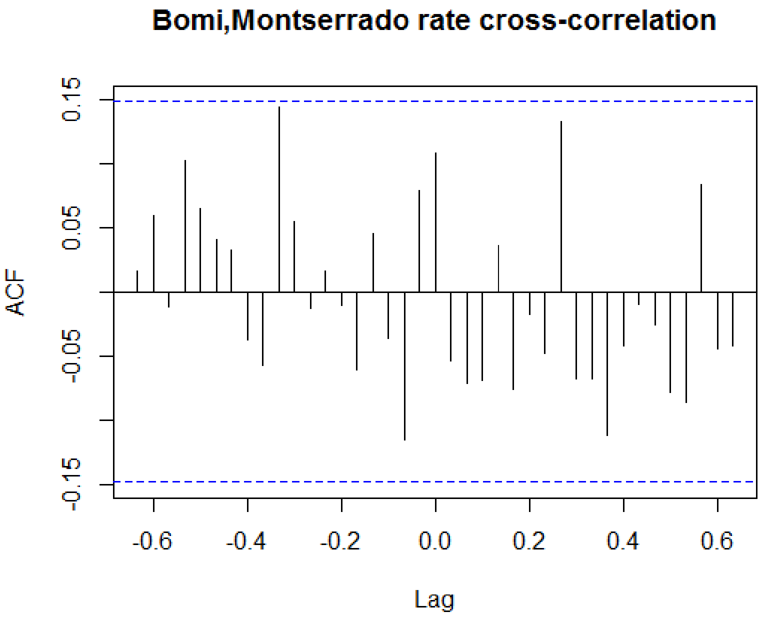


Figure 29 : Bomi and Montserrado Rate Cross-Correlation

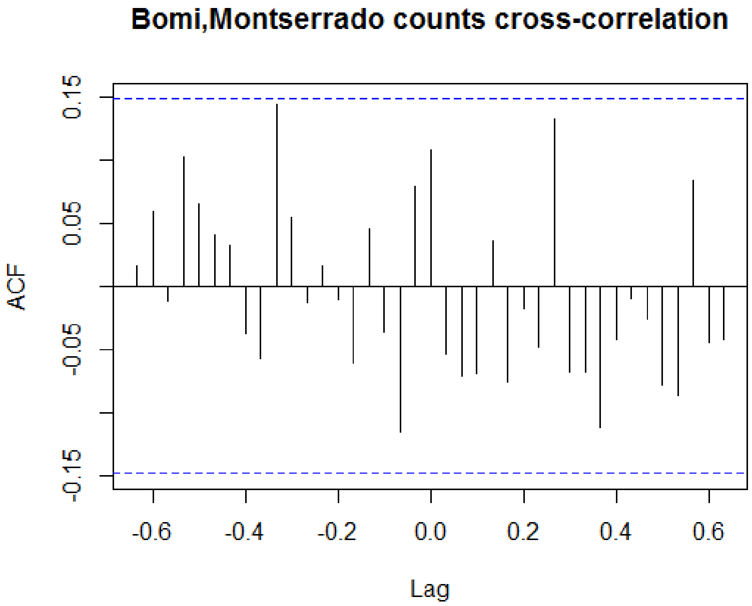


Figure 30 : Bomi and Montserrado Case Counts Cross-Correlation

After performing the cross-correlation function on the low rate counts, it was decided not to analyze them in this thesis as the case counts are too low and they do not seem to provide much evidence or information for analysis purposes.

Bomi County:

<u>ARIMA model:</u>

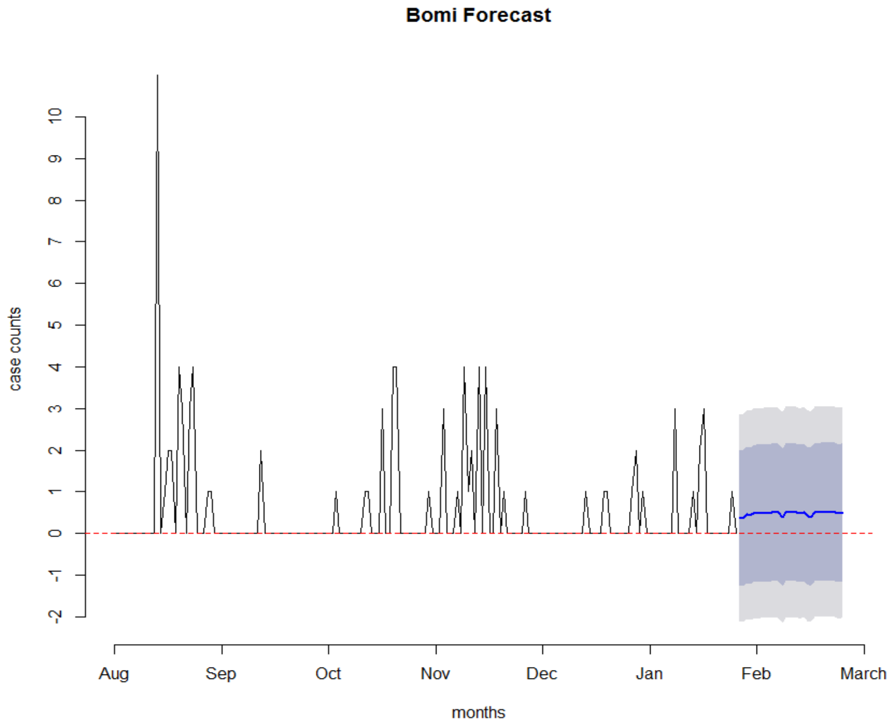


Figure 31 : Bomi Forecast

When I applied the ARIMA model to Bomi (1,0,1) model was used with seasonal (1,0,0) model, which means that the ARIMA model had first autoregressive model order , zero difference order where at the same time it had one moving average model order. The ARIMA model for Bomi forecast (Figure 31) showed that the case counts predicted for the 30 next days will remain between zero and two which means that for the next month the epidemic will continue to spread in this county but in a very low case count average.

CCF between Bomi and Other Counties:

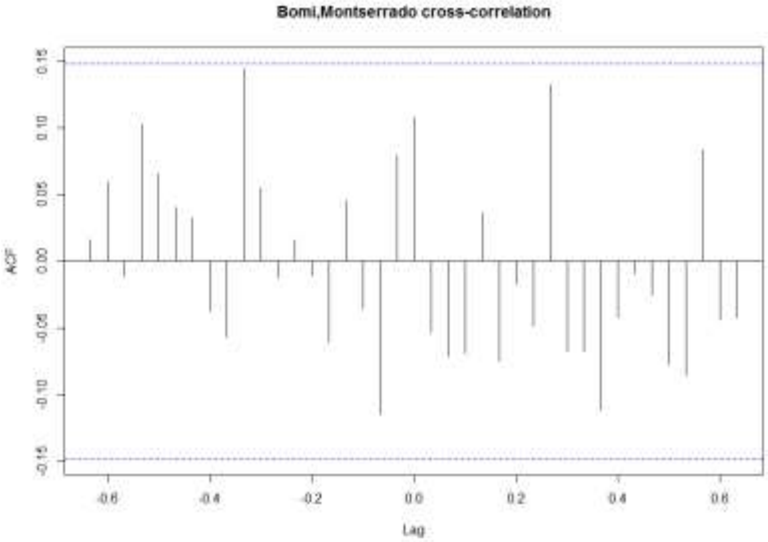


Figure 32: Bomi and Montserrado CCF

Bomi County and Montserrado County: Bomi borders Montserrado from the west. The case counts and rate plot showed that it had lower case counts than Montserrado. There are no shared characteristics between the two counties at any of the three PCA plots. The CCF between the two counties (Figure 32) did not show any significant level of correlation or anti-correlation.

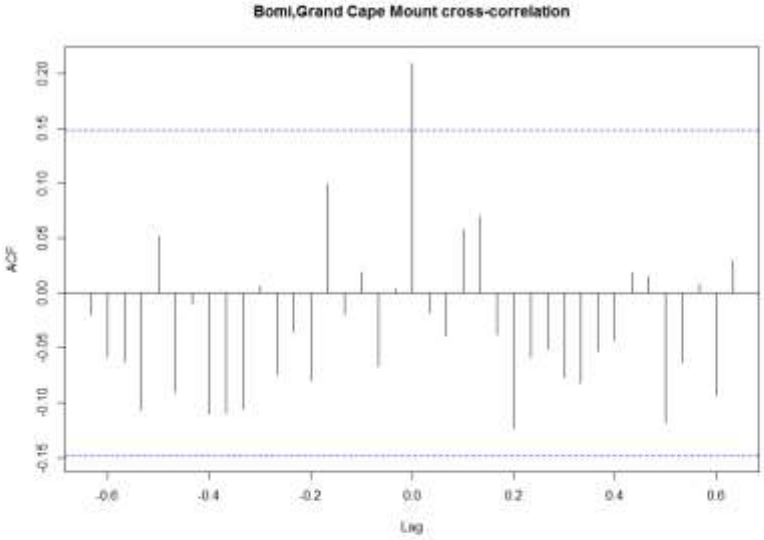


Figure 33: Bomi and Grand Cape Mount CCF

Bomi County and Grand Cape Mount County: Grand Cape Mount borders Bomi from the north and they had several characteristics in common. The ethnicity PCA showed that they shared similar ethnicity distribution of Gola and Vai. The religion PCA showed that they have similar distribution of Muslims. The social and political PCA showed that they are mostly have the same characteristics. The case counts and rate plot showed that Bomi and Grand Cape Mount have medium rate level. The CCF (Figure 33) showed that they are weakly correlated at lag zero indicating that Bomi has the same progress of case counts as Grand Cape Mount.

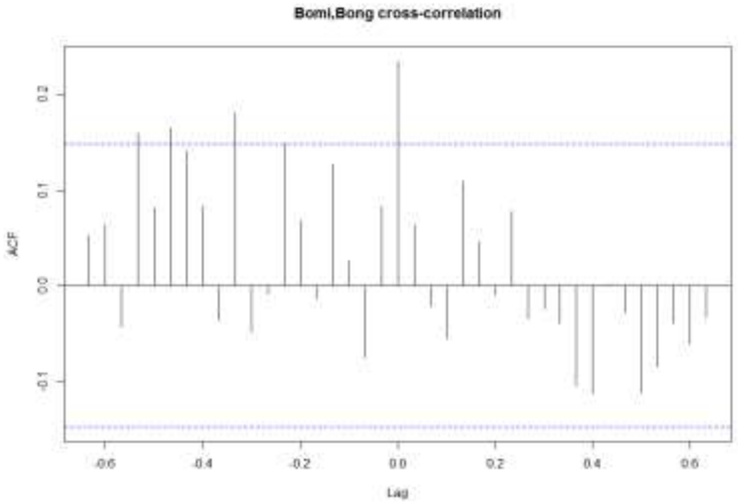


Figure 34 : Bomi and Bong CCF

Bomi County and Bong County: Bong was in the same rate category of Bomi, Bomi and Bong share a small border area. They do not have any common PCA characteristics in any of the three PCAS. After applying the CCF between them (Figure 34), it was found that Bomi and Bong are weakly correlated at zero.

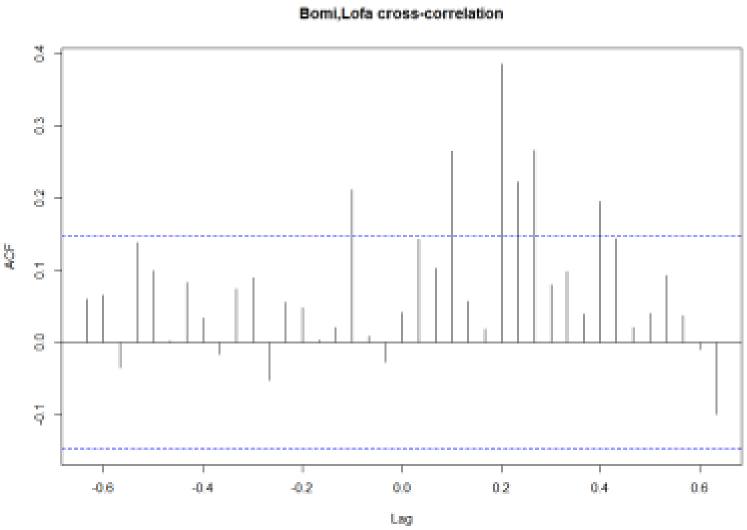


Figure 35: Bomi and Lofa CCF

Bomi County and Lofa County: Lofa was in the same rate category of Bomi, Bomi and Lofa does not share any borders. The ethnicity and social and political PCA showed that they do not share any characteristics in common. The religion PCA showed that Lofa lies between Muslims and Tradtional beliefs and Bomi is in the Muslim direction. The CCF (Figure 35) between the two counties showed that they are moderately correlated with Bomi leads Lofa once then they have the same epidemic progress and after that Bomi lags Lofa. The strongest correlation was at a lag 0.2, indicating that the outbreak in Lofa was 6 days ahead of Bomi.

Bong County:

<u>ARIMA model:</u>

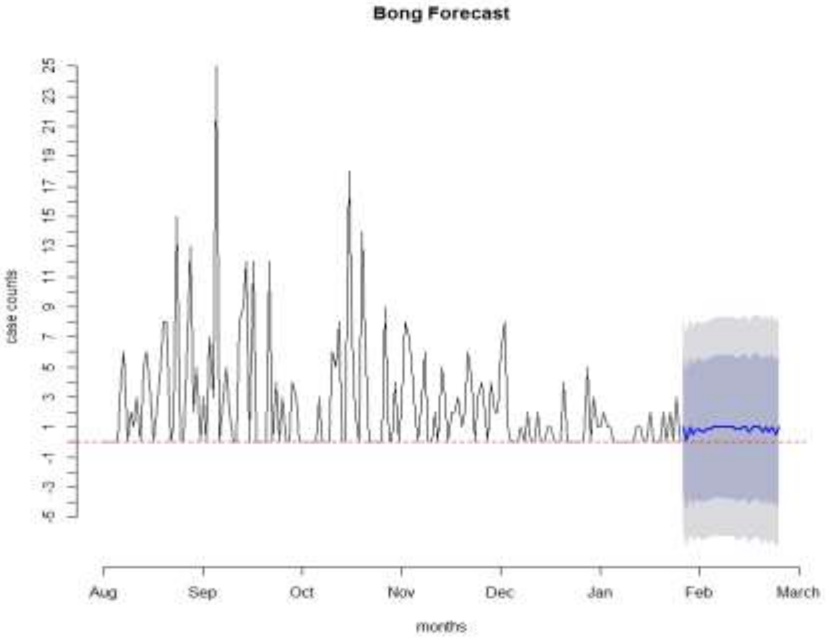


Figure 36: Bong Forecast

The ARIMA model for Bong was applied with (0,1,1) and seasonal (1,0,0), which means that it had to be stationary by applying the difference method once, the autoregressive model order is zero and the moving average model order equal one. The ARIMA forecast plot for Bong County (Figure 36) showed that the case counts predicted for the 30 next days will remain between zero and five which means that for the next month the epidemic will continue to spread in this county but in a very low case count average.

CCF between Bong and Other Counties:

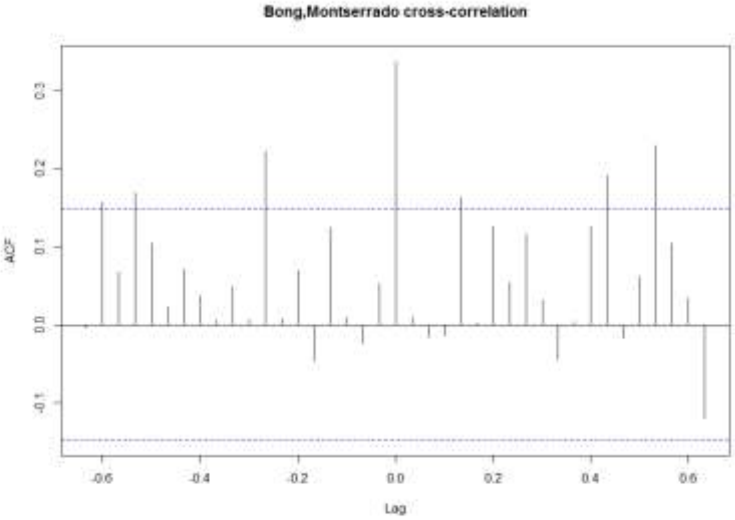


Figure 37: Bong and Montserrado CCF

Bong County and Montserrado County: Bong has borders with Montserrado and the rate plot showed that Bong has medium rate counts. They do not share any characteristics at any of the three PCAs. The cross-correlation between them (Figure 37) showed that they are moderately correlated at lag zero, which means that the epidemic was progressing at the same level in the two counties.

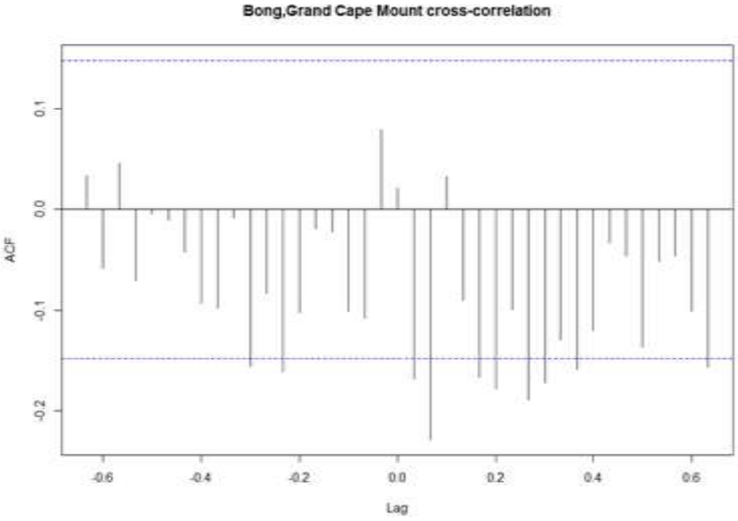


Figure 38: Bong and Grand Cape Mount CCF

Bong County and Grand Cape Mount County: These counties were in the same rate category. There were no common characteristics between them in the PCA plots. They have no common borders. The cross-correlation plot (Figure 38) showed that there are weakly anti-correlated at 0.36 which means that Bong leads Grand Cape Mount by 11 days.

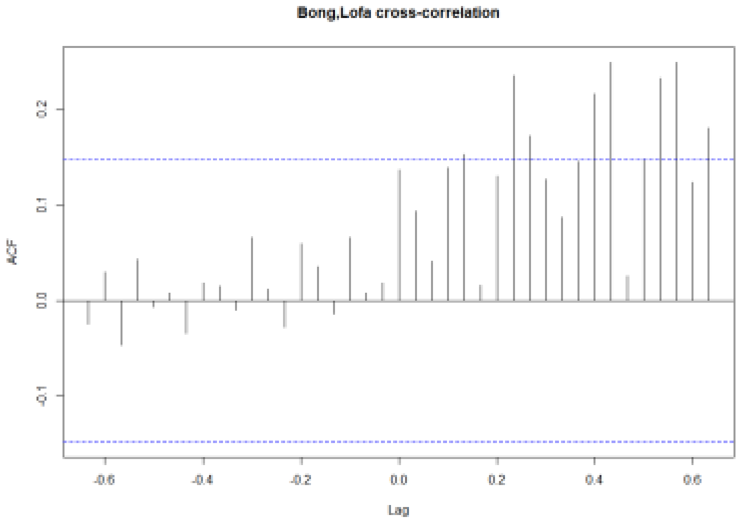


Figure 39: Bong and Lofa CCF

Bong County and Lofa County: Bong borders Lofa from the south. They were in the same rate level and they shared the same characteristics on the social and political PCA. The cross-correlation plot (Figure 39) showed that they are weakly correlated at 0.43 and 0.57, which means that Bong lagging Lofa by 13 days and 17.1 days.

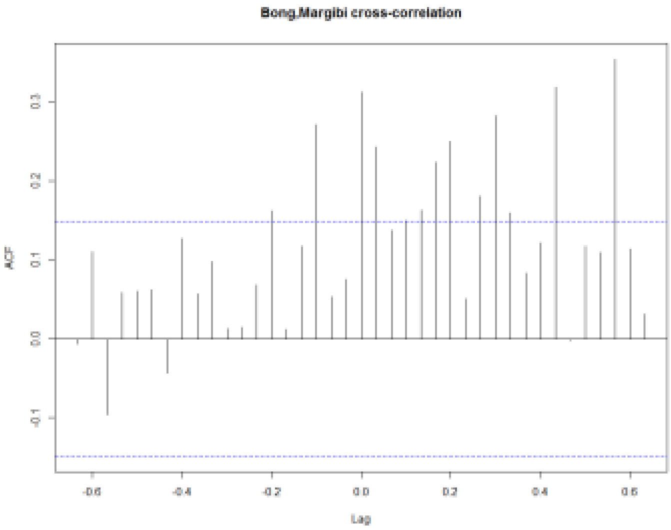


Figure 40: Bong and Margibi CCF

Bong County and Margibi County: Bong borders Margibi from the north. They did not share any PCA. The cross-correlation plot (Figure 40) showed that they are moderately correlated, where Bong lags Margibi at 0.56 which means 15 days.

Grand Cape Mount:

<u>ARIMA model:</u>

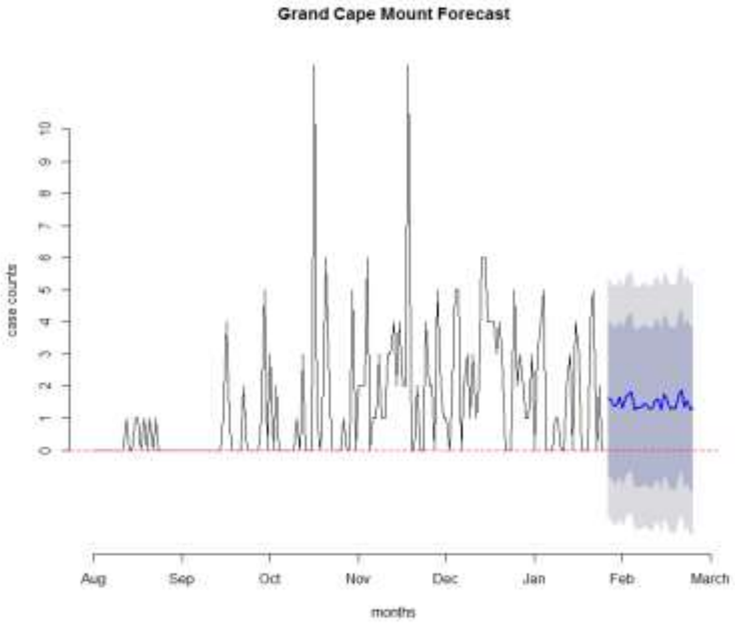


Figure 41: Grand Cape Mount Forecast

The ARIMA model for Grand Cape Mount was applied with (0,1,1), seasonal(0,0,1) which means that the autoregressive model order equals zero, to make the data stationary the difference had to be taken once and the moving average model order equal one. The ARIMA forecast plot (Figure 41) for Grand Cape Mount County showed that the case counts for the next month is going to stabilize between zero and five.

CCF between Grand Cape Mount and Other Counties:

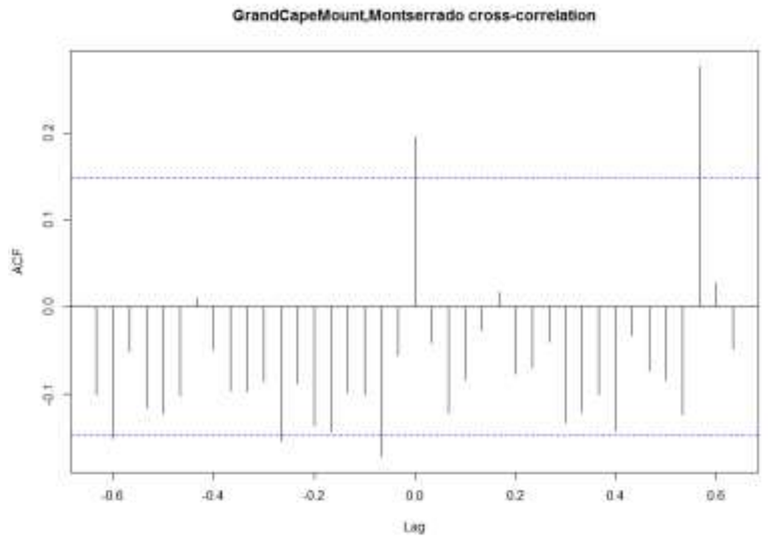


Figure 42: Grand Cape Mount and Montserrado CCF

Grand Cape Mount County and Montserrado County: Grand Cape Mount and Montserrado does not share any borders between them, none of the PCA plots showed any common characteristics between the counties and the rate counts and rates plots do not have the same average counts. The CCF between the two counties (Figure 42) showed that they are weakly positively correlated at lag 0.56 which means Grand Cape Mount lags Montserrado by 16.8 days.

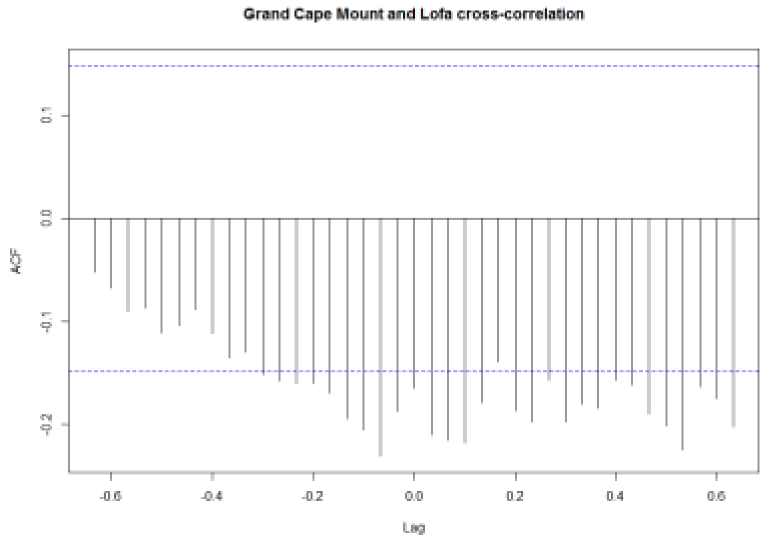


Figure 43: Grand Cape Mount and Lofa CCF

Grand Cape Mount County and Lofa County: does not share any borders or any PCA characteristics but they belong to the same rate level. The CCF (Figure 43) showed that they are weakly anti-correlated at -0.06 which corresponds to 1.8 days .which means there are no relationships between the two counties.

Lofa.

ARIMA model:

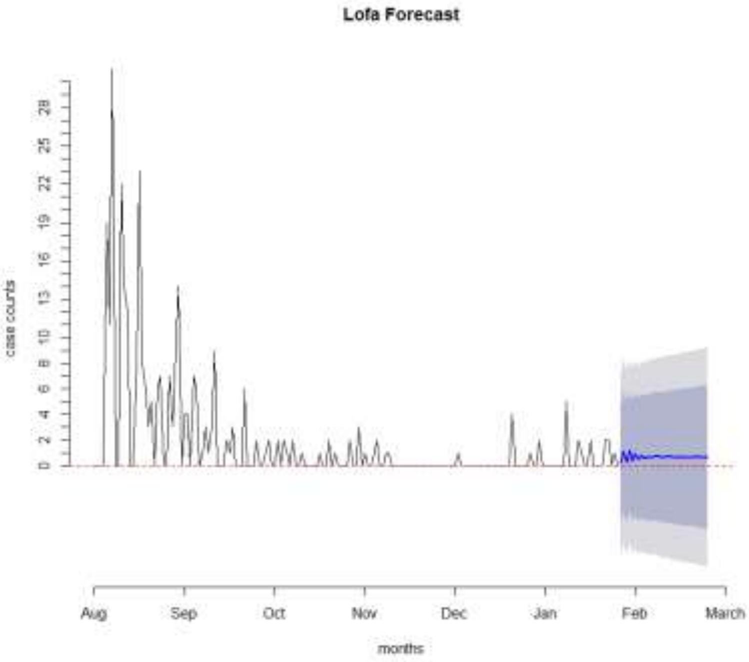


Figure 44: Lofa Forecast

The ARIMA model for Lofa was applied with (2,1,5), seasonal (1,0,0), which means that the autoregressive model order equal to two , the difference order equal to one and the moving average model order equal to five. The ARIMA forecast plot (Figure 44) for Lofa County showed that the case counts for the next month is going to stabilize between zero and five which mean that the epidemic in this county will continue spreading in this county in low rate.

CCF between Lofa and Other Counties:

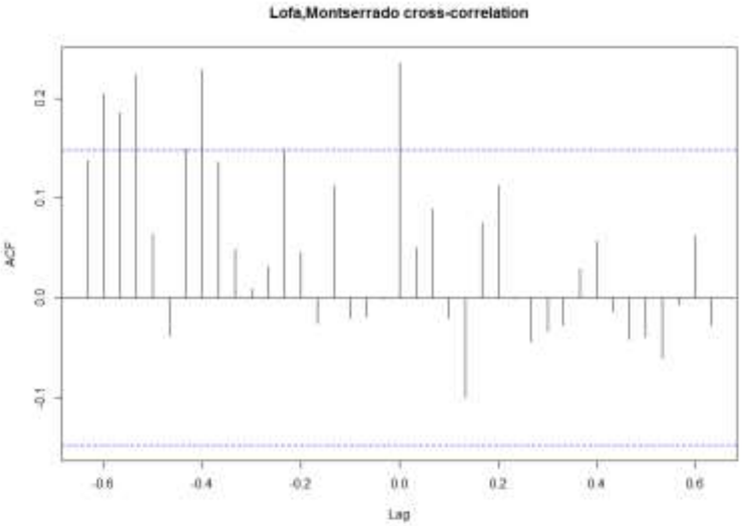


Figure 45: Lofa and Montserrado CCF

Lofa County and Montserrado County: There are no shared borders between Lofa and Montserrado. The three PCA plots showed that there are no common characteristics between the two counties. The CCF between the two counties (Figure 45) showed that they significantly, but weakly, correlated at lag 0 which means that the epidemic was progressing at the same level between the two counties. We see a similar correlation level at lag 0.4 as well, which may reflecting earlier wave infection of Lofa County at the beginning of the epidemic.

Margibi

ARIMA model:

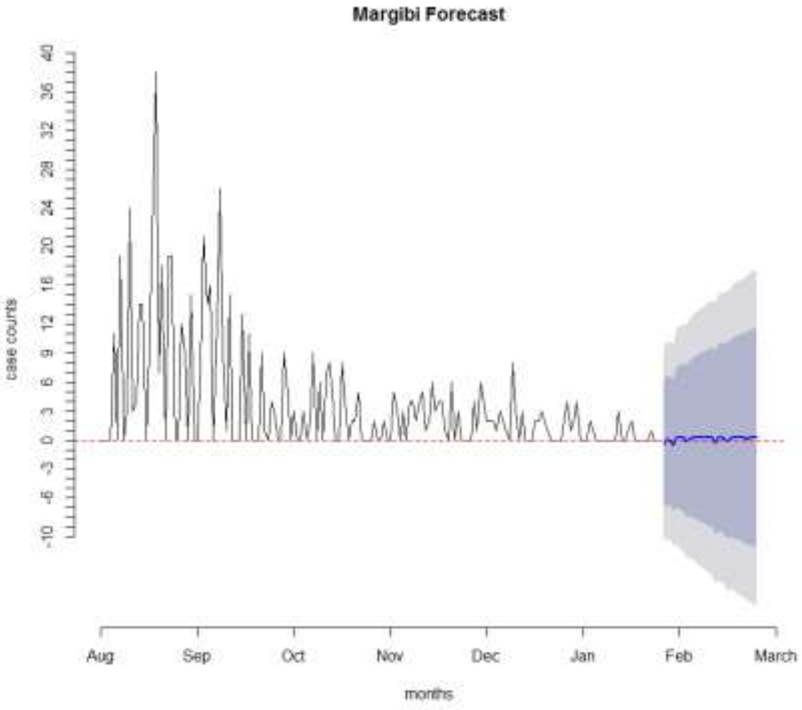


Figure 46: Margibi Forecast

The ARIMA model for Margibi was applied with (5,1,0), seasonal (1,0,0) which means that the autoregressive model order equal to five and the difference order equal to one and the moving average model order equal to zero. The ARIMA forecast plot (Figure 46) for Margibi County showed that the case counts for the next month is going to stabilize around zero, which means that the epidemic will eventually disappear from this county.

CCF between Margibi and Other Counties:

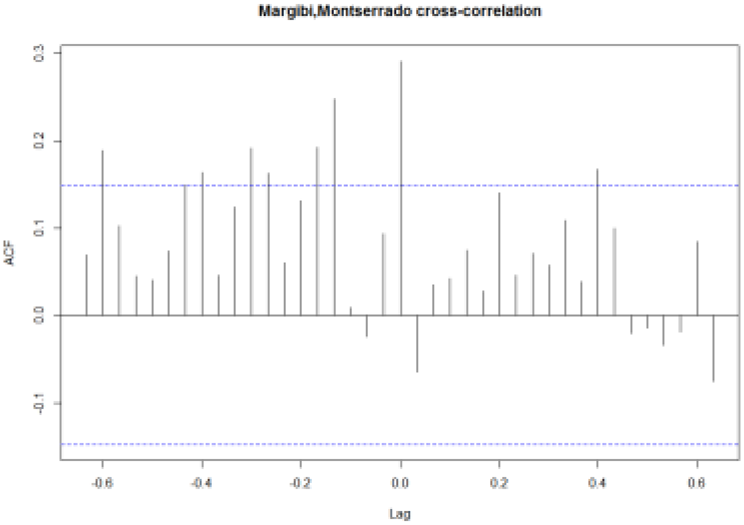


Figure 47: Margibi and Montserrado CCF

Margibi County and Montserrado County: Margibi borders Montserrado from the east. They have the same case rate, ethnicity and religion PCA but not the social and political PCA. The CCF between the two counties (Figure 47) showed that they significantly correlated at lag 0 which means that the epidemic was progressing at the same level between the two counties.

Montserrado

ARIMA model:

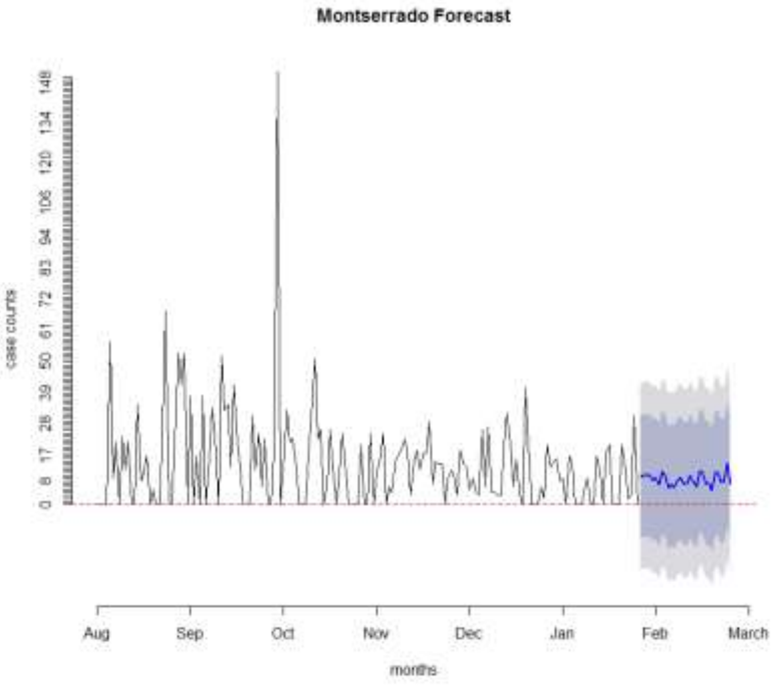


Figure 48: Montserrado Forecast

The ARIMA model for Montserrado was applied with (2,1,2),seasonal(1,0,1)which means that the autoregressive model order equal to two, the difference order equal to zero and the moving average model order equal to two. The ARIMA forecast plot for Montserrado County (Figure 48) showed that the case counts for the next month is going to stabilize between zero and fifty, which means that the epidemic will continue it is spread on this county at a high level.

Summary of cross-correlation analysis:

After performing the cross-correlation between the counties, the resulting relationships was divided between the counties into five categories: moderately correlated which has 0.4 or more correlation, weakly correlated which has less has less than 0.4 correlation, not correlated or anti-correlated, strongly anti-correlated which has -0.4 or less correlation, and weakly anti-correlated which has less has more than -0.4 correlation.

Most of the cross-correlations results were between the moderately correlated and weakly correlated categories. There were few cross correlations that were between not correlated neither anti-correlated, weakly anti-correlated categories. There were no cross correlation at the strongly anti-correlation category.

The first category: Four of the twelve cross correlations were moderate (Figure 35: Bomi and Lofa CCF, Figure 37: Bong and Montserrado CCF, Figure 40: Bong and Margibi CCF, Figure 47: Margibi and Montserrado CCF). They had some common features between them. Three of the four cross correlations were between two counties with common borders. Another two of the three cross correlations has two counties with common religion PCA.

The second category: Five of the twelve cross correlations were weak (Figure 33: Bomi and Grand Cape Mount CCF, Figure 34: Bomi and Bong CCF, Figure 39: Bong and Lofa CCF, Figure 42: Grand Cape Mount and Montserrado CCF, Figure 45: Lofa and Montserrado CCF). They had some common features between them. Three of the five cross correlations has two counties with common borders. There were no common sociodemographic characteristics common across the five cross-correlations.

The third category: Only one of the cross-correlations resulted in no significant cross-correlation (Figure 32: Bomi and Montserrado CCF). The two counties in this cross correlation shared borders and none of the three PCA were common between them.

The fourth category: weakly anti-correlated counties has two of the twelve cross correlations (Figure 38: Bong and Grand Cape Mount CCF, Figure 43: Grand Cape Mount and Lofa CCF). The main observation was that these two cross correlation between the counties has neither common borders nor common PCA between them.

Discussion:

Statement of Project's Significance:

The goal of this thesis was to fill the gap in understanding the dynamics of the epidemic spread in relation to social and demographic characteristics. The type of data compiled in this thesis is what distinguish this work from the previous work. Social and demographic data were used to help to analyze the spread of Ebola epidemic. The relationship between these data and the epidemic case counts was explored and search was conducted to look for any connection between them that might explain the huge spread of the epidemic in this area.

Main Findings:

What can be concluded from this simple classification of cross-correlation is that common borders might have played a role in the spread of EVD. This finding is not new as many papers have mentioned this reason before. The spread of the EVD in the neighboring counties is a direct result of the people's continuous movements between them. There were no consistent pattern on religious, ethnic social or political behavior that could have derived the spread of the EVD in the cross-correlations. However, attributes of single counties could have led to unique patterns.

Lofa was unique in its ethnicity PCA (Figure 8) where it has Lorma, Gbandi, Mandingo and Kissi ethnicity. The social and political PCA (Figure 10) showed that Lofa County has healthcare spending as priority and the people have their information through family and friends. Those unique characteristics, might be the main reasons behind the quick decrease of case counts in Lofa County. As awareness of the importance of admission to hospitals in case of suspected EVD case spread in the community through their connection with family and friends.

Gbarpolu was another unique county according to the social and political PCA (Figure 17). I found that it has the same characteristics as Lofa. The people in Gbarpolu consider spending on health care as priority and have access to a health care facility at the same time which apparently reduced the case counts on this county plus they have good relationships with their friends and family which means that they have a good connection between each other which might help them spread the awareness to protect themselves from the epidemic.

On the other hand, there was not any evidence of any uniqueness between the rest of the counties that had equal spread of the EVD, which means that they have good mixing of religion, ethnicity and social and political properties. This good mixing resulted in a perfect environment for the spread of EVD and no clear patterns that could be detected to perform a conclusion about the main reason that helped spread of the epidemic.

The cross correlation between the counties shows that the epidemic started and died quickly at Lofa. The cross correlation shows that Bomi and Lofa are moderately correlated at lag 0.2 which means that Bomi lags lofa by 6 days. Bong and Lofa are weakly correlated at 0.43 which means that Bong lags Lofa by 13 days. Then the EVD spread to Bomi, Bong, Margibi and Montserrado at the same time with the same rate. This can clearly be seen by the looking at the rate and case counts plots. The plots showed that the peaks of the four counties are moving together, as Lofa's only peak was at the beginning of the epdemic. Bong and Montserrado were correlated at lag 0, Margibi and Montserrado are moderately correlated at lag 0, Lofa and Montserrado weakly correlated at lag 0.

Another unique observation was that Grand Cape Mount is weakly correlated to Montserrado at lag 0 and then again at lag 0.56. This means that the cases in Grand Cape Mount were associate with the cases in Montserrado at the same time, but also lags Montserrado by 16.8 days. This supports the theory that once the epidemic reached Montserrado it spread quickly everywhere into the neighboring counties. At the same time Grand Cape Mount was anti-correlated with Lofa at lag -0.06 which means that Grand Cape Mount leads Lofa by 1.8 days. There is evidence from the cross correlation results that Grand Cape Mount was infected at the beginning of the epidemic then again after Montserrado got infected, Grand Cape Mount was hit by another wave of the infection.

Explanation of Limitations:

This thesis used the best available data for a social and demographic analysis. The major limitation of this projects was the scarcity of the sources and datasets as well as the uncertainty that surrounded the available ones. There were not any reliable sources for precise records of the EVD cases. The data that we used here were on county level only - not on a city or village level which will make it more accurate to observe the people's behavior from one area to another. The problem was not limited to the scarcity of data, the information, papers and journals were rare or limited giving that the epidemic is recent. New data, journals or papers about the EVD could be available in the future which could help for improved understanding and analysis for the social and demographic relationship between Ebola virus epidemics in West Africa.

Bibliography:

Centers for Disease Control and Prevention (2011). Viral Hemorrhagic Fever (VHF) 2011 Case Definition. .Retrieved November 7, 2015 from http://wwwn.cdc.gov/nndss/conditions/viral-hemorrhagic-fever/case-definition/2011/

Centers for Disease Control and Prevention (2015). 2014 Ebola outbreak in West Africa. Retrieved November 7, 2015 from http://www.cdc.gov/vhf/ebola/outbreaks/2014-west-africa/case-counts.html

Centers for Disease Control and Prevention (2015). Case Definition for Ebola Virus Disease (EVD). Retrieved November 7, 2015 from http://www.cdc.gov/vhf/ebola/healthcare-us/evaluating-patients/case-definition.html

Centers for Disease Control and Prevention (2015). Emerging Infectious Disease journal. Retrieved November 7, 2015 from http://wwwnc.cdc.gov/eid/article/21/11/et-2111_article#tnF2

Centers for Disease Control and Prevention (2015). Regional Spread of Ebola Virus, West Africa, 2014 - Volume 21.Retrieved November 7, 2015 from http://wwwnc.cdc.gov/eid/article/21/3/14-1845_article

Ebola: Mapping the outbreak - BBC News. (n.d.). Retrieved November 7, 2015.from http://www.bbc.com/news/world-africa-28755033

Etymologia: Ebola - Volume 21, Number 11-November 2015 - Emerging Infectious Disease journal - CDC. (2015, November 1). Retrieved November 7, 2015from http://wwwnc.cdc.gov/eid/article/21/11/et-2111_article#tnF2

Feldmann, H., & Geisbert, T. (2011). Ebola Haemorrhagic Fever. *The Lancet, 377*(9768), 849-862. Retrieved November 7, 2015, from http://www.sciencedirect.com/science/article/pii/S0140673610606678

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

Ki, M. (2014). What do we really fear? The epidemiological characteristics of Ebola and our preparedness. *Epidemiol Health Epidemiology and Health*, *36*. doi:10.4178/epih/e2014014

Kirch, W. (2008). *Encyclopedia of public health* (Vol. 1, p. 314). New York: Springer.

Lewnard, J., Mbah, M., Alfaro-Murillo, J., Altice, F., Bawo, L., Nyenswah, T., & Galvani, A. (2014). Dynamics and control of Ebola virus transmission in Montserrado, Liberia: A mathematical modelling analysis. *The Lancet Infectious Diseases*, 1189-1195. Retrieved November 7, 2015, from http://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099(14)70995-8.pdf

Rencher, A., & Christensen, W. (2012). Methods of multivariate analysis (Third ed.). Hoboken, New Jersey: Wiley.

Sanchez, A. (2003). Ebola Viruses. *Encyclopedia of Life Sciences*. Retrieved November 7, 2015, from http://onlinelibrary.wiley.com/doi/10.1038/npg.els.0001019/full

Seasonal ARIMA models. (n.d.). Retrieved December 1, 2015, from https://www.otexts.org/fpp/8/9

Shumway, R., & Stoffer, D. (2006). Time series analysis and its applications with R examples (2nd [updated] Ed.). New York: Springer.

Suffredini, A. (2015). Caring for Critically Ill Ebola Virus Disease Patients With One Hand Tied Behind Your Back*. Critical Care Medicine, 43(10), 2249-2250. Retrieved November 7, 2015, from http://journals.lww.com/ccmjournal/Fulltext/2015/10000/Caring_for_Critically_Ill_Ebola_Virus_Disease.26.aspx#P13

Vinck P, Pham PN, Kreutzer T (June 2011). Talking Peace: A Population-Based Survey on Attitudes about Security, Dispute Resolution, and Post-Conflict Reconstruction in Liberia. Human Rights Center, University of California, Berkeley. Available at www.peacebuildingdata.org

Walker, N., & Whitty, C. (2015). Tackling emerging infections: Clinical and public health lessons from the West African Ebola virus disease outbreak, 2014-2015. *Clinical Medicine*, *15*, 457-460. Retrieved November 7, 2015, from http://www.clinmed.rcpjournal.org/content/15/5/457.full

World Health Organization (2014). Ebola virus disease. (2015, August 1). Retrieved November 7, 2015 from http://www.who.int/mediacentre/factsheets/fs103/en/

World Health Organization (2015). Ebola in West Africa: Heading for catastrophe? Retrieved November 7, 2015 from http://www.who.int/csr/disease/ebola/ebola-6-months/west-africa/en/

World Health Organization (2015).Ebola Situation Reports. Retrieved November 7, 2015 from http://apps.who.int/ebola/ebola-situation-reports

World Health Organization (2015).Ground zero in Guinea: The Ebola outbreak smoulders – undetected – for more than 3 months. Retrieved November 7, 2015 from http://www.who.int/csr/disease/ebola/ebola-6-months/guinea/en/

World Health Organization (2015).The Ebola outbreak in Liberia is over. Retrieved November 7, 2015 from http://www.who.int/mediacentre/news/statements/2015/liberia-ends-ebola/en/