UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Spatio-temporal point process models for the spread of avian influenza virus (H5N1)

Permalink
https://escholarship.org/uc/item/8nc0r19n

Author
Kim, Harry

Publication Date
2011

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Spatio-temporal point process models for the spread of avian influenza virus (H5N1)

by

Harry Kim

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Bin Yu, Co-chair
Professor Cari Kaufman, Co-chair
Professor Peng Gong

Spring 2011

Spatio-temporal point process models for the spread of avian influenza virus (H5N1)

Copyright 2011
by
Harry Kim

Abstract

Spatio-temporal point process models for the spread of avian influenza virus (H5N1)

by

Harry Kim

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Co-chair

Professor Cari Kaufman, Co-chair

An outbreak of the devastating avian influenza virus (H5N1) was first observed in China in 1996. The explosive re-emergence of the virus after 7 years of its debut is estimated to be responsible for 14 million poultry deaths globally. Our research aims to identify the key factors (such as promixities to cities and roads and temperature) that are associated with the spread of H5N1 in Turkey and quantify their relationships to the virus dispersal. Our statistical model, the EAI (Epidemic Avian Influenza) model, is based on self-exciting point processes inspired by Hawkes [24] and Ogata [42]. A self-exciting point process can incorporate spatial and temporal dependencies of H5N1 outbreaks by specifying a branching structure among the outbreaks. In addition to quantifying the relationship between the virus spread and the key factors, the estimation result of the EAI model is used to predict future flu occurences.

To my family and friends

Acknowledgments

I would like thank my advisors, Professors Bin Yu and Kaufman, for the support and encouragement they have given me over the years. Without them, finishing this dissertation (and my Ph.D) would never have been possible. I also would like to thank Professor Peng Gong for introducing me to the subject and offering me his insight.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Introduction

1.1 The emergence of avian influenza virus (H5N1)

According to the World Health Organization, the devastating avian influenza virus (H5N1) which swept the world was first observed in China in 1996 [46]. After its debut, however, the outbreaks of influenza caused by the H5N1 virus had not been reported for seven years until 2003. The re-emergence of the virus raised major concern in Asia due to its lethal and explosive nature. Within four months after its reintroduction, the virus quickly spread to nine Asian countries damaging both domestic production and international trade of poultry products. The avian influenza scare continued on to Europe through Russia in 2004. Several North African countries also fell victim to the overwhelming epidemic disease shortly after.

The spread of H5N1 peaked in 2006; among the 115 human cases reported in 2006, 79 patients did not survive the deadly flu. This statistic—both number of cases and death—is about 25% cumulative statistic gathered from 2003 to 2010 [45]. It is estimated that 14 million birds were culled worldwide in effort to contain the virus. Thanks to increased awareness from the local governments and global organizations, the explosive spread of the fatal virus was significantly reduced in later years .

Despite the lack of coverage in the media in recent years, perhaps due to the emergence of infamous swine flu, the threat of H5N1 is still ongoing. There are a few countries, especially in South and South East Asia, where H5N1 is no longer considered to be an epidemic but an endemic disease. These countries suffer from recurring outbreaks of H5N1.

1.1.1 Characteristics of the H5N1 virus

In spite of its explosive spread, H5N1 infection is not caused by the usual means of flu dispersal, transmission through the air, which can be highly contagious. The avian influenza virus (H5N1) instead spreads strictly by direct contact. Susceptible birds can be infected when they come in contact with saliva, nasal secretions, and feces of infected birds [9]. Domesticated birds, therefore, have a higher chance of becoming infected and are more pathogenic due to their proximity to each other. Poultry farms often pack many birds in a limited space, which heightens the exposure to infected birds and contaminated materials.

For birds, researchers report that H5N1 has a mortality rate that can reach 90-100% within 48 hours [9].

The mortality rate for humans is also quite lethal; of the 507 infected humans from years 2003 to 2010 worldwide, about 60% (302) did not survive [45]. Fortunately, human-to-human transmission of H5N1 is virtually impossible as H5N1 strains are found to be attached only to the receptors on cells in the deepest regions of the lungs [56]. Although there are only two alleged cases of human-to-human transmission, H5N1 is a highly mutable virus [9] and researchers alert that this virus may trigger a devastating epidemic disease for humans similar to the Spanish influenza, which is responsible for more deaths than World War I [47].

1.2 Data

The lethal threat of avian influenza can be seen from the global data available on the H5N1 outbreaks. In this section, we describe these data and discuss their shortcomings originating from the collection method. We chose to focus our analysis on Turkey. In addition to the technical benefits which will be useful in our statistical analysis, Turkish H5N1 outbreak data appears to be more reliable than in other countries, although it is self-reported.

1.2.1 Data on the global spread of H5N1

To study the spread of H5N1, we use the data compiled by Declan Butler, a senior reporter for Nature magazine, gathered from FAO (Food and Agriculture Organization) and OIE (The World Organisation for Animal Health) [1] The data features 3206 H5N1 outbreaks in 54 countries from December 10^{th} 2003 to November 29^{th} 2006. For each outbreak, its location in latitude and longitude and the date of its occurrence were provided. Additional information for some outbreaks including the type of species infected, the number of dead, destoried, culled and vaccined birds were also given. Although it was unavailable at the time of our research, additional global data on spread of avian influenza is available for years 2006 and onward at the OIE website. The spatial distribution and temporal evolvement of all the outbreaks from December 10^{th} 2003 to November 29^{th} 2006 are shown in Figures 1.1 and 2.1 respectively.

Many of these outbreaks, especially the ones that occurred earlier, were self-reported, and the data suffer greatly from under-reporting and their details lack uniformity. For the majority of the cases, the recorded species of infected birds were unavailable and ambiguous even if they were provided. OIE has indicated that countries such as China and Indonesia have been consistently under reporting their cases [38]. Figure 1.1 suggests the possibility that the Chinese government might have only reported cases that resulted in more than 50 dead birds at each location. In comparison to other countries, where most of the outbreaks were reported to have less than 50 dead birds, the number of the outbreaks with death count less than this threshold in China was strikingly lower.

[1]The original data in KML (Google Earth XML) format was recompiled by Liang Lu, a graduate student at the Department of Remote Sensing, Chinese Academy of Science, into a excel spread sheet.

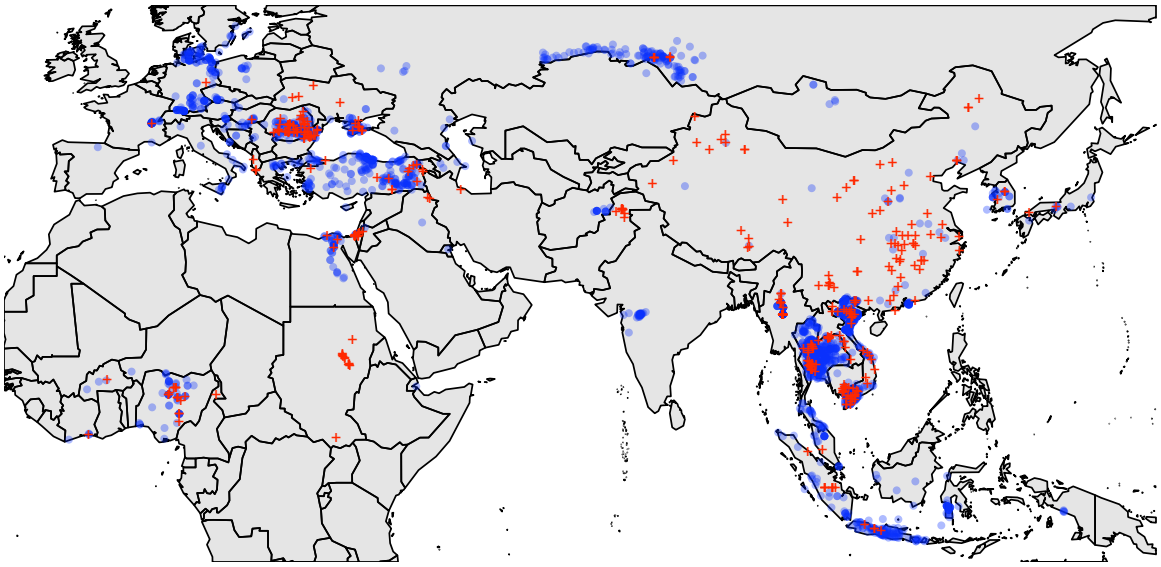

Figure 1.1: Plot of the global locations of H5N1 outbreaks. Outbreak locations with over 50 dead poultry are denoted with red crosses. The blue dots represent the rest of the outbreak locations.

1.2.2 Study Area: Turkey

While we can merely guess the reliability of the H5N1 outbreak data for each country, reported outbreaks in Turkey appear credible, judging from the consistency of provided data. In addition, the Turkish data has other technical advantages that will be explained later in this section.

For these reasons, we will focus our analysis on Turkey in this thesis. The Turkish spread of avian influenza, provided in the Declan's global data, lasted for 182 days from October 1, 2005 to March 31, 2006. During this period, there were total of 221 reported cases of H5N1 outbreaks. For each case, numbers of dead and destroyed birds and their species were given. The majority of reported cases (198) involved backyard poultry, but there were a few cases (9) associated with wild birds—including a case with a migratory bird, a mallard duck. Figures 1.2 and 2.2 illustrate the spatial and temporal distributions of the outbreaks that occurred in Turkey during this period.

There are several notable advantages of concentrating our analysis on Turkey: 1) Turkey has one of the highest poultry productions among the European and Western Asian countries [65]. 2) The domestic demand and production of poultry do not fluctuate according to religious holidays in November and December, as 99% of its population are Muslims [41]. Further, its main market for export of poultry products is the Middle East[2] [40]. Thus seasonality, which will be noted in Chapter 3, is not a result of changes in poultry production level. 3) Numerous data on its infrastructures—locations of railroad and highways for

[2]Recently Vietnam became a big importer of Turkish poultry products but the international trade was initiated after the study period

example—are readily available. 4) Turkey is relatively invariant under different geographical projections due to its proximity to equator. Its rectangular shape is attractive for our statistical approach. 5) This statistical approach also benefits from Turkey's traffic networks, which are not as dense as those of other European countries. Further details of its technical advantage will be discussed in Chapters 2 and 4.

Aside from the reported cases of H5N1, we require a few additional data sets for our study. In the chapters to follow, proximity from an outbreak location to the nearest highways, railroads, and major cities, and temperature time series will be instrumental in construction of our EAI (Epidemic Avian Influenza) model presented in Chapter 4. The locations of highways and railroads in Turkey were obtained from ESRI (Environmental Systems Research Institute) and are shown in Figure 1.2 with green and blue lines respectively. Likewise, locations of 96 cities in Turkey were obtained from the same source and are represented as red crosses in Figure 1.2. Information on air temperature was retrieved from NCEP (National Centers for Environmental Prediction) and is available for latitude and longitude coordinates ranging from $-180°$ and $180°$ by increment of $2.5°$ per day.

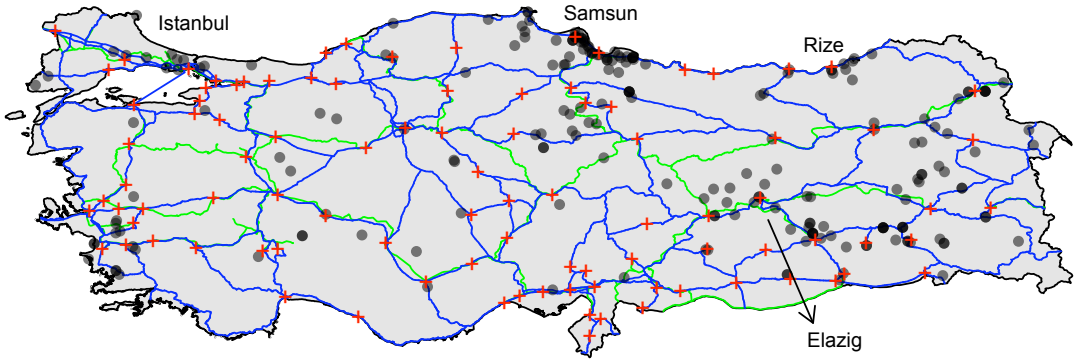

Figure 1.2: Plot of Turkey with outbreak locations marked black. The red crosses correspond to locations of cities in Turkey. The green and blue lines represent major highways and railways respectively.

1.3 Current hypothesis on spread of H5N1

Based on the globally observed outbreaks of avian influenza, many prior research have focused on determining the key factors that influence the H5N1 dispersal as a part of disease surveillance. Understanding the major contributors to the disease spread is a crucial first step towards the prevention of the deadly influenza.

In this section, we present three notable possibilities contributing to the spread of the H5N1 virus. They are movements of poultry and related products by humans, migratory birds that are capable of being healthy carriers of H5N1, and ecological factors . Because we

want to develop a general framework for modeling H5N1 spread, it is important to understand the factors associated with the dynamics of the disease in a global scale, although our late analysis will focus only on Turkey.

1.3.1 Movements of poultry and poultry products

Domestically, the spread of H5N1 is linked to movements of poultry, poultry manure, poultry by-products and accidental transfer of infected material from poultry farms, such as water, straw or soil on vehicles, clothes, and shoes [26].

International trade—both legal and illegal—may also have contributed to dispersal. In 2007, legally imported live poultry from Hungary, which was suffering from H5N1 spread at the time, was suspected to be a cause of avian influenza outbreaks in United Kingdom [33]. A year later, 17 H5N1 positive cases were reported in legally imported poultry seized at two Vietnamese ports of entry [39].

In 2004 and 2005, multiple illegal smuggling of exotic birds from Asia infected with H5N1 were intercepted in Europe [26]. This time frame coincides with the explosive dispersal of H5N1 outbreaks in Europe, although incidents of smuggling may have been completely irrelevant to the outbreaks.

Some researchers note that a possible avenue of disease spread from Asia to Europe in 2005 is the Trans-Siberian railway, a major trading route between the two continents [20]. As shown in Figure 1.3, the locations of H5N1 outbreaks were found to be scattered along the Trans-Siberian railway, suggesting an association between the H5N1 outbreaks and poultry trade.

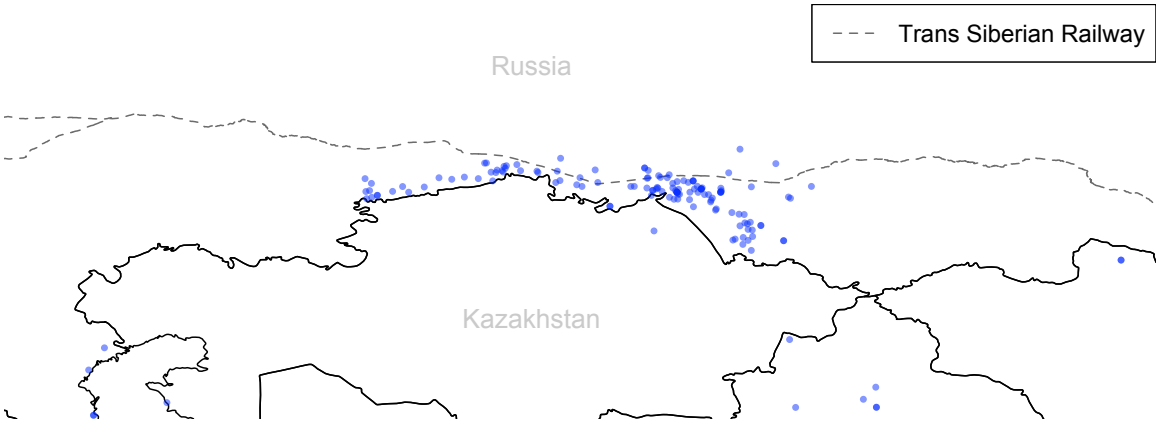

Figure 1.3: The Trans-Siberian railway is represented with a dotted line. Locations of outbreaks in Russia marked in blue.

1.3.2 The role of migratory birds

In addition to humans serving as a vector of the epidemic by transporting poultry and related goods, researchers suspect migratory birds to be another major contributor to the spread of H5N1. However, whether the migratory birds play a role in the virus dispersal is still heavily debated. While we cannot completely rule out the effect of migratory birds, many recent publications have been denouncing such a hypothesis for lack of evidence [20, 29, 17, 18, 62].

In the earlier cases, most of the wild birds found dead were in close proximity with the farms swept by H5N1 and were thought to have obtained the disease from birds at the farms. It was not until April 30th, 2005, when hundreds of bar headed geese were found to be infected and dead with H5N1 at Qinghai lake in China, that researchers started to seriously consider the possibility of wild birds as carriers of H5N1 [20]. Later that year, in November and December, 840 wild birds (mainly swans) were reported dead in two of the ten major regions affected by H5N1 in Russia. There were also swans reported dead in Romania (137 in number) and Croatia in October due to H5N1. The spread of the virus continued towards western Europe, and the great majority of reported cases from February of 2006 included dead wild birds. Although most of them were resident waterbird species, some were migratory, including Common Pochard and Tufted Duck.

In 2004, it was shown that domestic Mallard ducks could be healthy carriers of the virus. Researchers believed that the genus Anas (the family of bird specie that Mallard duck belongs to) is highly likely a carrier. The few numbers of deaths reported for this bird specie seemed to support this claim. However when the researchers surveyed tens of thousands of Anus ducks in Europe, Asia, Australia and Africa for verification, only 33 were found to be carrying H5N1 [17].

While the few healthy carriers of the virus may be able to infect other birds locally, researchers have found no evidence that migratory birds are responsible for long distance transmission of H5N1 [20, 17, 18]. Migratory birds are often capable of traveling several hundred kilometers in a single day. If they are the main agents of the virus dispersal, the locations of outbreaks should occur in jumps of several hundred kilometers, corresponding to their migratory sites. Instead, the observed global spread of avian influenza developed progressively without displaying long distance jumps.

Researchers who claim that migratory birds are not a main agent of H5N1 dispersal emphasize the role of humans in virus transmission through transportation of poultry and poultry products [20, 17].

1.3.3 Ecological factors

Besides humans and migratory birds, ecological factors can also impact the dynamics of H5N1 dispersal. Similar to any other viruses, the survival of avian influenza viruses depends on ecological factors. It is a well known fact that temperature is inversely related to the survival of all types of avian influenza viruses. Moreover, influenza viruses can prolong their life significantly in water. Laboratory experiments confirm these facts. Webster et al. [63] found that a subtype of H3N6 was able to survive up to 32 days at 4°C, while it only managed

4 days at 32°C in river water. Stallknecht et al. [58] produced similar results with another type of avian influenza, H6N2.

In more recent experiments, Brown et al. [6, 7] compare the effect of various ecological factors to the survival of H5N1 virus. They found that lower salinity, acidity, and temperature of the water containing H5N1 virus yield higher survival rates of H5N1.

While these favorable ecological factors may extend the presence of H5N1 virus, whether they directly contribute to the virus spread still remains uncertain. Infected birds can transmit the virus to nearby birds by contaminating a water source with increased persistence in cold weather. Wintering sites of migratory birds, Qinghai Lake for example, often provide these suitable environmental conditions as migratory birds flock together at wetlands for easy access to water and food. In reality, however, there is no adequate way to measure the relationship between the increased persistence of the virus and its lethality other than comparing the numbers of outbreak cases according to the temperature. The exploratory data analysis featured in Chapter 2 will examine this relationship and show that there is an association between the number of outbreaks and temperature, especially for our study area, Turkey.

1.4 Overview of the thesis

The goals of this thesis are threefold: first, through exploratory data analysis, to investigate the mechanisms of H5N1 spread and determine the key factors that contribute to its explosiveness; second, to develop a statistical model based on point processes to assess the past progression of the disease in Turkey; third, to build an algorithm using our statistical model to predict the future disease spread conditioned on past observations of Turkish H5N1 outbreaks.

This thesis addresses these goals within the following structure. Chapter 2 visually explores temporal and spatial patterns of H5N1 spread in Turkey and compares it to the global trend. Through exploratory data analysis, we aim to determine the contributing factors to the virus dispersal. Among the contributing factors to H5N1 outbreaks we considered, temperature and proximity to traffic networks and cities were found to be associated with the virus dispersal. Prior to constructing our statistical model using these factors as predictors, Chapter 3 discusses past statistical approaches used to model the spread of avian influenza, and they are based on spatial logistic regression and spatial autoregression. Their results and drawbacks are noted in the same chapter. These modeling approaches omit the analysis of temporal trend and do not incorporate spatio-temporal dependencies of H5N1 outbreaks appropriately. To address these shortcomings, we consider a different modeling approach in Chapter 4. The first part of Chapter 4 reviews self-exciting point process [24], which will serve as a starting point for our statistical approach. Further, an expansion of the self-exciting process, ETAS (Epidemic Type Aftershock Sequence) model [42], is discussed along with benefits of this modeling approach to our data. Drawing inspiration from the ETAS model, we propose our statistical model, the EAI (Epidemic Avian Influenza) model, along with three estimation methods: backfitting, Poorman's EM (Expectation - Maximization) and EM. We provide results from fitting the model using each method to the Turkish data.

Five variations of EAI models are considered and the best model is determined based on AIC (Akaike Information Criterion). Chapter 5 presents a validation for the best EAI model through a residual analysis using Stoyan-Grabanik weights [59]. In addition, the validation results for the second best EAI model will be compared against those of the best to examine the improvement in model fit. Chapter 6 presents simulation studies of the EAI model, and examines whether the three estimation methods are able to obtain accurate estimates when the truth is known. Performance of the three estimation methods is compared in terms of sensitivity to the starting values and accuracy of the estimates with and without an edge correction. Moreover, prediction results based on Monte Carlo simulation is provided for Turkey during a 60-day period before the last day of observed H5N1 outbreak. Finally, Chapter 7 summarizes our findings and discusses areas that can be improved in the future.

Chapter 2

Exploratory Data Analysis

In the first chapter of this thesis, we introduced and discussed three potential contributors to H5N1 spread. While the role of humans in transferring the disease is widely accepted, whether the migratory birds, which are capable of being carriers of H5N1, impact the virus dispersal is heavily debated. In addition, ecological factors are scientifically proven to affect the survival of the virus, but a clear linkage between these factors and the spread of avian influenza has not been determined.

Through exploratory data analysis (EDA), this chapter investigates the contributors to avian flu mentioned above, using the H5N1 data featured in Section 1.2. EDA is an important part of a statistical analysis, as asserted by the great statistician, John Tukey, since it provides candidates for formal statistical modeling. We want to identify the key factors that are highly likely to be associated with the dynamics of the disease spread. Identified factors will be included as predictors in our point process model, the EAI (Epidemic Avian Influenza) model, proposed in Chapter 4. Using the EAI model, we aim to quantify the relationship between the potential contributing factors and H5N1 outbreaks, as the quantified relationship will provide better understanding of the dynamics of the virus spread. Because our goal is to establish a general framework for modeling spread of avian influenza, when appropriate, we will examine the differences between patterns of global and Turkish H5N1 outbreaks for each factor, despite the fact that we later restrict our study area to Turkey.

The EDA presented in this chapter is organized into two parts: temporal and spatial analyses on the patterns of H5N1 outbreaks. First, we focus on the temporal aspect of the virus dispersal and investigate the viable causes of observed seasonality shown in Figure 2.1. Among many possibilities, we inspect how the outbreak occurrences vary according to an ecological factor, temperature, and migration patterns of wild birds in order to verify the corresponding hypotheses mentioned in Chapter 1. The relationship between temperature and number of H5N1 outbreaks in Turkey is explored in Figure 2.2. Moreover, we analyze migratory patterns for two bird species, whose breeding and wintering sites closely match the locations of Russian and European outbreaks, as shown in Figure 2.4.

Following the temporal analysis, we turn our attention to exploring the spatial distribution of H5N1 outbreaks. Similar to seasonality observed from the temporal distribution, the most noticeable spatial feature of the H5N1 outbreaks is their proximity to railroads

and major highways. This spatial feature supports the hypothesis that the transportation of poultry and poultry products is influential to the vector of the virus. Proximity of H5N1 outbreaks to the nearest traffic networks is analyzed for European countries with emphasis on our study area, Turkey, in Figure 2.5. In addition to the clustering along the traffic networks, the locations of the outbreaks are found to crowd near cities. To understand this phenomenon, we use Figure 2.6 to examine populations of the cities, which may be one of the key factors responsible for the observed clustering, as poultry farms favor close proximity to populated cities for easy access to markets. Lastly, we review the poultry density map, created by Wint and Robinson [65] and reproduced in Figure 2.7, to determine whether our analysis can benefit from incorporating their covariates to the EAI model. A few notable covariates considered in the production of the density map are proximity to traffic networks and cities, human population, vegetation index, and elevation.

2.1 Temporal Analysis on the observed seasonality

The most notable temporal feature of global spread of H5N1 is its seasonal variation. Figure 2.1 displays the number of globally reported H5N1 cases per day from December 10th, 2003 to November 29th, 2006[1]. From the plot, we can observe three waves of outbreaks peaking in the colder seasons, fall and winter. While it is apparent that there is an association between H5N1 outbreaks and seasonality, determination of the causes may be difficult. One possibility is that the temperature in cold seasons prolongs the life of the avian influenza virus. Another possibility is related to the poultry production cycle. Globally, the production of poultry increases just prior to winter, to accommodate high demand in poultry products for religious holidays. Higher poultry production leads to larger total susceptible population, thereby increasing the chance of virus spread[2]. In addition, some migratory birds can become hosts of the H5N1 virus and potentially infect other susceptible birds at their wintering sites, contributing to higher number of outbreaks in the colder months. We will further investigate these issues in the following sections.

2.1.1 The effect of temperature

As noted in Chapter 1, temperature has a crucial influence on the survival of avian influenza virus. Laboratory results conclude colder temperature aids the persistence of H5N1 virus when it is present in water [6, 7]. Although a correct analysis would include proximity to water sources or locations of wetlands in addition to temperature, determining the type of water sources that realistically contribute to the virus spread is arduous. Therefore, we will restrict our analysis to temperature. We will further focus on our study area, Turkey, as applying the same analysis to the global data is not only difficult but harder to interpret.

[1] There are a few extreme numbers of outbreaks in 2004 and one in 2005, which strongly suggests delayed reporting. Most of these highly concentrated number of outbreaks were reported from South Asian countries. A smoothed frequency plot using LOWESS was created to make a fairer comparison among the three waves.

[2] Our choice for the Turkish data prevents this possibility, and it will be excluded from further investigation.

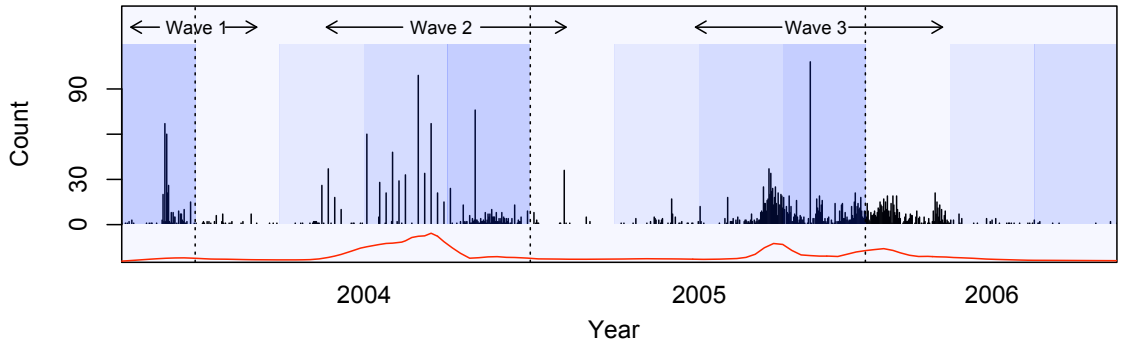

Figure 2.1: The global frequency of H5N1 outbreaks measured per day from December 10th, 2003 to November 29th, 2006. The different shades represent the change in seasons sequentially with the lightest corresponding to spring and the darkest to winter. The dotted lines mark the beginning of the years. The red curve on the bottom is smoothed frequency plot generated using LOWESS (Locally Weighted Scatterplot Smoothing).

We wish to confirm that the laboratory experiment results agree with the patterns of past outbreaks.

Figure 2.2 displays the frequency of outbreaks in Turkey from late 2005 to early 2006 with the corresponding temperature trend. The trend shown in the graph is an average of air temperatures from four locations in Turkey, taken from weather data, provided by NCEP (National Centers for Environmental Prediction). Although the temperatures differ according to location, the patterns of their changes are almost identical. Therefore, the average temperature should not be interpreted as an average over the entire country, but relative changes are interpretable.

The plot suggests that there is an association between temperature and number of outbreaks. In November and December, times when the temperature drops below 0°C roughly correspond to the times of outbreak incidences. We observe many more outbreaks when the temperature falls even further in January and February, with the number of cases peaking in early January and slowly declining until the end of March. The outbreak count appears be inversely related to temperature, as fewer outbreaks occur after the temperature returns to above 0°C in late February. Although these relationships over a single year are not conclusive, they are suggestive that colder weather can facilitate the spread of H5N1 outside of laboratory conditions.

Examining Figure 2.3, the scatter plot of number of reported H5N1 cases against the corresponding average temperature, we note the association noted in the previous paragraph does not display a linear relationship. The scatter plot indicates that the highest numbers of cases were reported on the days with temperatures around 0°C and the numbers decline as the temperature increases. The numbers of cases also decrease for extremely cold days, but in general, they are greater than those of the days with temperatures over 10°C. There

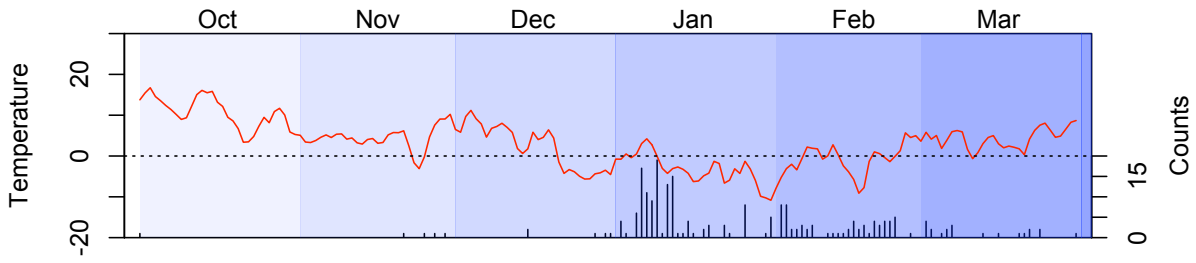

Figure 2.2: Plot of reported H5N1 cases per day in Turkey from October 1, 2005 to March 31, 2006. Average temperature taken from 4 locations in Turkey was also plotted in red. The dotted line marks 0°C, and different shades are used to denote the change in months.

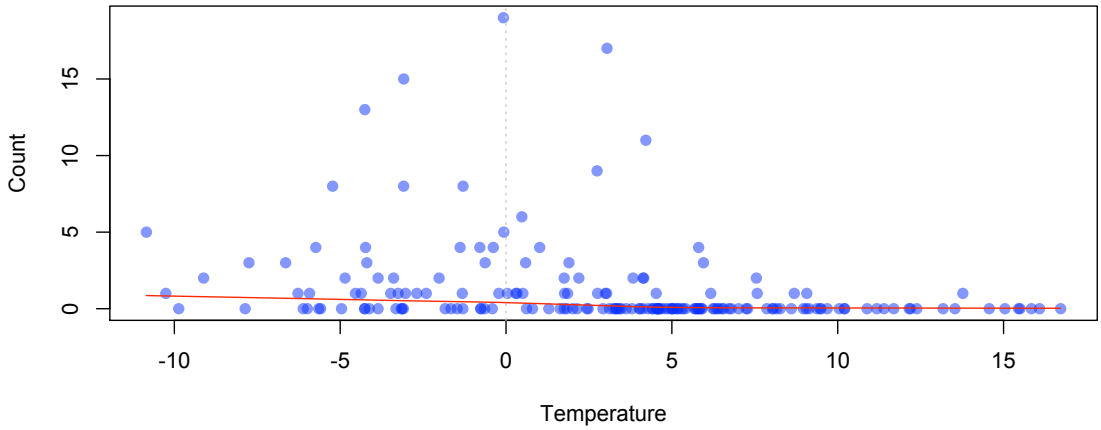

Figure 2.3: Scatter plot of number of reported H5N1 cases in Turkey per day from October 1, 2005 to March 31, 2006 against the corresponding average temperature. The red curve is generated using LOWESS (Locally Weighted Scatterplot Smoothing).

are much more days with temperatures above 0°C that did not have any reported cases compared to those with temperatures below the freezing point.

Furthermore, an interesting pattern can be noted in Figure 2.2, though it may be coincidental; the harsh temperature drops in late December, January, and February all precede the large number of outbreaks, and these temporal clusters occur as the temperature bounces

back. We have not found a reasonable explanation for this behavior, and it may deserve further research.

Lastly, we note, despite the association between the temperature and the number of outbreaks described above, the initial outbreak occurs when the average temperature is quite high (15°C). This outbreak is clearly an outlier, but it should be accounted for our analysis. Although its time of occurrence is unusual, it may provide important information about how an H5N1 outbreak triggers others. No H5N1 outbreaks were reported until 51 days after the first outbreak, presumably due to unfavorable ecological factors such as high temperature. If in fact temperature is a determining factor in variation of the virus spread, the initial outbreak will be useful in constructing a branching structure of the disease for our statistical model proposed in Chapter 4.

2.1.2 Migratory patterns of wild birds

In addition to ecological factors, another suspected cause of H5N1 dispersal is the contribution from migratory birds, which are potential carriers of the avian influenza. To justify such a claim, we hope to unveil temporal features that suggest a relationship between past progression of outbreaks and wild bird migrations. We first investigate the global migratory patterns, then discuss the implication of the results for our study region.

Gilbert et al. [21] identify 38 species of migratory birds as possible hosts of H5N1. Using the GROMS (Global Register of Migratory Species) database, we were able to retrieve the locations of the breeding, staging, and wintering sites of 29 identified species. Among the 29 species, we found migratory patterns of two species particularly interesting. Figure 2.4 shows the breeding, staging, and wintering sites for Slender-billed Curlew and White-headed Duck with H5N1 outbreak locations. The frequency plots are also provided according to the different migratory sites.

As illustrated in Figure 2.4, the breeding/wintering areas for both species match the locations of outbreaks in Russia, and their wintering/resident areas cover a large number of outbreak locations in Europe. In Section 1.3.1, it was mentioned that the virus spread continued on to Europe from Asia through Russia. The Trans-Siberian railway was speculated to be responsible for the transfer of the deadly virus, possibly offering a mode of transportation for infected poultry and poultry products [20]. Therefore, the discovery of migratory areas that coincide with the Russian and the East European H5N1 outbreaks can provide evidence for the role of wild birds in spreading H5N1.

However, it is difficult to verify that the two bird species contributed in introducing H5N1 virus to Europe. For the Slender-billed curlew, outbreaks in wintering areas in Russia appear to occur after the European outbreaks, which is the opposite of what we expected. On the other hand, the migratory pattern and temporal occurrences of H5N1 outbreaks are matched better with White headed ducks, as the outbreaks in their breeding area preceded those in their resident area, which covers Turkey and Eastern Europe.

Although the White headed ducks appear to be associated with European H5N1 spread, their contribution is likely to be small. The Bird Life International Organization lists both wild bird species as endangered, estimating less than 50 Slender-billed Curlew and 13,000

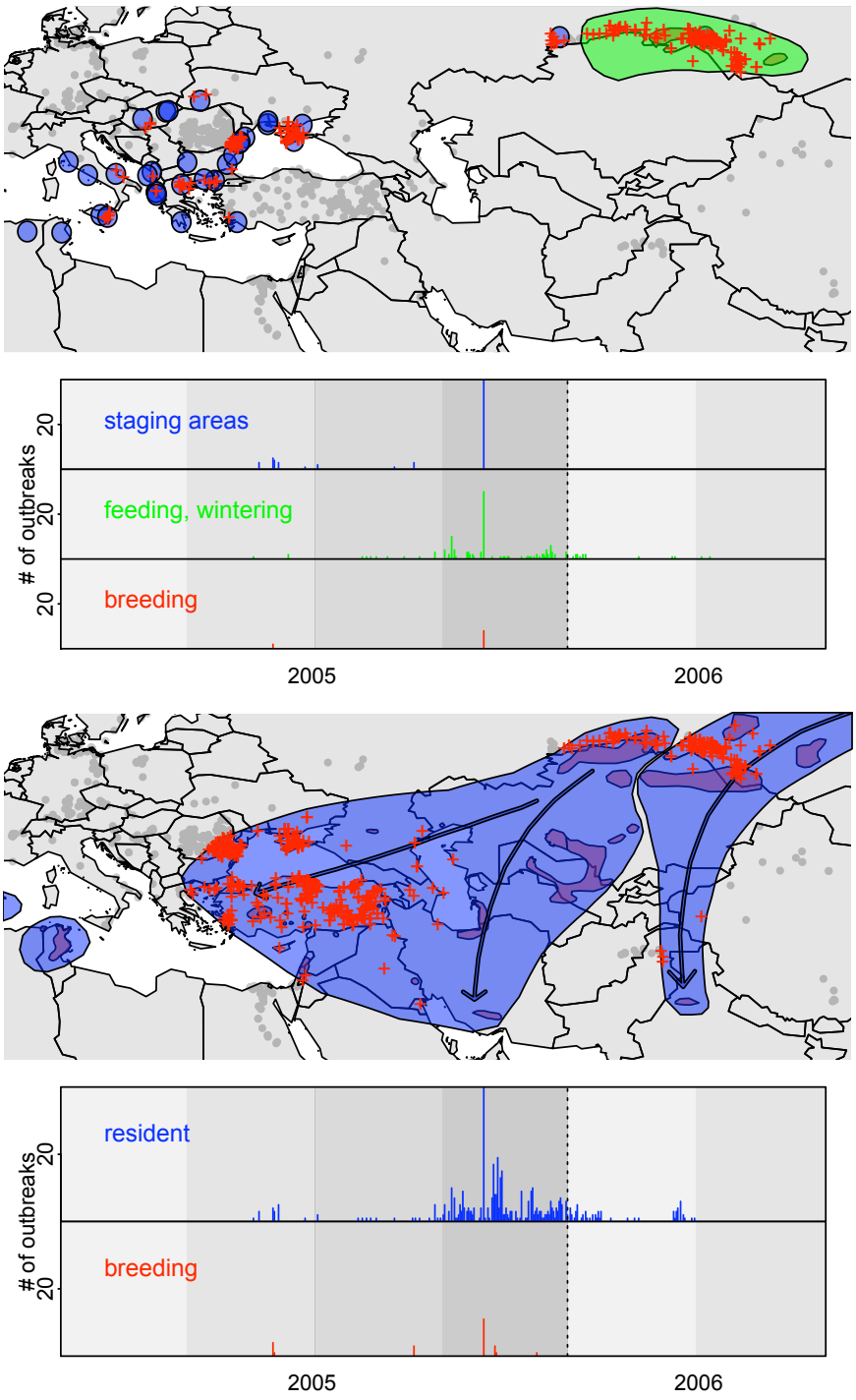

Figure 2.4: Plot of breeding, staging, and wintering sites for Slender-billed Curlew and White-headed Duck respectively. Locations of H5N1 outbreaks are marked with red crosses if they fall in the migratory sites and with grey dots otherwise. The frequency plots for both species are also shown according their migratory sites with matching colors. The different shades correspond to change in season with darkest shade representing winter. The dotted line indicates change in years.

White-headed Ducks remaining worldwide [27, 28][3]. Further, there have been no reported sightings of Slender-billed Curlew since 1999.

Other than the two species, we were unable to observe other migratory patterns closely matching the locations of outbreaks. While we cannot rule out the effect of migratory birds, our explanatory data analysis did not find a convincing association between the migratory sites of wild birds and H5N1 spread.

The same statement can be made about our study area, Turkey. Only one of the 221 reported Turkish H5N1 cases is known to have involved a migratory bird, a mallard duck. From Section 1.3.2, we know that mallard ducks are capable of being healthy carriers of H5N1. Therefore, one reported case pertaining mallard duck is not entirely unexpected. However, compared to other factors that are highly relevant to dispersal of Turkish H5N1 outbreaks, it is nearly impossible to gain concrete evidence that migratory birds impact the dynamics of the epidemic. In the text to follow, because of this uncertainty, the role of migratory birds will be excluded.

2.2 Spatial Analysis on the outbreak locations

Our analysis of the temporal distribution of H5N1 outbreaks revealed that temperature is highly associated with the dynamics of the virus spread. On the contrary, we decided to eliminate the impact of migratory birds from the future consideration due to lack of convincing evidence.

Continuing our investigation of factors in virus dispersal, we examine spatial features of H5N1 outbreaks. Despite the obvious economic and environmental differences among the countries that have suffered from the deadly virus, the outbreaks in these countries share a few of spatial patterns; the outbreaks tend to aggregate near major traffic networks, highways and railroads, and cities. The same pattern can be observed from the Turkish H5N1 data, as illustrated in Figure 1.2. These two features are now graphically explored in depth, and viable explanations for these behaviors are discussed in Sections 2.2.1 and 2.2.2.

In addition to analyses of proximity to traffic networks and cities, we review a global poultry density map prepared for FAO (Food and Agriculture Organization of the United Nations) by Wint and Robinson [65]. The map displays the estimated number of poultry at each given location, indicating susceptible population of poultry to avian influenza. We will investigate whether our analysis can benefit from incorporating the covariates used to generate their estimates. A few notable covariates are proximity to traffic networks and cities, human population, vegetation index, and elevation.

2.2.1 Proximity to traffic networks

The most universal spatial feature of H5N1 outbreaks is its proximity to traffic networks. Although our analysis is limited to the countries with available traffic network data, regard-

[3]In comparison, the population of Mallard ducks, which are capable of being carriers of H5N1, are estimated to be 8.5 Million.

less of the countries involved, the outbreaks consistently cluster along railroads and highways. While a global examination is complicated, we provide a comparison among the European countries to study the aggregation[4]. Using the European railroads and major highways data provided by ESRI (Environmental Systems Research Institute), the minimum distances from outbreaks to nearest traffic networks were measured and plotted in Figure 2.5[5].

The collection of box plots in Figure 2.5 confirms that the outbreak locations are close to traffic networks across all European countries. The calculated distances are a lot shorter in developed, mostly Western European, countries probably because of their denser traffic networks and fewer number of outbreaks. In comparison, Greece, Turkey, Ukraine, and Romania had relatively large spread of distance distribution with the maximum distance reaching as far as 68 km. The traffic infrastructure in these countries, which are not as well developed as the rest of the European countries, may be responsible for the observed spatial pattern. Moreover, lack of government response and required infrastructure to quarantine the disease could also have boosted the number of outbreaks, thereby contributing to the larger variation; the number of outbreaks from three countries, Turkey, Ukraine, and Romania, in fact, make up 52% of all European incidences.

Turning attention from all European countries to our study region, Turkey, we examine the distribution of distances from outbreaks to nearest traffic network in Turkey. The histogram in Figure 2.5 illustrates that locations of Turkish outbreaks are indeed closer to traffic networks than what we would expect if the locations were uniformly scattered. The distribution is heavily shifted to the left relative to the red density curve, produced by simulating locations in Turkey via poisson process with a homogenous rate[6]. The distribution of the distances are significantly higher than the simulated outcome within 10 km but falling below for longer distances.

A plausible explanation for this clustering is that the poultry farms tend to be located near traffic networks and cities to gain easy access to markets [65]. Unlike other livestock farming, poultry farming is less dependent on land resources for its feed, and the locations of poultry farms are determined accordingly. Moreover, transfer of infected poultry and associated products are likely to be linked with the major highways and railways as the traffic networks provide the mode of transportation.

2.2.2 Proximity to major cities and their populations

The second most prominent spatial pattern of the H5N1 outbreaks, following the proximity to traffic networks, is the vicinity to cities. Globally, H5N1 outbreaks are frequently found to be clustered around major cities. Our study area, Turkey, is no exception. Referring

[4]Reliable data on locations of traffic networks for other parts of the world are not readily available.

[5]The traffic network data obtained from ESRI is in shapefile format, comprised of groups of line segments forming curves, which represent major highways and railways. To calculate the minimum distances from observed outbreaks to the nearest traffic network, 20 equal spaced points were created between the two edges of each line segment. The geodesic distances from an outbreak location to all points created between the line segments are calculated, and the smallest value is returned as the minimum distance.

[6]The homogenous rate for the Poisson process is estimated using the ratio between number of outbreaks in Turkey, 221, and the its area, 783,562 km^2

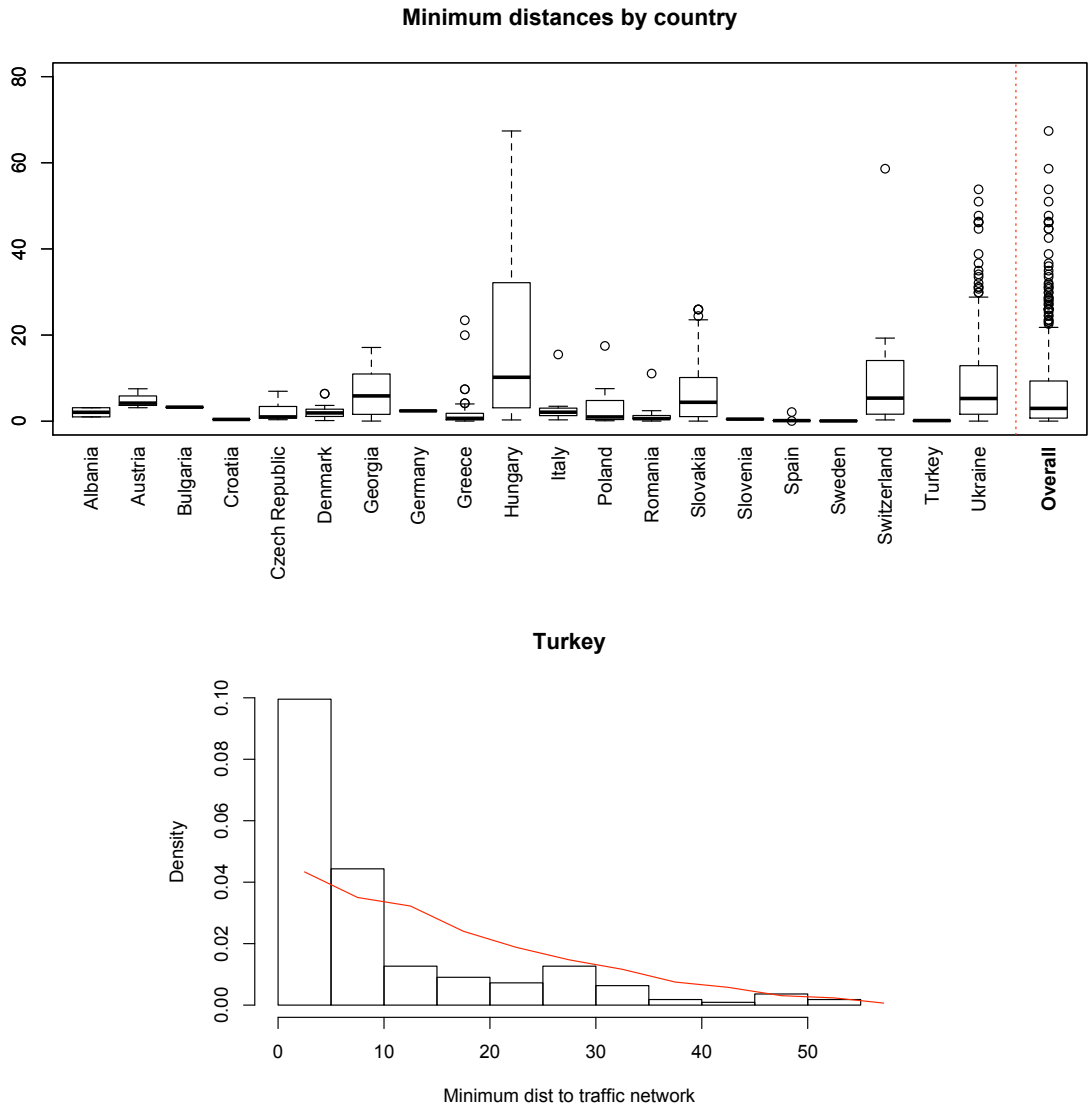

Figure 2.5: The first plot is a collection of boxplots of minimum distances from outbreaks to nearest traffic networks calculated for each European country. The distribution of overall distances is included next to the red dotted line. The second plot is the histogram of minimum distances in Turkey. The red density curve is a result from uniformly simulating the same number of locations in Turkey and measuring the distances to the nearest traffic networks. The unit for distance is kilometer.

back to Figure 1.2, we notice outbreaks in Turkey crowd especially around the three coastal cities, Samsun, Rize, and Istanbul, and an inland city named Elazig. Curiously, this rather extreme aggregate behavior is merely observed for these few cities and does not apply to other Turkish cities included in our analysis. Therefore, this section will not only investigate the clustering corresponding to proximity to cities but examine whether populations of the

cities are responsible for attracting the outbreaks.

We first analyze the distribution of distances from outbreak locations to the nearest cities in Turkey. The histogram in Figure 2.6 illustrates that the distribution appears to be constantly decreasing according to the length of distances—with an exception of a little hump at 40 km. Other than the hump, the distribution closely follows the red density curve generated from measuring distances between the outbreaks and the uniformly simulated locations featured in the previous section.

Judging by the dense outbreak clusters observed near cities, the lack of discrepancy between the actual and simulated distance distribution may seem unusual. Despite the heavy clustering in vicinity to urban areas, the skewness of the distribution shown in Figure 2.6 did not differ substantially from the simulated outcome, in comparison to that of proximity to traffic networks demonstrated in Figure 2.5. The deficit of skewness is not entirely unexpected because the dense cluster of outbreaks are mainly formed near the few coastal cities. The hump noted around 40 km is the result of sparse aggregation of inland outbreaks near Elazig. The outbreaks are loosely scattered around the inland cities relative to the heavy clusters found near the coastal cities, Istanbul, Samson, and Rize. Although the reason for this difference is unknown, it is possible that there are more poultry farms located near the coastal cities for easy access to ports. As mentioned in Chapter 1.2, Turkey is one of the major exporters of poultry and poultry products in the world.

While mapping the locations of poultry farms is a challenging task, population of an urban area can be a useful indicator for density of poultry farms. In developing countries, human population is a significant determinant of locations of animal farms [65], especially in Asia [19]. Because poultry farming is less dependent on the land resources for feeding, poultry farms favor proximity to cities even more to gain easy access to markets.

If the population is a differentiating factor for density of poultry farms, higher number of H5N1 outbreaks are expected occur close to heavily inhabited areas. The scatterplot in Figure 2.6 explores the relationship between the number of outbreaks that occurred in proximity to a city and its population. Only the 15 cities with the largest numbers of corresponding outbreaks are shown. Although the scatterplot indicates a clear outlier, Istanbul (the largest Turkish city), with a population of 8 million, the populations of other cities appear to remain under 1 million as the number of their corresponding outbreaks vary. Therefore, the scatterplot does not show convincing evidence that human population influences the aggregation of H5N1 outbreaks near urban areas, at least in Turkey[7].

The exploratory analysis in this section revealed that majority of the cities did not experience the heavy clusters of outbreaks formed near a few cities. While H5N1 outbreaks densely aggregated around the coastal cities, clusters close to the inland cities were relatively looser. We examined whether human population is the differentiating factor but found no convincing evidence. Although we were not able to determine why clusters of outbreaks form around certain cities, we believe that proximity to city influences the spread of bird flu, and it will be included in our modeling consideration presented in Chapter 4.

[7]This conjecture may be more applicable in developing countries in Asia

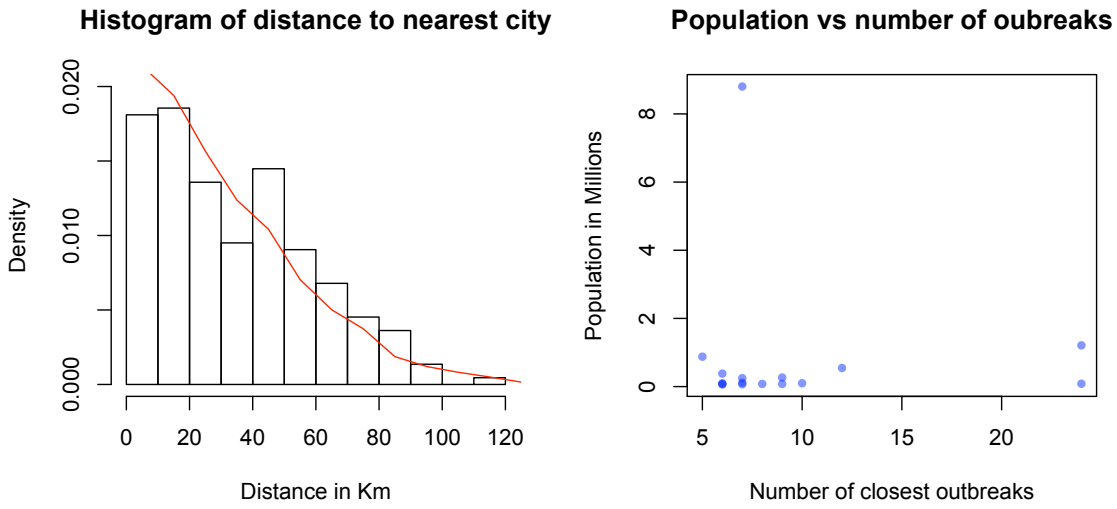

Figure 2.6: The plot on the left is a histogram of minimum distances from outbreaks to nearest cities. The red density curve is a result from measuring distances to uniformly scattered locations generated via simulation. The same simulated locations were used in Figure 2.5. For each outbreak location, its nearest city is determined, and the cumulative counts of the outbreaks were assigned to their closest city. The plot on the right is a scatterplot between the number of outbreaks corresponding to nearest city and their populations. Only the 15 cities with largest number of neighboring outbreaks are shown.

2.2.3 Poultry density at outbreak locations

In the previous section, we considered population of the urban areas as a possible indicator for density of poultry farms. Assuming that higher density of poultry farms equates to a larger susceptible poultry population exposed to avian influenza, we expected to observe more H5N1 outbreaks in heavily inhabited areas. Our analysis, however, suggested that population is not likely a factor that impacts the aggregate behavior of the virus spread.

Knowledge of poultry density distribution would aid our understanding of H5N1 dispersal. With such knowledge, we would be able to verify, for example, whether the three Turkish coastal cities that severely suffered from avian influenza had more poultry farms around them than the inland cities. Although estimation of global poultry population is a challenging task, Wint and Robinson [65] provide a global poultry density map, constructed using linear regression model. Among the many covariates considered in their analysis, some of them are proximity to roads and cities, vegetation index, human population, and elevation. .

Figure 2.7, a reproduction of Wint and Robinson's poultry density map, suggests that the outbreak locations—shown as blue dots—coincide with the areas high poultry density. The histogram in Figure 2.7, which compares the poultry density at outbreak locations with the overall distribution of poultry density drawn in red, confirms our observation. The

distribution of overall poultry density is highly skewed to the right with most of the poultry density less than 400, while the mode for poultry density at outbreak locations occurs around 400. Higher poultry density is observed at the outbreak locations relative to the rest of the world.

While this outcome is promising, we will adapt some of the explanatory variables Wint and Robinson considered, instead of incorporating the poultry density map into our analysis directly. Even though the density map may be a good predictor for locations of poultry farms susceptible to H5N1, we note that the spatial factors we have already addressed in this chapter are included in their analysis. Other than proximity to traffic networks, and cities, we will exclude the remaining explanatory variables such as vegetation index and elevation from our consideration. Vegetation index is not likely to be a strong indicator of farm locations susceptible to H5N1 because poultry farming depends less on land resources [65]. Moreover, variation in elevation is likely to be explained by the locations of cities and traffic networks as these locations tend to avoid high altitude. In addition, the map from Wint and Robinson [65] is an estimate of poultry density subject to uncertainty; we prefer to condition on variables known with near certainty such as cities and traffic networks.

2.3 Summary

Among the contributing factors to H5N1 outbreaks we investigated through exploratory data analysis, the most notable ones were temperature and proximity to traffic networks and cities. These factors correspond to the possible causes of H5N1 mentioned in the introduction. The temperature is scientifically shown to extend H5N1 persistence when the virus is present in water. Proximity to traffic networks is closely linked with transportation of infected poultry and poultry goods. Although the aggregation of outbreaks is only observed for only a few cities in Turkey, we will also include proximity to cities in our modeling consideration, since the same pattern is noted globally. The three variables, temperature, proximity to traffic networks and cities, will be incorporated to our EAI model as covariates in Chapter 4.

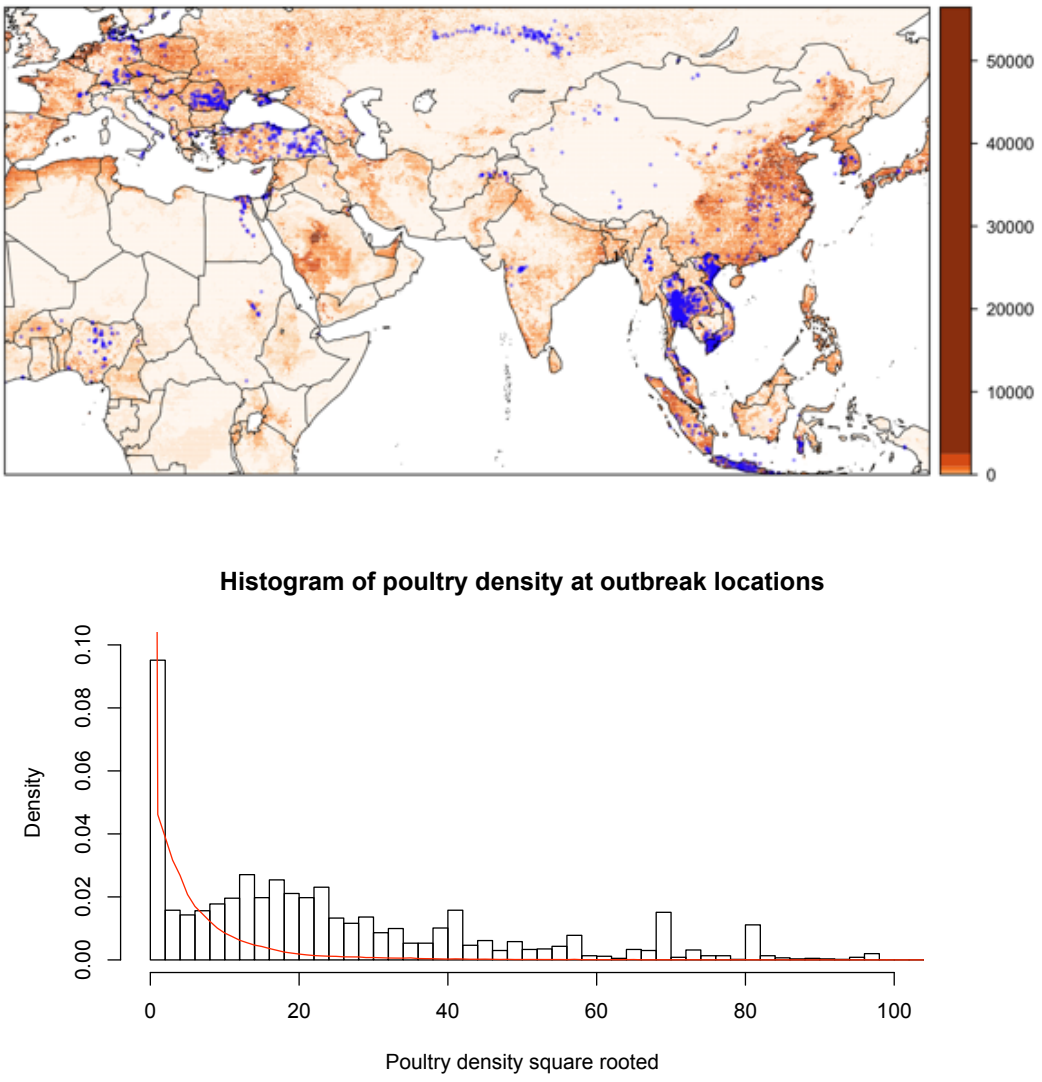

Figure 2.7: The plot on the top is the poultry density map with outbreak locations represented in blue. A unit grid width is approximately 111km, and the values denote the number of domestic poultry in each unit grid. It is plotted with logarithmic scale. The plot on the bottom is a histogram of poultry density at the outbreak locations. The red curve is a density curve of overall poultry density.

Part II

Modeling the global spread of avian influenza with spatio-temporal point process

Chapter 3

Previous Work

Prior to constructing our EAI statistical model with the factors mentioned above in Chapter 4, this chapter discusses past statistical approaches to modeling the global spread of H5N1 and their limitations. There have been numerous attempts to statistically model the spread of H5N1, but many were based on extreme assumptions and incorrectly applied statistical analysis. Among the past studies, Gilbert et al. [22] and Fang et al. [16] employ perhaps the most well thought-out statistical models: spatial logistical regression and SAR (Spatial Autoregression) respectively. The two models are similar; essentially SAR is an advanced spatial logistical regression that incorporates spatial dependencies among the nearby regions. Although these models feature different predictors and study areas, both models aim to produce a predictive risk map for the virus dispersal in an effort to prevent the explosive H5N1 spread.

3.1 Spatial logistic regression

Gilbert et al. [22] study the H5N1 outbreak patterns that occurred from 2004 to 2006 in China. Their analysis using spatial logistic regression strictly considers spatial patterns of Chinese avian influenza, involving 128 outbreaks and 640 randomly sampled locations from uninfected areas.[1]. Even though Gilbert et al. do not explicitly specify their statistical model, it can be defined with i ($1 \leq i \leq 768$), an index for all locations of reported outbreaks and random samples, the associated covariates for location i, X_i, and the corresponding binary response variable, Y_i, indicating whether location i suffered H5N1 infection:

$$logitP(Y_i|X_i) = X_i\beta. \tag{3.1}$$

X_i is a vector consisting of values of covariates at either outbreak or randomly selected location indexed i. The covariates include minimum distance to nearest highway, annual precipitation, and interaction between minimal distance to the nearest lake and wetland. Y_i

[1]The randomly sampled locations are meant to prevent bias introduced to model performance metrics due to sparsity of binary response variable, Y_i, at each location i.

takes two values, 1 and 0; Y_i is 1 if i indicates an outbreak of H5N1 and 0 if i represents a randomly selected location. β is the vector of parameters they wish to estimate. The hypothesis tests that $\beta = 0$ suggests the parameters corresponding to the covariates are statistically significant. In addition to the hypothesis test results, Gilbert et al. [22] assess the model fit using the Hosmer - Lemeshow goodness of fit test [25].

This modeling approach produces a predictive risk map of H5N1 spread, which displays the estimated probability of outbreak incidence. Although no uncertainty measure is given, the resulting map is easy to interpret, and it illustrates the locations of probable future outbreaks according to their model.

While the estimation and prediction results are easy to understand and may be useful, the assumptions of the model are unrealistic. The most critical assumption of this model is that Y_i's, the occurrence of an outbreak, are independent. This assumption ignores the fact that the time and the location of H5N1 outbreaks are likely to be dependent due their highly contagious nature. With this formulation, an outbreak has nothing to do with the chance of observing another outbreak in close proximity both spatially and temporally. In other words, the clusters of outbreaks noted in Chapter 2 occurred independently, according to this model. This is an unrealistic assumption, but prediction results might still be useful.

In addition to the independence assumption, their analysis omits a temporal trend. Gilbert et. al do note the seasonality in Chinese H5N1 outbreaks, but they do not include temporal pattern in their spatial logistic regression model.

3.2 SAR (Spatial Autoregression)

Fang et al. [16] extends Gilbert et al.'s statistical approach by specifying a dependence structure among the outbreaks. They examine H5N1 outbreaks in Southeast Asian countries, Vietnam and Thailand, from years 2004 to 2005, incorporating a SAR (Spatial Autoregression) model [2] into the linear predictors within the logistic model:

$$logitP(Y_i|X_i) = X_i\beta + \gamma \frac{\sum\limits_{j:i\sim j} w_{ij}Y_j}{\sum\limits_{j:i\sim j} w_{ij}}. \tag{3.2}$$

The spatial logistic regression model in equation (3.1) now becomes spatial autologistic regression (SAR) model due to the newly introduced second term. In this model, i is an index for subdistricts in Vietnam and Thailand with H5N1 outbreaks and an equal number of randomly sampled subdistricts that did not suffer from H5N1[2]. An equal number of random samples, in comparison to the number of substricts that suffered H5N1 infection, was chosen to produce receiver operating characteristic (ROC) [49] via bootstrapping, which was used to evaluate the predictive power of their model. X_i again is a vector representing the covariates for subdistrict i. The covariates included in the SAR model are human population size,

[2]The numbers of subdistricts included in their analysis for Thailand and Vietnam are not specified in their paper.

altitude, and numbers of estimated chicken, duck, and mean rice cropping intensity. The binary response variable, Y_i, is 1 for subdistrict i with reported outbreaks, and 0 for randomly sampled subdistricts.

The newly introduced term, $\frac{\sum_{j:i\sim j} w_{ij}Y_j}{\sum_{j:i\sim j} w_{ij}}$, denotes a weighted average of neighboring Y_i's where $j : i \sim j$ indicates indices of subdistricts defined as neighbors of subdistrict i. The weight, w_{ij} is inversely proportional to the distance from subdistrict i to j. The parameter γ for the weighted average captures the strength of dependencies between the neighboring subdistricts.

Additionally, Fang et al. [16] separate their data into three temporal "waves" and their SAR model is fit to each of them. The outcomes from estimating β and γ are compared among the three waves.

Similar to the spatial logistic regression model, the SAR model produces an easily comprehensible predictive risk map of H5N1 spread. This result improves upon the spatial logistic regression by considering the spatial dependencies of the neighboring subdistricts, but there is still more room for improvement.

Spatially, considering the data at a subdistrict level limits the flexibility of the model. Depending on the sizes of the subdistricts, their model may fail to recognize the clusters of outbreaks. Outbreaks occurring within the same subdistrict are grouped together and will only count as 1 in the response variable, Y_i. This formulation neglects the spatial clustering and is unable to describe the most prominent feature of H5N1 outbreaks, clustering along the traffic networks and around cities.

Temporally, their construction does not account for the outbreaks occurring at different times but at the same location. Furthermore, the apparent temporal trend is excluded from their analysis. While Fang et al. acknowledge the need for temporal analysis, they handle the temporal trend by simply separating the outbreaks into three groups of "waves." This approach merely models spatial patterns of H5N1 spread for each wave and fails to incorporate the seasonality and the temporal dependencies discussed earlier in Chapter 2.

Apart from the criticisms stated above, the SAR model assumes that the strength of the spatial dependencies is uniform. In other words, the impact of an outbreak in a subdistrict to its neighbors will be exactly the same as that of an outbreak in another subdistrict. It is hard to believe that the spatial dependency is uniform across the different locations of the outbreaks, even after adjusting for the covariates. There are many factors that could potentially alter the dependencies such as market access, transportation, etc.

Chapter 4

Model specification and estimation

Previously, we have reviewed two past statistical research on H5N1 employing a spatial logistic regression model and its variation, the spatial autologistic regression (SAR) model [22, 16]. Our main criticisms on the two modeling approaches were twofold. First, both approaches neglected analysis on the temporal spread of avian influenza, and only considered the spatial features of the virus dispersal. Fang et al. [16] splits its time window and applies SAR to three waves of H5N1 outbreaks, but such an analysis is hardly sufficient for seasonality and temporal dependencies of H5N1 outbreaks noted in Chapter 2. Second, although Fang et al. introduces an autoregressive term in their model to deal with spatial dependencies among the observed H5N1 outbreaks, both modeling approaches are in need of better dependence structures. The outbreaks in Fang et al. are grouped by their corresponding subdistricts and this arrangement may not be able to represent the clustering in vicinity to cities noted in Chapter 2, defeating the purpose of proposing an autoregressive term.

In this chapter, we provide a statistical framework that will cope with the drawbacks mentioned above. We consider a point process model capable of incorporating the observed temporal pattern and specifying the spatial and temporal dependencies of H5N1 outbreaks with a branching structure. We determine that Ogata's ETAS model [42], an extension of Hawke's self-exciting point process [24], is a suitable candidate. The ETAS model was designed to model the branching structure of earthquakes. An earthquake with a large magnitude triggering aftershocks is analogous to a contagious H5N1 outbreak with favorable conditions causing infection at other locations. Therefore, adapting a point process framework similar to the ETAS model can be beneficial to our analysis. We will further discuss the details of both self-exciting point process and the ETAS models in Section 4.1. Other prior research modeling disease spread pattern using point processes will also be mentioned in Section 4.1.2.

Drawing inspiration from Ogata's ETAS model, we formulate and introduce our EAI (Epidemic Avian Influenza) model in Section 4.2 to study the mechanism of Turkish H5N1 spread. The EAI model aims to quantify the relationship between H5N1 outbreaks and the potential contributing factors discussed earlier in Chapter 2. Furthermore, the fitted EAI model can be used to produce prediction results via Monte Carlo simulation, which will be

helpful in establishing an effective virus quarantine. The prediction results will be presented in Section 6.3.

Following the model specification, the likelihood of the EAI model is presented along with the computational challenges that arise from maximizing the likelihood for estimating the parameters of the EAI model. We consider three estimation methods to answer these challenges: backfitting, "poorman's EM" (Expectation - Maximization), and EM methods. While the backfitting method yields the fastest computational speed, the EM method is known to produce the most accurate parameter estimates [60]. Poorman's EM is an hybrid method of the two, intended to take advantage of the computational speed and the accuracy, but the performance was lower than what we expected, as shown in Chapter 6. The motivation and algorithm for each estimation method will be provided in detail in Section 4.4.

The results—parameter estimates, their SEs (Standard Error), and the corresponding maximized likelihoods—obtained using the backfitting method are compared for five competing variations of the EAI model[1]. The best fitting model is determined based on their AICs (Akaike Information Criterion), a measure of model performance which penalizes the maximized likelihood by the number of parameters. For the best model, surface plots of the likelihood with respect to the parameters around the estimates are inspected to verify that the estimated parameters, indeed, occur at the maximum.

Lastly, we assess the differences among the results of all three methods in terms of how and where their estimates converge. It will be shown later in Chapter 6 that this observed behavior is congruent with the simulation results.

4.1 Modeling ideas based on point process

Although both statistical approaches used in Gilbert et al. [22] and Fang et al. [16] produced useful prediction results, their models were not flexible enough incorporate a temporal component and spatio-temporal dependencies of H5N1 outbreaks. A more natural way to approach the outbreaks, instead of grouping them together according to their corresponding subdistricts as shown in Fang et al., is to treat each of them as an individual event. In this formulation, each outbreak event is allowed to influence others—and cumulatively, the spread of the disease—with its attributes. Temperature and proximity to traffic networks are among the few contributing factors that can potentially impact the virus dispersal, as discussed in Chapter 2.

Point process modeling is a popular framework for analyzing patterns of events in statistics. The heart of a point process model is its intensity, which governs the expected arrival rate of the events. We take this approach to avian influenza data, including both spatial and temporal factors influencing disease spread as components of intensity in order to explain the mechanism of virus dispersal.

In Section 4.1.1, we review a self-exciting point process model [24], which is designed specifically to formulate a branching structure according to the properties of an event. By

[1]Regardless of the method employed, the results from all methods are nearly identical, and they do not influence the process of model selection.

construction, each event in the self-exciting point process is allowed to trigger (or self-excite) additional events in the process. The number of events triggered by each event is varied by its properties and is usually set to decay according to time of its occurrence.

Subsequently, a spatio-temporal extension of Hawke's point process, the ETAS (Epidemic Type Aftershock Sequence) model, is discussed in depth, and we will use the ETAS model as a basis for our EAI (Epidemic Avian Influenza) model proposed in Section 4.2. The ETAS model was proposed to model the mechanism of earthquake and associated aftershocks according to their magnitude. In the model, the branching structure earthquakes are determined by location and time of their occurrences, along with their magnitude. The flexibility of the ETAS model in specifying the triggering mechanism is attractive for our statistical analysis, and it is a framework we can build our EAI model upon to analyze the spread of highly pathogenic H5N1 virus.

As we note in Section 4.1.2, point process models similar to the ETAS model have been used successfully to model disease spread patterns. While most of them focus on quantifying spatio-temporal associations between disease occurrence and the contributing factors [30, 10, 14, 52, 34], some aim to develop a tool for detecting abnormal epidemic dispersal [15] and to detect the source of the endemic disease [32]. The details of these modeling approaches will be provided in the same section.

4.1.1 Self Exciting Point Process

The self-exciting point process was introduced by Hawke [24] to structure temporal dependencies among events. Veen and Schoenberg [60] give a formal and concise definition of Hawke's self-exciting point process; suppose a simple temporal point process is represented by a random count measure, N, on $[0, \infty]$ adapted to filteration, H_t, and define the conditional intensity, $\lambda(t|H_t)$, as the unique, nondecreasing, H-predictable process such that $N([0,t)) - \int \lambda(t|H_t)dt$ is an $\mathbf{H}$-martingale. Then $\mathbf{H}$ must contain the history of the process up to time t, represented as $H_t = \{t_i : t_i < t\}$ with t_i corresponding to the time that event i occurs. As shown in Daley and Vere-Jones [11], it is sufficient to model this point process with $\lambda(t|H_t)$, because the finite-dimensional distributions of such a point process are uniquely determined by its conditional intensity. Hawke's self-exciting point process features an inhomogenous point process with an intensity conditioned on the past history which takes the form:

$$\lambda(t|H_t) = \frac{E[N(dt)|H_t]}{dt} = \mu(t) + \sum_{i:t_i<t} g(t - t_i). \tag{4.1}$$

The conditional intensity of the self-exciting point process, $\lambda(t)$, is the expected value of the number of points—or events—in an infinitesimal time window, $N(dt)$, given the past history, H_t divided by the length of dt. The latter expression, $\mu(t) + \sum_{i:t_i<t} g(t - t_i)$, provides details of the conditional intensity; $\mu(t)$ corresponds to the background intensity at time t, and the cumulative sum $\sum_{i:t_i<t} g(t - t_i)$ represents the contribution of past events occurring prior to time t—denoted $i : t_i < t$. Naturally if $g(t - t_i)$ takes a decaying functional form,

only the recent past events will contribute to the intensity at current time t and the strength of their contribution will die out over time.

By construction, the self-exciting point process has a branching structure. An event will trigger other events and the new generation of events will continue to produce their offspring events according to the conditional intensity stated above. Hawkes' [24] self exciting point process has been successfully applied to wide range of topics from modeling earthquakes to studying theft patterns [42, 35, 48].

ETAS (Epidemic Type Aftershock) model

Arguably, the most influential application of Hawkes' self-exciting point process is the Ogata's Epidemic Type Aftershock Sequence (ETAS) model. Ogata [42] takes the branching structure of self-exciting point process a step further by adding a spatial component to Equation (4.1). With his ETAS model, Ogata seeks to unveil how earthquakes trigger their aftershocks and to quantify the spatial and temporal dependencies. Since its first introduction [42], Ogata has improved the ETAS model over the years; Ogata [43] compares different types of functional forms for composing a triggering structure, and Zhuang et al. [67] develop advanced methods for fitting and validating a semi-parametric version of ETAS model. Stochastic reconstruction of earthquakes using the fitted ETAS model is featured in Zhaung et al. [66][2]. The conditional intensity of Ogata's ETAS model is defined as:

$$\lambda(t, x, y, M|H_t) = \frac{E[N(dt\,dx\,dy\,dM)|H_t]}{dt\,dx\,dy\,dM} \tag{4.2}$$

$$= \mu(x, y) + \sum_{i:t_i<t} k(M_i)g(t - t_i)f(x - x_i, y - y_i|M_i) \tag{4.3}$$

where

x, y, and t represent the x and y coordinates, and time respectively. Additionally M denotes the magnitude of an earthquake occurring at (x, y, t).Therefore, (x_i, y_i, t_i, M_i) denote the spatial location, time, and magnitude of an earthquake event i $(1 \leq i \leq n)$ with n corresponding to the total number of earthquakes considered in Ogata's study. As defined in Section 4.1.1, $N(dt\,dx\,dy\,dM)$ represents the random number of events in an infinitesimal space, $dt\,dx\,dy\,dM$.

$H_t = \{(x_i, y_i, t_i, M_i); t_i < t\}$ is the observational history of the location, time, and magnitude of earthquakes up to time t.

$\mu(x, y)$ represents the spatial variation of earthquakes that are not triggered by other earthquakes.

$k(M)$ is the expected number of events triggered from an event of magnitude

[2]Identifiability issues of Ogata's ETAS model have not been addressed in his research.

M–that is an aftershock–given by $k(M) = Aexp[\alpha(M - M_c)]$. A and α are constants. M_c is a threshold for magnitude. Any magnitude M less than M_c, therefore, would have very little effect in producing aftershocks.

$g(t)$ is the p.d.f of the occurrence times of the triggered events, taking the form $g(t) = \frac{p-1}{c}(1 + \frac{t}{c})$. This is a p.d.f version of Omori law which describes the decay of after shock frequencies with time.

$f(x, y|M)$ is the p.d.f of the locations of the triggered events, which is formulated as $f(x, y; M) = \frac{1}{2\pi De^{\alpha(M-M_c)}} exp\left(-\frac{x^2+y^2}{2De^{\alpha(M-M_c)}}\right)$ [3]

J(M) is probability density for magnitude for all events independent from other components (x, y, and t) and is derived from Gutenberg-Richer law: $J(M) = \beta e^{-\beta(M-M_c)}$.

The intensity of the ETAS model consists of two parts: the intensity describing the background events, $\mu(x, y)$, and that of the triggered events corresponding to the latter summand. The branching structure of this model allows the background events to trigger generations of offspring according to the decaying spatial and temporal dependencies.

In comparison to the spatial logistic regression and SAR models, this framework provides a greater flexibility in terms of specifying the dependencies of the outbreaks. However, flexibility of the model does not equate to a better model, if the functional forms of dependencies among the events are unknown. One of the major challenges in using this model is that it requires a knowledge of how events trigger other events. In Ogata's case study, the functional form of magnitude of an earthquakes and temporal triggering were dictated by the corresponding scientific theories, the Gutenberg-Richer and the Omori laws, respectively. Therefore, determination of the spatial triggering function was the only challenging task left for Ogata.

In our case, very little is known about the branching mechanism of H5N1. In Chapter 2, we have determined few key factors that may affect the virus spread, but the functional form of triggering has not been specified by scientific research. Specification of the branching structure of the disease would be a daunting task.

Even with these challenges, we will adapt this framework for our research. We believe that the self exciting point process model is capable of capturing disease spread in reality better than the models based on spatial logistic regression. It is logical to think that an outbreak of highly contagious disease like H5N1 would provoke another outbreak at a nearby location and in the near future. The strength of its infectivity will vary depending on the factors described earlier in Chapters 1 and 2. The ETAS model can be used as a basis for specifying the spatial and temporal relationship among the outbreaks.

Moreover, using a self exciting point process model provides solutions to the shortcomings of the two modeling approaches. Temporal trend—the seasonality and the variation of disease

[3] They also consider the following function for longer range decay: $f(x, y; M) = \frac{(q-1)D^{(1-q)}e^{\alpha(q-1)(M-M_c)}}{\pi[x^2+y^2+De^{\alpha(M-M_c]^q}}$

presence according to temperature—is easily incorporable to this framework. We can model the seasonality as an overall temporal variation and include temperature as a component in the triggering intensity. The issue of grouping outbreaks to be only counted once for the response variable is automatically solved because each outbreak will be treated as an event.

4.1.2 Point process models analyzing disease spread

To our knowledge, there is no past research on the avian influenza epidemic using point process as its modeling framework. However, there are studies applying point process models to analyze the dynamics of other types of diseases. Most of these models focus on quantifying spatio-temporal associations between disease occurrence and the contributing factors.

Lawson and Leimich [30] were one of the first to utilize point process to explain infectious disease dispersal. With their point process model, they explored the mechanism of measles spread in Hegelloch, Germany. The intensity of their point process model is conditioned on the past history of outbreaks, and it describes the spatio-temporal distribution of the susceptible population in relation to spatial and temporal lag from the infected. Their model is fitted by maximizing its partial likelihood [10] to bypass the complicated calculation of the integral required for the full likelihood[4].

Diggle [14] proposes a model similar to self-exciting point process to analyze the foot and mouth disease spread in the UK. His point process model features a conditional intensity of virus transmission from one farm to another given the past history. This conditional intensity incorporates the number of cows and sheeps at each farm, the distance from one farm to another, and whether a farm was infected by the deadly disease. Like Lawson and Leimich, Diggle [14] also advocates the use of partial likelihood to avoid the computationally expensive likelihood maximization to estimate the parameters of his model. Scheel et al. [52] applies Diggle's [14] modeling approach to infectious salmon anemia data gathered from Atlantic salmon farms.

Meyer et al. [34] provide a point process model resembling the ETAS model. Although their model is used to study outbreak patterns of invasive meningococcal disease, the purpose of their work is to design a general framework for modeling a spread of disease. Much like the ETAS model, the intensity of their model consists of two components to describe endemic and epidemic dispersal, which correspond to background and triggering intensity in Equation (4.3). Unlike other works cited above, Meyer et al. fit their model by directly maximizing the likelihood of point process, as a few of parameters in their model become unidentifiable via profile likelihood maximization under the model specification[5].

While the aformentioned researches focused on modeling the dynamics of epidemic spread, a couple of past statistical approaches based on point process had different goals. Moreover, these approaches do not rely on self-exciting point process as its framework. Diggle et al. [15] develop a tool for detecting abnormal disease spread by modeling the outbreaks of disease with non-stationary a log-Gaussian Cox process[6]. This point process model has an inten-

[4]The computational issue of likelihood maximization will be examined in detail in Chapter 5.

[5]Closed form solutions were not available and were computed via numerical maximization.

[6]An intensity of a Cox point process contains stochastic component.

sity consisting of two deterministic components, describing spatial and temporal patterns of outbreaks spread, and an additional unobserved stochastic component to represent the departures from the normal pattern. To detect abnormality, the authors calculate the probability that a realization of the stochastic component exceeds a certain threshold with the fitted model.

Martinez-Beneito et al. [32] investigate Legionnaire's disease in Alcoi, Scotland to find the source of the respiratory endemic. They are interested in whether ecological factors influence the dynamics of the disease dispersal. Using Ripley's K-function [51], Martinez-Beneito et al. construct a statistical hypothesis test comparing the distribution of cases and controls with random labeling. Based on the estimated intensities for cases and controls via kernel density estimation, the probability of observing a case at a given location is computed and plotted to indicate the variation in outbreak risk.

4.2 Proposed model: EAI (Epidemic Avian Influenza) model

Adapting the framework of the ETAS model discussed in the previous section, we propose our EAI model to analyze the dynamics of avian influenza (H5N1). The intensity of our self-exciting point process conditioned on the past history is a function of space—longitude (x) and latitude (y)—and time (t)[7]:

$$\lambda(x, y, t|H_t) = \frac{E[N(dt\, dx\, dy)|H_t]}{dt\, dx\, dy} \tag{4.4}$$

$$= \lambda_B(x, y, t) + \underbrace{\sum_{i:t_i<t} \alpha\, f(x - x_i, y - y_i)\, g(t - t_i)\, h_{\text{traff}}(x, y)\, k(T(t_i))}_{\lambda_T(x, y, t)} \tag{4.5}$$

$$\tag{4.6}$$

where i is an index for each of the 221 H5N1 outbreaks occurred in Turkey during 182 days between October 1, 2005 to March 31, 2006. Consequently, (x_i, y_i, t_i) represents the location and time of an outbreak i, and $H_t = \{x_i, y_i, t_i; t_i < t\}$ corresponds to the past history of outbreaks up to time, t. The conditional intensity in (4.5) is composed of two parts, $\lambda_B(x, y, t)$ and $\lambda_T(x, y, t)$ whose subscripts B and T stand for background and triggering respectively. The definitions of the components in (4.5) are provided below:

$\lambda_{\mathbf{B}}(\mathbf{x}, \mathbf{y}, \mathbf{t}) = ae^{-bR_{\text{city}}(x,y)}e^{-kT(t)}$ represents the background intensity and has two components—each of them corresponding to spatial and temporal patterns of outbreaks respectively. It is designed to describe the recurring outbreaks displaying seasonality and clustering near infrastructures such as major cities. The

[7]Note that only the general framework is provided here. For the list of models we considered, refer to Section 4.5.1

spatial aggregation is captured in $\lambda_B(x, y, t)$ with $R_{\text{city}}(x, y)$, which measures the closest distance from an outbreak location (x, y) to a major city. Moreover, the noted seasonality is expressed as a function of temperature $T(t)$ at time, t. a is a scaling parameter.

$\lambda_{\mathbf{T}}(\mathbf{x}, \mathbf{y}, \mathbf{t})$ denotes the intensity corresponding to triggered events from the background or other triggered events. It consists of several components:

α is a scaling parameter for the triggering process analogous to a in $\lambda_B(x, y, t)$. To avoid possible identifiability issues, we chose to have one scaling parameter for all functional forms in the triggering process.

$f(x - x_i, y - y_i) = e^{-\beta\sqrt{(x-x_i)^2+(y-y_i)^2}}$ describes the decay in intensity of triggered outbreaks according the spatial lag—a distance between locations of an outbreak and a past outbreak.

$g(t) = e^{-\gamma t}$ is a temporal version of $f(x - x_i, y - y_i)$, depending on the temporal lag between the occurrence times of an outbreak and a past outbreak.

$h_{\text{traff}}(x, y) = e^{-\beta_{road} R_{\text{traff}}(x,y)}$ corresponds to inverse relation between the minimum distance to traffic networks (major highways and railroads) and the chance of observing an outbreak. $R_{\text{traff}}(x, y)$ denotes the distance from (x, y) to the nearest traffic networks.

$k(T(t_i)) = e^{-\kappa(T(t_i))}$ is the triggering strength of past event varied by temperature at its occurrence.

The two parts of the intensity, $\lambda_B(x, y, t)$ and $\lambda_T(x, y, t)$, can be thought of as endemic and epidemic spread of disease, respectively. While $\lambda_B(x, y, t)$ describes the recurring outbreaks depending on the temperature and proximity to cities, $\lambda_T(x, y, t)$ corresponds to outbreaks that were triggered by other outbreaks, with their numbers varied by location and time of their parents, proximity to traffic networks, and temperature.

By construction, this model assumes the following: first, the times and the locations of the recurring outbreaks depend on the seasonality and proximity to the nearest city. Second, the duration of the presence of the H5N1 virus–which boosts infectivity—differs by the temperature. Third, the chance of observing an outbreak after an outbreak is inversely related to the both temporal and spatial lag. Forth, an outbreak is likely to be triggered near the traffic networks.

The presented model is the most complex EAI model and corresponds to Model (5) in Section 4.5.1. The estimation results for Model (5) and four simpler versions of EAI models featuring different combinations of components will be examined in the same section.

For the sake of simplicity, we chose all functional forms for components in intensity 4.5

to be exponential. We are much more interested in estimating the rate of decay than finding the perfect functional form for the components in the triggering process.

4.3 Maximum Likelihood Estimation (MLE)

To estimate the parameters of the EAI model, we employ a classical method in statistics, maximum likelihood estimation (MLE)[8]. Previously, Rathbun [50] has shown that the standard large sample theory of MLE holds under regulatory conditions for self-exciting point processes. A similar result for general point processes was provided in Schoenberg [55] with simpler assumptions[9]. The log likelihood function of our point process can be written as [11]:

$$logL(\theta) = \sum_{i=1}^{n} log\lambda_\theta(x_i, y_i, t_i|H_{t_i}) - \int_0^T \int\int_S \lambda_\theta(x, y, t|H_t)dxdydt, \tag{4.7}$$

where θ denotes the set of parameters, $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma, \kappa)$, in our model, and $i\,(1 \leq i \leq n)$ is an index for each event (outbreak) occurring in region S and time interval $[0, T]$. For our study, S represents Turkey, and T corresponds to 182 days it has suffered from H5N1 dispersal. The number of total H5N1 outbreaks, denoted n, in this time interval is 221. As before, the set, (x_i, y_i, t_i), indicates the occurrence time and geodesic coordinates—in terms of longitude (x) and latitude (y)—of an outbreak event i. $H_{t_i} = \{x_j, y_j, t_j; t_j < t_i\}$ denotes the past history of events that occurred before event i. $\lambda_\theta(x_i, y_i, t_i|H_{t_i})$ is the intensity function defined in Equation (4.5).

Since the closed form solutions for maximum likelihood are not available, the likelihood (4.7) must be numerically maximized. The calculation of the first term in Equation (4.7), the sum of the log intensity evaluated at each event, is quite simple. The calculation of the integral part, however, is complicated and computationally expensive. Diggle [14] advocates using partial likelihood instead of the full likelihood (4.7), which bypasses the calculation of the integral term. Cox [10] shows that estimates obtained by maximizing the partial likelihood inherit the general asymptotic properties of maximum likelihood estimators with a possible loss of efficiency. He also notes that the parameters in the original model may become unidentifiable in the partial likelihood. To avoid potential identifiability issues, we will use the full likelihood in our estimation and calculate the integral in Equation (4.7). Moreover, calculating the integral will be advantageous for the construction of the EM method and simulation of our point process.

A brief derivation and an outline for the numerical integration are now presented. Naturally, we can split the integral (4.8) into two parts with each integrand corresponding to background and triggering intensities respectively:

[8]Technically, only the backfitting method, one of the three estimation methods proposed in section 4.4, does MLE. The other two methods, poorman's EM and EM methods, target the likelihood 4.7 but do not maximize it directly.

[9]In general, the large sample theory of MLE is assumed for point process models.

$$\int_0^T \int\int_S \lambda_\theta(x,y,t|H_t)dxdydt = \int_0^T \int\int_S \lambda_B(x,y,t)dxdydt+$$
$$\int_0^T \int\int_S \sum_{i:t_i<t} \alpha\, f(x-x_i, y-y_i)\, g(t-t_i)\, h_{\text{traff}}(x,y)\, k(T(t_i))dxdydt \tag{4.8}$$

The first integral, the integral for the background intensity, is relatively easy to compute[10]. Therefore, we focus on the derivation of the second integral. The second integral can be separated further into temporal and spatial components:

$$\int_0^T \int\int_S \sum_{i:t_i<t} \alpha\, f(x-x_i, y-y_i)\, g(t-t_i)\, h_{\text{traff}}(x,y)\, k(T(t_i))\, dxdydt \tag{4.9}$$
$$= \alpha \int_0^T \sum_{i:t_i<t} k(T(t_i))\, g(t-t_i)\, dt \int\int_S f(x-x_i, y-y_i)\, h_{\text{traff}}(x,y)\, dxdy. \tag{4.10}$$

The above integral can be rearranged into:

$$= \alpha \sum_{i=1}^{n} \int_{t_i}^{T} k(T(t_i))\, g(t-t_i)\, dt \int\int_S f(x-x_i, y-y_i)\, h_{\text{traff}}(x,y)\, dxdy \tag{4.11}$$

and the derivation will be shown in the appendix.

The analytic computation of the integral (4.11) is nearly impossible, and consequently, it is numerically approximated by discretizing S. The time interval $[0,T]$ is recorded in days, therefore discretization is unnecessary. Since the earth is an ellipsoid, creating a spatial grid for region, S, using the raw longitude and latitude coordinates will not yield a consistent grid area. Therefore, we consider a projection that preserves area [57]. The resulting grid from area preserving projection will provide us a grid with equal areas, simplifying our calculations. We apply Mollweide projection to S and a demonstration of this projection is shown in Appendix A.1.

For integration of the spatial component in (4.11), the likelihood function is evaluated at the center of each grid and multiplied with the corresponding grid area[11]. We let g $(1 \le g \le G)$ indicate an index for each square in the grid; x_g and y_g denote the latitude and longitude at the center of a square g[12]. The integration of temporal component simplifies to a sum of the likelihood function evaluated at each day because the increment, dt, is one day. The numerical integration of (4.11) is

[10]Using notation that will be defined in the following paragraphs, the expression for numerical integration of $\int_0^T \int\int_S \lambda_B(x,y,t)dxdydt$ is $\sum_{t=1}^{T}\sum_{g=1}^{G} ae^{-bR_{\text{city}}(x_g,y_g)}e^{-kT(t)}$.

[11]This is equivalent to the mid point rule.

[12]The grid was constructed by dividing the range of Turkey in latitude to 200 segments. The range of Turkey in longitude was divided according to the resulting distances of the segments

$$\alpha \sum_{i=1}^{n} \sum_{t_i}^{T} k(T(t_i))\, g(t - t_i) \times \sum_{g=1}^{G} f(x_g - x_i, y_g - y_i)\, h_{\text{traff}}(x, y) \cdot \Delta_g. \tag{4.12}$$

Consequently, the numerical approximation of the log likelihood 4.7 becomes

$$\begin{aligned} logL(\theta) = &\sum_{i=1}^{n} log\lambda_\theta(x_i, y_i, t_i | H_{t_i}) - \sum_{t=1}^{n} \sum_{g=1}^{G} \lambda_B(x_g, y_g, t) \cdot \Delta_g \\ &- \alpha \sum_{i=1}^{n} \sum_{t_i}^{T} k(T(t_i))\, g(t - t_i) \times \sum_{g=1}^{G} f(x_g - x_i, y_g - y_i)\, h_{\text{traff}}(x, y) \cdot \Delta_g. \end{aligned} \tag{4.13}$$

This numerical approximation of log likelihood will be implemented in all three parameter estimation methods presented in the next section.

4.4 Parameter estimation methods for the EAI model

As Veen and Schoenberg [60] note, estimating parameters via MLE for self-exciting point process can be a daunting task. In practice, its log likelihood function is often flat near the maximum and multimodal[13]. Therefore, arbitrary starting values using conventional numerical optimization routines may lead to divergence.

To obtain accurate estimates for the parameters of EAI model, we introduce and implement three different estimation methods: backfitting, "poorman's EM" (Expectation Maximization), and EM. The idea behind the backfitting method—maximizing the likelihood in terms of a set of parameters while holding the others constant—will serve as a backbone for the other two. Poorman's EM method was intended to take advantages of the computational speed and accuracy of backfitting and EM methods respectively[14]. Poorman's EM is similar to the EM method in that it involves calculation of the probability of each event belonging to the background process. However, while the poorman's EM uses this probability to classify background events at each iteration, the EM method does not incorporate classification in its routine. The EM method is a popular method in estimating parameters for models with an unobserved latent variable. In our case, the unobserved latent variable is the branching structure of the EAI model, the information on how each event triggers other events. The likelihood involved in the EM method, called *complete data likeilhood*, incorporates this branching structure probabilistically and often is easier to maximize than that of MLE for parameter estimation [60]. The EM method estimates the parameters iterating between two

[13]The log likelihood used in Veen and Schoenberg [60] is a full likelihood including the integral shown in Equation 4.7.

[14]Despite our intention, the performance benchmark presented in Chapter 6 shows the Poorman's EM method does not improve accuracy of the estimates over that of backfitting method and it is also slower in computational speed.

steps: the E step where the method computes the probability of an event triggering another event, and the M step where the probability is integrated as a part of the maximization procedure.

The details of the three algorithms are provided in the following sections. Further, the estimates obtained using all three method for the best variation of EAI model, in terms of AIC (Akaike Information Criterion), is presented in Section 4.5. A performance benchmark of the three estimation methods using simulated results will be given in Chapter 6.

For all three methods, a quasi-Newton optimization routine, L-BFGS-B (Limited memory-BFGS-Bounded) [8] will be employed to ensure all resulting estimates of the parameters are positive[15]. The starting values for all methods were determined by the first step of Algorithm 2 in Section 4.4.2, and they are $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma, \kappa) = (1, 40, 0.1, 32, 40, 70, 0.1, 0.15)$.

4.4.1 Backfitting method

The aforementioned optimization issue concerning the flatness of the likelihood can be aggravated with an increased number of parameters. To cope with such a problem, we borrow an idea from the maximization procedure introduced by Breiman and Friedman [4], and popularized as an estimation procedure for GAM (Generalized Additive Model) called backfitting algorithm [23]. Backfitting algorithm is a rather simple iterative procedure. A likelihood gets maximized with respect to a set of parameters first holding the others constant. The algorithm then maximizes the likelihood with respect to the ones held constant previously while holding the first set of parameters constant. This procedure is repeated until the value of the likelihood converges.

It is natural to consider the backfitting algorithm to obtain the MLE for our model since the intensity function is conveniently divided into two parts: background and triggering processes. With θ_b and θ_t representing the sets of background and triggered parameters respectively, we employ the following algorithm to estimate the parameters[16]:

Algorithm 1: Backfitting.

Step 1. Set $k = 1$. Choose initial values for both θ_B and θ_T. As mentioned in the previous section, let $\theta_B = (1, 40, 0.1)$ and $\theta_T = (32, 40, 70, 0.1, 0.15)$. Denote them $\hat{\theta}_B^{(1)}$ and $\hat{\theta}_T^{(1)}$ respectively.
Step 2. Compute the log likelihood (4.7) numerically using Equation (4.13) with the initial values, $l(\hat{\theta}_B^{(1)}, \hat{\theta}_T^{(1)})$.
Step 3. Obtain estimates for θ_B via MLE with L-BFGS-B routine holding $\theta_T = \hat{\theta}_t^{(k)}$. Denote these estimates $\hat{\theta}_B^{(k+1)}$
Step 4. Obtain estimates for θ_T via MLE using L-BFGS-B routine with the

[15] In addition, we rescale the unit of geodesic distance used in our calculation from km to $\frac{1}{1000}$km to avoid estimates of scaling parameters a and α lying near the bound, 0.

[16] For example, the parameter sets for model featured in section 4.4 are $\theta_B = (a, b_{\text{city}}, k)$, $\theta_T = (\alpha, \beta, \beta_{\text{road}}, \gamma, \kappa)$.

estimated parameters from step 4, $\hat{\theta}_B^{(k+1)}$, holding them constant. Denote the updated estimates $\hat{\theta}_T^{(k+1)}$

Step 5. Set $k = k + 1$

Step 6. Repeat steps 3, 4, and 5 until the difference between the log likelihood values with estimated parameters from $(k-1)^{th}$ and k^{th} iterations become smaller than 10^{-3}.

The resulting parameter estimates from this algorithm for five EAI models will be presented in Section 4.5 along with the corresponding SEs and AICs.

4.4.2 Poorman's EM method

The poorman's EM method is intended to be a hybrid of backfitting and EM methods, taking advantages of the computational speed and accuracy of the two methods respectively. The estimation results in Veen and Schoenberg [60] show that the EM method yields more accurate parameter estimates than MLE employing a conventional maximization routine, Newton–Raphson. They suggest that the improvement in accuracy may be due to the limited number of observations, as most of the theoretical results relating to maximum likelihood only hold asymptotically. On the contrary, the number of observations for the EM method could be enough to produce accurate results because it incorporates the information on the branching structure at the E-step. However, the E-step is responsible for the slow computational speed of the EM method. The calculation of probabilities of an event triggering another event at each iteration requires longer computational time in comparison to the backfitting method.

The poorman's EM method aims to simplify this probability calculation to gain computational efficiency. It utilizes the probability of an event being a background event calculated from parameters estimated in each iteration. This probability can be obtained from the following formula using the intensity function of our model:

$$P(u_i = 0) = \frac{\lambda_B(x_i, y_i, t_i)}{\lambda(x_i, y_i, t_i | H_{t_i})} \tag{4.14}$$

where u_i is an unobserved quantity such that $u_i = 0$ indicates the event i belonging to the background, and its probability is simply the ratio between its background and the overall intensities. Each event i will be chosen as a background event if its estimated probability, $P(u_i = 0)$, is greater than $\frac{1}{2}$[17].

Unlike the backfitting algorithm, the estimation of background parameters, θ_B, will involve only the chosen background events. Thus the MLE of θ_B is estimated from the likelihood of inhomogeneous point process with only the background intensity, $\lambda_B(x, y, t)$:

$$l(\theta_B) = \sum_{i;u_i=0} log\lambda_{B_{\theta_B}}(x_i, y_i, t_i) - \int_0^T \int\int_S \lambda_{B_{\theta_B}}(x, y, t) dxdydt. \tag{4.15}$$

[17]We thought that experimenting with other choices of threshold may be too data specific. We want to be build a general framework that can be readily applied to H5N1 occurrences in other nations.

The estimated values of θ_T are then produced via MLE with L-BFGS-B maximization routine using the likelihood (4.7). The procedure of updating $P(u_i = 0)$ and classifying background events is repeated until the values of the current and the precedent log likelihoods converge.

The poorman's EM algorithm is explained in detail below:

Algorithm 2: Poorman's EM.

Step 1. Choose a subset of events to be initially categorized as background events. We chose an event to be a background event if more than 3 following events occurred within 100 kilometers and a week.
Step 2. Set $k = 1$. Select initial values for both θ_B and θ_T. As before, let $\theta_B = (1, 40, 0.1)$ and $\theta_T = (32, 40, 70, 0.1, 0.15)$. Denote them $\hat{\theta}_B^{(1)}$ and $\hat{\theta}_T^{(1)}$ respectively.
Step 3. Maximize the log likelihood with background intensity, $\lambda_B(x, y, t)$, shown in Equation (4.15) with respect to θ_B only using selected background events. Use L-BFGS-B as the maximization routine. Denote the new estimates for background parameters, $\hat{\theta}_B^{(k+1)}$.
Step 4. Maximize the log likelihood with both both background and triggering intensity (4.7) with respect to θ_t holding the estimates of $\theta_B = \hat{\theta}_B^{(k+1)}$. Use L-BFGS-B as the maximization routine. Denote the new estimates for triggering parameters, $\hat{\theta}_T^{(k+1)}$.
Step 5. With the updated estimates, calculate the $P(u_i = 0)$ for each $1 < i < n$, employing Equation 4.14. Classify as background event if its estimated probability is greater $1/2$.
Step 6. If the difference between the log likelihood values, $l(\theta_B^{(k+1)}, \theta_T^{(k+1)})$ and $l(\theta_B^{(k)}, \theta_T^{(k)})$ become smaller than 10^{-3}, stop. Otherwise set $k = k + 1$ and repeat steps 3, 4, 5, and 6.

The results obtained from applying poorman's EM method will be compared to those of other two methods in Section 4.5.3. Further comparison of the performance using simulation among the three methods will be given in Chapter 6.

4.4.3 Expectation-Maximization (EM) algorithm

The Expectation-Maximization (EM) developed by Dempster et al [12] is one of the widely used techniques to cope with probabilistic models that depend on latent variables. The idea is to average over the likelihood with respect to current estimates of the distribution for the latent variables and iteratively maximize this average to obtain better and updated estimates for the parameters. The EM algorithm has two steps: the E step to take an expectation of a complete data log likelihood—a log likelihood for the data assuming the latent variables are known—and the M step which maximizes the likelihood from E step and yields updated estimates.

Veen and Schoenberg [60] apply the EM algorithm to estimate parameters for an ETAS (Epidemic Type Aftershock Sequence) model congruent with Ogata [43]. Under the model and also in reality, it is unknown whether an event is a background or a triggered event. While the determination is impossible, we can indeed calculate the probability of an event triggering another event and integrate this information into our EM routine.

Consider u_i, an unobserved quantity defined in Section 4.4.2, and let $u_i = j$ additionally indicate whether event i was triggered by event j for $1 \leq i < j \leq n$. As earlier, $u_i = 0$ denotes whether event i belongs to the background. With this formulation, u_i will define the branching structure of our model, indicating whether event i is a background event or was triggered by another event j. The triggering probability $P(u_i = j)$ is

$$P(u_i = j) = \frac{\lambda_T(t_i - t_j, x_i - x_j, y_i - y_j)}{\lambda(t_i, x_i, y_i | H_{t_i})}. \tag{4.16}$$

This probability is the essential component in the E step of EM algorithm which will be explained in detail in the following section.

The E (Expectation) step

Assuming the complete branching structure and the u_i's are known for all i's, we can construct the complete data log likelihood for our model[18]:

$$l_c(\theta) = log(n_B!) - \Lambda_B + n_b log(\Lambda_B) + \sum_{i;u_i=0} log(\lambda_B(x_i, y_i, t_i)) - log(\Lambda_B) + \tag{4.17}$$

$$\sum_i log(n_{T_i}!) - \Lambda_{T_i} + n_{T_i} log(\Lambda_{T_i}) + \sum_{i;u_i \neq 0} log(\lambda_{T_i}(x_i - x_{u_i}, y_i - y_{u_i}, t_i - t_{u_i})) - log(\Lambda_{T_i}), \tag{4.18}$$

where n_B and n_{T_i} denote the number of background events and the number of events triggered by event i respectively. $\Lambda_B = \int\int\int \mu(x, y, t) dx dy dt$ is the expected number of inhomogeneous background point process. Similarly Λ_{T_i} is the expected number of events triggered by event i obtained from evaluating the integral of $\lambda_{T_i}(\cdot)$ over the given space and time:

$$\Lambda_{T_i} = \int_{t_i}^{T} \int\int_S \alpha\, f(x - x_i, y - y_i)\, g(t - t_i)\, h_{\text{traff}}(x, y)\, k(T(t_i)) dx dy dt. \tag{4.19}$$

The sums $\sum_{i;u_i=0} log(\lambda_B(x_i, y_i, t_i)) - log(\Lambda_B)$ and $\sum_{i;u_i \neq 0} log(\lambda_{T_i}(x_i, y_i, t_i)) - log(\Lambda_{T_i})$ are conditional log likelihood of spatial and temporal distribution of events given the number of background and triggered events respectively.

[18] The complete data likelihood for an inhomogeneous point process is $P(x_1, ..., x_n, N(B) = n) = P(N(B) = n) P(x_1, ..., x_n | N(B) = n) = \frac{\Lambda^n}{n!} e^{-\Lambda} \prod_i \frac{\lambda(x_i, y_i, t_i)}{\Lambda}$

Taking the expectation over the unobserved quantity, u_i, we have the following expected complete data likelihood:

$$E[l_c(\theta)] = log(n_B!) - \Lambda_B + n_B log(\Lambda_B) + \tag{4.20}$$

$$\sum_i P(u_i = 0) \left[log(\lambda_B(x_i, y_i, t_i)) - log(\Lambda_B) \right] + \tag{4.21}$$

$$\sum_i log(n_{T_i}!) - \Lambda_{T_i} + n_{T_i} log(\Lambda_{T_i}) + \tag{4.22}$$

$$\sum_{i \geq 2} \sum_{j=1}^{i-1} P(u_i = j) \left[log(\lambda_{T_i}(x_j - x_i, y_j - y_i, t_j - t_i)) - log(\Lambda_{T_i}) \right]. \tag{4.23}$$

At the beginning of each iteration, the values of triggering probabilities given in 4.16 are updated with the estimates obtained from the previous iteration. With the updated triggering probabilities, we can estimate the expected number of background events, n_B, and expected number of events triggered directly from event i, n_{T_i}. At k^{th} iteration, estimates of n_B and n_{T_i} are

$$\hat{n}_{T_i}^{(k)} = \sum_{j > i} P(u_j = i) \tag{4.24}$$

$$\hat{n}_B^{(k)} = \sum_i P(u_i = 0) = n - \sum \hat{n}_{T_i}^{(k)}. \tag{4.25}$$

Incorporating $\hat{n}_{T_i}^{(k)}$ and $\hat{n}_B^{(k)}$, the expected complete data likelihood at k^{th} iteration becomes:

$$E_{\hat{\theta}^{(k)}}[l_c(\theta)] = log(\hat{n}_B^{(k)}!) - \Lambda_B + \hat{n}_B^{(k)} log(\Lambda_B) + \tag{4.26}$$

$$\left[\sum_i P^{(k)}(u_i = 0) log(\lambda_B(x_i, y_i, t_i)) \right] - \hat{n}_B^{(k)} log(\Lambda_B) + \tag{4.27}$$

$$\sum_i log(\hat{n}_{T_i}^{(k)}!) - \Lambda_{T_i} + \hat{n}_{T_i}^{(k)} log(\Lambda_{T_i}) + \tag{4.28}$$

$$\left[\sum_{i \geq 2} \sum_{j=1}^{i-1} P(u_i = j) log(\lambda_{T_i}(x_j - x_i, y_j - y_i, t_j - t_i)) \right] - \sum_i \hat{n}_{T_i}^{(k)} log(\Lambda_{T_i}). \tag{4.29}$$

Canceling $\hat{n}_B^{(k)} log(\Lambda_B)$ and $\sum_i \hat{n}_{T_i}^{(k)} log(\Lambda_{T_i})$ and noting $log(\hat{n}_B^{(k)}!)$ and $log(\hat{n}_{T_i}^{(k)}!)$ are constants in the M-step, essentially maximization of $E_{\hat{\theta}^{(k)}}[l_c(\theta)]$ further simplifies to:

$$\sum_i P^{(k)}(u_i = 0) log(\lambda_B(x_i, y_i, t_i)) - \Lambda_B + \tag{4.30}$$

$$\sum_{i \geq 2} \sum_{j=1}^{i-1} P^{(k)}(u_i = j) log(\lambda_{T_i}(x_j - x_i, y_j - y_i, t_j - t_i)) - \Lambda_T. \tag{4.31}$$

As expected, the resulting likelihood function resembles the original conditional likelihood 4.7 we wish to maximize. The only difference is that the estimated background and triggering probabilities for each event are multiplied to the intensity evaluated at the corresponding event. The sum of intensities in (4.7) are now weighted with the associated background and triggering probabilities.

The M (Maximization) step

At the M-step, the likelihood computed in the previous section will be maximized using the backfitting procedure described in Section 4.4.1 with the same starting values. We first maximize the likelihood with respect to the background parameters first holding the triggering parameters constant. Then the triggering parameters are estimated in the same fashion. This iterative procedure will continue until the difference between the values of the likelihood becomes smaller than 10^{-3}.

Combining the E and the M steps, the algorithm for the EM method is outlined below:

Algorithm 3: Expectation - Maximization.

Step 1. Set $k = 1$. Choose initial values for both θ_b and θ_t. As before, let $\theta_B = (1, 40, 0.1)$ and $\theta_T = (32, 40, 70, 0.1, 0.15)$. Denote them $\hat{\theta}_B^{(1)}$ and $\hat{\theta}_T^{(1)}$ respectively.
Step 2. E-step Compute the background and triggering probabilities, $P^{(k)}(u_i = 0)$ and $P^{(k)}(u_i = j)$ for all events.
Step 3. M-step Obtain $\hat{\theta}_B^{(k)}$ by maximizing the likelihood (4.31) with L-BFGS-B holding $\theta_T = \hat{\theta}_T^{(k)}$.
Step 4. Obtain estimates for $\hat{\theta}_T^{(k)}$ by maximizing the likelihood (4.31) with L-BFGS-B holding $\theta_B = \hat{\theta}_B^{(k)}$.
Step 5. Set $k = k + 1$
Step 6. Repeat steps 2, 3, 4, and 5 until the difference between the log likelihood values with estimated parameters from $(k-1)^{\text{th}}$ and k^{th} iterations becomes smaller than 10^{-3}.

The computed EM estimates for the best EAI model are compared with the results from other estimation methods in Section 4.5.3. Using simulation, performance of the EM method in terms of accuracy and robustness against starting values will be examined in Chapter 6.

4.5 Result and model comparison

Finally, we present the results from employing the three estimation methods. We first compare five competing EAI models using the Akaike Information Criterion (AIC) and determine the most suitable model [1]. After the comparison, we assess surface plots of the log likelihood (4.7) near the parameter estimates and examine whether the estimation procedure can potentially suffer from its flatness. Lastly, we compare the estimates and the values of log likelihoods calculated by the three maximization algorithms for our best model.

4.5.1 Comparison among the five proposed models

The Akaike Information Criterion proposed by Akaike [1] is an estimated measure of prediction error for an estimated statistical model. The AIC provides a useful guideline in comparing statistical models, due to its form, which penalizes a likelihood of a model with the number of estimated parameters. While more complex models surely yield higher values of maximum likelihood, the AIC compensates the simpler models by penalizing the complex models with the increased number of parameters. The AIC is defined as:

$$AIC^{(m)} = 2\rho - 2logL(\hat{\theta}^{(m)}_{ML}) \tag{4.32}$$

where ρ is the number of parameters estimated and m indicates the AIC score for Model (m), $(1 \leq m \leq 5)$. The second term in (4.32) denotes log likelihood of the fitted model at $\hat{\theta}_{ML}$, an estimate of θ obtained via MLE under Model (m). The statistical model with lower AIC is preferred. The AIC is a common in statistical model comparison and it was previously employed by Ogata [42] to compare his ETAS models.

Table 4.1 presents the parameter estimates and their AICs for five proposed models obtained using backfitting method[19]. The definition for the most complex model, Model (5), was presented in Section 4.2. The Models (1), (2), (3), and (4) are simpler versions of Model (5), featuring different combinations of Model (5)'s components corresponding to the parameters listed in Table 4.1. As mentioned earlier, all functional forms of components in the EAI model are exponential.

The first two models, Models (1) and (2), assume that all events, H5N1 outbreaks in Turkey, were purely background events. In other words, under their configurations, the outbreaks would not trigger other outbreaks. The difference between the two models is the spatial component in the background process. While Model (1) relies on the proximity to traffic network, Model (2) depends on the distance to the nearest cities. These are the simplest models out of the five as they do not possess the branching structure of the other three, Models (3), (4), and (5). Model (3) is an extension of Model (2) with basic triggering structure determined only by the spatial and temporal lags from an outbreak to a past outbreak. Both Models (4) and (5) have an additional component, proximity to traffic

[19]The estimates produced by the EM methods were almost identical to those of backfitting method.

	Background			Triggering				
Models	Scale	Spatial	Temporal	Scale	Spatial lag	Temporal lag	Temperature	AIC
1	$\hat{a} = 3.33$	$\hat{b}_{road} = 52.3$	$\hat{k} = 0.113$	-	-	-	-	103
2	$\hat{a} = 4.74$	$\hat{b}_{city} = 24.7$	$\hat{k} = 0.113$	-	-	-	-	102
3	$\hat{a} = 4.74$	$\hat{b}_{city} = 24.7$	$\hat{k} = 0.113$	$\hat{\alpha} \approx 0$	$\hat{\beta} = 49$	$\hat{\gamma} = 61$	-	108
4	$\hat{a} = 1.21$	$\hat{b}_{city} = 19.2$	$\hat{k} = 0.0873$	$\hat{\alpha} = 72$	$\hat{\beta} = 42.4$ $\hat{\beta}_{road} = 65.3$	$\hat{\gamma} = 0.159$	-	-346
5	$\hat{a} = 1.21$	$\hat{b}_{city} = 19.2$	$\hat{k} = 0.0872$	$\hat{\alpha} = 71.9$	$\hat{\beta} = 42.3$ $\hat{\beta}_{road} = 65.3$	$\hat{\gamma} = 0.159$	$\hat{\kappa} \approx 0$	-344

Table 4.1: Comparison of models in terms of their estimated parameters and AICs

networks, embedded in the triggering process. However, in Model (5), the strength of an outbreak triggering another outbreak is varied by the temperature[20].

Surprisingly, Model (2) had a slightly better AIC score than Model (1), albeit the spatial distribution of outbreaks were found to be much closer to the traffic networks than to the cities in Chapter 2. Although the better fit might be under noise level, Model (2) could have benefitted from the tight clusters of outbreaks found near the cities—especially Samsun and Rize. Model (3), an extension of Model (2) with basic triggering process, performs worse than both Model (1) and (2) in terms of their AICs. The scale parameter for triggering process, α, is estimated to be 0 and, therefore, produces the same maximum log likelihood for Model (3) in comparison to Model (2). The AIC of model 3 is penalized by the three additional parameters estimated. This shows that simply including a triggering process dictated by the spatial and temporal lags does not improve the fit of the model.

The AIC score substantially improves with Model (4) as it introduces proximity to the traffic networks to the triggering process. This result suggests that the locations of triggered outbreaks are jointly determined by the proximity to other outbreaks and to the nearest traffic network. Model (5) incorporates the temperature as one of the components that affects the strength of triggering ability of an outbreak. By including the temperature in the triggering process, we wished to examine whether the relationship between persistence of H5N1 and temperature, noted in Chapter 1, has any impact on the branching structure of the disease. Similar to the scale parameter in Model (3), the additional parameter, κ, in Model (5) was estimated to be approximately 0. Although the value of $\hat{\kappa}$ may suggest that the temperature does not alter the triggering ability of an outbreak, it is also possible that the model is not able to estimate the parameter properly due to its construction. In chapter 6, we will examine the estimation issue for k via simulation.

Of the five proposed models, Model (4) was found to be the most suitable with the smallest AIC. Model (5) produced a similar AIC score, but it is is essentially Model (4) with estimated parameter for κ approximately equal to 0. We will focus on this model throughout

[20]By including temperature component in the triggering process, we would like to see if the epidemic of H5N1 is affected by the variation of temperature in addition to the seasonality modeled in the background process.

the rest of this thesis.

4.5.2 Surface plots of the likelihood around the estimates

Often models based on self exciting point process will suffer from flat and multimodal likelihood functions. It has been noted in the past that the scale parameter of the triggering process, in our case, α, is notoriously difficult to estimate correctly [60, 44]. The likelihood with respect to this scaling parameter is commonly found to be flat. Therefore, this section inspects the surface of likelihood function for our best model, Model (4), in order to verify whether the obtained estimates occur at the maximum. These surface plots will also give us an idea of which parameters would be hard to estimate via MLE for our EAI model.

Figures 4.1 and 4.2 illustrate the surfaces of the log likelihood 4.7 for Model (4) with different combinations of parameters around their estimates obtained using the backfitting algorithm[21]. All surface plots were produced by varying one or two parameters from their estimated values while holding the other parameters at their estimates. All deviations from the estimates were scaled relative to themselves for a purposes of visualization.

The surface plots in Figure 4.1 for the background parameters generally look promising. The first plot in the top left corner is the one dimensional surface plot for all three background parameters, $\theta_B = (a, b_{city}, c)$. Among the three parameters, the parameter for the temperature component in the background process, k, had the flattest surface, followed by b_{city}. The rest of the surface plots show the behavior of the log likelihood when two parameters are jointly varied around the MLEs. With an exception of surface plot generated for parameters a and b_{city}, the other two dimensional surface plots demonstrate that maximums of parameters are moderately well defined. The surface plot of a and b_{city} suggests obtaining MLE with a numerical optimization routine may encounter difficulties as the surface of the log likelihood is flat along parameters a and b_{city}.

The flatness of the log likelihood (4.7) is more severe with the triggering parameters, $\theta_T = (\alpha, \beta, \beta_{road}, \gamma)$, as shown in Figure 4.2. The results from one dimensional surface plots indicate that log likelihood is the flattest with β_{road} and α. Of the six possible arrangements, five two dimensional surface plots which had the flattest surfaces fill up the rest of figure 4.2. The two dimensional surface plots concur with our findings from one dimensional plot as they are usually flat along β_{road} and α.

From the plots, we can hypothesize that the estimation of background parameters, θ_B, will generally be easier than that of the triggering parameters, θ_T. Although these surface plots can serve as a useful diagnostic tool for assessing the flatness of log likelihood, they are certainly limited because they are merely one or two dimensional manifolds of the seven dimensional log likelihood. When all parameters are jointly estimated, flatness of likelihood may affect the results of MLE differently in higher dimensions. In Chapter 6, it will be shown using simulated data that the parameters that all estimation methods struggle the most with are α and β, the parameters for scaling and spatial lag in the triggering process.

[21] We use the estimates from the backfitting method because the estimates generated from other methods were roughly the same.

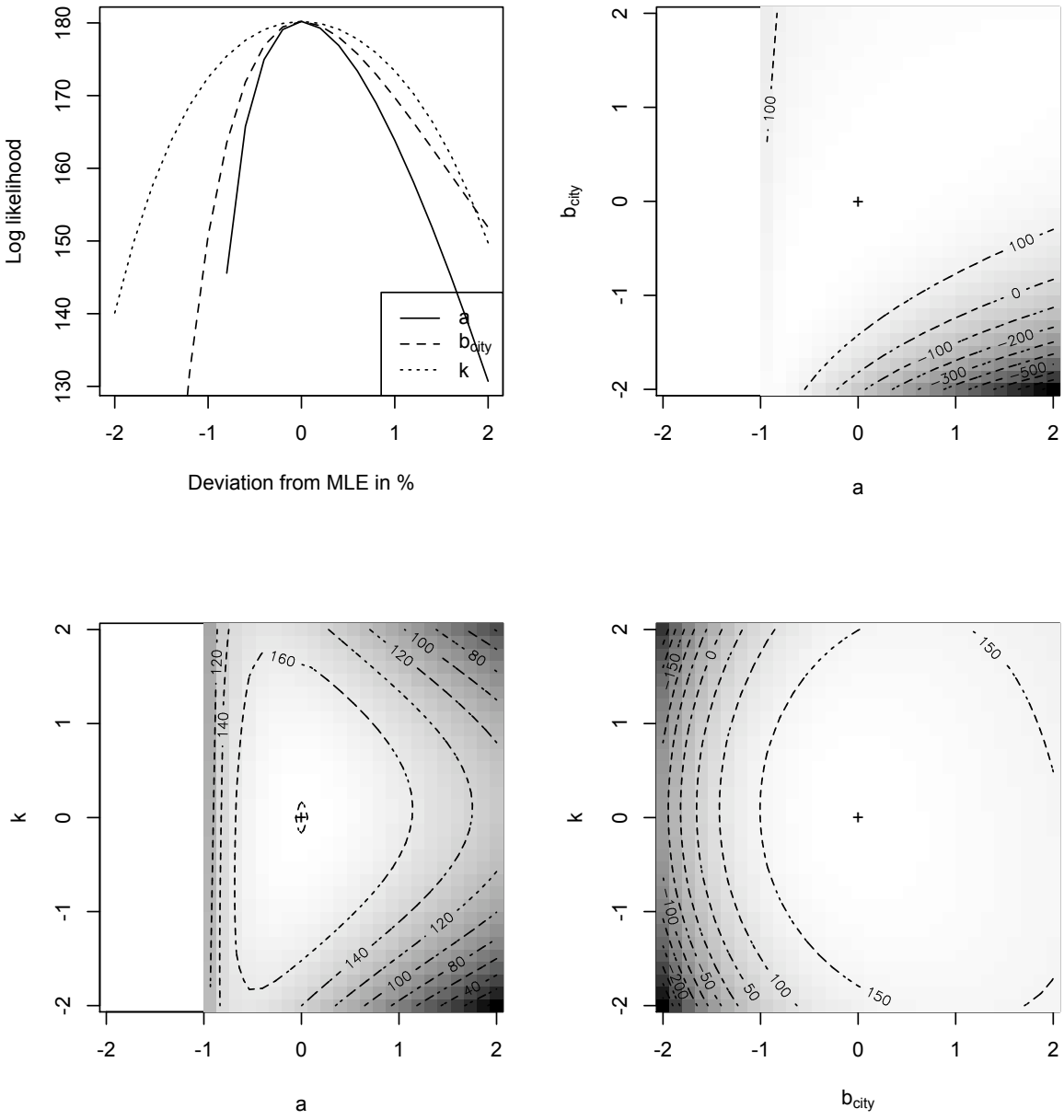

Figure 4.1: Surface plots for background parameters in Model (4). The crosses in the middle for the 2D surface plots mark the locations of the MLEs. In the 2D surface plots, the lighter shade corresponds to higher value in log likelihood.

4.5.3 Comparison of results from the three estimation methods

We now present the results from the three estimation methods—backfitting, Poorman's EM, and EM—for our best EAI model in terms of AIC score, Model (4). As mentioned earlier, the results from the three estimation methods are comparable. Their estimates and the corresponding likelihood values at each iteration of the algorithm are shown in Figure 4.3 with red, green, and blue indicating backfitting, Poorman's EM, and EM respectively. The EM algorithm reached convergence the slowest with 29 iterations, whereas backfitting and Poorman's EM reached convergence at 4 and 7 iterations respectively[22]. The parameter estimates using backfitting and the EM algorithms tend to agree, but the estimates for α and β using the EM is slightly below those of backfitting. This result is expected because the EM estimates should be the same as ML estimates. All parameter estimates from Poorman's EM algorithm were smaller than estimates from the other two methods and produced the smallest value of maximum log likelihood. The under-estimation is possibly due to misclassification of background events, but the exact reasons are unknown.

The backfitting method reached convergence the fastest and yielded estimates similar to that of EM method. The hybrid method, Poorman's EM, had relatively fast convergence, but its estimates do not agree with the results from the other two. We are uncertain whether the poorman's EM methods produces more accurate estimates. Given that the other two methods provides estimation results disagreeing with those of poorman's EM, it is unlikely that poorman's EM method works better in terms of accuracy. This issue will be investigated further by simulating the EAI model and re-estimating the parameters in Chapter 6.

[22]The convergence criterion is the difference between the likelihoods of last and the previous iterations becoming less than 10^{-3}.

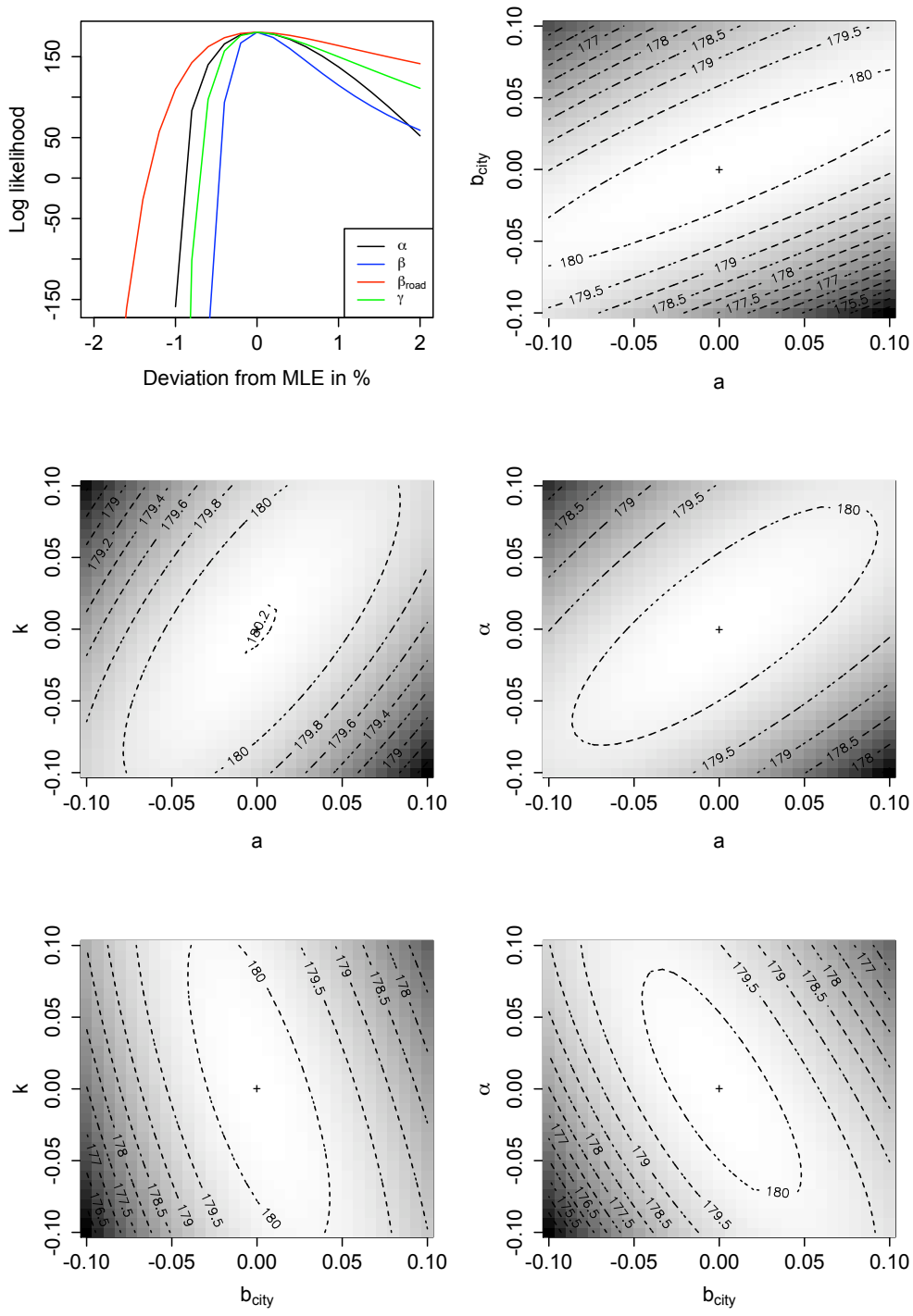

Figure 4.2: Surface plots for triggering parameters in Model (4). The crosses in the middle for the 2D surface plots mark the locations of the MLEs. In the 2D surface plots, the lighter shade corresponds to higher value in log likelihood.

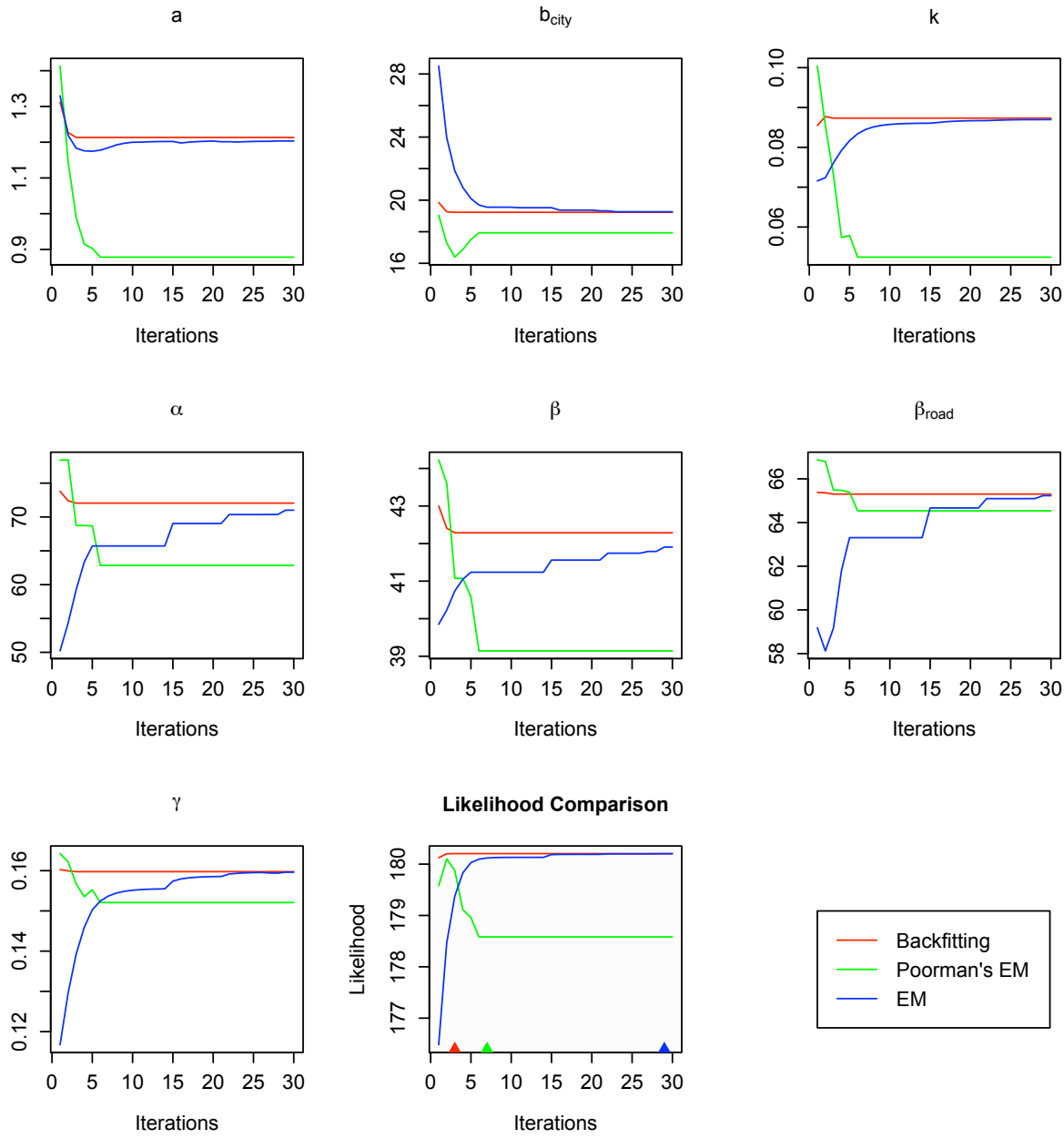

Figure 4.3: A plot of the parameter estimates for all parameters, $\theta = (a, b_{city}, c, \alpha, \beta, \beta_{road}, \gamma)$ at each iteration. The red, green, and blue triangles mark the numbers of iterations each method required to reach convergence. The last plot shows the log likelihood calculated at each iteration for all three algorithms.

Part III

Validation of the EAI model

Chapter 5

Model validation through residual analysis

In the previous chapter, we established our EAI model based on self exciting point process drawing inspirations from Ogata [42]. We devised three parameter estimation methods to deal with its flat and potentially multimodal likelihood and fitted five competing EAI models to determine the best model. Model (4) from Section 4.5 was shown to provide the best fit as its AIC was the smallest among the five.

Following the model fitting, this chapter aims to assess the fit of our best EAI model through residual analysis. The residuals of Model (4) will be compared against those of our second best model, Model (2), which assumes that H5N1 outbreak patterns are independent and dictated purely by proximity to the city and temperature. As explained earlier, Model (5) has the second best AIC score, but this model is essentially the same as Model (4) with its estimated κ approximately equal to 0. Therefore, we consider Model (2), a model without a branching structure, to be the second best model. This comparison will gauge the improvement of Model (4) over Model (2) from introducing the branching structure described by proximity to the nearest traffic networks, in addition to spatial and temporal lags.

The procedure for computing residuals for an inhomogeneous point process is outlined in the next section, followed by the comparisons of results for Models (2) and (4). The residual analysis will reveal that while Model (4) improves the fit substantially in terms of longitude compared to Model (2), the improvement observed for latitude is only marginal. Model (4) demonstrates better fit temporally especially for periods with temporal clusters, but the observed difference in comparison to Model (2) is not as remarkable as that of longitude.

5.1 Residual analysis with Stoyan-Grabarnik weights

Baddeley et al. [3] introduce the most comprehensive tool for performing residual analysis on inhomogeneous spatio-temporal point process. Their method is analogous to previously proposed methods [42,13,53,54], which transform a point process into a Poisson process with uniform intensity on a given interval. The deviation of the transformed point process from a homogenous Poisson process will indicate a possible model misfit.

The method presented in Baddeley et al. [3] applies Stoyan-Grabarnik weights [59] to the intensity of a point process. Consider the following zero martingale as a function of time, t, where $N(dxdydt)$ denotes a count measure for number of events and $\lambda(x, y, t)$ is an intensity function of inhomogeneous point process $\{N(x, y, t)\}$,

$$\int_0^T \int \int_S N(dxdydt) - \lambda(x, y, t)dxdydt. \tag{5.1}$$

Then

$$E\left(\int_0^T \int \int_S h(x, y, t)[N(dxdydt) - \lambda(x, y, t)dxdydt]\right) = 0, \tag{5.2}$$

where $h(x, y, t)$ is a weight function. With $h(x, y, t) = \frac{1}{\lambda(x,y,t)}$, we have

$$E\left(\sum_{i:(x_i,y_i,t_i)\in S\times T} w_i\right) - |S \times T| = 0 \tag{5.3}$$

where w_i denotes the Stoyan-Grabarnik weight, $\frac{1}{\lambda(x_i,y_i,t_i)}$, and $|S \times T|$ is the volume of the given space and time. The Equation (5.3) should intuitively make sense as it is essentially the same as transforming a point process into a Poisson process with intensity 1. The other popular variation of Stoyan-Grabarnik weight uses a weight function, $h(x, y, t) = \frac{1}{\sqrt{\lambda(x,y,t)}}$, which is analogous to Pearson residuals in linear regression. In our residual analysis, we will use $\frac{1}{\lambda(x_i,y_i,t_i)}$ as our weight function because the calculation of expectation term in Equation (5.3) for $h(x, y, t) = \frac{1}{\sqrt{\lambda(x,y,t)}}$, $\int_B \frac{1}{\sqrt{\lambda(x,y,t)}}dxdydt$, can be quite complicated.

With data, the intensity of inhomogeneous point process employing the estimated parameters, $\hat{\lambda}_{\hat{\theta}}(x, y, t)$, is used in place of the true intensity, $\lambda(x, y, t)$ for computing the weight, w_i. Let $\hat{w}_i$ denote the estimated Stoyan-Grabarnik weight for event i. Assuming that the model is right, we would still expect the difference in Equation (5.3) to remain 0 even with the estimated weights, $\hat{w}_i$'s.

Equation (5.3) can be modified to examine the fit of our EAI model in terms of three dimensions, longitude, latitude and time. For each dimension, we calculate the difference between the cumulative sum of the estimated Stoyan-Grabarnik weights and the corresponding volume:

$$s_1(x) = \sum_{i:x_i<x} \hat{w}_i - |x, Y \times T| \tag{5.4}$$

$$s_2(y) = \sum_{i:y_i<y} \hat{w}_i - |X, y \times T| \tag{5.5}$$

$$s_3(t) = \sum_{i:y_i<t} \hat{w}_i - |S \times t| \tag{5.6}$$

In our analysis, the estimated weights, $\hat{w}_i$, are computed with $h(x, y, t) = \frac{1}{\hat{\lambda}_{\hat{\theta}}(x,y,t)}$. For each $s_r(\cdot)$ for $1 \le r \le 3$, $s_r(\cdot) > 0$ indicates that the model expects more points than the

observed for the given dimesion. The results from computing residuals for Model (2) and (4) will be discussed in the following section.

5.1.1 Results

Plots of $s_1(x)$, $s_2(y)$, and $s_3(t)$ calculated for the two best models, Models (2) and (4), from Table 4.1 are shown in Figures 5.1 and 5.2. The black and blue curves correspond to the residuals from Model (2) and (4) respectively.

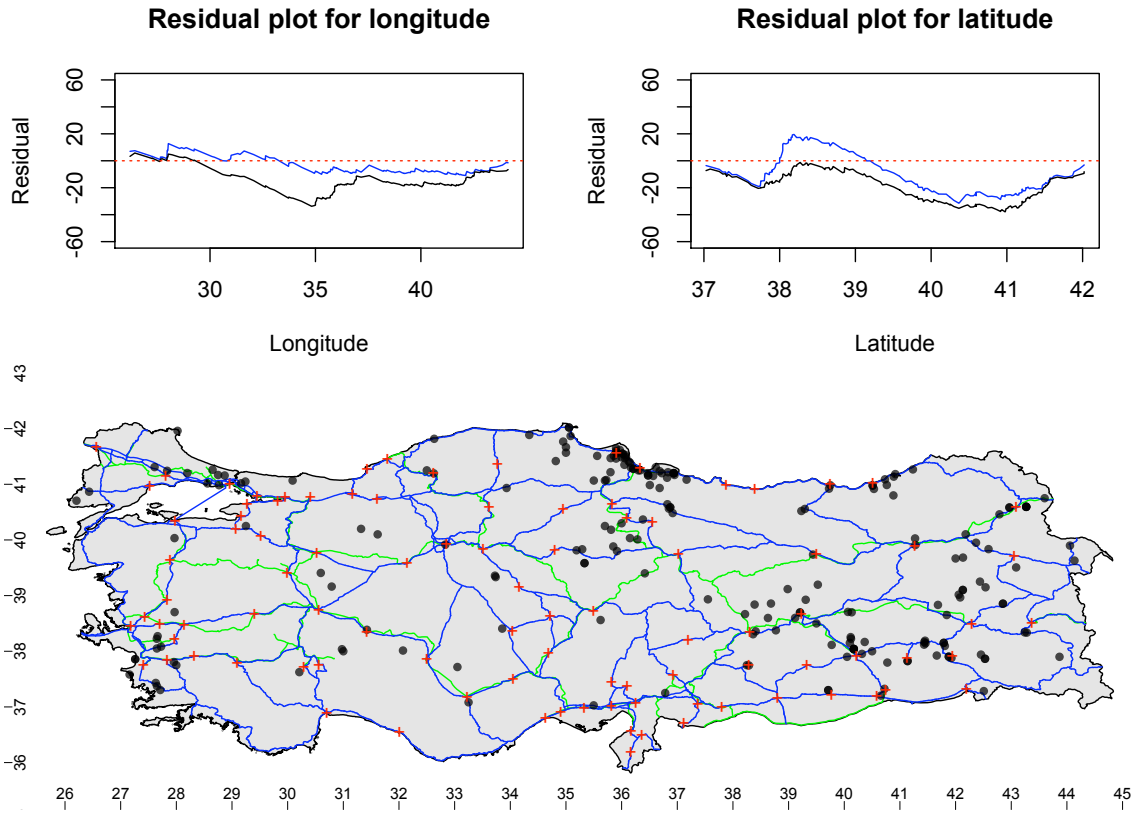

Figure 5.1: Residual plots for longitude, $s_1(x)$, and latitude $s_2(y)$. Plot of Turkey in Figure 1.2 is also provided below as a reference. The black dots represent the H5N1 outbreak locations. The green and blue lines correspond to Turkish railroads and highways respectively. The cities in Turkey are marked with red crosses.

Examining the residual plot for longitude in Figure 5.1, we observe that Model (4) has dramatically improved its fit over Model (2). While Model (2) only consists of background process with proximity to city and temperature as its components, Model (4) extends Model (2) with additional factors in the triggering process: spatial and temporal lag, and proximity to traffic networks.

The residuals from Model (2) indicate that Model (2) underestimates the longitudinal variation from 30° to 43°. This underestimation is most severe between 30° and 35°, where the outbreak locations do not show the tight clustering observed near cities such as Istanbul, Samsun, and Rize. The residuals from Model (2) approach 0 between 35° and 37° due to the heavy cluster of outbreaks near the city of Samsun but do not improve for longitude greater than 37°. Model (2) still expects more points from 37° and onward, as the inland clusters are looser than what the model dictates.

In comparison, residuals from model (4) demonstrate much better fit, staying along zero throughout. In addition to proximity to cities incorporated to background process, Model (4) aims to explain the spatial clusters of outbreaks with proximity to the traffic network and the distances among outbreaks. Thanks to its spatial branching structure, Model (4) successfully describes the variation between 30° and 40°, where Model (2) fails to provide an adequate fit. While the background intensity of Model (4) captures the loose clusters formed around cities, its triggering intensity explains the linear aggregation observed near Istanbul, Samsun, and Rize. The spatial triggering structure is able to describe the linear clusters of outbreaks near the cities much better than Model (2).

Despite its excellent fit for variations in longitude, Model (4) does not improve its fit—at least substantially—in terms of latitude compared to Model (2). From 40° to 42°, where the most of the tight clusters of outbreaks are observed, Model (4) only is able to improve the fit of Model (2) slightly. We suspect that the triggering structure of Model (4) unsuccessfully accommodates the vertical variation in this region as most of the clusters are formed horizontally along the traffic networks. Moreover, it overestimates the spatial trend between 38° and 39° in latitude. Under Model (4), we expect to observe more outbreaks aggregating in this area because there are many outbreaks that could potentially trigger other outbreaks. Instead, they are loosely scattered, which is the opposite of what Model (4) dictates. As mentioned in Section 2.2.2, a possible reason for the loose cluster of outbreaks observed near inland cities may be because poultry farms are more densely populated near the port cities for ease of trade.

Temporally, both residuals from Models (2) and (4) show that the models are heavily affected by the outlier, an outbreak that occurred on the first day. Both models expect fewer outbreaks from day 1 to day 100—perhaps no outbreaks at all. With their configurations, they are unable to explain the observed temporal variation, until the sudden surge in the outbreak counts around to day 100. The temporal lag component in the triggering process enhances the fit of Model (4), where temporal clusters are eminent. From day 100 and onward, Model (4) provides slightly better fit during the periods with a large number of outbreaks.

Overall, introducing the branching structure featured in Model (4) successfully improved the fit of Model (2). Model (4) performs substantially better in fitting longitudinal variation due to the new spatial component, proximity to traffic networks, added to spatial lag in the triggering process. As noted in table 4.1, under the configuration of Model (3), spatial lag in the triggering process alone was not able to capture the linear outbreak clusters formed along the traffic networks. While Model (4) with its spatial branching structure provides much better fit in terms of longitude, it only marginally improves fit of latitudinal variation.

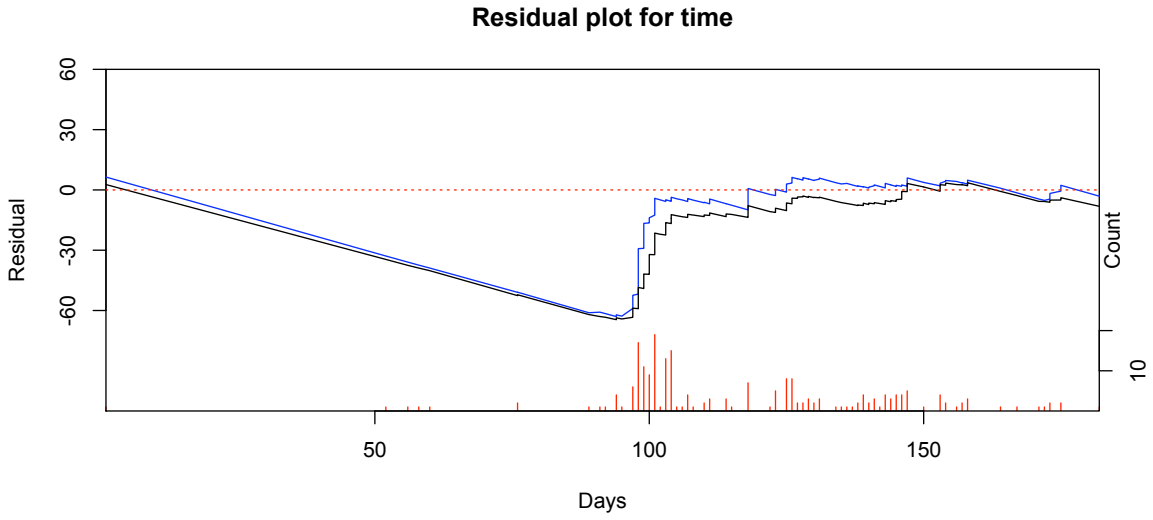

Figure 5.2: Residual plot for time, $s_3(t)$. The frequency of number of outbreaks corresponding to the dates of their occurrences is shown at the bottom as a reference.

This contrasting result is an aftermath of the fact that most of the clusters formed along the traffic networks were horizontal[1]. Temporally, the residuals of Model (2) and Model (4) are quite similar, but Model (4) provides a slightly better fit when temporal aggregation is present, thanks to the temporal component in the triggering process.

The next chapter on simulation of the EAI model will demonstrate how our validated EAI model can be applied using predicted results, along with performance benchmarks for different estimation methods.

[1]The horizontal aggregation may have been purely due to chance. Since our aim is to provide a general framework for modeling H5N1 spread, tailoring the model to fit only the horizontal variation would be too data specific.

Part IV

Simulation

Chapter 6

Simulating the EAI model

So far, we have focused on fitting, validating, and examining possible applications of the EAI model. The results presented in Chapter 5 for validating the model for the Turkish data were based on the estimates produced via backfitting method featured in Chapter 4. Both chapters did not employ the estimates from two other methods, poorman's EM and EM. Section 4.5.3 presented results from all postulated methods, but no formal comparison among the three methods was provided. As noted previously, the estimates obtained from poorman's EM method were consistently different for all seven parameters for Model (4). The comparison of computational speed indicated that the backfitting method reached convergence (within 10^{-3}) the fastest, followed by Poorman's EM and EM in that order.

This chapter aims to gain deeper insight on the performance of the three methods than the brief comparison given in Section 4.5.3. Through simulation, we can not only gauge the performances of the three estimation methods but make a simple prediction of future outbreaks in Monte Carlo fashion.

The first part of this chapter constructs an algorithm for simulating our EAI model inspired by Lewis [31] and Zhuang et al. [66]. Further, adopting the simulation strategy of Zhuang et al. [66] and Vere-Jones [61], we present a modified simulation algorithm with an edge correction method to mitigate any boundary effects, commonly encountered when simulating a point process.

Following the simulation algorithm, the performance of the estimation methods will be measured in two different categories: robustness against starting values and accuracy. These benchmarks involve simulating the EAI model with a given set of parameters and estimating them based on the simulated data. The deviations of the estimated parameters from the truth will gauge the performance of each estimation method. With this comparison, we hope to understand how effectively each method deals with the flat likelihood of EAI model with respect to the parameters.

The robustness of our procedures will be inspected first against a wide range of starting values to verify that all estimates converge to the same values at the end of the iterations. Then the accuracy of each method will be examined by estimating the parameters from simulated data both with and without edge correction. These performance benchmarks identify two parameters that are particularly difficult to estimate and determine the best

performing estimation methods in terms of accuracy and computational speed.

Further, we can gauge the performance of the EAI model itself in lieu of the estimation methods. Using simulation results, we can construct hypothesis tests on the parameters of the EAI model and calculate the corresponding empirical powers to assess whether our EAI model is able to correctly detect the components of the triggering process when in fact they are present. We are especially interested in the EAI model's ability to capture the presence of temperature variation in the branching structure. The empirical power will assist in determining if the configuration of the model is responsible for failing to detect the temperature component included in Model (5) from Table 4.1.

Lastly, we extend the proposed simulation algorithm to predict the occurrences of H5N1 in Turkey with the fitted EAI model. Given the past progression of H5N1 outbreaks, we can repeatedly simulate the future H5N1 incidences, obtaining a distribution of predicted outbreaks—their locations and times of occurrences. Prediction of H5N1 outbreaks via simulation can serve as a useful guide for responding to the explosive spread of H5N1.

6.1 Simulation algorithm for EAI model

Lewis and Shedler [31] outline the most widely accepted method for simulating an inhomogeneous point process. Their idea is rather simple; after simulating a homogenous point process in a desired space, each point will be "thinned out" if the ratio of the intensity at the given point and the maximum intensity is less than a number generated from a uniform distribution. For an inhomogeneous point process with rate $\lambda(x, y, t)$, Lewis' simulation algorithm is provided below:

Algorithm 4 (Lewis): Simulation algorithm for an inhomogeneous point process with rate, $\lambda(x, y, t)$.

Step 1. Simulate a homogeneous poisson process in a given space with rate, $\max \lambda(x, y, t)$

Step 2. Take a simulated point, evaluate the intensity at its position using the inhomogenous rate, compare the ratio with randomly generated number from uniform distribution on [0,1]. If ratio is less than the randomly generated number, delete the point.

Step 3. Repeat step 2 to all points simulated from step 1 to obtain inhomogeneous point process.

In our case, this algorithm will be only used to generate the background point process because simulating the triggering process, corresponding to $\lambda_T(x, y, t)$ in Equation (4.7), using Lewis' algorithm might be difficult; depending on the past progression of events, the

calculation of maximum intensity (4.5) at given time, t, can be quite complicated when triggering intensities of outbreaks, $\lambda_T(x, y, t)$, overlap amongst each other spatially.

Therefore, we adopt a simulation procedure proposed by Zhuang et al. [66], which takes advantage of the branching structure of our point process. This approach was also noted in Brix and Kendall [5] and Moller et al [37, 36]. Their simulation algorithm keeps track of generations of triggered events, denoted $G^{(l)}$ for each generation l. The events in the background process are generated first with algorithm 4 and are labeled as generation 0, $G^{(0)}$. The next generations of events will continue to be produced, based on the expected number of events triggered by each event from the previous generation, until the most recent generation ceases to contain any events.

Combining the two simulation approaches with slight modifications, the algorithm for simulating the EAI model is outlined below:

Algorithm 5: Simulation algorithm for the EAI model

Step 1. Simulate background point process with intensity $\lambda_B(x, y, t)$ using Algorithm 4 over $S \times T$. Call this generation 0 and denote it $G^{(0)}$.

Step 2. Set $l = 0$

Step 3. For each event i, (x_i, y_i, t_i) in $G^{(l)}$, simulate the number of its offspring first using λ_{T_i} 4.19 as a rate of a homogenous point process. Consequently, draw spatial and temporal lags from probability density forms of $s(x - x_i, y - y_i)$ and $g(t - t_i)$. Assign them to the corresponding offspring.

Step 4 For each offspring from step 3, assign an angle drawn from a uniform distribution on $[0, 2\pi]$. With the computed spatial lag from the location of its parent event and the angle, determine the latitude and longitude of the offspring. Evaluate $s(x, y)$ at the obtained location and redraw the location of the offspring until $s(x, y) > unif[0, 1]$[1]. Assign the resulting location in longitude and latitude to the offspring.

Step 5 Add the temporal lag—computed in step 3—to the occurrence time of its parent and assign it to the offspring.

Step 6 Denote the offsprings generated from an event i in $G^{(l)}$, $O_i^{(l)}$. For each $O_i^{(l)}$, delete offsprings of $O_i^{(l)}$ if $O_i^{(l)}$ that do not belong in $S \times T$. Set $G^{(l+1)} = \cup_{i \in G^{(l)}} O_i^{(l)}$.

Step 7 If $G^{(l+1)}$ is not empty, set $l = l + 1$ and go to step 3. Otherwise return the resulting catalog, $\cup_{j=0}^{l} G^{(j)}$.

[1]This is a rejection sampling.

We will use Algorithm 5 to simulate the H5N1 outbreaks in Turkey for 182 days with parameters $\theta = (a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma) = (1, 20, 0.1, 50, 40, 70, 0.15)$. If otherwise, the parameters will be specified. Algorithm 5 will also serve as a backbone for predicting future outbreaks of H5N1 in Section 6.3.

6.1.1 Edge correction method

Edge effects are a common problem in the point process literature and are even more severe for an inhomogeneous point process with a branching structure. The study region for our self-exciting point process, the three dimensional space spanned by Turkey and 182 days, is finite, and the continuous intensity 4.5, $\lambda(x, y, t|H_t)$, will get cut off at the boundary. As a result, the corresponding intensity of events located near the boundaries, both spatially and temporally, will be under-estimated. Although our region of interest, Turkey, is quasi-convex and the edge effect is less severe in comparison to other countries, our analysis may suffer greatly from the edge effect introduced by the dense cluster of outbreaks observed at the coastal cities such as Samsun and Rize.

The edge effect will plague both estimation and simulation results. Fitting a self-exciting point process is highly likely to produce biased estimates: the parameters corresponding to spatial and temporal variation, $(b_{city}, \beta, \beta_{road}, k, \gamma)$, are expected to be overestimated, as the cut-off of intensity function (4.5) at the boundary is apt to yield faster decaying functional forms.

Likewise, the simulation will suffer from the bias introduced by the finite space of interest. With Algorithm 5, the events located within the boundary are unable to produce offspring outside. In addition, it will simply exclude the events potentially triggered by other events beyond the boundary even if they lie within the given space.

Zhuang et al. [66] and Vere-Jones [61] addresses this issue by simulating the events in a bigger space—denoted $S_{big} \times T_{big}$ —than the study region. This configuration sets up a buffer zone in $S_{big} \times T_{big}$, which spatially surrounds and temporally follows the study space $S \times T$. Simulating the self-exciting process in $S_{big} \times T_{big}$ and including it, at least partially, in the estimation procedure is likely to improve the parameter estimates, as noted in Vere-Jones [61].

In our case, however, considering occurrences of H5N1 outbreaks outside Turkey is unrealistic. More than half of the country is surrounded by Black and Mediterrian Sea and it is physically impossible to observe any outbreaks beyond the country's border. Although the southeastern part of Turkey is connected to the Middle East, we cannot assume the similar disease spread mechanism in this area; the neighboring countries possess different infrastructures, administration, and ecological features that may result in disparate spread of the H5N1 virus.

Despite these drawbacks, we will adopt the simulation strategy of Zhuang et al. [66] and Vere-Jones [61]. Self-exciting point process is notorious for its difficulty in parameter estimation [60, 44], and we wish to measure the improvement in accuracy of the estimated parameters with the introduction of edge correction method—at least in theory.

Let S_{big} represent a rectangular area spanning from 25° to 46° and from 35° to 43° in

longitude and latitude respectively. Temporally, we do not define a bigger time interval but retain offspring occurring later than T and prevent these offspring from triggering a future generation. The simulation algorithm with edge correction is presented in Algorithm 6:

Algorithm 6: Edge correction (Addendum to Algorithm 5)

Step 1. Simulate the background point process with intensity $\lambda_B(x, y, t)$ using Algorithm 4 over $S_{big} \times T_{big}$. Call this generation 0 and denote it $G^{(0)}$.

Step 2. Set $l = 0$ and initialize $P^{(0)} = G^{(0)}$. $P^{(l)}$ will represent group of events generated at l^{th} iteration, which are allowed to produce offsprings in the next iteration.

Step 3. For each event i, (x_i, y_i, t_i) in $P^{(l)}$, simulate the number of its offsprings first using λ_{T_i} 4.19 as a rate of a homogenous point process. Consequently, draw spatial and temporal lags from probability density forms of $s(x - x_i, y - y_i)$ and $g(t - t_i)$. Assign them to corresponding offsprings.

Step 4 For each offspring from step 3, assign an angle drawn from a uniform distribution on $[0, 2\pi]$. With the computed spatial lag from the location of its parent event and the angle, determine the latitude and longitude of the offspring. Evaluate $s(x, y)$ at the obtained location and redraw the location of the offspring until $s(x, y) > unif[0, 1]^2$. Assign the resulting location in longitude and latitude to the offspring.

Step 5 Add the temporal lag—computed in step 3—to the occurrence time of its parent and assign it to the offspring.

Step 6 Denote the offsprings generated from an event i in $P^{(l)}$, $O_i^{(l)}$ and set $G^{(l+1)} = \cup_{i \in G^{(l)}} O_i^{(l)}$. For every offspring in $G^{(l+1)}$, check if the its occurrence time greater than T and only include the offsprings that occurred before T to $P^{(l+1)}$.

Step 7 If $P^{(l+1)}$ is not empty, set $l = l + 1$ and go to step 3. Otherwise return the resulting catalog, $\cup_{j=0}^{l} G^{(j)}$.

In Algorithm 6, the background process, $G^{(0)}$, is simulated over $S_{big} \times T_{big}$. Spatially, the future generations will not spawn events far away from S as the intensity of the point process (4.5) decreases exponentially according to the distance from the traffic networks located near the border. The traffic networks serve as a natural bound preventing the simulated outbreaks to drift away from S. With our best model, Model (4), such a natural bound does not

[2]This is a rejection sampling.

exist temporally because the outbreaks will be generated purely based on the temperature component in the background process. For periods with favorable temperature, Algorithm 5 will continue to produce outbreaks without a temporal limiting factor in the triggering process. Therefore, we limit the events generated after T by retaining but preventing them from triggering any offsprings in the future generations.

In the following section, simulated data produced from Algorithm 5 will be used to assess performances of the three estimation methods in terms of sensitivity to starting values and accuracy of the estimates. The accuracy results obtained from Algorithm 5 will be compared with those of Algorithm 6 to determine whether our edge correction method yields more accurate estimates.

6.2 Comparison of the three estimation methods

In Chapter 4, we proposed backfitting, Poorman's EM, and EM methods to resolve common estimation issues arise from maximizing likelihood of self exciting point process. It was briefly mentioned that the main benefits of backfitting and EM methods are computational speed and accuracy respectively. Poorman's EM was designed to take advantages of the two estimation methods by simplifying complicated probability calculation involved in EM method for faster convergence.

This section will examine the performances of the three methods through simulation. First, we explore how each estimation method manages wide range of starting values to reach convergence by estimating parameters for data generated from Algorithm 5. The result will indicate which methods fail to produce concurring estimates for the troublesome parameters of EAI model. Consequently, the accuracy of the estimates produced by the three estimation methods are compared using simulated data sets with and without edge correction. We discuss the differences in precision of parameter estimates and the computational speed among the estimation procedures and inspect the their changes from introducing the edge correction.

6.2.1 Sensitivity to starting values

Prior to the comparison in accuracy, we examine how the starting values may affect the convergence of the estimates produced by the three estimation methods. The likelihood of a self-exciting point process is infamous for being flat and multimodal, which can lead to wrong parameter estimates [60, 44].

We simulated 10 data sets using Algorithm 5 with parameters $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma) = (1, 20, 0.1, 50, 40, 70, 0.15)$. For each data set, 100 starting values for the parameters were drawn randomly from a uniform distribution, whose range is $\frac{1}{5}^{th}$ of the true value to 5 times the true value, per parameter. The results from estimating parameters with the 100 starting values for one of the simulated data are illustrated in Figures 6.1, 6.2, and 6.3. The three figures correspond to the results obtained using backfitting, poorman's EM and EM methods respectively. The individual plots show the values of estimated parameters at each iteration

until their convergences. The estimated parameters of the other simulated data agree with the results provided and therefore are omitted.

Overall, all three methods seem to yield similar parameter estimates and maximum likelihoods, excluding the two parameters from the triggering process, α and β. While all other estimates were within 0 to 100% in range relative to the true parameter values, the estimates for the two troubling parameters were severely biased, converging to roughly 3 to 5 times their true values. The estimated values of the two parameters are generally in the neighborhood of 4 and 3 times their true values respectively, but their distributions vary depending on the estimation methods.

The backfitting method is able to achieve convergence the fastest with all estimation runs terminating within 4 iterations, and it produces the most consistent parameter estimates. On the other hand, poorman's EM method struggles to produce consistent estimates for α and β, although it reaches convergence faster, with fewer required iterations less than for the backfitting method for these two parameters on average. The majority of estimates of α produced by poorman's EM do not converge to the same value, with quite a few of them falling in the range of 4 to 6 times its true value. While the estimated values of β are more consistent, the two divergent estimates for α and β lead to conflicting maximum likelihood values as illustrated in the last plot of figure 6.2.

Similarly, the EM algorithm struggles with the estimation of α and β, especially in the early stage, producing a wide range of estimates. Despite the highly variable estimated values observed in the early iterations, the estimates of α and β converge to acceptable ranges; the range of EM estimates of α are similar, in comparison to the poorman's EM, whereas the EM estimates of β are closer to the truth with less spread[3].

In summary, all algorithms produced comparable parameter estimates and maximum likelihood except for parameters α and β. The estimates for the two parameters were severely biased, and Poorman's EM and EM algorithms yielded disagreeing results with Poorman's EM performing worse than EM. In the next section, we will assume different starting values will yield consistent results and compare their accuracy, although we acknowledge that the estimated values of α and β may be highly variable.

One hundred starting values for the parameters were drawn randomly from a uniform distribution, whose range is $\frac{1}{5}^{th}$ of the true value to 5 times the true value, per parameter.

6.2.2 Accuracy of the estimates for data simulated without edge correction

To compare the accuracy of the estimates produced by the three methods, we simulated 100 data sets using Algorithm 5 with the same parameter values mentioned earlier in Section 6.2.1. For each simulated data set, a set of starting values were drawn randomly from

[3]We also note that the EM algorithm takes much longer to achieve convergence with median of 11 iterations. The maximum iteration required for convergence was 29. While a fairer comparison can be made if the progression of estimates from 1 to 29 iterations were shown, the parameter estimates generally agree within 10 iterations for all parameters except α and β

a uniform distribution, whose range is $\frac{1}{2}^{th}$ of the true value to 2 times the true value, per parameter. These starting values were employed by all three methods in estimating θ.

Figure 6.4 and Table 6.1 display the results from estimating parameters with all three methods. We first refer to plots in Figure 6.4 to illustrate the general pattern of the performances. The first seven plots correspond to the distributions of obtained estimates for parameters, $\theta = (a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma)$, and the last plot displays the distributions of number of iterations required to reach convergence by the three estimation methods.

Overall the distributions of estimates are similar across all three estimation methods. It is apparent from the plots in figure 6.4 that the parameters for the background process, a, b_{city}, and k are most accurately estimated; these parameter estimates range approximately 70 % below and above their true values regardless of the methods employed. Among the three parameters, estimates of k demonstrate the greatest accuracy with the medians of estimates from all three methods matching the true value of k. In contrast, parameters a and b_{city} are over and under estimated respectively. This behavior is probably due to the fact that the values of two parameters are linked, as the functional form of proximity to city, $e^{-b_{city} \cdot r_{city}}$, is not standardized; a decrease in b_{city} causes an increase in the background scale parameter a and vice versa.

There are no clear winners amongst the estimation methods in terms of accuracy for the background parameters. As shown in Table 6.1, while poorman's EM yields the most accurate result for a, EM produces more precise estimates than the other two algorithms for b_{road} in terms of bias and RMSE. For k, the Poorman's EM estimates edges out the other two in terms of bias, though with a larger RMSE. The estimates of k from the other two methods have smaller RSME in comparison, but it is slightly more biased than those of Poorman's EM. Although no method distinguished itself with its performance, the estimates produced by all three methods showed relatively good accuracy; all computed biases and RSMEs for (a, b_{road}, c) are less than 13% and 34%.

On the contrary, estimates of parameters in the triggering process, $(\alpha, \beta, \beta_{road}, \gamma)$, are found to be heavily biased across all estimation methods. Figure 6.4 demonstrates that estimates of α and β suffer from severe bias with their medians deviating from their true values by roughly 200%. The range of obtained estimates for α is atrociously large spreading from 0 to 8 times its true value. This behavior was expected as the scaling parameter of the triggering process is known to be difficult to estimate for self exciting point process models [60, 44]. The distribution of estimates for β does not have the immense spread in comparison, but none of the estimates agreed with its true value. This bias observed for β raises concerns because the estimates were consistently inaccurate for all employed methods. It is also interesting to note that surface plots of likelihood (4.7) in Section 4.5.2 suggested α and β_{road} to be potentially troublesome parameters in estimation. In practice, the two parameters, α and β, noted for producing divergent estimates in the previous section, were found to be difficult to estimate.

Although the medians of estimates for β_{road} and γ deviate from their truth, these parameter estimates are only slightly biased relative to α and β. β_{road} and γ are both underestimated perhaps due to the overestimation of α and β. Among all the triggering parameters, estimates of γ were the most accurate with their medians roughly 25% away from the truth,

and their distribution ranging from 50% below and 5% above. The underestimation of β_{road} was also relatively minor as the distribution of its estimates were centered about 70% below the expected, spanning from -90% to 10% of the true value.

Examining the accuracy of the estimates according the estimation methods with Table 6.1, the EM method produced the best results, especially for the parameters giving the most trouble, α and β. The EM estimate for α exhibits about 30% less bias over the estimates produced by other two methods along with roughly 40% reduction in RMSE. For β, the improvement in accuracy is relatively smaller than that of α, but the EM estimates surpass the performances of the rest nonetheless, reducing 10% both in bias and RMSE. Unlike α and β, all procedures yield estimates with comparable precision for β_{road} and γ.

In summary, while no procedure separates itself from the rest for estimation of background parameters, the EM method clearly outperforms the two, producing much more accurate estimates for some of the triggering parameters. Although all estimated parameters from the three methods are heavily biased, the EM method improves the precision of the estimates for problematic parameters, α and β in comparison to the other methods. A possible reason for the improvement was discussed earlier in Section 4.4.2. With limited observations, the asymptotic properties of MLE may not hold in comparison to EM, which could produce more accurate estimates due to the incorporated branching structure.

Despite its better accuracy, the EM method is computationally the most expensive. As demonstrated in the last plot of Figure 6.4, the maximum number of iterations required for the backfitting and the poorman's EM methods were 5 and 7 respectively. The longest run of estimation procedure using the EM algorithm, however, took 27 iterations. On average, the EM method took 12 iterations to converge resulting in roughly 4 and 2.5 times longer computational time than the backfitting and Poorman's EM method.

6.2.3 Accuracy of the estimates for data simulated with edge correction

In the previous section, the background parameters, (a, b_{city}, c), were shown to be correctly estimated with good accuracy regardless of the employed methods. All procedures, however, struggled in estimating the parameters of the triggering process $(\alpha, \beta, \beta_{road}, \gamma)$, producing heavily biased results. Because the inaccuracy in parameter estimates is only observed for $(\alpha, \beta, \beta_{road}, \gamma)$, this bias may have originated from the edge effect. As mentioned in Section 6.2.3, edge effect, a common problem among point process models, is aggravated for our EAI model due to its branching structure. This section investigates whether the edge correction implemented in Algorithm 6 is beneficial to parameter estimation of the EAI model. We are particularly interested in its effect to estimation of the triggering parameters. The outcomes from using simulated data with and without edge correction will be compared according to the three estimation methods, to assess the improvement in bias, if there is any.

To obtain the new parameter estimates, one hundred data sets were generated with the same parameter values as before employing Algorithm 6, the simulation algorithm for EAI models with edge correction. The starting values from Section 6.2.1 were reused for each simulated data set. The results from estimating parameters with backfitting, poorman's

EM and EM methods are provided in Table 6.1, and they are graphically represented in Figure 6.5. The layout of the plots in Figure 6.5 is identical to Figure 6.4, except for the new results shown next to the ones produced without the edge correction.

Figure 6.5 illustrates that the distribution of estimates for θ obtained using simulated data with edge correction generally agree across all employed estimation methods. The reduction in bias that we hoped to achieve for all parameters is found to occur for successful in most parameters for all methods, although this reduction is modest. While we observe major improvement in accuracy for some parameters such as b_{city}, α, β, and β_{road}, the parameters, a and γ, are slightly more underestimated compared to the outcomes without edge correction. The distribution of new estimates for k is similar to the previous result but with bigger spread. Because the parameters are closely related, sharing the same scaling parameter, the minor bias introduced to the three parameters are likely to be an outcome of obtaining notably more precise estimates for the other parameters.

Referring to Table 6.1, we note a substantial improvement in accuracy for α. Depending on the method employed, the bias is reduced to at least 60% after simulating data using the edge correction. The corresponding RMSEs also decrease more than 100% for all methods. Another parameter that benefits from edge correction is β with sizable improvement in bias at least by 35% for all procedures. The changes in distribution for other parameter estimates were marginal in comparison.

The performance of the three methods in terms of their precision remains the same even with the new results. Despite using the simulated data with edge correction, there is still no clear winner for the background parameters. The estimates obtained using the EM algorithm produced better estimates for b_{city} and c than those of other methods, but the differences are negligible.

Similarly, the EM method still produces the most accurate estimates for the triggering parameters, though in a lesser degree than the prior outcome. Without the edge correction, the RMSE—measured in percentage deviation from the true value—of EM estimates are roughly 40% and 15% smaller than the next smallest RMSEs for α and β respectively. Applying the edge correction reduces these differences to 11% and 12%, but they are still major improvement over changes observed in estimates for other parameters.

However, the EM method again takes the longest time to reach convergence. Using the updated triggering probabilities at every iteration leads to more precise estimates, but it also requires more computation time than the other two methods.

Overall, using the simulated data with edge correction leads to extended computational time for all methods perhaps due to the increased number of events in each simulated data set. Simulating our EAI model in the expanded target space, $S_{big} \times T_{big}$, will certainly yield more events with the same intensity, thereby increasing the processing time for estimation. With edge correction, the processing times for obtaining estimates for each data set increases by 90%, 137% and 80% on average for backfitting, poorman's EM, and EM methods respectively. Poorman's EM algorithm may have required much longer time relative to other methods since its background selection procedure at each step takes longer for larger data sets. The processing time of the EM method was the least affected and generally required fewer number of iterations until its convergence, as demonstrated in Figure 6.5. Although

the exact reasons for this improvement are not known, it is probable that the edge correction aided in determining estimates for α and β, the parameters that are responsible for longer computational time for the EM algorithm. On the contrary, the number of iterations required for backfitting and poorman's EM algorithms have increased, illustrated by the fatter tail in their distributions from Figure6.5. It is also interesting to note that these shifts in their distributions do not exceed the maximum number of iterations observed without applying the edge correction.

Our investigation, so far, has shown that simulating data with edge correction does improve the accuracy of estimates for some parameters—most notably α and β—for all methods. Parameters representing the temporal components of the EAI model, k and γ, did not benefit from the simulated data with edge correction, possibly due to only allowing events generated in $[0, T]$ to produce offspring in Algorithm 6. Therefore, a modification of temporal edge correction may improve the results for these parameters. A viable remedy would be permitting events falling out of $[0, T]$ to spawn offspring for another generation or more.

Upon reviewing the estimation results, the EM method is shown to consistently outperform other estimation methods in terms of accuracy, especially for α and β. However, it is computationally the most expensive, reaching convergence three times more slowly on average than the fastest method, backfitting. In practice, the computational time difference is negligible as we do not aim to provide real time results. Therefore, we recommend the use of the EM method for parameter estimation of the EAI model.

6.2.4 Power to detect components in triggering process

The past sections have focused on comparing the performances of the three estimation methods through simulation. Our analysis based on robustness against starting values and accuracy revealed that the EM method stands out above the rest, producing the most precise estimates.

In this section, we turn our attention to testing performance of the EAI model itself. We wish to examine whether our model is capable of detecting the presence of various triggering components. Using simulation, we can construct a likelihood ratio test by comparing the maximum likelihoods estimated from the simulated data for two models: one assuming the null hypothesis, that the data is generated without the influence of the component of interest, and the other assuming the negation, the alternative hypothesis. By fitting the two models to the data simulated assuming the alternative hypothesis, we can calculate the empirical power of a test, the probability that test correctly reject the null hypothesis given that the alternative is true for a particular parameter setting. In other words, the empirical power of the test will give an idea of how well our model can correctly detect the presence of added triggering component.

In Chapter 4, the outcome of fitting Model (5), which includes the effect of temperature in the triggering process, indicated no sign of influence in the triggering process corresponding to the temperature. From the introduction, we know that the colder temperature extends the life of H5N1 virus in water. Although the parameter for temporal variation, κ, was

			Background			Triggering			
	Method	Statistic	$\hat{a}$	$\hat{b}_{road}$	$\hat{k}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\beta}_{road}$	$\hat{\gamma}$
	Backfitting	bias	12.3	-11.2	1.38	232	183	-62.0	-19.9
		RMSE	32.5	33.3	23.4	279	194	66.9	24.6
without edge	Poorman's EM	bias	7.47	-11.8	0.219	230	181	-62.0	-20.0
correction		RMSE	29.3	33.5	24.0	276	191	67.0	24.7
	EM	bias	11.3	-10.9	1.54	**202**	**170**	-62.5	-20.3
		RMSE	31.6	32.9	23.5	**239**	**178**	67.3	24.6
	Backfitting	bias	-14.5	-3.14	5.25	144	141	-59.9	-26.4
		RMSE	26.2	21.2	26.8	165	147	63.0	28.4
with edge	Poorman's EM	bias	-17.2	-5.56	-0.233	148	143	-59.8	-26.3
correction		RMSE	28.2	21.8	24.9	169	149	62.9	28.3
	EM	bias	-15.4	-3.38	6.01	**126**	**131**	-60.1	-26.8
		RMSE	26.8	21.4	25.2	**144**	**135**	63.2	28.7

Table 6.1: Table of biases and root mean squared errors (RMSE) of the estimates obtained by backfitting, poorman's EM, and EM algorithm. All statistics are shown in terms of the percentage deviation from the true values. The results—separated according to the simulation method—are grouped by the employed algorithms.

estimated to be 0, we suspect that the configuration of the model may not be able to detect the presence of such an effect even if is present in reality. The computation of the empirical power mentioned above will provide a statistical measure for assessing the performance of our model in terms of its ability to detect the presence of triggering component.

The test statistic of a likelihood ratio test for nested models is [64]:

$$D = -2ln\left(\frac{L_{simple}(\hat{\theta}^s_{ML})}{L_{complex}(\hat{\theta}^c_{ML})}\right) \tag{6.1}$$

where L_{simple} and $L_{complex}$ are the likelihood of simple and complex models, and their corresponding maximum likelihood estimates are denoted $\hat{\theta}^s_{ML}$ and $\hat{\theta}^c_{ML}$ with $\theta^s \subset \theta^c$. For our analysis we have two sets of simple and complex models which will be presented in the next paragraphs. The first set of simple and complex models are Model (4) and Model (5) respectively. The second set consists of a simple model, Model (3), and a complex model, Model (5). The null hypothesis of the statistical test is $H_0 : \phi = 0$ with ϕ representing the set of parameters belonging to θ^c excluding θ^s. The two models are nested in the sense that the complex model can be transformed into the simple one by imposing set of linear constraints on the parameters. The test statistic, D, asymptotically follows χ^2 distribution with degrees of freedom corresponding to the number of extra parameters, ϕ, under the condition that

the asymptotic properties of MLE holds for both simple and complex models [64].

To construct our test, we simulate 100 data sets with parameters $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma, \kappa)$ $= (1, 20, 0.1, 50, 40, 70, 0.15, 005)$ assuming the alternative hypothesis, that the effect of temperature is present in the triggering process. In this case, Model (4) is the simple model representing the null hypothesis with $\theta^s = (a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma)$. The more complex model, Model (5), includes an extra parameter, κ, in θ^c compared to θ^s. The results from fitting Models (4) and (5) to the 100 simulated data are illustrated in Figure 6.6. The scatterplot in figure 6.6 confirms that the complex model indeed yields higher maximum likelihood values as all results lie either on or over the 45 degree line. Intuitively, the added temporal variation in triggering process can only improve the model fit if not the same.

The computed test statistics approximately follow χ^2 distribution with one degrees of freedom, as the number of parameters differ by one between the two models. With the test statistics, we can calculate the empirical power of our test. At 5% significance level, marked with a blue triangle on the histogram in Figure 6.6, we are only able reject the null hypothesis, that there is no temporal effect, 58% of the time, although the data sets were simulated assuming the negation, the alternative hypothesis. The empirical power of the test, 58%, does not look promising. Even if temperature impact the triggering process of the avian influenza virus, our model is able to detect this component only 58% of the time, under the setting that the true parameters are as above.

Although the exact reason for the low empirical power is unknown, under Model (5), the temperature components for both background and triggering processes take the same functional form, and considering the difficulty of estimating α, the scaling parameter for triggering, the estimation methods may not be able to distinguish the two temperature components in the background and triggering processes.

Next, we compare this result to another component in the triggering process, proximity to traffic networks. In Chapter 5, the fit of Model (3), which contains spatial and temporal lags as parts of the triggering process, was greatly enhanced by introducing proximity to nearest traffic network to the branching structure. Calculating the empirical power for this factor will not only measure the ability of our model in determining its presence, but allows us to interpret the previous result for temperature in comparison.

The new statistical test for proximity to traffic network is constructed in the same manner as before. The complex model representing the alternative hypothesis is now Model (4). Based on this model, we simulate 100 data sets with parameters, $\theta^c = (a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma)$ $= (1, 20, 0.1, 50, 40, 70, 0.15)$. The simple model is Model (3), which excludes the distance from the nearest traffic networks as a factor in the branching structure. The set of parameters for Model (3) is $\theta^s = (a, b_{city}, k, \alpha, \beta, \gamma)$. The null hypothesis for this statistical test is, $H_0 : \beta_{road} = 0$, which assumes no effect of proximity to traffic networks in the triggering process. The resulting test statistics will again follow χ^2 distribution with one degree of freedom.

The two plots in Figure 6.7 illustrate the results from fitting Models (3) and (4) to the data simulated assuming the alternative hypothesis with the given parameters. The scatter plot of the estimated maximum log likelihoods for the two models indicates that distribution of the test statistics is likely to shift, compared to the results in Figure 6.6, as more points are

scattered to the right. The distribution of the test statistics confirms the shift, and majority of them are larger than the value of $\chi^2_1(0.95)$. At 5% significance level, the empirical power, the probability of rejecting the null hypothesis when the alternative hypothesis is correct and under the specified parameters, is computed to be 89%, which is remarkably higher than 58%, obtained for temperature variation.

This result suggests the configuration of our model may be limited to correctly detect the change in the branching structure of H5N1 according to the temperature. For our study, understanding how temperature affects the spread of virus is crucial. Although laboratory experiments have shown colder temperature prolongs the life of H5N1 in water, the implication of this result to the mechanism of the virus spread has not been found. A reasonable solution may be to consider a different functional form for the temperature component, since the impact of temperature in background and triggering processes may differ.

6.3 Prediction

In addition to benchmarking performances of estimation methods and examining the EAI model's capability to detect the components in the triggering process, the simulation algorithms presented earlier can be used to make predictions of the future H5N1 outbreaks in Turkey. With simulation, virtually any feature of our point process can be forecasted [61][4]. A value of interest, such as the number of future outbreaks, can be obtained from simulating the EAI self-exciting point process model over a desired time interval in Turkey. By repeatedly simulating the future outbreaks based on the observed, we can obtain a distribution for the value of interest. We are particularly interested in three features of the future outbreaks, their number along with the spatial and temporal distributions of their occurrences. Knowledge of these features would be greatly beneficial to H5N1 surveillance and the prevention of future disease spread. The prediction results for these features will be provided, following the introduction of the prediction algorithm in the next paragraph.

We propose Algorithm 7 to generate the predicted H5N1 outbreaks in Turkey over $(T-p, T]$, where T represents the last day with an observed H5N1 outbreak, and p is a positive integer, denoting the length of the interval. We chose p to be 60, a third of observation period, $[0, T]$, because the predictability will suffer for a long prediction period. The observed numbers of outbreaks in $[0, T-60]$ and $(T-60, T]$ were 130 and 91 respectively. The prediction will use the parameter estimates for Model (4), produced from fitting backfitting method to data observed during $[0, T-60)$: $(\hat{a}\,, \hat{b}_{city}\,, \hat{k}\,, \hat{\alpha}\,, \hat{\beta}\,, \hat{\beta}_{road}\,, \hat{\gamma}\,) = (0.834\,, 25.8\,, 0.0612\,, 16.5\,, 16.9\,, 46.2\,, 0.215\,)$. Comparing the predicted results with the actual outcome will give us a sense of the performance of our prediction algorithm. Algorithm 7 is a modification of Algorithm 5, and we will generate 300 sets of future outbreaks with Algorithm 7. The details of Algorithm 7 is provided below:

[4]Although our research did not take the following approach, Vere-Jones [61] additionally notes that uncertainty in parameter values can be allowed for by adopting a Bayesian framework and selecting the parameters from the posterior distribution before starting the simulation.

Algorithm 7: Prediction algorithm for the EAI model

Step 1. Simulate background point process with intensity $\lambda_B(x, y, t)$ using the parameter estimates for Model (4), produced from fitting backfitting method to data observed during $[0, T-p)$, over $S \times (T-p, T]$. Call this generation 0 and denote it $G^{(0)}$.

Step 2. Simulate one generation of offsprings from the H5N1 outbreaks observed in $[0, T-p]$ using steps 3, 4, and 5 in Algorithm 5. Delete offsprings that do not belong to S. Additionally remove offsprings that occurred during $[0, T-p]$. Include the remaining offsprings to generation 0, $G^{(0)}$.

Step 3. Simulate the future generations of offsprings over $S \times (T-p, T]$ using steps 2, 3, 4, 5, 6, and 7 in Algorithm 5.

The results for the three features produced from 300 simulations are shown in Figures 6.8, 6.9, and 6.10. Figure 6.8 illustrates the distribution of the predicted numbers, while Figures 6.9 and 6.10 show the spatial and temporal distributions of future outbreak occurrences.

Examining Figure 6.8, we found the prediction result for the number of future outbreaks to be satisfactory. Both the mean and the median of the predicted numbers, 89 and 80 respectively, were close to the actual number of outbreaks observed in $(T-60, T\,]$, 91. In addition, the actual number of outbreaks fell between the 25^{th} and 75^{th} percentiles of the distribution, shown as red empty triangles in Figure 6.8 . In general, the distribution of predicted number of outbreaks is skewed to the right with its mode, 65, corresponding to 15% of simulated data sets. The predicted outcome suggests less number of outbreaks than the observed will occur under our model, but this difference is acceptable.

The spatial pattern of future outbreaks shown in Figure 6.9 indicates that the majority of the predicted outbreaks may have been triggered by the few last observed outbreaks in $[0, T-60]$. The areas with high density of outbreak occurrences, especially near Elazig, overlap with the locations of the past outbreaks from the observed data, marked with blue crosses in Figure 6.9[5] The plot also suggests that our prediction approach based on Monte Carlo simulation suffers from the edge effect, unable to predict the majority of the outbreaks occurred near the coastal cities of Samsun and Rize situated on the border. On the contrary, the predicted result matches better with the outbreaks occurring inland.

Turning our attention to the temporal distribution of the predicted outbreaks in Figure 6.10, we notice that the distribution is skewed to the left and is gradually increasing over time. This pattern is an opposite of the temporal trend of outbreaks observed in $(T-60, T]$. In reality, the temporal distribution of outbreaks peaks around 127^{th} and 143^{th} days and tails off as time elapses. A viable explanation for such a difference is the absence of a temporal

[5]Only the past outbreaks occurred in $[T-90, T-60]$ are shown because the triggering function that involves temporal lag tails off to 0 at 30 days.

component in Model (4), which can adjust the strength of triggering in addition to the temporal lag. With Model (4), depending on the rate of decay for the temporal lag component, it is possible for outbreaks to continue triggering other outbreaks for a long time period. Model (5), in comparison, is designed to limit the triggered number of outbreaks with the additional temperature component in the triggering process, although the estimate of the corresponding parameter turned out to be approximately equal to zero.

Overall, the prediction results were satisfactory. Both the mean and the median of predicted number of future outbreaks were close the actual observed number of outbreaks, 91. The observed number of outbreaks also fell between the 25^{th} and 75^{th} percentiles of the predicted distribution. Spatially, future outbreaks simulated under model (4) did not agree with the observed dense clustering near the boundary, due to edge effect. The predicted locations matched better with outbreaks that occurred inland. Temporally, the predicted results were opposite of the observed, exhibiting skewness to the left and increasing over time. The predicted outcome indicated that the risk of future disease spread could last for a long time, under the configuration of model (4).

Lastly, we would like to make a remark regarding comparison of our prediction results to those of the past research, such as Gilbert et al. [22] and Fang et al. [16]. While the prediction map based on the spatial logistical regression models is temporally invariant, our prediction results are highly dependent on the outbreaks observed in the past, as shown in Figures 6.9 and 6.10. We believe that in practice, the prediction results produced from the EAI model are more useful, because it provides an idea of how the virus will spread based on the past progression of outbreaks. The prediction results produced from the spatial logistical regression models could analogous to predicting future outbreaks using only the background process of the EAI model. Besides the prevention of the recurring virus spread, which can be predicted by the background process of our model, prediction results from the triggering process will be helpful in containing the future disease dispersal, after H5N1 outbreaks are observed. The predicted spatial distribution of outbreaks can serve as a guide for establishing an effective disease quarantine.

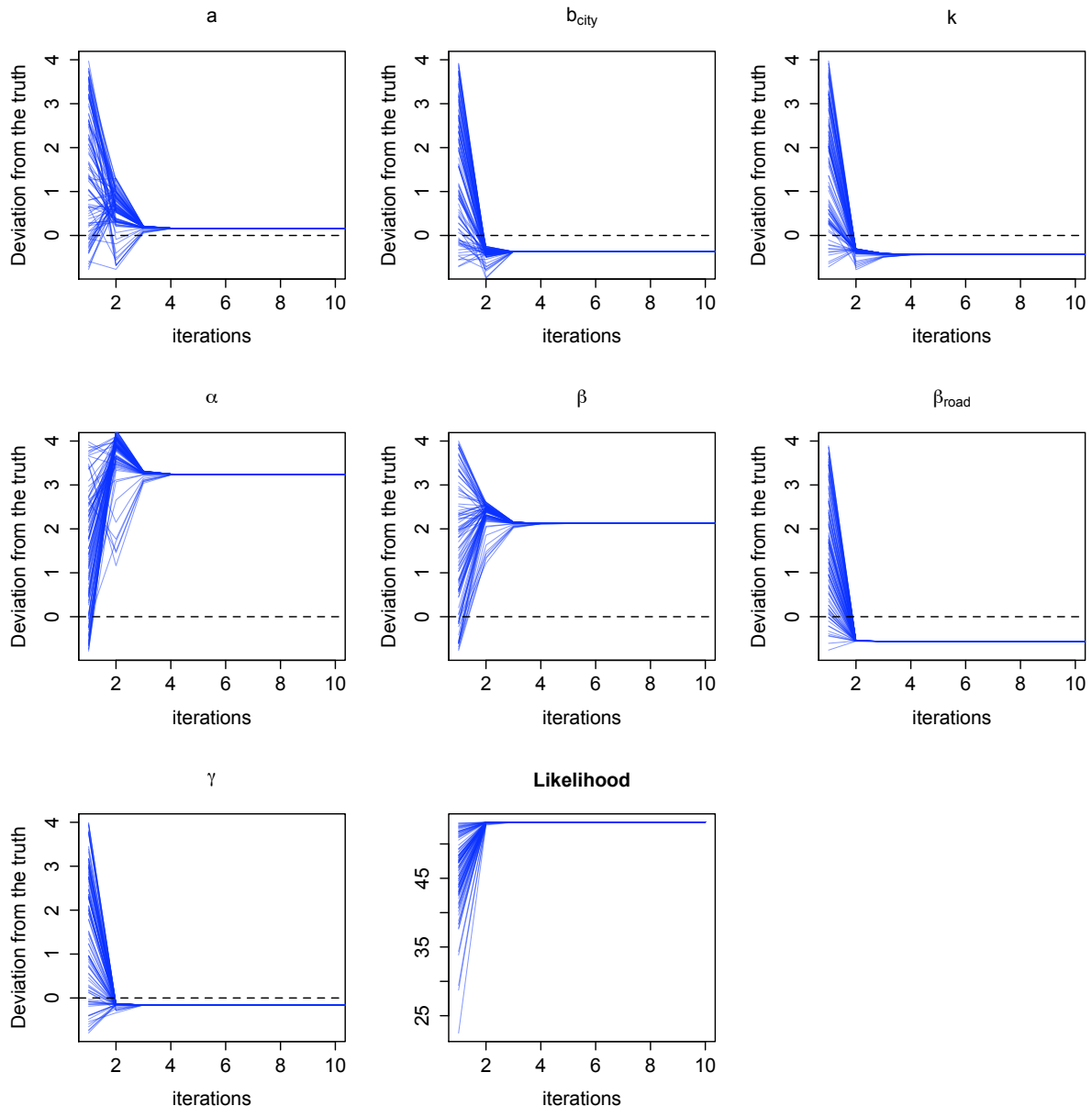

Figure 6.1: Progression of estimates of θ and their associated likelihood values at each iteration using the backfitting algorithm. The data was generated without the edge correction and the starting values were drawn from uniform distribution, whose range is $\frac{1}{5}^{th}$ of the true value to 5 times the true value, per parameter. The black dotted lines mark the true values of the parameters. The scale on y-axis is proportion to the true value of the parameter.

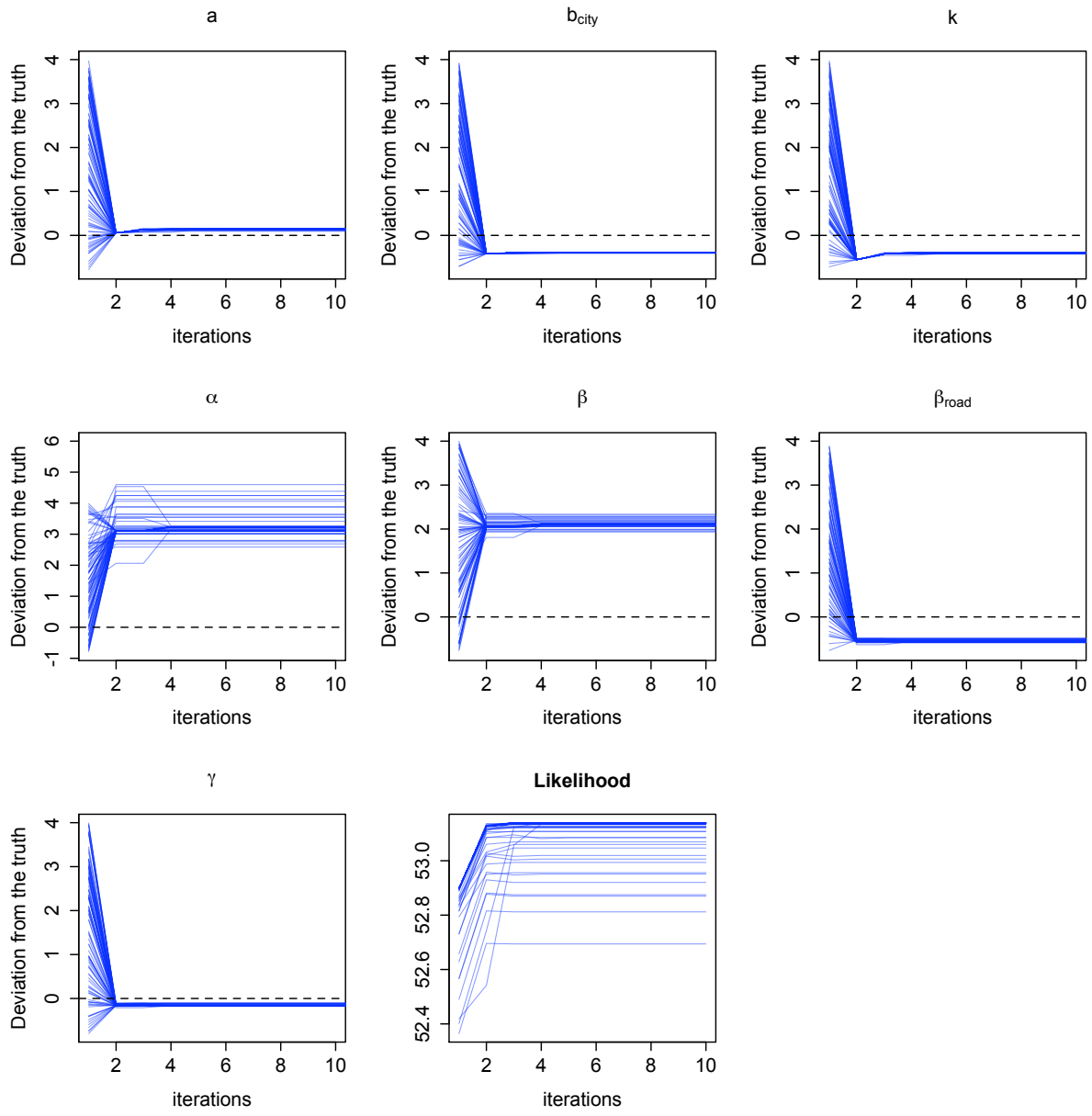

Figure 6.2: Progression of estimates of θ and their associated likelihood values at each iteration using the Poorman's EM algorithm. The data was generated without the edge correction and the starting values were drawn from uniform distribution, whose range is $\frac{1}{5}^{th}$ of the true value to 5 times the true value, per parameter. The black dotted lines mark the true values of the parameters. The scale on y-axis is proportion to the true value of the parameter.

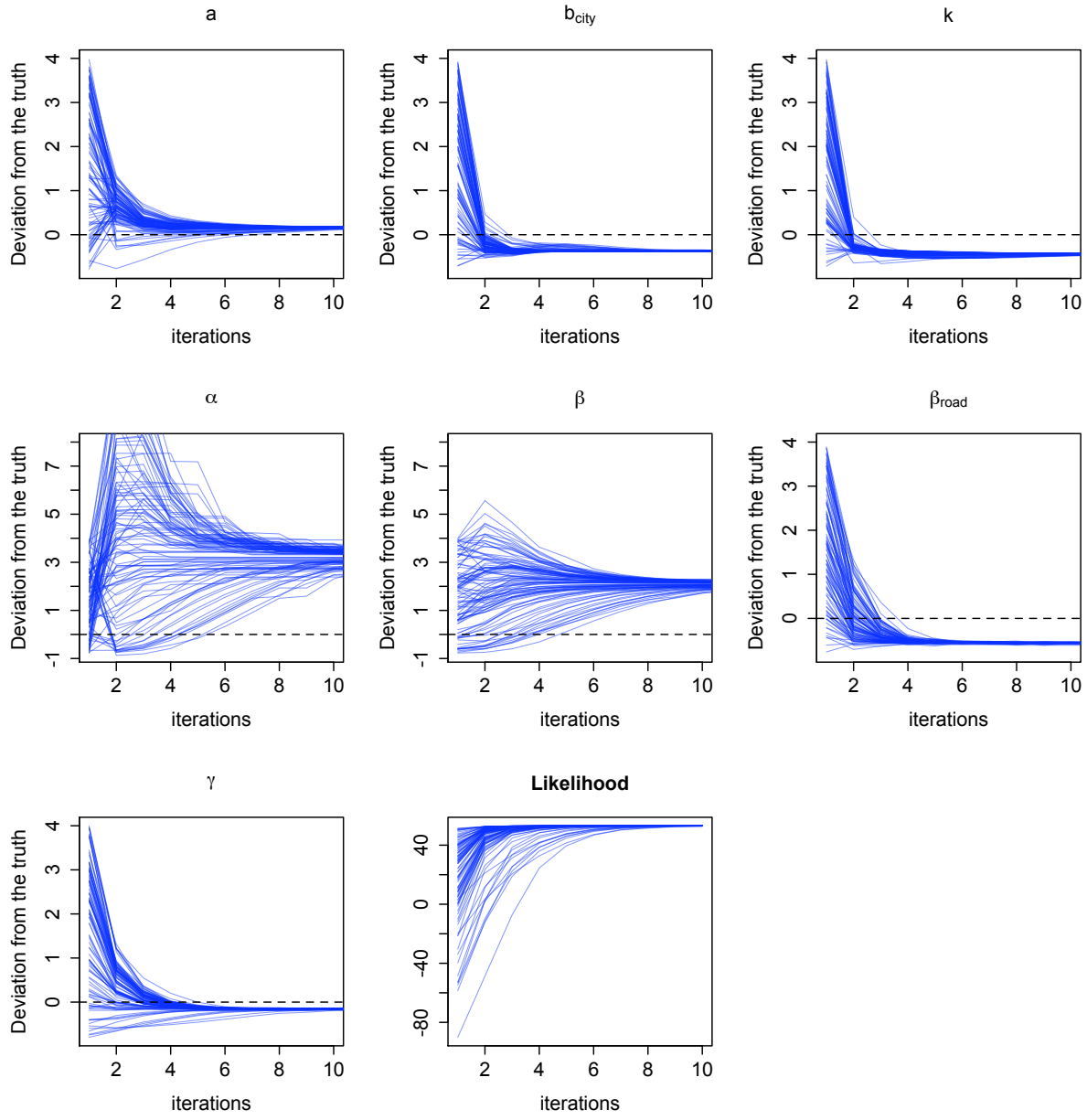

Figure 6.3: Progression of estimates of θ and their associated likelihood values at each iteration using the EM algorithm. The data was generated without the edge correction and the starting values were drawn from uniform distribution , whose range is $\frac{1}{5}^{th}$ of the true value to 5 times the true value, per parameter. The black dotted lines mark the true values of the parameters. The scale on y-axis is proportion to the true value of the parameter.

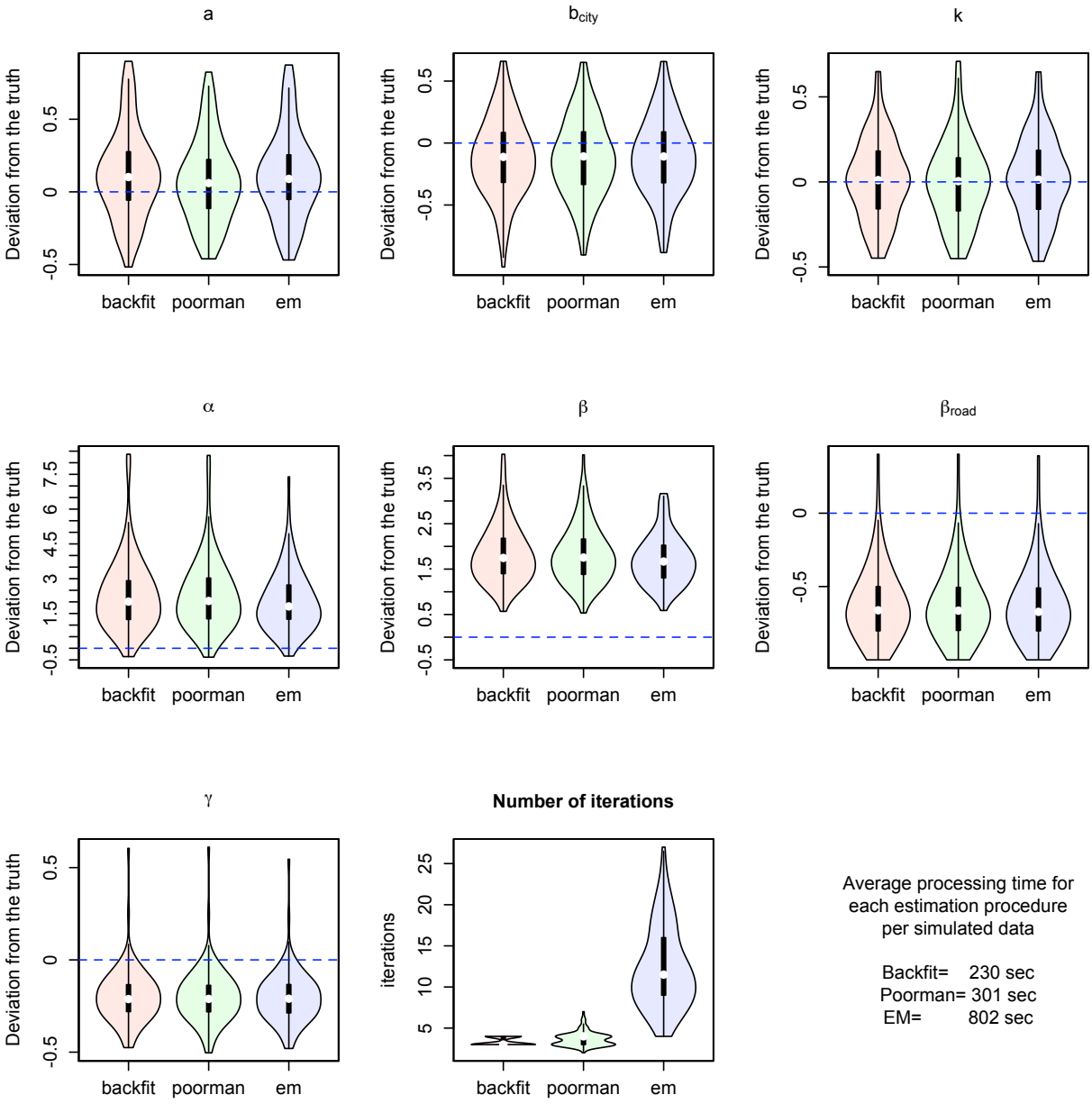

Figure 6.4: Comparison of parameter estimates and the numbers of iterations required according to the three estimation methods, backfitting, poorman's EM, and EM. The plots were generated using a violin plot, a modification of boxplot with density plots of the corresponding distribution shown on the sides. The white dot represents the median and the black box illustrates the IQR (Inner Quartile Range). The first seven plots show the distributions of estimates obtained for $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma)$. The remaining plot displays the distribution of number of iterations took for each run grouped by the three methods.

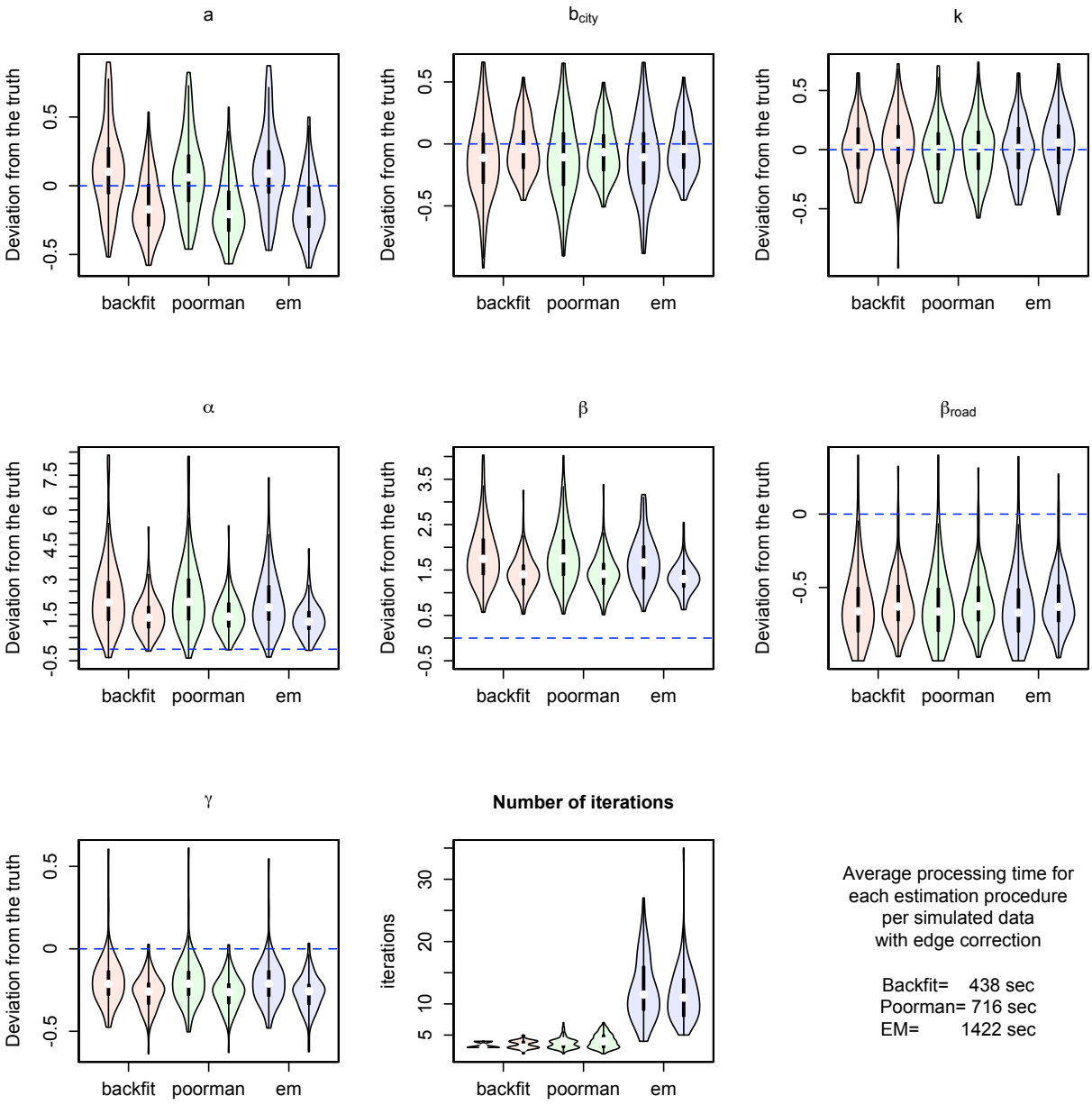

Figure 6.5: Comparison of parameter estimates and the numbers of iterations required according to the three estimation methods: backfitting, poorman's EM, and EM with edge correction. The plots were generated using a violin plot, a modification of boxplot with density plots of the corresponding distribution shown on the sides. The white dot represents the median and the black box illustrates the IQR (Inner Quartile Range). The first seven plots show the distributions of estimates obtained for $(a, b_{city}, k, \alpha, \beta, \beta_{road}, \gamma)$. The new results with edge correction were plotted next to previous outcome represented in figure 6.4. The remaining plot displays the distribution of number of iterations took for each run grouped by the three method both with and without the edge correction.

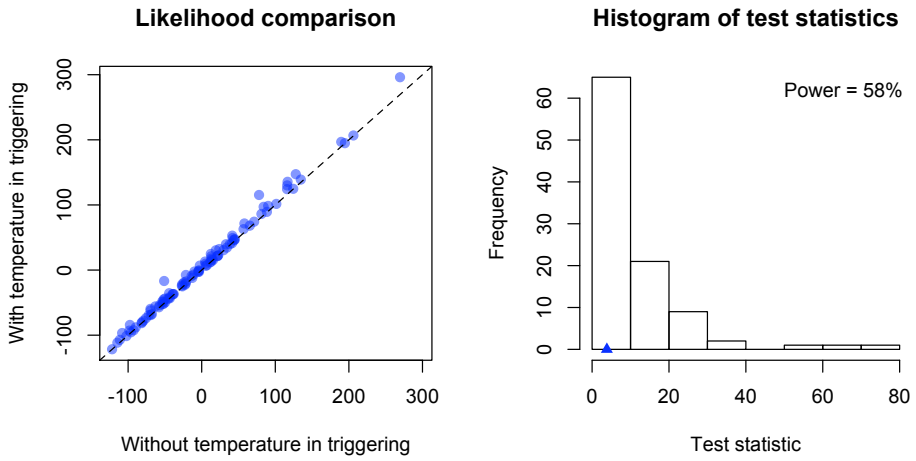

Figure 6.6: The scatter plot on the left shows the relationship between the maximum log likelihood values from fitting 100 simulated data sets including temperature variation in the triggering process, using two models: with and without temperature component in the triggering process from the intensity (4.5). The histogram on the right displays the computed test statistics for the likelihood ratio test. The blue triangle marks $\chi^2_{1(.95)}$.

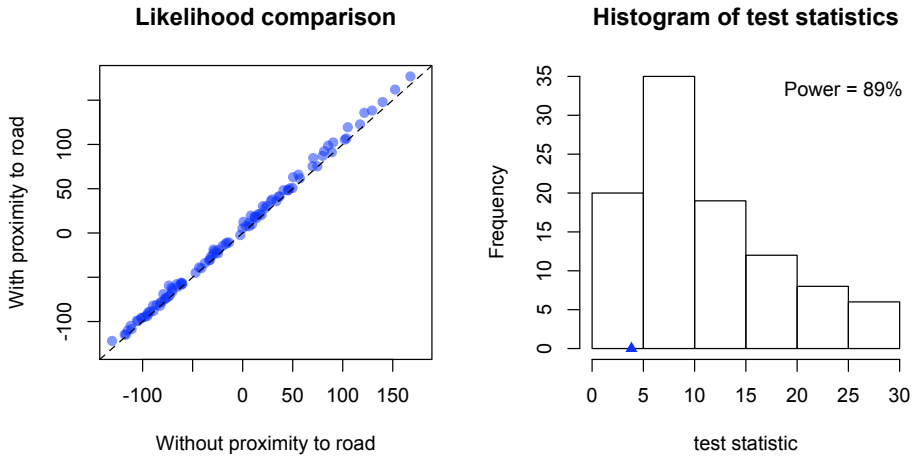

Figure 6.7: The scatter plot on the left shows the relationship between the maximum log likelihood values from fitting 100 simulated data sets assuming model 4, using two models: with and without proximity to nearest traffic network in the triggering process from the intensity (4.5). The histogram on the right displays the computed test statistics for the likelihood ratio test. The blue triangle marks $\chi^2_{1(.95)}$.

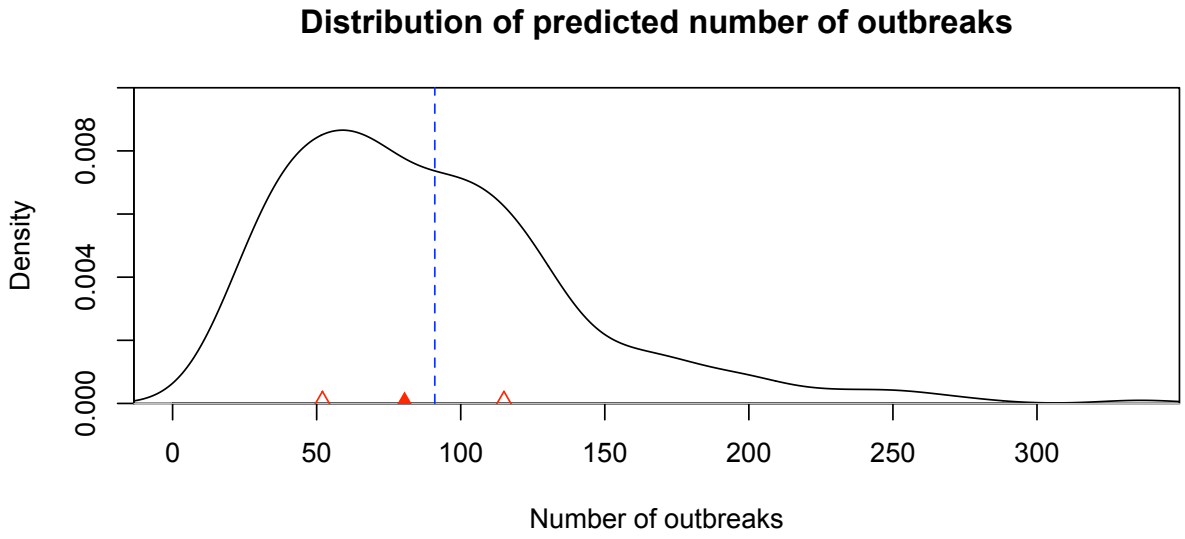

Figure 6.8: Density plot of predicted number of outbreaks from 300 simulated data over $S \times (T-60, T]$. The blue dotted line marks the number of H5N1 observed in reality, 91. The mode of the distribution, 65, is less than the observed. The red filled and empty triangles indicate the median, 80, and the 25^{th} and 75^{th} percentiles, 52 and 115, of the distribution respectively.

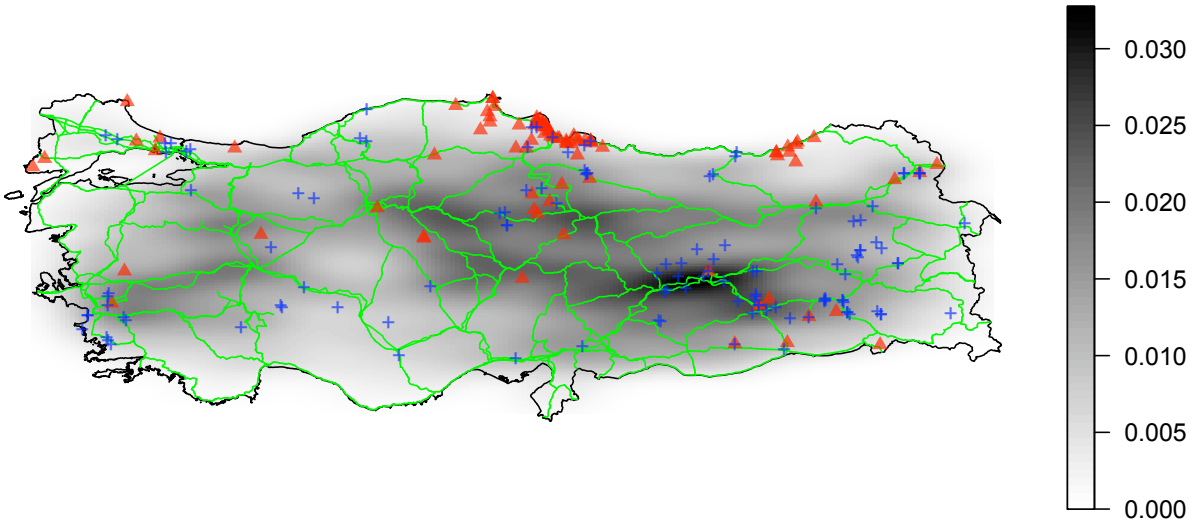

Figure 6.9: Kernel density estimation plot for spatial patterns of predicted outbreaks from 300 simulated data over $S \times (T-60, T]$. The darker areas on the map corresponds to higher density of predicted outbreaks. The red triangles mark the locations of the outbreaks observed in $S \times (T-60, T]$. On the other hand, the blue crosses represent the past outbreak locations occurred in $S \times (T-90, T-60]$. The railroads and major highways are shown in green.

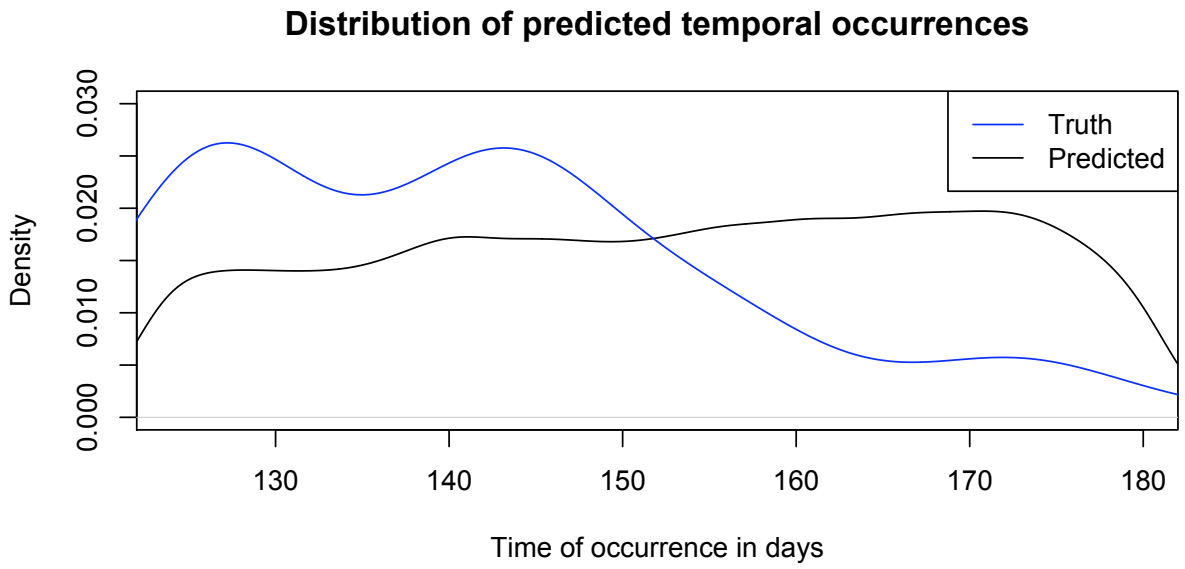

Figure 6.10: Density plot of predicted occurrence times of outbreaks from 300 simulated data over $S \times (T-60, T]$. The blue line corresponds to the density of the actual time of occurrences observed in $(T-60, T]$. The black line represents the density of predicted temporal occurrences.

Part V

Conclusion

Chapter 7

Conclusion

As mentioned in the first chapter of this thesis, our work on spread of avian influenza (H5N1) had three goals: first, using exploratory data analysis, to investigate the mechanism of the H5N1 spread and determine the key factors that contribute to its explosiveness; second, to develop a statistical model based on point process to model the past progression of the disease in Turkey; third, to develop an algorithm using our estimated statistical model to predict the future disease spread conditioned on the past observations of the Turkish H5N1 outbreaks.

Our explanatory data analysis in Chapter 2 demonstrated that spatially, proximity to the nearest traffic networks and cities are both influential factors of the disease spread mechanism. This observation agrees with the fact that poultry farms tend to be located near traffic networks and cities for easy access to markets [65]. Temporally, the global seasonality was also observed in Turkey, and temperature was found to be associated with this trend. Laboratory results confirm temperature as one of the key factors of H5N1 spread, noting colder temperature extends the infectivity of the avian influenza virus in water [6, 7]. Due to insufficient evidence, we decided to exclude the role of migratory birds in our temporal analysis.

Using these identified factors as predictors, we proposed our EAI (Epidemic Avian Influenza) model in Chapter 4 to quantify the relationship between the factors and the Turkish H5N1 spread. The EAI model is inspired by the ETAS (Epidemic Type Aftershock Squence) model, an extension of the self-exciting point process [24], which allows one to incorporate temporal trend and spatio-temporal dependencies of the H5N1 outbreaks, excluded from statistical analyses in past studies [22, 16] discussed in Chapter 3. Among the five variations of EAI models we considered, shown in Table 4.1, Model (4) demonstrated the best fit in terms of the AIC (Akaike Information Criterion) score. Model (4) assumes that the outbreaks tend to occur near the cities and in a colder temperature, and these outbreaks trigger other outbreaks at nearby locations and times, forming clusters of outbreaks along traffic networks. Analyzing the residuals of the fitted models in Chapter 5, Model (4) was shown to improve upon the second best model, Model (2), which does not include a triggering process.

In Chapter 6, we used simulation to assess the performance of the three proposed estimation methods for the EAI model: backfitting, poorman's EM (Expectation - Maximization),

and EM. The results from simulating and estimating the parameters of Model (4) indicated that the backfitting algorithm is able to achieve the fastest convergence, closely followed by the poorman's algorithm. While the EM algorithm was remarkably slow in comparison, it was found to produce more accurate estimates than the other two methods, especially for the two troublesome parameters, α and β. The resulting estimates for α and β—corresponding to scale and spatial lag parameters in the triggering process—were heavily biased compared to the other parameter estimates. In order to cope with this bias, we implemented edge correction in our simulation, and while the accuracy greatly improved for these parameter estimates, the edge correction introduced a slight bias to other parameter estimates.

Moreover, using the simulation results, we investigated whether our model is able to successfully detect the individual component in the trigger process by constructing a likelihood ratio test. The empirical power calculated for effect of temperature to the triggering process suggested that our model struggles to determine its presence.

A further application of the simulation algorithms is prediction of the future outbreaks, discussed in Section 6.3. Based on the past progression of H5N1 outbreaks, we simulated and obtained distribution of number of predicted outbreaks, their locations and times of occurrences via Algorithm 7, for 60 days before the last observed H5N1 outbreak in Turkey. The prediction results agreed with the actual H5N1 spread observed during this time period with the mean and the median of numbers of predicted outbreaks closely matching the observed.

While the overall results of our analysis are satisfactory, there are areas in need of further research. The functional form of the temperature component in the triggering process can be reconsidered with possible addition of a threshold, similar to the magnitude threshold for the ETAS model described in Section 4.1.1. As shown in Figures 2.2 and 2.3, number of outbreaks seemes to increase when the temperature drops below 0°C. In addition, temperature may not have log-linear relationship with the number of outbreaks, and the empirical power calculation in Section 6.2.4 confirmed that the current model is not able to properly detect the temperature component in the triggering process, even when it is present.

In Chapter 6, fitting our EAI model to simulated data using edge correction generally produced more accurate parameters than simulated data without. Likewise, we could further improve our estimation results by applying edge correction in the estimation procedure. Fitting the EAI model to the data in extended space and time may produce better estimation results.

Another area of potential improvement is prediction. As mentioned before, uncertainty in parameter values can be allowed by adopting a Bayesian framework and selecting the parameters from the posterior distribution before starting the prediction procedure [61].

This thesis is a product of our effort to contribute to the ongoing H5N1 surveillence. We hope our results provide a new perspective on analysis of avian influenza (H5N1) and statistical modeling strategies involving self-exiciting point process.

Appendix A

Maximum Likelihood estimation

A.1 Mollweide Projection

Mollweide projection is a popular area preserving projection method for depicting the surface of the Earth. Our study area, Turkey, with Mollweide projection applied is shown in Figure A.1. The locations of centers of the squares in the grid is first determined using the Mollweide projection. The coordinates of the centers are projected back into longitude and latitude, which allows to calculate the geodesic distance between a center of a grid and a given location in Turkey. This re-projection is illustrated in Figure A.2.

With this construction, the grid has 10,537 centers with corresponding area of 74.98 km^2. The approximated area of Turkey computed using the grid is 790,076.3 km^2. This result is very close to the actual area of Turkey, 783,562 km^2.

Figure A.1: Molleweide projection with the dots indicating the center of squares in the grid.

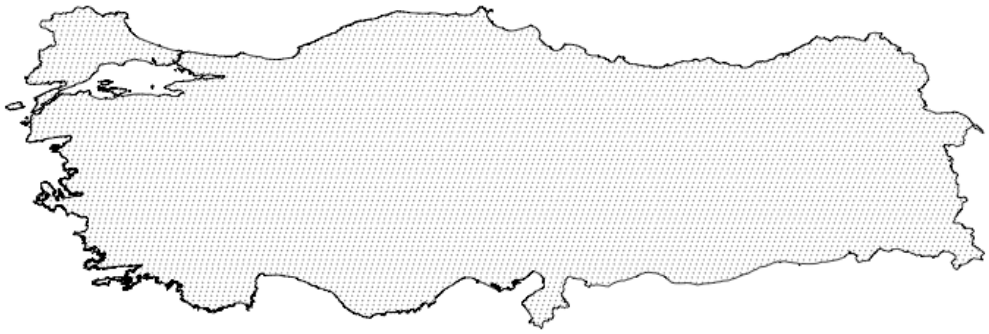

Figure A.2: Re-projection of figure A.1 into longitude and latitude.

A.2 Derivation of integral 5.7

In this section, we present the derivation of integral shown in equation 5.7. The expression

$$\int_0^T \sum_{i:t_i<t} k(T(t_i))\, g(t-t_i)dt \tag{A.1}$$

can be broken into integrals with smaller time intervals, $(t_i, t_{i+1}]$, for $(1 \leq i \leq N-1)$:

$$\int_0^{t_1} 0dt + \int_{t_1}^{t_2} k(T(t_1))\, g(t-t_1)dt + \int_{t_2}^{t_3} k(T(t_1))\, g(t-t_1) + k(T(t_2))\, g(t-t_2)dt \tag{A.2}$$

$$+ \cdots + \int_{t_{n-1}}^{t_n} [k(T(t_1))\, g(t-t_1) + \cdots + k(T(t_{N-1}))\, g(t-t_{N-1})]dt. \tag{A.3}$$

Gathering $k(T(t_i))\, g(t - t_i)$ for each i yields:

$$= \left[\int_{t_1}^{t_2} k(T(t_1))\, g(t - t_1) dt + \cdots + \int_{t_n}^{t_n - 1} k(T(t_1))\, g(t - t_1) dt \right] \tag{A.4}$$

$$+ \left[\int_{t_2}^{t_3} k(T(t_2))\, g(t - t_2) dt + \cdots + \int_{t_n}^{t_n - 1} k(T(t_2))\, g(t - t_2) dt \right] \tag{A.5}$$

$$\vdots \tag{A.6}$$

$$+ \int_{t_N - 1}^{t_N} k(T(t_{N-1}))\, g(t - t_{N-1}) dt \tag{A.7}$$

$$= \int_{t_1}^{T} k(T(t_1))\, g(t - t_1) dt + \int_{t_2}^{T} k(T(t_2))\, g(t - t_2) dt \tag{A.8}$$

$$\vdots \tag{A.9}$$

$$+ \int_{t_{N-1}}^{T} k(T(t_{N-1}))\, g(t - t_{N-1}) dt \tag{A.10}$$

$$= \sum_{i=1}^{N} \int_{t_i}^{T} k(T(t_i))\, g(t - t_i) dt. \tag{A.11}$$

Bibliography

[1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716 – 723, December 1974.

[2] N. H. Augustin, M. A. Mugglestone, and S. T. Buckland. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33(2):339–347, 1996.

[3] A. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes (with discussion). *Journal Of The Royal Statistical Society Series B*, 67(5):617–666, 2005.

[4] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):pp. 580–598, 1985.

[5] Anders Brix and Wilfrid S. Kendall. Simulation of cluster point processes without edge effects. *Advances in Applied Probability*, 34(2):pp. 267–280, 2002.

[6] Justin D. Brown, Ginger Goekjian, Rebecca Poulson, Steve Valeika, and David E. Stallknecht. Avian influenza virus in water: Infectivity is dependent on ph, salinity and temperature. *Veterinary Microbiology*, 136(1-2):20 – 26, 2009.

[7] Justin D. Brown, David E. Swayne, Robert J. Cooper, Rachel E. Burns, and David E. Stallknecht. Persistence of h5 and h7 avian influenza viruses in water. *Avian Diseases*, 51(s1):285–289, 2007.

[8] Richard H. Byrd, Peihuang Lu, Peihuang Lu, Jorge Nocedal, Jorge Nocedal, Ciyou Zhu, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208, 1994.

[9] Center Disease Control and Prevention. Key facts about avian influenza (bird flu) and avian influenza a (H5N1) virus, May 2007. `http://www.cdc.gov/flu/avian/gen-info/facts.htm`.

[10] D. R. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[11] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes, Volume 1*, volume I. Springer, 2002.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):pp. 1–38, 1977.

[13] Peter Diggle. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 153(3):pp. 349–362, 1990.

[14] Peter Diggle. Spatio-temporal point processes, partial likelihood, foot and mouth disease. *Statistical Methods in Medical Research*, 15(4):325–336, 2006.

[15] Peter Diggle, Barry Rowlingson, and Ting-li Su. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics*, 16(5):423–434, 2005.

[16] Li-Qun Fang, Sake J. de Vlas, Song Liang, Caspar W. N. Looman, Peng Gong, Bing Xu, Lei Yan, Hong Yang, Jan Hendrik Richardus, and Wu-Chun Cao. Environmental factors contributing to the spread of H5N1 avian influenza in mainland china. *PLoS ONE*, 3(5):e2268, 05 2008.

[17] Chris J. Feare. The role of wild birds in the spread of HPAI H5N1. *Avian Diseases*, 51(s1):440–447, 2007.

[18] Chris J. Feare. Role of wild birds in the spread of highly pathogenic avian influenza virus H5N1 and implications for global surveillance. *Avian Diseases*, 54(s1):201–212, 2010.

[19] Food and Agriculture Organization of the United Nations. Asian livestock to the year 2000 and beyond, 1999. `http://www.fao.org/docrep/003/x6624e/x6624e00.htm`.

[20] M. Gauthier-Clerc, C. Lebarbenchon, and F. Thomas. Recent expansion of highly pathogenic avian influenza H5N1: a critical review. *Ibis*, 149:202–214(13), April 2007.

[21] M. Gilbert, X. Xiao, J. Domenech, J. Lubroth, V. Martin, and J. Slingenbergh. Anatidae migration in the western palearctic and spread of highly pathogenic avian influenza H5N1 virus. *Emerg Infect Dis*, 12(11):1650–6, 2006.

[22] M. Gilbert, X. Xiao, D. U. Pfeiffer, M. Epprecht, S. Boles, C. Czarnecki, P. Chaitaweesub, W. Kalpravidh, P. Q. Minh, M. J. Otte, V. Martin, and J. Slingenbergh. Mapping H5N1 highly pathogenic avian influenza risk in southeast asia. *Proc.Natl.Acad.Sci.U.S.A*, 105(12):4769–4774, March 2008.

[23] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*. London: Chapman & Hall, 1990.

[24] Alan Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[25] D. W. Hosmer, T. Hosmer, S. Le Cessie, and S. Lemeshow. A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in Medicine*, 16(9):965–980, 1997.

[26] BirdLife International. Birdlife statement on avian influenza, August 2007. `http://www.birdlife.org/action/science/species/avian_flu/`.

[27] BirdLife International. Slender-billed curlew numenius tenuirostris, July 2010. `http://www.birdlife.org/datazone/speciesfactsheet.php?id=3011`.

[28] BirdLife International. White-headed duck oxyura leucocephala, 2010. `http://www.birdlife.org/datazone/speciesfactsheet.php?id=359`.

[29] Elsa Jourdain, Michel Gauthier-Clerc, and Philippe Sabatier. Ecoregional dominance in spatial distribution of avian influenza (H5N1) outbreaks: In response. Emerging Infectious Diseases, Aug 2007.

[30] Andrew B. Lawson and Petra Leimich. Approaches to the space-time modelling of infectious disease behaviour. *Mathematical Medicine and Biology*, 17(1):1–13, 2000.

[31] P. A. W. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3):403–413, 1979.

[32] Miguel A. Martnez-Beneito, Juan J. Abellán, Antonio López-Qulez, Hermelinda Vanaclocha, Óscar Zurriaga, Guillermo Jorques, and José Fenollar. Source detection in an outbreak of legionnaires disease. In P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin, S. Zeger, Adrian Baddeley, Pablo Gregori, Jorge Mateu, Radu Stoica, and Dietrich Stoyan, editors, *Case Studies in Spatial Point Process Modeling*, volume 185 of *Lecture Notes in Statistics*, pages 169–182. Springer New York, 2006.

[33] Susan Mayor. Link between H5N1 in UK and recent outbreaks in hungary is investigated. *BMJ*, 334(7589):335, 2007.

[34] Sebastian Meyer, Johannes Elias, and Michael Hhle. A space-time conditional intensity model for infectious disease occurence. Technical report, University of Munich, 2010.

[35] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106(493):100–108, 2011.

[36] Jesper Møller and Jakob Rasmussen. Approximate simulation of hawkes processes. *Methodology and Computing in Applied Probability*, 8:53–64, 2006. 10.1007/s11009-006-7288-z.

[37] Jesper Møller and Jakob G. Rasmussen. Perfect simulation of hawkes processes. *Advances in Applied Probability*, 37(3):pp. 629–646, 2005.

[38] BBC News. China, Indonesia and African nations are under-reporting incidences of bird flu, according to the world organization for animal health (OIE), May 2006. http://news.bbc.co.uk/2/hi/science/nature/5034276.stm.

[39] Tung Nguyen, C. Todd Davis, William Stembridge, Bo Shu, Amanda Balish, Kenjiro Inui, Hoa T. Do, Huong T. Ngo, Xiu-Feng Wan, Margaret McCarron, Stephen E. Lindstrom, Nancy J. Cox, Cam V. Nguyen, Alexander I. Klimov, and Ruben O. Donis. Characterization of a highly pathogenic avian influenza h5n1 virus sublineage in poultry seized at ports of entry into vietnam. *Virology*, 387(2):250 – 256, 2009.

[40] United States Department of Agriculture Foreign Agriculture Serivice. Global agricultural information network report: Turkey, January 2010. http://gain.fas.usda.gov/Recent%20GAIN%20Publications/Poultry%20update_Ankara_Turkey_1-29-2010.pdf.

[41] United States Library of Congress. Country profile - Turkey, January 2006. http://www.unhcr.org/refworld/docid/46f9135d0.html.

[42] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, pages 9–27, 1988.

[43] Yosihiko Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, June 1998.

[44] Yosihiko Ogata and Jiancang Zhuang. Space-time etas models and an improved extension. *Tectonophysics*, 413(1-2):13 – 23, 2006. Critical Point Theory and Space-Time Pattern Formation in Precursory Seismicity.

[45] World Health Organization. Cumulative number of confirmed human cases of avian influenza A/(H5N1) reported to who, July 2010. http://www.who.int/csr/disease/avian_influenza/country/cases_table_2010_07_29/en/index.html.

[46] World Health Organization. H5N1 avian influenza: timeline of major events, October 2010. http://www.who.int/csr/disease/avian_influenza/ai_timeline/en/index.html.

[47] J. S. Malik Peiris, Menno D. de Jong, and Yi Guan. Avian Influenza Virus (H5N1): a Threat to Human Health. *Clin. Microbiol. Rev.*, 20(2):243–267, 2007.

[48] Roger D. Peng, Frederic Paik Schoenberg, and James Woods. A space-time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, 100:26–35, 2005.

[49] Margaret Sullivan Pepe. Receiver operating characteristic methodology. *Journal of the American Statistical Association*, 95(449):pp. 308–311, 2000.

[50] Stephen L. Rathbun. Asymptotic properties of the maximum likelihood estimator for spatio-temporal point processes. *Journal of Statistical Planning and Inference*, 51(1):55 – 74, 1996.

[51] B. D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):pp. 255–266, 1976.

[52] Ida Scheel, Magne Aldrin, Arnoldo Frigessi, and Peder A Jansen. A stochastic model for infectious salmon anemia (ISA) in Atlantic salmon farming. *Journal of The Royal Society Interface*, 4(15):699–706, 2007.

[53] Frederic P. Schoenberg. Transforming spatial point processes into poisson processes. *Stochastic Processes and their Applications*, 81(2):155 – 164, 1999.

[54] Frederic P. Schoenberg. Multidimensional residual analysis of point process models for earthquake occurrences. *Journal of the American Statistical Association*, 98(464):789–795, 2003.

[55] Frederic P. Schoenberg. Consistent parametric estimation of the intensity of a spatial-temporal point process. *Journal of Statistical Planning and Inference*, 128(1):79 – 93, 2005.

[56] Kyoko Shinya, Masahito Ebina, Shinya Yamada, Masao Ono, Noriyuki Kasai, and Yoshihiro Kawaoka. Avian flu: Influenza virus receptors in the human airway. *Nature Publishing Group*, 2006.

[57] John P. Snyder. *Map Projections: A Working Manual*. Geological Survey (U.S.), 1987.

[58] D. E. Stallknecht, M. T. Kearney, S. M. Shane, and P. J. Zwank. Effects of ph, temperature, and salinity on persistence of avian influenza viruses in water. *Avian Diseases*, 34(2):412–418, 1990.

[59] Dietrich Stoyan and Pavel Grabarnik. Second-order characteristics for stochastic structures connected with gibbs point processes. *Mathematische Nachrichten*, 151(1):95–100, 1991.

[60] Alejandro Veen and Frederic P. Schoenberg. Estimation of spacetime branching process models in seismology using an emtype algorithm. *Journal of the American Statistical Association*, 103:614–624, June 2008.

[61] David Vere-Jones. Some models and procedures for space-time point processes. *Environmental and Ecological Statistics*, 16:173–195, 2009.

[62] China View. UN: Migratory birds not major cause of flu transmission, November 2006. `http://news.xinhuanet.com/english/2006-11/02/content_5283058.htm`.

[63] Robert G. Webster, Maya Yakhno, Virginia S. Hinshaw, William J. Bean, and K. Copal Murti. Intestinal influenza: Replication and characterization of influenza viruses in ducks. *Virology*, 84(2):268 – 278, 1978.

[64] S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):pp. 60–62, 1938.

[65] William Wint and Timothy Robinson. Gridded livestock of the world, 2007. `http://www.fao.org/docrep/010/a1259e/a1259e00.htm`.

[66] Jiancang Zhuang, Yosihiko Ogata, and David Vere-jones. Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research*, 109, 2004.

[67] Jiancang Zhuang, Yosihiko Ogata, and David Vere-Jones. Diagnostic analysis of space-time branching processes for earthquakes. In Adrian Baddeley, Pablo Gregori, Jorge Mateu, Radu Stoica, and Dietrich Stoyan, editors, *Case Studies in Spatial Point Process Modeling*, volume 185 of *Lecture Notes in Statistics*, pages 275–292. Springer New York, 2006.