MODELING LONGITUDINAL INTERRUPTION DATA FROM TURKISH ELECTRICITY DISTRIBUTION COMPANIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZÜLFİYE EBRU KORKMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

MAY 2019

Approval of the thesis:

MODELING LONGITUDINAL INTERRUPTION DATA FROM TURKISH ELECTRICITY DISTRIBUTION COMPANIES

submitted by **ZÜLFİYE EBRU KORKMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Ayşen Akkaya
Head of Department, **Statistics** _____

Prof. Dr. Özlem İlk Dağ
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Prof. Dr. İnci Batmaz
Statistics, METU _____

Prof. Dr. Özlem İlk Dağ
Statistics, METU _____

Assoc. Prof. Dr. Ceylan Talu Yozgatlıgil
Statistics, METU _____

Assist. Prof. Dr. Fulya Gökalp Yavuz
Statistics, METU _____

Assoc. Prof. Dr. Könül Bayramoğlu Kavlak
Actuarial Science, Hacettepe University _____

Date: 28.05.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Zülfiye Ebru Korkmaz

Signature:

ABSTRACT

MODELING LONGITUDINAL INTERRUPTION DATA FROM TURKISH ELECTRICITY DISTRIBUTION COMPANIES

Korkmaz, Zülfiye Ebru
Master of Science, Statistics
Supervisor: Prof. Dr. Özlem İlk Dağ

May 2019, 145 pages

In recent years, many developments have been implemented by the players of the sector to provide sustainable energy flow in Turkey. One of them is the obligation of recording electricity interruption statistics. The Turkish energy regulatory compels new rules to local electricity distribution companies about recording their interruption statistics, including the reasons for electricity interruption, after 2003. However, all of the local distribution companies do not use the same standard to record these statistics. This situation causes complexities for decision makers and researchers for modeling electricity interruptions. In this study, we aimed to find appropriate longitudinal models for the dataset of electricity interruptions. However, the observed data in this study is discrete count type and most of them are zero. Markov Chain Monte Carlo Generalized Linear Mixed Models (abbreviated MCMCglmm), especially the type of zero-inflated and hurdle could be appropriate for these type of data. Therefore, Poisson, zero-inflated Poisson, and hurdle-Poisson distributed models were implemented to a real electricity interruption count dataset belonging to Çankırı in this study. The models have been implemented by using MCMCglmm package in R. To compare the models, Deviance Information Criteria (DIC) and posterior predictive checks were used. Geweke-Halfwidth and Heiderberger-Welch diagnostic tests were used to detect convergence and stationary status of the models. Despite the excessive

zero in the dataset, it was observed that Poisson MCMCglmm estimates were better than the models of zero-inflated Poisson and hurdle Poisson MCMCglmm. Furthermore, Poisson MCMCglmm gave better estimation results in shorter computational time as well.

Keywords: Markov Chain Monte Carlo Generalized Linear Mixed Models, Poisson, Zero-Inflated, Hurdle, Turkish Electricity Distribution Companies, Interruption Statistics.

ÖZ

TÜRK ELEKTRİK DAĞITIM ŞİRKETLERİNE AİT UZUNLAMASINA ELEKTRİK KESİNTİSİ VERİLERİNİN MODELLENMESİ

Korkmaz, Zülfiye Ebru
Yüksek Lisans, İstatistik
Tez Danışmanı: Prof. Dr. Özlem İlk Dağ

Mayıs 2019, 145 sayfa

Türkiye'de, enerji sektörü oyuncuları tarafından son yıllarda kesintisiz enerji akışını sağlamak için geliştirilen birçok yeni uygulama bulunmaktadır. Bunlardan biri elektrik kesinti istatistiklerinin tutulması zorunluluğudur. Enerji Piyasası Denetleme Kurumu, Elektrik Dağıtım Şirketlerine 2003 yılı sonrası için, nedenleriyle birlikte, kesinti istatistiklerinin tutulması zorunluluğunu getirmiştir. Ancak, tüm elektrik şirketleri, bu kesinti istatistiklerinin tutulmasında aynı standardı kullanmamaktadır. Bu durum, karar verici ve araştırmacılar için modellemeyi karmaşık hale getirmektedir. Bu çalışmada, elektrik kesintisi verileri için uygun uzunlamasına modellerin bulunmasını amaçladık. Ancak, bu çalışmada, gözlenen veri kesinti-sayı tipindedir ve verinin çoğu da sıfırdır. Bu tip veriler için Monte Carlo Markov Zinciri Genelleştirilmiş Doğrusal Karma modeller, özellikle de sıfır arttırılmış ve engelli modeller uygundur. Bu çalışmada, Poisson, sıfırı arttırılmış Poisson ve engelli-Poisson dağılım modelleri Çankırı'ya ait gerçek bir kesikli elektrik kesinti verisine uygulanmıştır. Modeller R'da bulunan MCMCglmm paketi kullanılarak uygulanmıştır. Modellerin karşılaştırılması için, sapma bilgi kriteri (DIC) ve sonraki tahmin kontrolleri kullanılmıştır. Modellerin yakınsama ve durağanlığını tespit edebilmek için Geweke- Halfwidth ve Heiderberger-Welch tanımlama testleri de ayrıca kullanılmıştır. Veri kümesinde aşırı sıfır gözlemlenmesine rağmen, Poisson

MCMCglmm, sıfır arttırılmış Poisson MCMCglmm ve engelli Poisson MCMCglmm modellerinden daha iyi tahminler vermektedir. Ayrıca, Poisson MCMCglmm'in daha iyi tahminleri daha kısa bir sürede verdiğini de belirtmek gerekir.

Anahtar Kelimeler: Monte Carlo Markov Zinciri, Genelleştirilmiş Doğrusal Karma Modeller, Türkiye Elektrik Dağıtım Şirketleri, Arıza İstatistikleri, Sıfırı Arttırılmış Poisson model, Engelli Poisson model.

To my lovely, pretty and unique family, Havva, Mehmet and Caner KORKMAZ

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my unique advisor Prof. Dr. Özlem İLK DAĞ for the continuous support of my master study and related research, for her endless patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my master study.

I would like to thank to my thesis committee: Prof. Dr. İnci Batmaz, Assoc. Prof. Dr. Ceylan Yozgatlıgil, Assist. Prof. Dr. Fulya Gökalp Yavuz and Assoc. Prof. Dr. Könül Bayramoğlu Kavlak. Their good comments provided me to see different perspectives of my research.

I am thankful to Jarrod Hadfield and Fitzmaurice Garret for sharing their knowledge about related academic topics of this thesis.

I would like to thank to my previous company, Eltemtek A.Ş. I was able to find this topic with my experiences in there. Also, I am grateful to my current company Havelsan A.Ş. and the leader of my department: Hale Yağlıcı for valuable understanding.

I am also grateful to the following my colleagues: Nilüfer Öktem, Fatıma Nur Çolakoğlu, Nurettin Özdemiroğlu and Aysun Bal for their unfailing support and their motivation.

I would like to express my thanks to my precious friends: Cande Kurt, Gülşah Serdar and Özge Tüzer. The most valuable things of my master study is to find them. I am so lucky to have your friendship.

I want to mention about my grandmother: Zülfiye KORKMAZ. Unfortunately, I could not have had any chance to see her. However, I would like to thank her for giving me a best father ever of the world and her name. My family and I always remember her in our prays. I dedicate this study for her dear soul as well.

Finally, I must express my very profound gratitude my beloved family, especially to my parents Havva, Mehmet, my wonderful brother Caner and to my husband Selçuk for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. I love and thank you to all.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

AIC: Akaike Information Criterion

BEDAŞ: Başkent Electricity Distribution Company

BIC: Bayesian Information Criterion

DIC: Deviance Information Criterion

EİEİ: Elektrik İşleri Etüt İdaresi

EMO: Elektrik Mühendisleri Odası

EMRA: The Energy Market Regulatory Authority of Turkey

ENS: Energy Not Supplied

GLM: Generalized Linear Model

GLMM: Generalized Linear Mixed Model

HBM: Hierarchical Bayesian Model

KS: Kolmogorov-Smirnov

LMM: Linear Mixed Effect Model

LV: Low Voltage

MC: Monte Carlo

MCMC: Markov Chain Monte Carlo

MCMCglmm: Markov Chain Monte Carlo Generalized Linear Mixed Model

ML: Maximum Likelihood

MLMC: Multilevel Monte Carlo

MTTF: Mean Total Time to Failure

MTTR: Mean Total Time to Repair

MV: Medium Voltage

REML: Restricted Maximum Likelihood

RT: Restoration Time

SAIDI: System Average Interruption Duration Index

SAIFI: System Average Interruption Frequency Index

SDE: Stochastic Differential Equation

std: Standardized

TEAŞ: Türkiye Elektrik Üretim İletim A.Ş.

TEDAŞ: Türkiye Elektrik Dağıtım A.Ş.

TLP: Total Loss of Power

VIF: Variance Inflation Factor

ZIP: Zero-Inflated Poisson

CHAPTER 1

INTRODUCTION

Authorities acknowledge that energy and especially electrical energy is the critical power all over the world. Significance of this critical role is increasing day by day. Governments not only need to provide the necessary electricity to their citizens anymore but also they need to provide using natural resources effectively and need to decrease the loss of electricity as well.

When we look at the history of the electricity sector in Turkey, history had begun before The Republic of Turkey. In the first quarter of the 20th century, the Ottoman Empire could provide electricity to main and important cities of the country such as Istanbul, Izmir, Thessaloniki, Beirut, Sam. After the War of Independence was ended and The Republic of Turkey was established, Turkey could have produced only 32.7 MW of electric power, which can be obtained from just one wind power plant today, in all over the country. In 1930, the total electricity power of Turkey increased to 78.8 MW. However, 85.200 MW of electric power can be produced in all over the country in 2017 (Resources, 2017). In 1935, the government took a new decision to establish the Authority of Electricity Business (”Elektrik İşleri Etüt İdaresi”, which is abbreviated as EİEİ in Turkish) for controlling the energy flow and production. After 1950, production, transmission, and distribution, which are the three main vein of Electricity, were divided into different private companies. However, this attempt was not successful because of the economic situation of private companies. Next, another decision was taken for regulating electricity activities by the government: Turkish Electricity Institution was established by the Law no.1312 in 1970 (EMO Energy Comission, 1981).

After the 1980s, Turkish governments have begun to deregulate on the electricity sector and encourage privatization. Turkish Electricity Institution continued to carry out electricity regulations and activities until 1994. In the same year, it was divided into two main sub-institutions: Turkish Electricity Production-Transmission Company ("Türkiye Elektrik Üretim İletim A.Ş.", which is abbreviated as TEAŞ in Turkish) and Turkish Electricity Distribution Company ("Türkiye Elektrik Dağıtım A.Ş." which is abbreviated as TEDAŞ in Turkish) . On the other hand, the most important and radical progress has been realized in the electricity sector with an electricity market Law of 2001. The main aim of the Law of 2001 is that it is allowed to vertical disintegration of production, transmission, and distribution, competition into production and retail sale, privatization of public production plants and distribution institutions and entities. Also, it was designed to make competitive electricity market conditions and encourage entrepreneurs to invest electricity components for boosting the efficiency of production and distribution of electricity (Özkıvrak, 2005).

In 2004, the privatization of TEDAŞ was initiated. At the time, TEDAŞ had 28 billion customers and was selling a total of 93 million kWh of electric power in 21 separated electricity distribution regions. The privatization of TEDAŞ for all electricity distribution regions was completed in 2013. Starting from 2013, each of electricity distribution region belongs to different private companies and electricity distribution activities are conducted by them. TEDAŞ performs to follow and control the performances of the private electricity distribution companies with the Energy Market Regulatory Authority (EMRA, 2008; Ertilav & Aktel, 2015). EMRA was established according to the Law of 2001 (Özkıvrak, 2005). It conducts the regulations of the energy market and orders the market conditions and quality standards in Turkey.

In 2008, EMRA published a regulation article for 21 electricity distribution private companies to control electricity supply continuity (EMRA, 2008). According to this regulation, all of the electricity distribution companies have to publish their statistics related to data for electricity supply continuity. These datasets include the same

variables such as interruption time, location and reason, etc. Each of the distribution companies has to publish this type of data for each month and year. Any customer or researcher can look at this type of data whenever s/he wants on the website of one of the distribution companies. These datasets show the performances of each electric distribution company according to EMRA's standardized indicators.

1.1 Objective and Significance of the Thesis

In this study, we aim to estimate the possible interruption counts with the data for electricity supply continuity by using Markov Chain Monte Carlo Generalized Linear Mixed Models. At the beginning of the study, all of the datasets which belong to different 21 distribution companies were intend to be used in this study, but it has been realized that terminology of the dataset of each distribution company is different from each other in spite of EMRA's standardized regulation. Therefore, data for one local area was used which belongs to Başkent Electricity Distribution Company (abbreviated BEDAŞ) for the location of Çankırı and its neighborhood. Başkent Electricity Distribution Company has been selected from 21 distribution companies since the data conditions are more regular and confidential than others. The reason of selecting the city of Çankırı is that it has medium size compared to other cities in the distribution region, which are Ankara, Bartın, Çankırı, Karabük, Kastamonu, Kırıkkale, and Zonguldak.

These kind of studies are necessary to measure the performances of private Electricity Distribution Companies and make the right decisions to invest in the right areas which are needed for electricity infrastructure. To the best of our knowledge, it is a first implementation study for the dataset of electricity supply continuity which has features of the longitudinal type of Turkish electricity distribution companies. Therefore, results need to be checked and compared by different researchers in the future.

1.2 Organization of the Thesis

This thesis has five main chapters. Chapter 1 introduces the objective of the thesis, study's background and thesis structure. Chapter 2 gives similar studies briefly from the literature. Chapter 3 presents the methodology of Bayesian data analysis, panel data analysis, and Markov Chain Monte Carlo Generalized Linear Models in detail. Chapter 4 shows data used in this study and its main points with some data analysis and specification and conditions of MCMCglmm, their results, and comparison. Finally, Chapter 5 clarifies conclusion of this study and possible future studies.

CHAPTER 2

LITERATURE REVIEW

Research into electricity as a scholarly field generally focused only on modeling demand, consumption or cost and price of electricity, with a particular emphasis on its production and effects on residents. Nowadays, even though electricity production and related questions of demand, consumption and cost continue to be important fields of research, distributing the produced electricity in an efficient way, without any loss, is becoming much more important than the production process itself. Nearly 90 % of total failures are recorded to have occurred in the distribution process (Meeuwsen, 1997), and this fact testifies to the intense interest in the distribution process, maintenance of which is thus as equally important as the production process.

In order to understand the efficiency of electricity distribution, nowadays companies and government institutions try to analyze electricity interruption counts in an electricity network, and the duration, length and location of these interruptions. Consequently, researchers all over the world have begun to improve different methods to explain the behavior of electricity interruption and failure rate estimation. Unfortunately, to the present day, there are no studies about interruption or failure rate modeling in Turkey to the best of our knowledge. The statistics of interruption which have been published by electricity distribution companies have been the only data that could be obtained so far. Having said that, it should be noted that some political actions of the government show that there is a tendency to support research into this area in the future. Moreover, there are not any studies on the application of Markov Chain Monte Carlo Generalized Linear Mixed Models (MCMCglmm) to estimate the electricity interruptions. Having indicated this point, it should be underlined that this thesis will hence function as a first implementation of this approach on electricity

interruptions. With the aim of contextualizing this debate, this literature review section will offer an overview of the different approaches used in some of the models for electricity interruptions.

2.1 Density Estimation using Kernel Method and Approximation by the Least Squares

In 2002, Tatietse et al. published an article on interruption modeling in medium voltage electrical networks. In order to find the probability of the electric interruption, this article used statistical methods to estimate density with Kernel method and approximation by the least squares. The study also included an experiment which had been conducted in Yaounde urban region in Cameroon, and the data referred to two stations, Ngousso and Melen. The article also presented information about their feeders' interruptions, which were collected monthly, over a period of two years.

Because of existing a strong linear correlation between intermittent interruption and permanent interruption in the same feeder, and considering the number of observations is too small for the empirical distribution, enough information could not be obtained for determining the probability law of interruptions. This situation compelled the researchers to use a probability technique depending on the Kernel Method (Silverman, 1986). Also, least square estimation method was appropriate for the model's estimation. Then, the interruptions of Ngousso were modeled as shifted and truncated Gamma distribution and interruption of Melen was modeled as a truncated Normal probability law. Overall, the article explained these differences because of the different qualities of maintenance at these stations. According to the article, maintenance of quality was found to be a lot poorer for the station of Melen. The station had a greater accumulation of random factors which were human, material and environmental and these factors were affecting the current service quality of the station mostly.

2.2 Hierarchical Bayesian Failure Rate Estimation

In their article, Moradkhani et al. (2014) discussed how to model the lack of appropriate outage data belonging to 34 electrical distribution feeders in Alborz Power Distribution company in 2010 with Hierarchical Bayesian failure rate estimation model and other Bayesian models, which were Bayesian estimation pooled failure rate and empirical Bayesian failure rate estimation, through a comparison by using Deviance Information Criterion (DIC). The solution offered for handling the lack of data was to use a shrinkage estimator for failure rate estimation of overhead lines.

The framework of electrical distribution maintenance consisted of these main levels. These were components level, network level, and utility level. Providers of electrical power, who were owners of electrical distribution companies and transmission authorities etc., needed to implement different plans to determine optimal asset maintenance, which consisted of these three main levels. This optimal plan required especially sophisticated reliability models at levels of components and network. Component's reliability was one of the practical and statistical methods used in order to determine the high number of installed components in an electrical network. Data deficiency, population variability, data censoring, and poor quality data were among some of these practical problems.

Moradkhani et al. (2014) discussed that in order to overcome data deficiency and population variability, constant failure rate estimation of medium voltage overhead lines in the presence of data was needed. This approach depended on Hierarchical Bayesian Model (HBM). There were some advantages of Bayesian modeling such as handling deficiency of data, as well as allowing for the combination of data with domain knowledge, providing possible information about causal relationships between variables, avoiding overfitting of data, giving good accuracy even with rather small sample size of data, and finally, combining with decision analytic tools (Li & Shi, 2012). As a result, the HBM was used for modeling the feeders' failure rate. Prior

distribution depended on hyperparameters which were obtained through a two-stage hierarchical model. In the first stage, the failure rates were assumed to be distributed the conjugate Gamma. Then, in the second stage, non-informative prior distributions were used for the prior distribution of hyperparameters. Also, each failure rate had independent prior distribution according to the state of hyper-parameters. The posterior distribution of HBM was analytically challenging for researchers. For this reason, Metropolis within Gibb's algorithm was used to calculate the distribution of hyperparameters.

The model comparison showed that the DIC value of the HBM gave a slightly better result than the values of the empirical Bayesian and pooled models noticeably. The pooled model had the worst result. In spite of using data pooling technique for eliminating the data deficiency problem, the estimation values of the pooled model were not totally reliable for getting the precise value of failure rates for the feeders when the real failure rates of which were considered high or low. On the other hand, in the empirical Bayesian failure rate estimation and hierarchical Bayesian failure rate estimation, the shrinkage estimator was used as the expected value of failure rates of the feeder. Different from HBM, the parameters of the empirical Bayesian model were estimated by the prior information from last year, but the component's failure rate depended on the current condition of components. Using last year's information as prior could not have given the current status of components. However, in HBM, Metropolis within Gibb's algorithm was used to obtain the shrinkage parameter. Also, the initial value of z_0 was selected as a value of which did not affect the sensitivity of posterior distribution. Depending on these results, HBM could be said to be the best model when adequate data about failures could not be obtained by researchers (Moradkhani, 2014).

2.3 Maximum Likelihood Estimation with Different Density Distributions

Prieto et al. (2014) stated that upper tail distribution (also known as Pareto distribution), suggested by empirical researches about the reliability of the electricity transmission networks with indicators such as Energy Not Supplied (ENS), total loss of power (TLP) or restoration time (RT), cannot be valid in the whole range of major events. In accordance with this, in the article, Prieto and his friends aimed to come up with a probability distribution suitable for those indicators. They hypothesized that a two-parameter model could be used to fit this type of data. Two-parameter models such as Pareto II, Fisk, Lognormal, Pareto, Weibull and Gamma distributions were alternatively analyzed on the European power grid data which included reliability indicators of ENS, TLP, and RT between 2002 and 2012 by maximum likelihood estimation (MLE). Model's estimation results were given according to datasets ENS, TLP and RT in the whole range, in both periods: (2002-2009) and (2002-2012). Results of the models were compared to each other according to AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) and goodness of fits were tested by empirical KS (Kolmogorov-Smirnov) statistic based on bootstrap resampling. Pareto II model was the most preferable model for the datasets of ENS and RT according to AIC and BIC. Moreover, Pareto II (Lomax), Fisk (log-logistic) and lognormal models were the preferable models in TLP dataset according to AIC and BIC.

In 2013, Alwan et al. published an article about reliability measurement for mixed mode failures in 33/11 kilovolt electric power distribution stations average time between electrical failures in Iraq. Duration of the failures of electric power distribution stations was manually collected through a period of five years. According to the article, reliability was important to consumers who affected the electrical cost of failure, repair, and maintenance. Acceptability of fit test showed that Dagum distribution fitted with the data very well. MLE was used to estimate the parameters of Dagum distribution. According to the article, 8 out of 14 components cause this low

reliability level because of the age of the components. These components had to be changed as soon as possible for eliminating electrical interruptions which occurred because of components' reliability.

2.4 Time-Varying Load Models and Estimation with Monte Carlo Simulation

In their survey, Huda & Živanović (2018) used Energy Not Supplied (ENS) index, which is an index of measures of the expected amount of energy unreached to customers in a specific time period because of failures in the distribution system. Analytical and MC methods were used to estimate the ENS before (Billinton & Wang, 1999). However, MC simulation approach was evaluated to be a time-consuming approach due to its requiring a large number of iterations to achieve acceptable accuracy. In this study, Multilevel Monte Carlo (MLMC) performed the simulation faster, and this method was used to estimate ENS with considering time-varying nature of different load models. MLMC was the advanced Monte Carlo method used to improve computational performance, and it incorporated Stochastic Differential Equation (SDE) of system variables (Giles, 2015). Besides, MLMC approach with Euler-Maruyama method was used for the estimation. Then, why did load models use it here? The answer was easy. In practice, most of the utilities only had records of the load demand data for a certain electricity distribution region. In other words, customer variations of the load data for an individual load point during 24 hours in a day for a year were not reachable.

In the ENS estimation, two variables of mean were considered: Time to Failure (MTTF) and Time to Repair (MTTR). Assuming that the randomness of MTTF of a component/ element j ($MTTF_j$) was modeled by SDE with standard Brownian motion, B_t on the time period (Kingman & Harrison, 1987) Stochastic model of $MTTF_j$ was solved through the use of Euler Maruyama discretization scheme.

Finally, Huda and Zivanovic (2018) stated that the ENS can be decreased by a small percentage in places where the failure of the electricity system occured during the peak loads of weekly and daily usage.

2.5 Multivariate Linear Regression Models on Panel Data

Eto et al. (2012) evaluated electricity reliability information collected over a period of 10 years from 155 different U.S. electric utilities, which were altogether responsible for approximately 50% of total U.S. electricity sales. The results of the survey showed that annual average electricity duration and annual average frequency of electricity interruptions have been increasing by 2% each year. An earlier study by Eto and LaCommare (2008) had indicated that more than 90% of average number of interruptions generally stemmed from electricity distribution systems. For this reason, it could be claimed that ill-conditioned electricity distribution systems caused the increase on the average frequency of interruptions and average electricity of duration. The dataset was unbalanced since it did not include reliability metrics for each year. Multivariate log-linear regression models with fixed effects and random effects were used in the survey separately. They enable us to point out the differences in the outcomes, which were caused by the correlations, and also the differences in the sources such as utility reported reliability metrics: System Average Interruption Frequency Index (SAIFI) and System Average Interruption Duration Index (SAIDI). Application of the models included four different steps respectively:

- Transforming reliability metrics to natural algorithms,
- Conducting F-test on the transformed reliability metrics,
- Applying Hausman specification test to understand the appropriateness of the estimation either for fixed effects model or random effects model,
- Estimating two sets of models: fixed effects and random effects.

Hausman specification test results showed that random effect models were consistent and more efficient than the fixed effects version. In accordance with this, Eto et al. (2012) found that temporal trends were significant for SAIFI and SAIDI. In addition, SAIFI and SAIDI were found to be increasing at the rate of 2 % each year, which showed that reported reliability gets worse over time.

2.6 Contribution of this Study to the Literature

In this chapter, different methods used in this area has been given as such:

- Density Estimation using Kernel Method and Approximation by the Least Squares (Tatietse et al., 2002)
- Hierarchical Bayesian Failure Rate Estimation (Moradkhani et al., 2014)
- Maximum Likelihood Estimation with Different Density Distributions (Prieto et al., 2014 ; Alwan et al., 2013)
- Time-Varying Load Models and Estimation with Monte Carlo Simulation (Huda & Živanović, 2018)
- Multivariate Linear Regression Models on Panel Data (Eto et al., 2012)

Different approaches and methods using for estimation of electrical interruption were given. When these studies are examined in detail, most of them had a continuous response in the data and just some of the data is longitudinal data type. In this study, to the best of our knowledge, different from the literature, it is the first time that count and longitudinal type of electrical interruption's data are analyzed with MCMCglmm. It is not encountered any article like this study in the literature so far by us. Also, this study is the first implementation of the R package of MCMCglmm, which has been used mostly in the area of biostatistics, for this type of electrical interruption's data.

We hope that this study will bring a new way to the analysts who want to investigate the electrical interruption's data type which is count and longitudinal. Also, this study will gain a new point of view to the decision makers who manage the electrical

network according to possible reasons of electrical interruptions. On the other hand, since the improving process of the package of MCMCglmm is continued, we hope that this study will be also given enlightening answers to the statisticians who study in this area.

CHAPTER 3

METHODOLOGY

The aim of the study, its background and the studies using in the area has been explained in Chapter 1 and Chapter 2. Useful methodological framework of the study will be laid out in this chapter. In the first place, longitudinal methods will be explained in detail, and then, secondly an examination of Bayesian approach will follow. Finally, the chapter will conclude with the analysis of the main structure MCMCglmm.

3.1 Longitudinal (Panel) Data Analysis

Longitudinal or panel data analysis used interchangeably has been improved in the early 1950s during the time the U.S. government shifted a substantial part of its research support from military to medical research. At that time, the main concern of researchers was to decrease morbidity and mortality, and as a consequence of this, early research focused on treating the diseases and eliminating the risk factors causing the diseases. Later, researchers tried to identify the risk factors which cause diseases in adult age, and could be detected in childhood. For example, researchers began to investigate childhood blood pressure level of a patient who has hypertension (Friedman et al., 1988). Through such studies, databases which include many factors belonging to a patient were formed. Following this, a new type of data structure, known as longitudinal or panel data was born.

This new data structure is based on taking measurements of the same individuals repeatedly throughout time, and thereby it allows the direct study of change of a certain factors over time. The main aim of a longitudinal study is to observe the change

in response over time and the changing influences of factors to response (Rowell & Walters, 1976). Then, with repeated measurements, one characteristic of the individual can be detected within individual-change. To illustrate, in contrast to cross-sectional data, which provides information for more than one unit in only an exact time period, or time series data, which gives information only for one unit through different periods, the evaluation of within subject influences on response over time can be detected in the type of longitudinal data. However, same entities (subjects) such as individuals, locations, countries, patients etc. and their differences at multiple time points can be observed through using longitudinal or panel data analysis.

3.1.1 Specifications and Advantages of Longitudinal Data Analysis

- One of the most important features of longitudinal data is that data is actually clustered. Repeated measurements of the individuals constitute clusters. Moreover, observations in each cluster have positive correlation between them.
- Differing from cross-sectional studies, longitudinal data analysis can demonstrate the differences and influences of response within individual changing over time. Cross-sectional studies cannot provide any information about how individuals change during the time period.
- Longitudinal data analysis yields more accuracy regarding inference of model parameters compared to cross-sectional studies.
- Longitudinal data analysis gives more information to researchers with more variability.
- Controlling ability on variables is easy even if on un-known measurements or un-observed variables.
- In longitudinal analysis, by comparing each response to one another, a longitudinal analysis may dispose noises which affect the response and by eliminating them, more accuracy of estimation can be obtained easily.

- Longitudinal studies allow certain predictions through collecting information from all individuals, and thus researchers can better predict individual change through time for a certain individual (Fitzmaurice, 2004).

Another specification in the longitudinal data analyses is “balanced” or “unbalanced” datasets. Balanced data means measurements are recorded just in time for every individual or subject. However, in real life, these conditions may not always be implemented. To illustrate, glucose level of patients may not have been measured every 30 days, or individuals might sometimes miss or forget their scheduled visit and participants might drop the study suddenly. With such impediments, recording individual-changes of the response in the study becomes harder for the researchers, and consequently mistimed measurements or random measurement might be obtained by researchers. This type of data is called “unbalanced” data. There are various reasons as to why the researches might wish to obtain unbalanced data. Researchers might use it to reduce cost of study while at the same time increasing overall participation of the individuals. This type of design is called as “rotating panel” design, where researchers determine before the study which measurement will be obtained. On the whole, while habitually balanced data structure is always preferred by the researchers, unfortunately sometimes it might be unreachable. For these occasions, it might be useful to remember that there are some assumptions and methods to overcome the disadvantages of unbalanced data structures. At this point, it should be noted that the data of this study is fortunately balanced dataset (Diggle et al., 2002).

3.1.2 Linear Models for Longitudinal Data

Linear model specifically means that the mean response indicates linear behavior in the regression parameters whereas longitudinal data structure depends on the assumption that a sample of N subjects are measured repeatedly over time (Ware, 1985). Both of these contents are combined in the same class for linear models on

longitudinal data. The expression of Y_{ij} denotes the response variable for the i^{th} subject on the j^{th} measurement occasion. In this case, the mean responses have a vector (1) which includes Y_i as simply a time-ordered collection of the n_i x1, which means, it consists of n_i rows and 1 column of elements for the i^{th} subject.

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \dots \\ Y_{in_i} \end{pmatrix}, \qquad i= 1,\dots,N. \tag{1}$$

One assumption is that the vectors of responses (1) for the N subjects are independent from each other. However, the repeated measurements of the same subject are not assumed to be consisting of independent observations. In longitudinal data analysis, this assumption or the correlation among repeated measurements adds positive effect to the analysis because correlated subjects provide more accurate estimates of the effect of covariates. This assumption resembles cluster data features: The observations that come from different clusters are assumed to be absolutely independent from each other, while observations in one cluster are not assumed to be independent from each other. In panel or longitudinal data analysis, every subject constitutes one cluster, such as one patient's health results. For example, one patient gives his blood to have his glucose level checked periodically for each month over one year. The output of the glucose level analysis in January is absolutely dependent on the output of the analysis in June, since these two analyses show the same patient's glucose level even if one factor in analysis might have changed (For example, patient stops taking sugar in his daily diet.)

In addition to one response of the data, there is a vector of covariates (2) with number of p rows:

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \dots \\ X_{ijp} \end{pmatrix}, \qquad i= 1,...,N. \ ; \ j=1,...,n_i. \tag{2}$$

The number of p rows of X_{ij} equal to different covariates of the analysis, which means one vector of covariates (2) exists corresponding to one of the n_i repeated measurements for the i^{th} subject. There are two main covariates:

- Covariates having unchangeable values during the study.
- Covariates having changeable values during the study.

When a need to compound structures of responses and covariates arises, a linear regression equation (3) becomes (Ware, 1985):

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \cdots + \beta_p X_{ijp} + e_{ij} \qquad j=1,\ldots,n_i. \tag{3}$$

With this regression equation (3), it is possible to observe what kind of relation is there between the responses and corresponding covariates in each occasion. Of course, there needs to be the number of n_i as well as separate equations for modeling each response variable. If all of parameters are grouped with each other, the model (4) appears as follows:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & & X_{i2p} \\ \vdots & \vdots & & \vdots \\ X_{in_i1} & X_{in_i2} & & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix} \tag{4}$$

e_{ij} is random errors for the responses of the i^{th} subject with mean zero. However, the errors at different time points are assumed dependent and hence we have a variance-covariance structure, Σ.

3.1.2.1 Main Assumptions of Linear Models: Mean, Variance, Covariance and Correlation Structures

In longitudinal data analysis, the main focus is directed to the mean of response. Mean response or expectation of each response is weighted as the average of all possible

values of the response and mean response for varying responses from individual to the other is denoted (5) as:

$$\mu_{ij} = E(Y_{ij}). \quad (5)$$

The mean response (5) gives the location of the center of the distributed Y_{ij}. Second important measurement is variance (6) which provides the measurement of spread of the response and variance (Fisher, 1925). It is denoted (6) as:

$$\sigma_j^2 = E\{Y_{ij} - E(Y_{ij})\}^2 = E(Y_{ij} - \mu_{ij})^2. \quad (6)$$

In addition, variance may vary from occasion to occasion while it may also be a function of selected covariates.

Another concept of the longitudinal analysis is the dependence according to responses, which is called covariance. Covariance for responses in the two different occasions can be (Y_{ij} and Y_{ik}) denoted (7) as:

$$\sigma_{jk} = E(Y_{ij} - \mu_{ij})\, E(Y_{ik} - \mu_{ik}). \quad (7)$$

Covariance (7) for these two responses in different occasions gives the relation of linear dependence. The covariance of responses (7) might have positive or negative values, but usually expected to be positive. Like other types of regression analysis, when the covariance becomes zero, no linear relation between these two responses exists. The covariance result is affected not only by the degree of dependence between two variables, but also by their units of measurement. Indeed, any change in the scale of the measurement affects the covariance's value. For instance, when a scale of variable changes from kilometer per hour to mile per hour, the result of covariance also changes. Therefore, the covariance value is not really informative. Covariance

always needs to be interpreted with the value of variance as the magnitude. To detect the measurement of linear dependence between two responses, a measurement which is free of units of measurement is more suitable. This measurement is correlation of two variables (8):

$$\rho_{jk} = \frac{E\{Y_{ij} - \mu_{ij})(Y_{ik} - \mu_{ik})}{\sigma_j \sigma_k}. \quad (8)$$

Correlation (8) is the linear dependence between two variables and it takes the range of 1 to -1. When correlation takes 1, it implies that if one variable increases, other variable also increases. However, when correlation takes -1, when one variable increases, then other variable decreases. On the other hand, if covariance is 0, the correlation is also zero. The most interesting thing here is that when two variables are statistically independent from each other, then they can be uncorrelated, but they do not need to be independent when the variables are uncorrelated. Statistical independence implies that there is no dependence between these two variables. However, correlation (8) shows only the linear dependence between the variables.

In longitudinal data, repeated measurements of the same individual are seen to be positively correlated with each other. In this situation, variance-covariance matrix (9) could be defined as below:

$$Cov\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & \dots & Cov(Y_{i1}, Y_{in}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & \dots & Cov(Y_{i21}, Y_{in}) \\ \vdots & \vdots & & \vdots \\ Cov(Y_{in}, Y_{i1}) & Cov(Y_{in}, Y_{i2}) & \dots & Var(Y_{in}) \end{pmatrix}. \quad (9)$$

It is necessary to remember that variance and covariance have a symmetry. As indicated in the example of $\text{Cov}(Y_{ij}, Y_{ik}) = \sigma_{jk} = \sigma_{kj} = Cov(Y_{ik}, Y_{ij})$ (9) and also this

aspect is the same with correlation matrix (10) $\text{Corr}(Y_{ij}, Y_{ik}) = \rho_{jk} = \rho_{kj} = Corr(Y_{ik}, Y_{ij})$. Correlation matrix (10) can be seen below:

$$Corr\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \dots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} 1 & Corr(Y_{i1}, Y_{i2}) & \dots & Corr(Y_{i1}, Y_{in}) \\ Corr(Y_{i2}, Y_{i1}) & 1 & \dots & Corr(Y_{i21}, Y_{in}) \\ \vdots & \vdots & & \vdots \\ Corr(Y_{in}, Y_{i1}) & Corr(Y_{in}, Y_{i2}) & \dots & 1 \end{pmatrix} \quad (10)$$

The diagonal values are equaled to 1 because it indicates correlation of a variable with itself.

Many articles have exposed that longitudinal data are correlated, furthermore, they are positively correlated (Diggle et. al., 2002). When the behavior of empirical observations about correlation in longitudinal studies was analyzed, it was necessary to revisit the correlations. Deriving from there, Fitzmaurice (2004) list these behaviors as follows:

- have positive relation,
- generally decline by time separation,
- correlation between repeated individuals (subjects) are barely close to zero,
- between two repeated measurements which are close to each other does not approach one.

In the process of longitudinal data estimation, ignoring the correlation and assuming that measurements are independent from each other might lead to apparent overestimation of variance. At the end of the analysis, this situation will lead to a bad estimate of precision, will cause larger standard errors and p-values as well as wider confidence intervals. Since, independence status of covariates of the model is

important for obtaining true estimation and interpretation for the longitudinal data analysis, a specific assumption, multi-collinearity must be checked.

Multi-collinearity is a statistical phenomenon which has a strong relation with predictor variables. Obtaining reliable estimates of coefficient for each variable is very difficult when multi-collinearity is encountered. For this reason, it cannot provide true interpretation based on the outcome of predictor variables. Multi-collinearity affects variances of the parameter estimates by inflating. Inflated variances may lead to insufficient significance of predictor variables. Therefore, significant variable acts like an insignificant variable. To conclude, multi-collinearity problem can cause serious problems for the variable coefficients and it may lead to wrong conclusions when researcher wrongly interprets outputs of a model (O'Hagan & McCabe, 1975).

To detect the multi-collinearity problem in a data analysis, it may be looked up:

- Correlation matrix of coefficients
- Variance Inflation Factor
- Eigenvalues Analysis

Correlation matrix; large correlation coefficients might be an evidence for multi-collinearity. If the correlation is high and close to 1 or -1 between coefficients of two predictor variables, it is possible to suspect a multi-collinearity problem in these variables.

Variance Inflation Factor (VIF); measures multi-collinearity situation in ordinary least-square analysis (Mansfield & Helms, 1982). VIF indicates the level of multi-collinearity by measuring the variance of the estimated regression coefficient (11):

$$VIF_j = \frac{1}{(1 - R_j^2)}. \quad (11)$$

Let R_j^2 is the coefficient determination for the cases when X_j is regressed on all other variables in the model. When there is no linear relation between j^{th} variable and other predictor variables, then $VIF_j = 1$.

The result of VIF (11) exceeds 5 or 10 indicates that j^{th} regression coefficient is estimated poorly because of multi-collinearity (Montgomery, 2001). We selected VIF to check the multi-collinearity problem in this study. VIF code was taken from (MCMCglmm-utils.R, 2019).

Correlation matrix with Eigenvalues; if eigenvalues are small or near to zero and corresponding condition number is large, it may be existed multi-collinearity problem in one or more predictor variables might occur.

3.1.2.2 Estimation Methods for Linear Models: Maximum Likelihood and Restricted Maximum Likelihood

Maximum Likelihood Method (ML); is a common approach used to estimate covariate of the model which is β and covariance parameter of the model which is θ. The method of ML depends on finding the most probable values of β and θ in the observed data. To find maximum values of β and θ, joint probability of the response variable is maximized in the observed data. The fixed set of observed values of response variables are regarded as the functions of β and $\sum_i(\theta)$ (In multivariate normal distribution, the covariance matrix of covariance parameters, θ, are presented with $\sum_i(\theta)$). Also, these functions are known as likelihood functions (Laird & Ware, 1982; Lindstrom & Bates, 1988).

To understand the mechanism of the method, standard linear regression model which has univariate normal distribution with all the observations are independent and also uncorrelated will be regarded as the simple case. For this, a cross-sectional type of data which has repeated at n different occasions is being analyzed. Data has a sample of N subjects at each occasion. Observations are independent from each other and data has constant variance which is shown as σ^2. The mean response function of the linear regression model (12) is:

$$E(Y_{ij}) = X_{ij}^T \beta. \tag{12}$$

The estimation of the covariates for linear regression model can be obtained from all the observations which maximize the joint normal density function. In the first place, the univariate normal probability density function (13) is shown below:

$$f(y_{ij}) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y_{ij} - \mu_{ij})^2 \Big/ \sigma^2\right\}. \tag{13}$$

The log likelihood function of the univariate normal probability density function (14) is that:

$$\begin{aligned} l &= \log\left\{\prod_{i=1}^{N}\prod_{j=1}^{n} f(y_{ij})\right\} \\ &= -\frac{K}{2}\log(2\pi\sigma^2) - \frac{\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij} - X_{ij}^T\beta)^2}{\sigma^2}, \end{aligned} \tag{14}$$

where the K is a matrix with the dimension of nxN.

In order to obtain the estimate of β, ignore the first term of the likelihood function (14) and take the derivative of the log-likelihood function. The ML estimator of β equals to ordinary least square estimate of β:

$$\hat{\beta} = \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n} (X_{ij} X_{ij}^{T}) \right\}^{-1} \sum_{i=1}^{N} \sum_{j=1}^{n} \left(X_{ij} y_{ij} \right). \tag{15}$$

Restricted Maximum Likelihood Estimation (REML); The ML estimator of β and $\sum_i(\theta)$ have large sample properties. To illustrate, $\sum_i(\theta)$ has big bias in the finite small samples. Also, the diagonal elements of $\sum_i(\theta)$ are underestimated. Therefore, REML might be a good option to overcome these problems. For example, assume that observations are independent and variance is constant which is σ^2 for a cross-sectional data, and ML estimator is given above (15). When the ML estimator of σ^2 (16) is that:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{n} (Y_{ij} - X_{ij}^{l} \hat{\beta})^2 / K \tag{16}$$

On the other hand the mean of ML estimator of σ^2 (17) is that:

$$E(\hat{\sigma}^2) = \left(\frac{K-p}{K} \right) \sigma^2, \tag{17}$$

where the p is the dimension of β.

To conclude that the MLE of σ^2 is biased in small samples and σ^2 is underestimated. An unbiased estimator needs to be obtained via using K-p (which is also residual degrees of freedom). Then the REML estimator of σ^2 (18) becomes:

$$\hat{\sigma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{n} (Y_{ij} - X_{ij}^{T} \hat{\beta})^2 / (K - p). \tag{18}$$

The main thought behind the REML estimation is that the data which has been used for estimating β is ignored and rest of data is used for estimating σ^2. Hence, relevant

parts of data is used for the process of estimation of $\sum_i(\theta)$, and it is unbiased. In this thesis we will use Bayesian inference, which will be explained in section 3.2.

3.1.3 Generalized Linear Models

Linear model and estimation concept of the longitudinal data have been examined in the previous section. However, when the response of a longitudinal data is discrete, the concept of linear models can no longer be appropriate. Generalized Linear Model (GLM) is a framework which combines discrete and continuous response variables of independent observations for regression analysis, and hence it comes to mind when the response variable is discrete in a longitudinal dataset, and when the methods of linear models which analyze mean response to covariates are being used (Nelder & Wedderburn, 1972). The concept of Generalized Linear Models (Liang & Zeger, 1986) has been improved by the researchers to handle these problems. However, due to correlation among observations of the same individual in the longitudinal data, GLM may not be so easy to implement. The main feature of GLM is its nonlinear-transformation of the mean response, which is a linear function of covariates (Mccullagh & Nelder, 1989). This non-linear transformation causes concern regarding the interpretation of the regression coefficients in longitudinal data analysis. This concern arises as a result of the different approaches to the sources of within–subject association in the longitudinal data. Nevertheless, GLM offers a unified approach for all univariate responses (binary, counts, continuous). Besides, GLM can be said to be a collection of regression models and analysis of variance models (ANOVA) for;

- •A normally distributed continuous response
- •Logistic regression models for binary or dichotomous response
- •Log-linear or Poisson regression models for counts

Generally, a response of GLM has three main specifications:

- Distributional assumption,
- One systematic component,

- Link function

(Mccullagh & Nelder, 1989).

Distributional Assumption; GLM is an extended version of the concept of standard linear regression analysis with settings where the response variable is discrete or categorical. GLM concept depends on the probability distribution of response variable, which is a member of the "Exponential Family". Exponential family has many distributions such as Normal, Bernoulli, Binomial and Poisson.

Random component is also included in the design of distributional assumption concept. Random component brings a probabilistic mechanism to the responses in accordance with its belonging to distributional assumptions. The members of the exponential family have the same statistical properties of the models. For example, the variance of the response for Binomial, Bernoulli and Poisson has a dispersion parameter which is expressed with ϕ and variance function which is expressed with $v(\mu_i)$ which is derived from the known function of the mean (μ_i) (19):

$$Var(Y_i) = \phi v(\mu_i), \textit{ where } \phi > 0 \qquad (19)$$

Variance function $\nu(\mu_i)$ describes how the variance of the response is related to the mean of the response. ϕ is a parameter which needs not to be estimated in most of distributions for discrete data, since it is a known constant (ϕ is 1 for Bernoulli and Poisson distributions.) but for other distributions, ϕ might be an unknown parameter which needs to be estimated. Moreover, ϕ could be bigger than 1 in the case of overdispersion, and less than 1 for underdispersion, even in Bernoulli or Poisson distribution. It will be discussed in details in section 3.1.3.1.

Variance depends on the mean of the distribution in Poisson and Bernoulli distribution. This feature is identical in most distributions of discrete responses.

However, the variance does not depend on the mean for normal distribution ($Var(Y_i)$ = ϕ and the variance function is $\upsilon(\mu_i)$=1). This is called "homogeneity of variance". Because of this feature for normal distribution, homogeneity of variance is maintained with the standard linear regression assumptions for normally distributed responses.

Systematic Component; Generalized Linear Models also have a common regression formulation such as having a common family of distribution. Linear regression component is an important feature of the linear regression model and this component also maintains its existence in Generalized Linear Models. Linear regression component is called as "systematic component" in the notation of GLM. Systematic component is denoted as the effects of covariates on the mean of response:

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}. \quad (20)$$

η_i is called as "linear predictor" and β_1 will also be called as intercept if the X_{i1} equals to 1.The linear predictor is described as a linear combination of unknown regression coefficients and covariates. The linear predictor is always denoted as "linear" since the mean response is explained as a straightforward weighted sum of regression parameters even if covariates of regression were not linear. Since the linear predictor must be linear and if the linear assumptions on mean response are not provided, some transformations have to be implemented to the mean response.

Link Function is a transformation of the mean response and it links the transformed mean response with covariates through using linear predictor. Link function g(.) assumes that the transformed mean response moves linearly with changing covariates. The use of certain non-linear functions like $\log(\mu_i)$, on the other hand, guarantees that predictions of mean response are located between suitable ranges.

Another important point to note is the concept of canonical or non-canonical link functions. Canonical link functions are unique and they can be derived from the specific distributions. To illustrate, logit link function ($\log(\frac{\mu}{1-\mu})$) is canonical link function of Bernoulli and binomial distributions In addition, the canonical link function of Poisson is the log link. However, non-canonical link functions such as probit link function can also be used. Canonical ones simplify the computational burden.

3.1.3.1 Log-Linear regression for Counts

Log-linear regression is generally defined as Poisson regression which is regression analysis of counts in a specific time interval. In this case, the response of the data is a count type and log-linear regression analysis helps to link the mean response with the set of covariates. The probability function of the Poisson distribution (21) is defined as such:

$$\Pr(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \text{ where } y_i = 0,1,2,\ldots . \quad (21)$$

y_i is the observed number of events and Poisson distribution is determined with a parameter which is the mean or expected number of events ($\mu_i = E(Y_i) \geq 0$). Besides, mean and variance of Poisson distribution is identical ($E(Y_i) = \mu_i = Var(Y_i)$). In addition, the expected rate is μ_i/T_i. T_i a measurement of the "time at risk" which is known and can be observed. The aim of log-linear regression is to detect the "positive" effect of set of covariates on the expected rate. In this case, expected rate can never take negative values. Then, logarithmic transformation is implemented to the regression model (22) as follows:

$$\log(\mu_i/T_i) = \beta_1 + \beta_2 X_i , \quad (22)$$

and also it can be referred as:

$$\log(\mu_i)=\log(T_i) + \beta_1 + \beta_2 X_i. \quad (23)$$

$\log(T_i)$ at the equation (23) is known and does not require estimation. Thus, it can be concluded that log-linear regression is a log rate of some events and covariates.

Overdispersion; A common failure of Poisson distribution needs to be mentioned here. Firstly, it should be remembered that the dispersion parameter is assumed constant (ϕ =1). However, in some applications, count data has great variability when the predicted values are far from observed values. This assumption is defined as "over-dispersion", which is a common failure assumption. It might sometimes occur in Poisson and binomial responses. In other words, the variance of response is greater than the response mean. The over-dispersion problem can be detected with the model deviance divided by degrees of freedom. When researchers suspect over-dispersion problem in data analysis, the main evidence they can use is that the Pearson chi-square statistic equals to the residual degrees of freedom which is calculated with the difference of a number of observed values and model parameters (Agresti, 1996).

3.1.4 Linear Mixed Effects Models

In brief, Linear Mixed Effects Model (in general LMM) can be defined as a model which combines random and fixed effects. In GLM, three main specifications existed: Distributional assumption, one systematic component and link function. In addition to these specifications, there exists an additional assumption which is called "conditional distribution" of each response, Y_{ij}, in the concept of linear mixed effects models. According to this assumption, vector of random effects which is b_i has normal distribution and conditional variance of Y_{ij} (24) is that:

$$Var(Y_{ij}|b_i) = \sigma^2 \quad (24)$$

Moreover, given the random effects, each response of the data is independent from one another. Hence, it can be said that distributional assumption is completed on the response, Y_{ij}.

Following the conditional variance of the LMM, the conditional mean of Y_{ij} is defined with the linear predictor (25) as such:

$$E(Y_{ij}|b_i) = \eta_{ij} = X_{ij}^T\beta + Z_{ij}^T b_i. \quad (25)$$

where Z_{ij} is a subset of X_{ij}.

Linear predictor of LMM has both population and individual effects, as a result of using fixed and random effects differing from the linear predictor of GLM. In addition, conditional mean of response is also the identical link function of LMM.

In this case, simple expression for the conditional mean response for any individual (26) is that:

$$E(Y_i|b_i) = X_i\beta + Z_i b_i. \quad (26)$$

And the marginal mean response for the population in average for all individuals (27) is that:

$$E(Y_i) = X_i\beta. \quad (27)$$

Another component of the model is "within subject measurement error" which is defined with e_{ij}. Normally, the within subject measurement error is also distributed independently with zero mean and variance σ^2. On the other hand, the covariance between observations, focused on the mean response of any individual is defined as $Cov(e_i) = R_i = \sigma^2 I_{n_i}$ given conditional independence assumption. Then, the within subject measurement is collected in a vector: $e_i \sim N(0, R_i)$.

To sum up, the linear mixed effect model (28) is defined as:

$$Y_i = X_i\beta + Z_i b_i + e_{i.} \tag{28}$$

3.1.5 Generalized Linear Mixed Models

The notion of fixed effects, regression covariates, coefficients and three main specifications of GLM were told in the last section. In this section, Generalized Linear Mixed Models (GLMM) which is the extended version of Generalized Linear Models (Skellam, 1948) will be examined.

The concept of GLMM (Gibbons & Hedeker, 1994) can be explained as the model in which the effect of regression coefficients is allowed to be deployed randomly to the individuals in the longitudinal data analysis. GLMM for longitudinal data provides the assumption of heterogeneity between individuals in the population of the study via using random effects. Due to the presence of unmeasured factors, random effects can be assumed the maintain natural heterogeneity.

GLMM still completely preserves the assumptions that come from Linear Mixed Models and Generalized Linear Models. Firstly, it is known that the distribution of random effects are multivariate normal distribution according to mathematical and computational convenience (Breslow & Clayton, 1993). Secondly, according to the features of exponential family, it is assumed that the responses for any individual are independent observations from the distribution. Thirdly, the assumption of "conditional independence", which means $R_i = \sigma^2 I_{n_i}$, is completely similar to Linear Mixed Models. Briefly, it can be said that GLMM is a general version which compounds linear mixed effects models and GLM.

To sum up, GLMM can be defined through 3 main specification:

- Assume that the conditional distribution of each Y_{ij} given a vector of random effects b_i is a member of exponential family, and given the random effects, each of Y_{ij} is independent from each other, pursuant to conditional independence assumption.
- Assume that the conditional mean of Y_{ij} depends on fixed and random effects by the linear predictor with a known link function.
- Assume that in principle, there exists a vector of b_i in any multivariate distribution and the vector of b_i is distributed multivariate normally with zero mean and qxq covariance matrix of G and b_i's are independent of covariates.

3.1.5.1 Generalized Linear Mixed Model for Counts

Suppose that the response of the data, Y_{ij}, is a count. Then, the three main specifications which are defined below exist in GLMM analysis:

- Conditional on a vector of random effects, Y_{ij} is the response which has independence assumption and Poisson distribution with $E(Y_{ij}|b_i) = Var(Y_{ij}|b_i)$.
- The linear predictor of the model depends on both fixed and random effects both and it is defined (29) as:

$$\log\{E(Y_{ij}|b_i)\} = \eta_{ij} = X'_{ij}\beta + Z'_{ij}b_i. \qquad (29)$$

Also, this is the conditional mean of the response by log-linear link function.

- The random effects are assumed to be bivariate normally distributed with zero mean and a x a covariance matrix G.

This model is called also a log-linear regression model with random intercepts and slopes. The model provides natural heterogeneity between individuals (Gardner, Mulvey, & Shaw, 1995).

3.2 Bayesian Inference & Markov Chain Monte Carlo

The concept of the Generalized Linear Mixed Models (GLMM) actually emerges between Restricted Maximum Likelihood (REML) as pros and Bayesian Markov Chain Monte Carlo (MCMC) Bayesian methods as cons (Hadfield, 2010). There are many differences between two methods. To begin with, REML is a fast and straightforward theory in application whereas MCMC is slower and more challenging in technical analysis due to selection of a sensible prior. However, analytical results cannot be obtained for non-Gaussian GLMM in REML because of its procedures. REML has basic approximate likelihood methods but it may not work well for non-Gaussian models. On the other hand, MCMC is also an approximation but the accuracy of approximation increases when the analysis is run for all type of models. Therefore, it can be said that Bayesian MCMC can offer more accuracy in the analysis even if it may be more challenging and slower than REML.

The concept of Bayesian statistic depends on combination of the prior belief and likelihood theory. To illustrate when several random deviates (y) from a normal distribution σ^2 have been observed, the conditional probability of the model parameters (30) can be shown to be proportional to:

$$\Pr(y \mid \mu, \sigma^2)\Pr(\mu, \sigma^2) \tag{30}$$

The first term of the equation (30) is the likelihood function and the second term is prior belief which the model parameters could take.

For the first term, the likelihood of the data is calculated on a grid of possible parameter values to obtain data from the likelihood surface by using the Maximum Likelihood or Restricted Maximum Likelihood Methods separately in order to obtain posterior distribution (Section 3.1.2.2). However, for non-Gaussian distribution, obtaining the derivative of the likelihood function is a more challenging process. To

cope with this mathematical challenge, Markov Chain Monte Carlo Methods, which are a class of algorithms, were improved and began to be used by the researchers as an alternative way (Geyer,1991; Kass, Gilks, Richardson, & Spiegelhalter, 2006; Marjoram, Molitor, Plagnol, & Tavare, 2003).

For the second term, the choice of prior is extremely important for the analysis since it entails the beliefs of researchers about the values, which the model parameters might take. In general, it may not take a suitable prior especially in the early stages of the analysis and this situation is called "improper priors" since knowledge of the researcher is restricted.

In the following section, the general framework of these two main parameters of the Bayesian data analysis will be given. First, the concept of MCMC and its algorithms will be told. Then, in the second part, the concept of prior function will be explained.

3.2.1 MCMC and Its Algorithms and Diagnostic Tests

Actually, Markov Chain Monte Carlo (MCMC) simulation techniques were developed in 1950's by the physicist (Metropolis et al., 1953). After that, the statisticians (Hastings, 1970; Geman & Geman, 1984; Gelfand, Hills, Racine-Poon, & Smith, 1990) have discovered the method and they improved it to obtain posterior distribution for model parameters and latent variables of the complex models as well.

MCMC generates a sample or likelihood surface stochastically. The first stage of implementing MCMC is to select the initial values which start the chain truly. These initial values should not be far away from the set of parameters. For instance, the values should be selected from sets where the posterior density is high. If the initial values are selected far away from the point where the posterior density is low, it will inevitably require a lot more iterations before being converged.

A chain of values x_t, with t = 1, …, T (T is the total number of iterations), is a Markov chain if x_t depends on only x_{t-1}. This dependency is provided by a model that includes a stochastic component. There are many MCMC algorithms used in Bayesian analysis. The most famous algorithms can be listed as:

- Metropolis-Hasting algorithm
- Gibbs Sampling

3.2.1.1 Metropolis-Hasting Algorithm

The Metropolis- Hasting algorithm is a framework of MCMC simulations which are suitable for constituting samples from Bayesian posterior distributions. For example, Gibbs sampling is a special case of Metropolis-Hasting algorithm. In this section basic Metropolis algorithm will be provided first (Metropolis et al., 1953), and then it will be generalized as Metropolis-Hasting algorithm (Hastings, 1970) as well.

The Metropolis algorithm (Metropolis et al., 1953) is an adaptation of random walk. It uses the rule of acceptance/rejection to converge to a specific distribution. The algorithm includes the following steps:

1- Select a starting point θ^0 to be the first sample from the starting distribution, $p_0(\theta)$.

2- For every t, t=1,2,…

   a- Sample a proposal θ^* from a proposal (or jumping) distribution $J_t((\theta^*|\theta^{t-1})$ at time t. The proposal or jumping distribution is symmetric in the Metropolis algorithm, $J_t((\theta^*|\theta^{t-1}) = J_t((\theta^{t-1}|\theta^*)$.

   b- Figure out the ratio of densities (31) :

$$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}. \tag{31}$$

   c- Then, set the ratio (31) to the accept/reject step:

$$\theta^t = \begin{cases} \theta^* & with\ probability\ \min(r,1) \\ \theta^{t-1} & otherwise \end{cases} \quad (32)$$

Generating (31) uniform random number u on [0,1].

If $u \leq r, then\ accept\ the\ proposal\ \theta^t = \theta^*$,

If $u > r, then\ reject\ the\ proposal\ \theta^t = \theta^{t-1}$.

3- Stop if the converge is satisfied.

The next step is **Metropolis-Hasting Algorithm** which is defined as the generalized version of Metropolis algorithm (Hastings, 1970). The Metropolis-Hasting Algorithm is generalized in two ways:

2- Jumping distribution does not need to be symmetric.

3- The correction on the jumping rule, the ratio of r is different:

$$r = \frac{p(\theta^*|y) \Big/ J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y) \Big/ J_t(\theta^{t-1}|\theta^*)}. \quad (33)$$

4- Stop if the converge is satisfied.

Generalization boosts the speed of the random walk with these two changes. Then,

If $r \geq 1$, $then\ accept\ the\ proposal\ \theta^t = \theta^*$,

else

$$\theta^t = \begin{cases} \theta^* & with\ the\ probability\ r \\ \theta^{t-1} & with\ the\ probability\ 1-r \end{cases} \quad (34)$$

The Markov Chain is begun at the starting point, and the algorithm is run to obtain iterations when the starting value's effect is decreased, or forgotten as well. These

samples which are called as burn-in are eliminated. The remaining accepted values of θ^t provides a sample from generated target distribution $p(\theta|y)$ The procedure of the converge to the target distribution is the same with Metropolis algorithm (Gelman et al., 2014).

3.2.1.2 Gibbs Sampling

The simplest special case of Metropolis-Hasting Algorithm is Gibbs sampling. Gibbs Sampling, which is also called alternating conditional sampling, is explained with in terms of sub-vectors. Assume that θ is the parameter vector, and it is divided into number of d components or sub-vectors. Hence, each iteration of the Gibbs sampling contains sub-vectors and is cycling through the sub-vectors. Each subsets is drawn as conditional on the values of all others. Therefore, each iteration of t has the number of d steps. In each iteration of t, θ is selected into an ordering sub-vectors of d, and θ_j^t provides a sample from the conditional distribution given all the other components of θ, $p\left(\theta_j \middle| \theta_{-j}^{t-1}, y\right)$.

The procedure can be summed up with the following steps:

1- Select an initial value of θ^0
2- Next sample after θ^0 is θ^{i+1}. θ^{i+1} has sub-vectors which are sampled in a vector, $\theta^{i+1} = \left(\theta_1^{i+1}, \theta_2^{i+1}, \ldots., \theta_d^{i+1}\right)$. Each sub-vector is conditioned on the other sub-vectors so far.
3- Repeat the above steps to reach the desired sample size.

3.2.1.3 MCMC Diagnostic Test for Checking Converge and Stationary Status of Posterior Distribution

3.2.1.3.1. Geweke Diagnostic Test

Geweke (1992) developed a convergence diagnostic for Markov Chains. The diagnostic depends on the assumption that equality of the means of the first and last part of the chain.

If the samples are from the stationary distribution or with another words, the samples (X_1 and X_2) reach the target distribution, then the means of the first (10% by default) and the last part (50% by default) of the chain equal to each other and posterior distribution converged. Equation of the statistic (35) is that:

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}. \tag{35}$$

Geweke statistic (35) is an asymptotically standard normal distribution. The sample variances s_1^2 and s_2^2 need to be adjusted and the samples are not independent. According to the diagnostic, if the Geweke's statistic (p-value) is less than 0.05 or greater than 0.95, or with other words, z-score is less than -1.96 or greater than 1.96 , then this is an evidence against converge.

3.2.1.3.2. Heidelberger and Welch Diagnostic Test

Heidelberger & Welch, (1981) diagnostic proposed a test statistic based on the Cramer-von Mises test statistics. According to the test, the null hypothesis that the chain is approximately estimated from a stationary/target distribution.

The test has two parts:

First part:

It defines with Heidelberg,

1- Generate a Markov chain with N iteration and identify a ρ level
2- Figure up the test statistic for the whole chain. Accept or reject the null hypothesis. If the null hypothesis is accepted then, the chain generated from a stationary posterior distribution.
3- If the null hypothesis is rejected, then remove the first 10% of the chain and figure up test statistic again.
4- Repeat the third step until 50% of the chain is removed. Then, if the null hypothesis is still rejected, the test result of chain is failed.

Second part:

If the chain pass successfully in the first part, the test will continue with the un-removed part in the second part. The halfwidth test figures up half of the width the (1-ρ) % reliable interval around the mean.

If the ratio of the halfwidth and the mean is lower than some ϵ, then chain passes the test.

3.2.2 Prior Belief / Function

As it has been mentioned before, Bayesian statistic or specifically Bayes theorem contains two terms/components in order to calculate the posterior probability distribution. One of them is prior belief. Prior belief is the probability distribution which represents the uncertainty about the parameter. A Bayesian data analysis cannot be carried out without using the prior distribution. Therefore, there are several types of prior distribution in the literature. The type of priors can be divided into four main subjects:

- **Non-informative Priors:** Non-informative priors can be described as “flat” relative to the likelihood function. The flat prior means that it assigns equal likelihood to all possible values of the parameters. When a non-informative priors are used, the effect of the prior on the posterior distribution is minimum. Many researchers prefer using non-informative priors because non-informative priors appear more objective. However, it cannot be claimed that these kind of priors do not give any information about parameter of interest. Sometimes, for example, they might lead to obtaining “improper posteriors“ which are non-integrable posterior densities as well. (Kass & Wasserman, 1994) gives more information about derivation of the non-informative priors in a detailed way.
- **Informative Priors:** If a prior distribution dominates the likelihood, then it can be said to be an informative prior. Bayesian methods state that information which includes past experience, previous studies and expert’s opinion can be gathered and have an impact on the data analysis. In this case, informative priors function as a key to reach this purpose.
- **Improper Priors:** Improper priors are generally used in Bayesian inference since they produce non-informative priors and proper posterior distributions, which means that non-informative priors do not include any subjective effect of researcher’s opinion or any assumption came from past. They cannot effect Bayesian conditional probability of the model. Proper posterior distribution affects directly to the model. However, improper prior distributions can cause improper posterior distributions, which means that improper prior affect the conditional probability of the model wrongly with researcher’s improper opinion or wrong assumptions.
- **Conjugate Priors:** If the prior and posterior distributions come from the same family, which means that the form of the prior distribution and the form of the posterior distribution is the same, then we call these priors as conjugate priors.

3.3 Multi-Response Generalized Linear Mixed Models

Multi-response models are not widely used in general except for quantitative genetics and some other related areas. However, they allow for assumptions of single models, and thus can be used and can be an effective way handling with data missing problems and other related difficulties.

A new data structure needs to be improved for this type of dataset that has more than one response. Responses are arranged as a matrix. According to this matrix, each of the rows is indexed by reserved variable called as “units” and each column is indexed by reserved variable called as “trait”. Responses are stacked as column-wise and other variables stacked are duplicated respectively (Table 3.1).

Table 3.1 *Multi-Response Data Structure with Reserved Variables Trait and Unit.*

#	Response 1	Response 2	id
1	0.75351	1.036808	1
2	0.622868	1.150577	1
3	0.568975	1.231025	
…			
800	1.568974	0.231026	200

#	Response	Trait	id	Unit
1	0.75351	Response 1	1	1
2	0.622868	Response 1	1	2
3	0.568975	Response 1	1	3
…				
800	1.568974	Response 1	200	800
801	1.036808	Response 2	1	1
802	1.150577	Response 2	1	2
803	1.231025	Response 2	1	3
…				
1600	0.231026	Response 2	200	800

3.3.1 Zero- Inflated Models

When researchers face to the dataset which have extra zeros, the first type of models that comes to mind are the Zero- Inflated models. These models have been proposed

by Lambert (1992) and the main aspect of the models is assuming that data come from a mixture of a regular count distribution (to illustrate: Poisson) and a degenerate or untruncated distribution of zero. ZIP models assume that response for subject i (36) is:

$$Y_i \sim \begin{cases} Poisson(\lambda_i) & with\ probability\ \phi_i \\ 0 & with\ probability\ 1-\phi_i \end{cases} \quad (36)$$

The quantity of $1-\phi_i$ denotes the probability of structural zero. On the other hand, ϕ_i=1 denotes that probability of zero-inflation equals to zero. It means that zero-inflation is not necessary to be modeled and model turns to in an ordinary Poisson distribution. Except for these two conditions, zeros of the data are inflated.

The probability distribution of Zero-Inflated Model (37) is:

$$P(Y_i = 0) = (1-\phi_i) + \phi_i e^{-\lambda_i}. \quad (37)$$

where $0< \phi_i <1$

$$P(Y_i = j) = \phi_i \frac{e^{-\lambda_i}\lambda_i^j}{j!} \quad \text{j=1,2,….} \quad (38)$$

Logistic regression zero-inflated process (40) is:

$$\log(\lambda_i) = x_{1i}^T \beta_1, \quad (39)$$

$$logit(\phi_i) = x_{2i}^T \beta_2\ . \quad (40)$$

With Expectation Maximization (EM) Algorithm or Newton-Raphson method, the model parameters can be estimated (Min & Agresti, 2005).

Zero-Inflated model has two latent variables. These latent variables are estimated by The EM algorithm or MCMC methods. In zero-inflated models, the first latent variable is associated with the named distribution and second latent variable is associated with zero inflation. The model's aspect depends on modeling a mixture distribution of zeros originating form of the named distribution (for example Poisson) and zeros originating from zero-inflation. It is actually a probability on the logit scale (37) and this probability is that a zero-inflation process with the second latent variable.

To provide overall population mean (41), combine $v_i(x_i) = \lambda_i\ (x_i)[1 - \phi_i\ (x_i)]$:

$$E(Y_{ij}|x_{ij}) = \frac{\exp(x_{1i}^T\beta_1)}{(1+\exp(x_{2i}^T\beta_2)} . \qquad (41)$$

(Preisser et al., 2012).

Important notes for the zero inflated models:

- If zeros of the data is expected to be around 30%, we expect zero-inflation to be a problem (Hadfield, 2016).
- Any residual variance cannot be observed in zero-inflated process and in addition, the residual covariance between the zero-inflated and the named distribution cannot be estimated because these processes cannot be observed in one data point.
- Especially, compared with Hurdle Models, the parameters of the zero-inflated models converge poorly.
- Model allows only zero-inflation process.
- Poor mixing of the parameters might arise when either distribution is not zero-inflated or the model is over-dispersed.

3.3.2 Hurdle Models

Hurdle model, which has been proposed by (Mullahy, 1986) a type of models for count data which can cope with excess zero and over-dispersion. It is also very similar with Zero-Inflated models. The difference of hurdle model from the zero-inflated model is that in the former zero-deflation can be used in addition to zero-inflation. Hence, hurdle models mix much better than Zero-Inflated models.

Hurdle model has two latent variables like Zero-inflated models. However, the first latent variable is the mean parameter of a zero-truncated named distribution (for example Poisson) and this model explains the observations bigger than the hurdle. On the other hand, second latent variable in Zero-Inflated models is the probability of observing zero because of zero-inflation, but in hurdle models, second latent variable is the probability (on the logit scale) of the model response which is zero or not. The probability mass function of Hurdle model (42) is:

$$\mathrm{P}(Y_i = y_i) = \begin{cases} w_0 & for\ \ y_i = 0 \\ (1 - w_0)\dfrac{e^{-\lambda_i}{\lambda_i}^{y_i}}{(1 - e^{-\lambda_i})y_i!} & for\ y_i > 0 \end{cases}, \qquad (42)$$

where $0 < w_0 < 1$.

If the probability of observed values which are bigger than 0 ($\mathrm{P}(Y_i > 0) = 1 - w_0$ and the probability of observations which equals to 0 ($\mathrm{P}(Y_i = 0) = w_0$, the logistic regression model of w_0 and a log-linear model for μ_i=λ_i of the truncated Poisson distribution (44) :

$$\log(\lambda_i) = x_{1i}^T \beta_1, \qquad (43)$$

$$logit(w_0) = \log\left(\frac{w_0}{1-w_0}\right) = \ x_{2i}^T \beta_2. \qquad (44)$$

Then, expected value of response in Hurdle model (45) is given as such:

$$E(\hat{y}_i) = \sum_{y_i=0}^{\infty} y_i \, Pr(Y_i = y_i) \tag{45}$$

Hurdle Models:

- It has better mixing properties than Zero-Inflated models.
- While it is assumed that zero observations come from the origins of "sampling" or from the "structural" in Zero-Inflated models, in hurdle models, it is assumed that zero observations come from only one "structural" source.

3.4 Markov Chain Monte Carlo Generalized Linear Mixed Models (MCMCglmm) Package in R

Until this point, methodological background of GLMM and MCMC algorithms were told. In this study, MCMCglmm Package, which is a package in R, combining all the methods was used and whose name is MCMCglmm Package. The package exists in the statistical software of R and its development process is still continued by the researchers.

As a closing remark, a few reasons as to why we select the MCMCglmm packages can be revealed here. In the first place, MCMCglmm package is suitable for longitudinal-discrete data that has been used. Next, it provides a flexible framework for modeling with GLMM and it can be used for non-Gaussian response variables for which we cannot be obtained likelihood in a closed form as well. Then, the package can be used in multi-response models for the distributions of Gaussian, Poisson, exponential, zero-inflated and censored distributions. Last, it allows for complicated variance structures as well (Hadfield, 2010).

CHAPTER 4

DATA DESCRIPTION, MODEL APPLICATION AND EMPRICAL RESULTS

Chapter 3 gave the methodological background of the data analysis related to this study. In this chapter, first section will explain the structure and details of the electricity interruption dataset used in the study. Then, the results from the applied models are going to be explained and their results will be compared with each other in the second section.

4.1 Data Description

The dataset used in this study provides information about the quality of the electricity supply continuity in Çankırı province and its neighborhoods. Çankırı is a city in the distribution region of Başkent Electricity Distribution Company. Entity of the dataset includes electricity interruption counts and variables of Çankırı for the year of 2015. These kind of datasets are published by 21 of the distribution companies in Turkey, in accordance with EMRA's regulation articles (EMRA, 2008). Dataset includes interruptions':

- Location (city, town, etc)
- Network component,
- Resource type (Low Voltage, Medium Voltage) ,
- Zone type of location (Zoned or not zoned) ,
- Duration,
- Count,
- Time (month, day, year),
- Reason (Operator, security, out of the distribution region),
- Explanation of reason (power switch breakdown, electrical fuse breakdown etc.)

After investigating the literature and evaluating the variables, location (Location), time (Month), counts of electricity interruption (Ycount), location status of zone type (Xin_out), resource type (Xlv_mv) and reasons (Xreason1 and Xreason2) have been used as covariates in this study. A part of dataset can be seen in Appendix-E.

Each row shows interruption's location, (city and town), resource type (low voltage or medium voltage), location status of zone type (zoned or not zoned) and its reasons, and the time when the interruption occurs (month), and the number of interruption counts at that particular time. Why were these variables selected for the models? Firstly, location, and location status of the zone give hints regarding the percentage of the electrical usage and the development of electrical network in a specific location. To illustrate, location status of the zone provides information about the density of the population. If the area is located in the zoned land, it means that density of population is higher than un-zoned land. Secondly, the variable of month gives the variation of interruption's count or duration according to time. Thirdly, the variable of reason gives the information about which kind of fault or misusage of the network might be causing the electrical interruption.

The variable of reason includes three main options: external, operator and security. One of these options is external, which means that problems from out of the region, which do not stem from the local distribution company, cause electrical interruptions in one particular location. Second option is the operator. Operator indicates the presence of problems due to the fault of the operator who works for the local distribution company. For example, in a construction process that is carried out to improve electrical network one operator might be the cause of power failure. On such occasions, the reason for electrical interruption is the operator. Another option is security. Security implies possible unauthorized interventions in the network. To illustrate, a resident might be building a new house and when he/she fails to get permission to intervene in the local network system, interruptions might occur because of his / her fault. These options are defined by the electrical distribution company and

these definitions differ from one electrical distribution company to the others. These three options represented as Xreason1 and Xreason2, and can be explained together such as:

- When Xreason1 and Xreason2 equal to 0, it indicates that interruption's reason is external,
- When Xreason1 equals to 0 and Xreason2 equals to 1, it indicates that interruption's reason is security,
- Finally, when Xreason1 equals to 1 and Xreason2 equals to 0, it indicates that interruption's reason is operator.

The variables of Xlv_mv and Xin_out have the same representations like Xreason1 and Xreason2. They take 0 and 1 value. They are coded as follows:

For Xlv_mv:

- 0 indicates that electrical interruption occurs in the network of medium voltage.
- 1 indicates that electrical interruption occurs in the network of low voltage.

For Xin_out:

- 0 indicates that electrical interruption occurs in un-zoned land and effects mostly the resident who lives there.
- 1 indicates that electrical interruption occurs in zoned land and effects mostly the resident who lives there.

4.2 Exploratory Data Analysis

With the aim of summarizing and evaluating the data used in the study, this section will present some descriptive data analysis.

4.2.1 Descriptive Statistics

All the analysis given in this section for the description of the data has been made by using the software of R (Hadfield, 2010).

In first place, frequency tables of the variables will be given. (Table 4.1)

Table 4.1 *Frequency Tables of the Variables*

		Xin_out	
		un-zoned	zoned
Xlv_mv	medium	53	23
	Low	131	80

Reasons		
External	Security	Operator
97	5	185

Ycount	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*≥10*
#	1441	124	49	27	13	21	16	8	7	4	18

The total numbers of observation are 1728. According to the Table 4.1, the dataset includes excessive zeros where total number is 1441. After eliminating zeros from the dataset, the electricity interruptions occurred in medium voltage vs un-zoned area is 53, in medium voltage vs zoned area is 23 and in low voltage vs un-zoned area is 131, in low voltage vs zoned area is 80. On the other hand, the electricity interruption occurred because of the external reason is 97, because of the reason of security is 5 and because of the operator is 185. It is understood that most of electricity interruptions occurred due to the reason of operator.

On the other hand, it can be seen that 83.4% of total number of observations are zero. The data conditions are suitable with general assumptions except of being observed

zero counts (see section 3.3.1 and 3.3.2). After the observed value of 10, the number of observations are getting decreased. Their total number of observed value which equals and be bigger 10 is 18. The maximum observed value is 52 which was observed only one time.

At this point, it would be beneficial to look at some useful plots, to understand the data (Figure 4.1 and Figure 4.2).

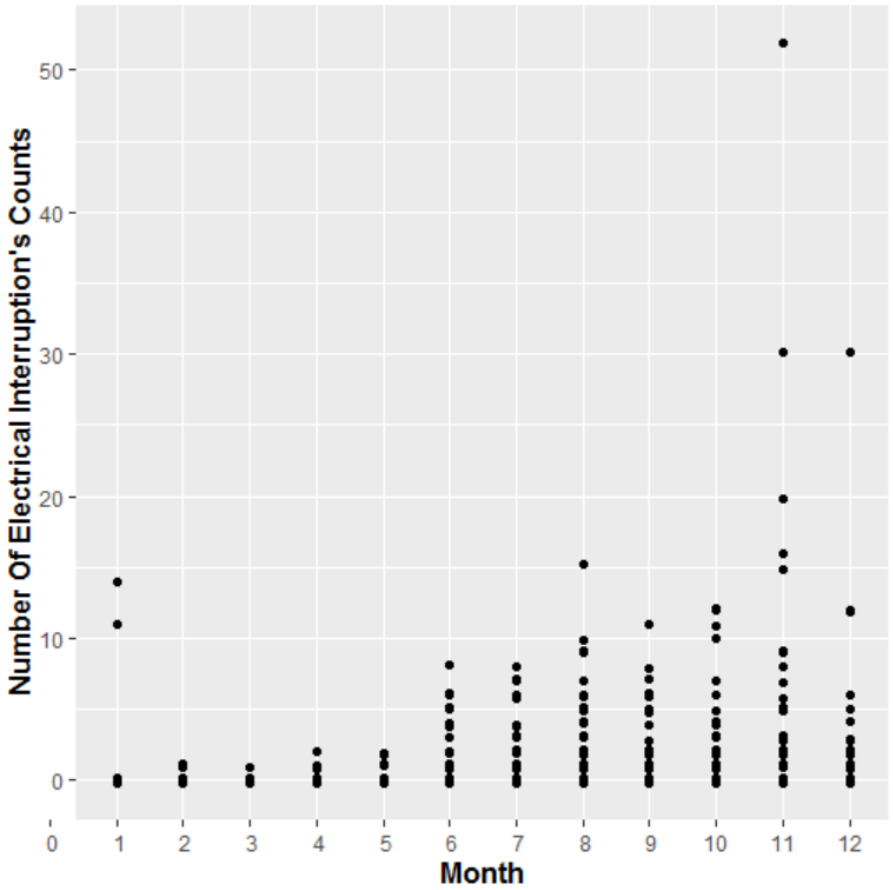

Figure 4.1 *Plot of Number of Interruption Counts vs Month*

According to Figure 4.1, the number of interruption counts rises from January to December. Nevertheless, it is possible to observe excess zero counts for every month. On the other hand, another important point is to see the highest interruption count,

which is 52, in November. These observed values are extremely different than others because most of the observed values are generally lower than 10. If BEDAŞ had failed to solve the problems, probably, a failure either in an important component or in any other element of the network might have occurred repeatedly. For this reason, more than one interruption might have occurred at that time.

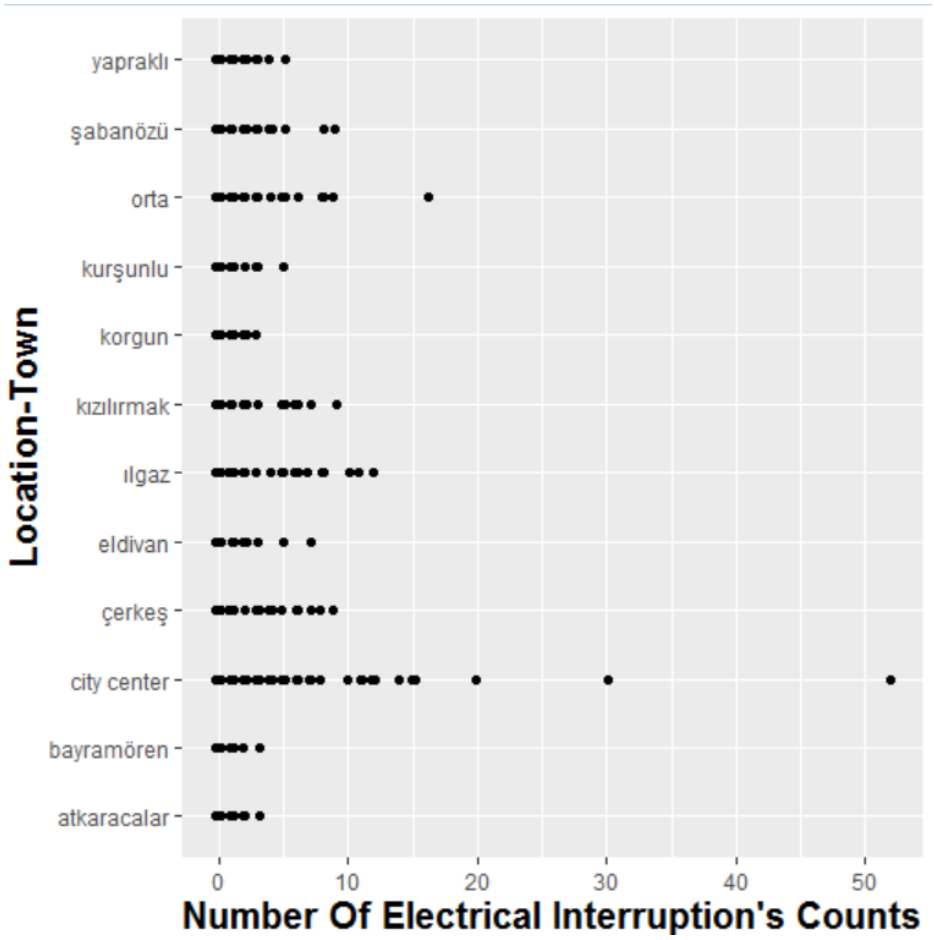

Figure 4.2 *Plot of Number of Interruption Counts vs Location-Town*

Figure 4.2 shows interruption counts in towns of Çankırı. It can be observed that majority of electricity interruptions occurred in the city center. This situation can be explained with continuing process of development and resident's intervention in the

electrical network mostly. Interruption count of town of Orta and town of Ilgaz follow this score. The observed value which occurred in the city center in November can be defined as an outlier.

The data includes excessive zeros. However, Figures 4.1 and 4.2 cannot show this situation precisely. Therefore, bar chart graph can be assisted here to better observing for the density of zero. (Figure 4.3)

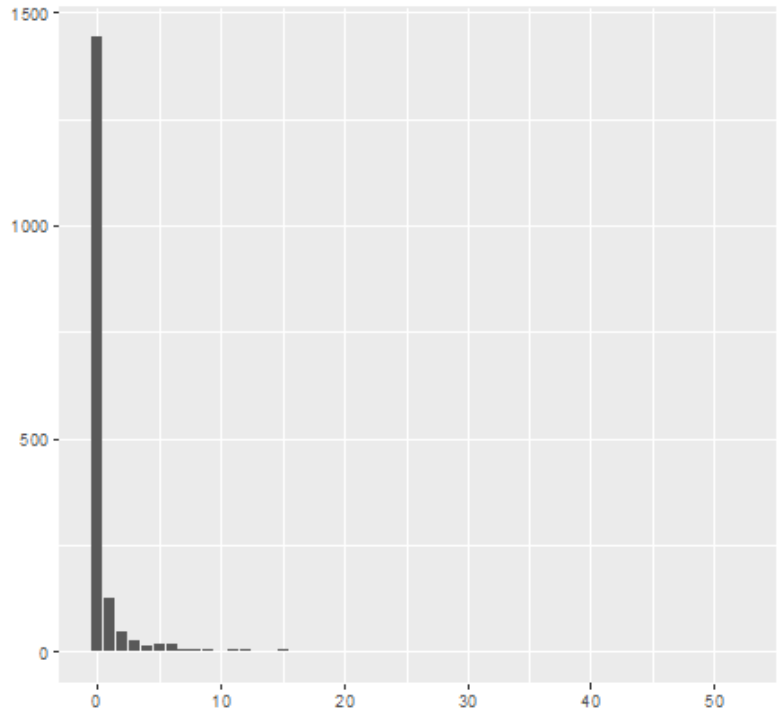

Figure 4.3 *Bar Chart for Density of Number of Electrical Interruption's Count*

Figure 4.3 shows the density of interruption counts. According to the bar chart, data has obviously excessive zeros. Number of electrical interruptions take zero value mostly when no electrical interruption is observed in any location, during that month. On the other hand, the electrical interruption count of one location can be measured periodically (each month) in different occasions. These specifications show that data structure is compliable with longitudinal panel data. Also, the data regarding the

response of the number of electrical interruption's counts is discrete. As these specifications suggest Poisson or zero-inflation models might be employed for using while modeling this type of data.

4.3 Empirical Results of the Data Analysis

In Sections 4.1 and 4.2, data analysis and important points were offered. This section presents the implementations of Poisson, zero-inflated Poisson and hurdle Poisson MCMCglmm, their important points and outputs will be given in detail.

It has been mentioned before that the data structure used in this study, which is called electricity interruption data, is a count-longitudinal type of data. It is clear that the model conditions are not suitable for modeling with the linear models because of having longitudinal, discrete and non-Gaussian properties. Also, it is assumed that the location of the electricity interruption impacts randomly on the response of the data, which is defined as Ycount. In this case, the variable of location includes unobserved variables such as numbers of consumers or residents, total investments made by local distribution company etc. Some heterogeneity in electricity interruptions is expected among different towns. In order to account for this heterogeneity, we include the location variable as a random effect in the models. Longitudinal-count data with random effects are handled by Generalized Linear Mixed Models (see section 3.1.5).

Although GLMM provides a flexible framework for modeling non-Gaussian response variable, because of longitudinal type of data, the likelihood function cannot be obtained easily. To cope with this problem, MCMC techniques were used in this study. The package of MCMCglmm (Hadfield, 2010) uses MCMC techniques which are combined with Metropolis- Hasting algorithm and Gibbs sampling when it generates the posterior distribution.

In this section, in order to get best explanatory estimation model, Poisson, zero-inflated Poisson and Hurdle Poisson MCMCglmm models have been applied separately. The results of each model will be given with this following alignment:

1- Results of the models with the first implementation on Poisson and Zero-Inflated Poisson MCMCglmm,
2- Results of the models with adding interaction effects and piecewise indicator variable to the Poisson and zero-inflated Poisson MCMCglmm
3- Results of the models with final implementation on Poisson, zero-inflated Poisson and Hurdle Poisson MCMCglmm
4- Posterior predictive checks and comparison of the final models of Poisson, zero-inflated Poisson and hurdle Poisson MCMCglmm

4.3.1 Results of the Models with the First Implementation on Poisson and Zero-Inflated Poisson MCMCglmm

In this section, the first implementation of Poisson and zero-inflated Poisson MCMCglmm will be told. First of all, the conditions and the model inputs, which are prior function, number of iterations, thinning interval etc, needs to be explained.
Both Poisson and zero-inflated Poisson models are ran by:

- Only the fixed effects of the covariates
- The number of iteration has been set to 50,000.
- Burn-in period is taken as 3,000,
- The thin value has been equaled to 10,
- Sample size has been 4,700 (after eliminating 3,000 of burn-in period the number of iteration is divided by thin value).
- The variable of the location is used in the models as random effect parameter.
- The default prior function was used for the models which were given in this section. The default prior function contains three elements: B, R and G

structures. B structure defines the fixed effects' prior distribution which is multivariate normal with mean vector (mu) 0 and variance-covariance matrix of a diagonal with large variances (10^{10}). Also, priors for the variance structure of R and G elements have inverse-Wishart with expected covariance equals to 1 and degree of belief parameter is 0 (Hadfield, 2010).

4.3.1.1 First Implementation of Poisson MCMCglmm

Due to the p-value of MCMCglmm (Table 4.2) all of the model coefficients are significant.

Table 4.2 *Summary of the First Implementation of Poisson MCMCglmm.*

	Post Mean	l-95% CI	u-95% CI	Efficient Sample	p-MCMC	
(Intercept)	-6.26	-7.18	-5.35	232.5	2x 10^{-4}	***
Xlv_mv	2.15	1.50	2.83	330.7	2x 10^{-4}	***
Xin_out	-0.99	-1.59	-0.40	633.7	4.26 x10^{-4}	***
Month	0.34	0.30	0.39	692.4	2x 10^{-4}	***
Xreason1	1.74	1.04	2.36	765.4	2x 10^{-4}	***
Xreason2	-3.75	-4.91	-2.61	98.4	2x 10^{-4}	***

Below is the open form of the model (46), as offered in the summary:

$$\log\left(\left(\widehat{Y_{ij}}\middle|b_{Location}\right)\right) = -6.26 + 2.15Xlv_{mv} - 0.99Xin_{out} + 0.34Time + 1.74Xreason1 - 3.75Xreason2 + \widehat{b_{Location}}. \quad (46)$$

This open formula illustrates that intercept, the covariate of Xin_out and the covariate of Xreason2 have the negative behavior in the model, while the others have the

positive behavior for the estimation of log(E(Y_{it})). Since Generalized Linear Mixed Model of Poisson has canonical link function, the response of the model acquires log link function in the open form (see section 3.5.1). On the other side, covariates of the model except the covariate of month are to be considered as the binary variables explained in the previous chapter. In addition, it should be noted that covariates which have negative behavior lead to the response where (log($\widehat{Y_{ij}}$)) is getting lower values than the value it gets in other covariates. Perhaps, it could be better expressed through a scenario such as this: The estimated interruption value which occur due to a fault of the operator in low voltage in a non-urban area in January is to be calculated. Covariates of this scenario are as follows : Xlv_mv equals to 1 , Xin_out equals to 0, Month equals to 1, Xreason1 equals to 1 and Xreason2 equals to 0:

$$\widehat{Y_{ij}} = e^{(-6.26+2.15(1)-0.99(0)+0.34(1)+1.74(1)-3.75(0)+b_{\widehat{Location}})} \qquad (47)$$

$$\widehat{Y_{ij}} = e^{(-2,03)} \qquad (48)$$

When $b_{Location} = 0$ (i.e. for an "average" location), the estimated value of the imagined scenario (48) is ≅0.13, which is quite close to 0. It is a probable estimate in the presence of the data which has many zeros. If the effects of covariates are observed more closely, it is to be seen that while the fitted value which is in low voltage and non-urban area, and in January due to a fault of Operator (Xlv_mv = 1) is 0.13, the fitted value with the same covariates except now in medium voltage (Xlv_mv = 0) becomes 0.015, which is even closer to 0. The results of this scenario thus show that the likelihood of an interruption is higher in a low voltage area than it is in the medium voltage area. This might sound strange, but in fact it is understood to be more reasonable since in low voltage areas, electricity systems are opened to intervention and thus an interruption is more likely to occur in this system than a in a high voltage area system. High voltage system is used for transfering the electricity power in long distances like between cities or regions. TEIAS is responsible for this type of transfer and the data is used in this study does not include high voltage area situation. However,

low or medium voltage systems is used for short distances such as between houses, neighborhoods or districts. Low voltage systems are open for intervention of people. This kind of intervention can occur for construction activities, infrastructure occupations etc. Therefore, probability of electricity interruption can be seen higher in low or medium areas than in a high voltage area system.

Let's focus on other covariates now. The result of the model lays bare that when other conditions are kept stable, interruption statistics on urban area (i.e. Xin_out=1) are closer to zero compared to the statistics in non-urban areas should be underlined. Non-urban area cover the villages or small districts which are close to the cities. Electricity lines and infrastructure are secured, and also voltage fluctuation is not seen very often in urban areas. However, electricity lines and infrastructure cannot be secured in non-urban area since construction and improvement of infrastructure continue there. Since it is usually the decreasing of the voltage that is to be blamed for the type of electricity interruptions mentioned here, it can be concluded that probability of interruptions in urban areas is lower than the probability in non-urban areas according to the model.

After explaining the model, and parameters, it needs to be looked at the estimated/fitted values and their comparison with the observed values.

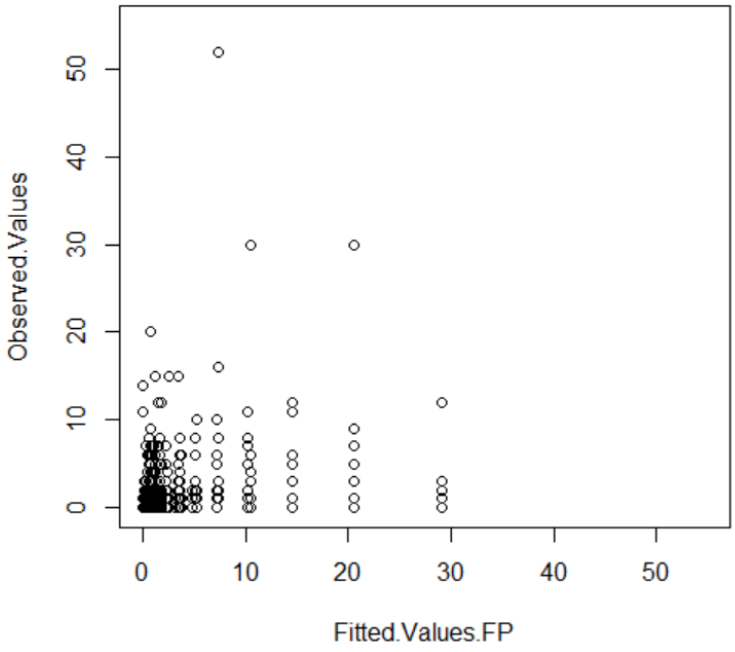

Figure 4.4 *Observed vs. Fitted Values of the First Implementation of Poisson MCMCglmm.*

In Figure 4.4 comparison of observed vs. fitted values is shown. According to the plot, fitted and observed values appear between 0 and 5 frequently. This indication is reasonable since 0 values in observed data are more frequent than other values. However, model is not good enough to explain observed values completely. To illustrate, fitted values of the model may estimate 5, 10, 15, 20 and bigger than 25, when observed value is 0. On the other hand, although the model's convergence seems reasonable, the autocorrelation is not good enough to be trustworthy (see Appendix-A section 1). For this reason, it is favorable to evaluate residuals status. At the end of the residual's checking, the model needs to be modified with some extra techniques probably.

This situation can also be seen in the residual vs fitted values and covariate plots (Figure 4.5). In the residual plots, the residuals are located between 40, which is upper bound, and -20 which is lower bound; hence the residual interval is so wide.

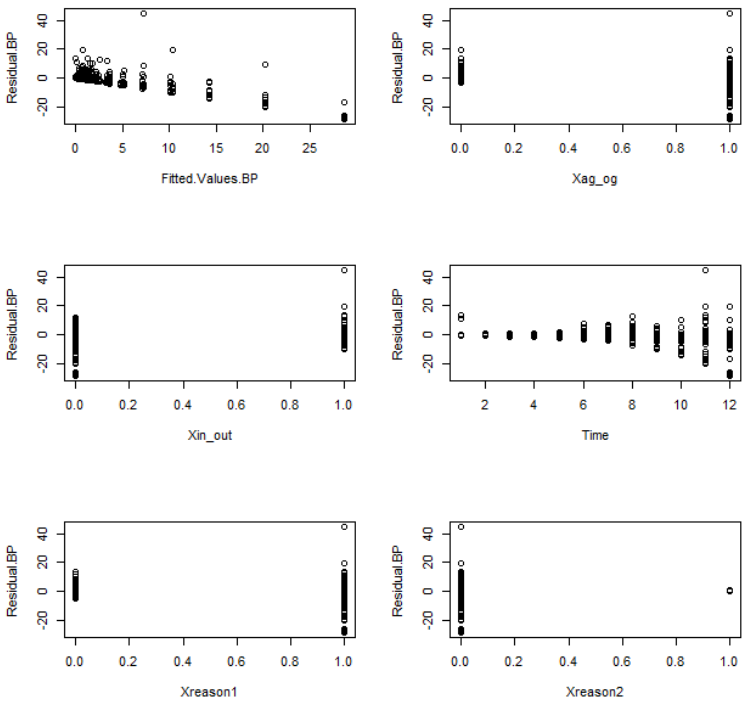

Figure 4.5 *Residual vs Fitted Values and Covariates of the First Implementation of Poisson MCMCglmm*

As justified by the presence of above mentioned problems, it is acknowledged that this is not the best model and it needs to be updated. The prediction fails to give the fitted values truly according to observed data. In the following parts, interaction effects and adding a slope term can be tried as a solution. Besides, the zero-inflated Markov Chain Monte Carlo Generalized Linear Mixed Models could be implemented to the data, as the data has excessive zero value in general.

4.3.1.2 First Implementation of Zero-Inflated Poisson MCMCglmm

The dataset of the study and its specifications have been expressed in section 4.1. The dataset is a specific longitudinal data with many zeros. At this point, it should be remembered that the response of the data used in this study has 83.4% of zeros. This situation brings to mind zero-inflation methods. In this section, application and results of zero-inflated Poisson model will be provided.

First of all, zero-inflated Poisson model used in this section has only the fixed effects like Poisson MCMCglmm model that was explained in the previous section. That is to say, there are not any interaction effects in the model. Besides, Zero- Inflated MCMCglmm is also a type of Multi-Response Model. Main structural differences in the model are those: The mean of reserved variable which is called "trait" can be added. Then, the unit shows response values in each row of "traits", which emerge as one type of response in MCMCglmm package for multi-response models. (see section 3.4). Considering these, it can be claimed that the first major difference from the first implementation of Poisson model is the residual covariance matrix. In Zero-Inflated Poisson MCMCglmm, the model works with heterogeneous residual variance. In other words, the residual (co)variance matrix allows each unit of the model to have different residual variances. In addition, the residual (co)variances between zero inflation and the Poisson process cannot be estimated because processes cannot be simultaneously observed in one data point.

The summary table of the model is given in Table 4.3.

Table 4.3 *Summary of the First Implementation of Zero-Inflated Poisson MCMCglmm*

	Post Mean	l-95% CI	u-95% CI	Efficient Samples	p-MCMC	
Intercept_Poisson	-5.48	-6.29	-4.38	7.59	$2x\ 10^{-2}$	**
Intercept_ZeroInflated	-2.36	-2.78	-1.73	6.67	$2x\ 10^{-2}$	**
Xlv_mv	2.40	1.67	2.89	4.58	$2x\ 10^{-2}$	**
Xin_out	-1.17	-1.53	-0.67	20.67	$2x\ 10^{-2}$	**
Month	0.35	0.30	0.39	23.43	$2x\ 10^{-2}$	**
Xreason1	1.55	0.88	2.31	7.56	$2x\ 10^{-2}$	**
Xreason2	-4.62	-5.41	-3.35	3.22	$2x\ 10^{-2}$	**

Predicted values of the model can be generated through two process which are Poisson process and zero-inflated process (49):

$$\log(\lambda_i) = -5.48 + 2.40Xlv_{mv} - 1.17Xin_{out} + 0.35Month + 1.55Xreason1 - 4.62Xreason2 + \widehat{b_{Location}}. \quad (49)$$

Second process (50) is:

$$logit(w_0) = \log(\frac{w_0}{1 - w_0}) = -2.36 \quad (50)$$

In the first step (49), $\log(\lambda_i)$ shows the usual generalized linear mixed effect regression model of Poisson process, whereas $logit(w_0)$ in equation (50) shows regression model of Zero-Inflated Poisson process, which is a logit model by default (see section 3.4.1). In zero-inflated process, there exists only the coefficient of intercept.

The summary of the model suggests that the parameters of the MCMCglmm of Zero-Inflated Poisson have the same behavior as the first implementation of Poisson MCMCglmm: Intercept, the covariate of Xin_out and the covariate of Xreason2 have the negative behavior in the model. However, others have the positive behavior on estimation. The results of the model are close to the results of the first implementation of Poisson MCMCglmm. However, before comparing the prediction results from two different models, it would be more appropriate to look at the trace and density plots of the first implementation of zero-inflated Poisson Model (see Appendix-A section 2).

Even though the autocorrelation problem exists according to autocorrelation plots (see Appendix-A section 2) for the first implementation of Zero-Inflated Poisson, it is still beneficial to look at the plot of Observed vs. Fitted Values to compare the models (Figure 4.6).

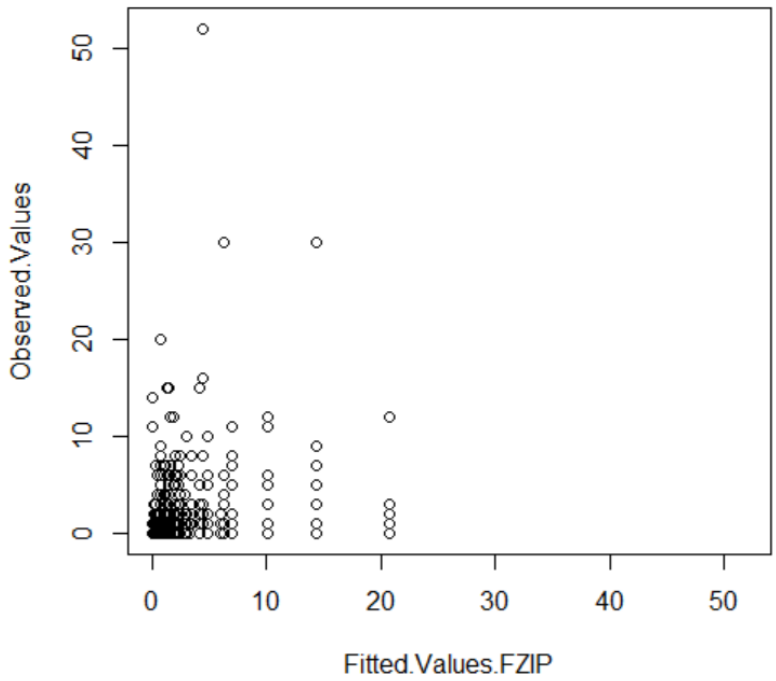

Figure 4.6 *Fitted vs Observed Values of the First Implementation of ZIP MCMCglmm*

Figures 4.6 and 4.7 make it clear that the same problem in the first implementation of Poisson MCMCglmm exists here, too. The model cannot effectively estimate the observed values. In the zoomed plots, this situation can be observed clearly. The first zoomed plot indicates the fitted vs observed values which are smaller than 10 and second zoomed plot also demonstrates the values which are bigger than 10.

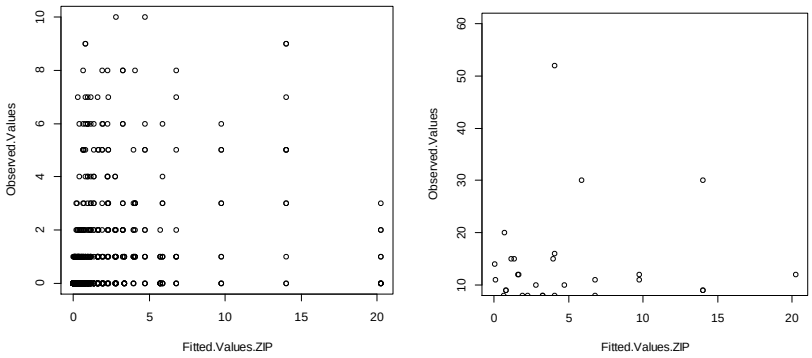

Figure 4.7 *Zoomed Plots for Fitted vs Observed Values of the First Implementation of ZIP MCMCglmm*

The model gives slightly a better estimation than the first implementation of Poisson MCMCglmm. This result can be observed in first zoomed plot clearly since the range of fitted values is between 0 and 20, while the range of fitted values of the first implementation of Poisson MCMCglmm takes values bigger than 25. On the other hand, in second zoomed plot gives when 10 or bigger values are observed in real dataset, fitted values take different values, the range of which changes between 0 and 20. This situation is an indication of the need for modifications to the model, especially when the observed values bigger than 10.

Residuals can be plotted versus observed covariates and fitted (Figure 4.8).

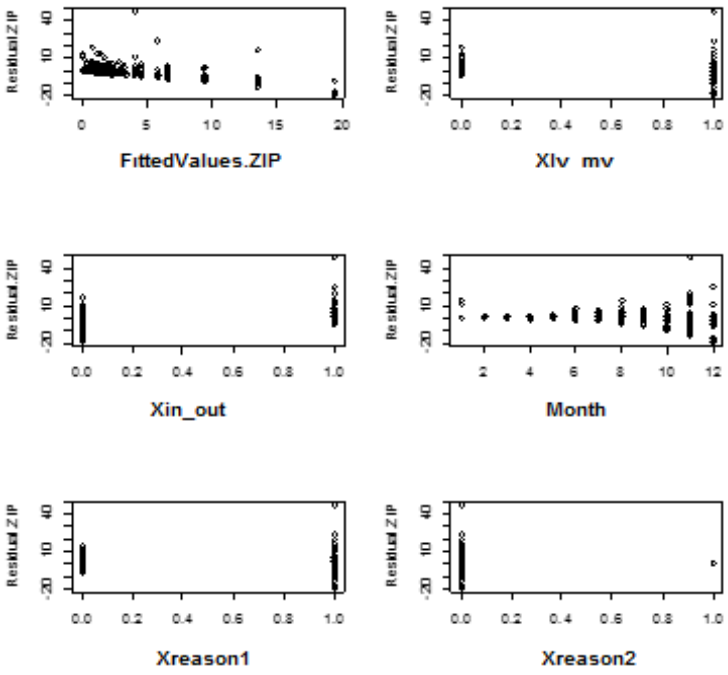

Figure 4.8 *Residual vs Fitted Values and Covariates of the First Implementation of ZIP MCMCglmm*

Residual plots show that fitted values of the first implementation of Zero-Inflated Poisson model are not entirely compatible with observed values. Thus, it is understood that ZIP does not fully solve the problems which were encountered in the use of first implementation of Poisson models. It is thought that problems generally arise due to the observed values which are bigger than 10. With the aim of effectively handling this situation, modifications to the models will be applied through next sections.

4.3.2 Results of the Models with Added Interaction Effects and Piecewise Indicator Variable to the Poisson and Zero-Inflated Poisson MCMCglmm

In Section 4.3.2, the first implementation of Poisson MCMCglmm and Zero-Inflated Poisson MCMCglmm were discussed. These models had only fixed effects. However, outputs of estimated values showed that especially the values which were bigger than 10 did not fit very well. This situation brings to mind options such as adding interaction effects and piecewise indicator variable to the models. Therefore, in this section, the versions of models with added significant interaction effects and piecewise indicator variable are to be discussed. The newly developed models and their differences are going to be mentioned as well.

A modification to the first implementation of Poisson and ZIP MCMCglmm were necessary because the model did not fit well especially to the observed values bigger than 10. If the data summary of the model is evaluated, it can be easily noticed that observed values which are bigger than 10 are seen after June (i.e. Month > 6) (Section 4.2). The approach offered by Piecewise Linear Regression Technique could be used one for this situation.

This method advocates using different intercept and/ or slope parameters to the model. In this way, after one point is taken as origin, the intercept and/or slope of the model changes. While applying this method, a new variable, “newMonthx”, was added to the model. For this, we first define Monthx variable: If Month is smaller than 6 or equals to 6, Monthx is 0, or else 1. The newMonthx is defined as the interaction of centered Month, i.e. (Month-6), and Monthx. These variables may be formulated as below:

Monthx = if else(Month<=6,0,1)
newMonthx= (Month-6) x Monthx

Both Poisson and zero-inflated Poisson MCMCglmm with being added interaction effects and piecewise indicator variable are run by:

- Using fixed and interaction effects,
- The number of iteration has been set to 50,000.
- Burn-in period is taken as 3,000,
- The thin value has been equaled to 10,
- Sample size has been 4,700 (After eliminating 3,000 of burn-in period the number of iteration is divided by thin value).
- The variable of the location is used in the models as random effect parameter.

The default prior function was used for the models which were given in this section. The default prior function contains three element: B, R and G structures. B structure defines the fixed effects' prior distribution which is multivariate normal with mean vector (mu) 0 and variance-covariance matrix being a diagonal with large variances (10^{10}). Also, priors for the variance structure of R and G elements have inverse-Wishart with expected covariance equals to 1 and degree of belief parameter is 0 (Hadfield, 2010).

In the following section, Piecewise Linear Regression Technique and significant interaction effects will be used to develop the models of Poisson MCMCglmm and Zero-Inflated Poisson MCMCglmm.

4.3.2.1 Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable

To begin with, it should be noted that the output of first implementation of Poisson model had shown that model was not fitting well for the values bigger than 10. Firstly, the method of adding interaction coefficients was tried to solve this problem. The interaction effects of Xlv_mv vs Month , Xlv_mv vs Xreason1 and Xin_out vs Month become significant in the model (Table 4.4).

Table 4.4 *Summary of Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable*

	Post Mean	l-95% CI	u-95% CI	Efficient Samples	p-MCMC	
(Intercept)	-9.68	-11.64	-7.93	79.06	$2x\ 10^{-4}$	***
Xlv_mv	4.05	2.26	5.93	56.85	$2x\ 10^{-4}$	***
Xin_out	-2.92	-4.10	-1.83	240.40	$2x\ 10^{-4}$	***
newMonthx	-0.72	-0.93	-0.51	322.05	$2x\ 10^{-4}$	***
Xreason1	5.21	3.75	6.92	46.71	$2x\ 10^{-4}$	***
Xreason2	-4.29	-5.58	-3.00	82.56	$2x\ 10^{-4}$	***
Month	0.67	0.51	0.86	292.14	$2x\ 10^{-4}$	***
Xlv_mv : Month	0.17	0.07	0.27	586.61	$2.13x\ 10^{-6}$	**
Xlv_mv:Xreason1	-4.71	-6.55	-3.09	47.40	$2x\ 10^{-4}$	***
Xin_out: Month	0.22	0.11	0.33	216.25	$2x\ 10^{-4}$	***

After adding interaction effects and slope parameter to the model, model's open form (51) changes as such:

$$\begin{aligned}\log(Y_{ij}) = -9.68 + 4.05Xlv_{mv} - 2.92Xin_{out} - 0.72newMonthx \\ + 0.67Month + 5.21Xreason1 - 4.29Xreason2 \\ + 0.17Xlv_{mvxMonth} - 4.71Xlv_{mvxXreason1} \\ + 0.22Xin_{outxMonth} + \widehat{b_{Location}}.\end{aligned} \quad (51)$$

Under the condition of $\widehat{b_{Location}} = 0$, the fitted value (51) is ≅0.014, when the same scenario is given: interruption due to a fault of operator in lower voltage and non-urban area in January (Xlv_mv equals to 1, Xin_out equals to 0, Month equals to 1, Xreason1 equals to 1 and Xreason2 equals to 0). On the other hand, newMonthx coefficient is invalid here because Month variable is 1 and value is smaller than 6. To understand the effect of piecewise indicator variable, Month variable should be

changed from 1 to 7. Thus newMonthx variable takes the value of (7-6) x1 and equals to 1. The result generating from this condition is ≅0.03. While the value of Month variable is increasing, predicted counts are expected to be increasing, too. The solution subsides with this approach completely. Yet, before proceeding with the next stage, a check for autocorrelation is needed (see Appendix-B section 1).

According to autocorrelation plot, autocorrelation problem gets higher here compared to the first implementation of Poisson MCMglmm. Xreason2 variable is still autocorrelated but, intercept, Xreason1 and Xlv_mv variables have also autocorrelation problem (see Appendix-B, section-1).

While developed model cannot fix the problems of autocorrelation, the fitted values are expected to be better than the first implementation of Poisson MCMCglmm. In order to confirm this, it is essential to study the observed vs fitted plot and residual plots as presented below (see Figures 4.9 and 4.10).

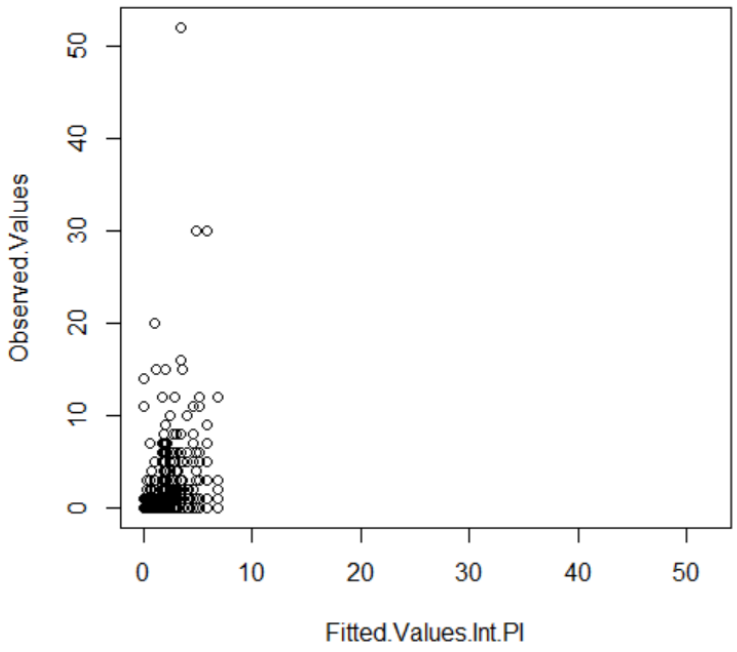

Figure 4.9 *Observed vs Fitted Values of Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable*

The plot of observed vs fitted values shows that the difference of observed /fitted values decreased from 25 to 6. It is obvious that developed model has increased the strength of estimation compared to the first implementation of Poisson MCMCglmm. However, still the model cannot rightly estimate the observed values bigger than 10. The possible causes might be autocorrelation problem, over-dispersion problem or using Poisson distribution. With the aim of overcoming this, ZIP MCMCglmm model is to be developed with interaction effects and piecewise indicator variable.

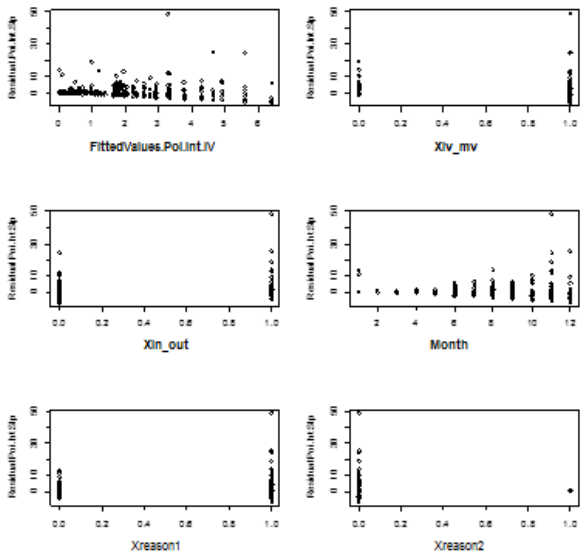

Figure 4.10 *Residual vs Fitted Values and Covariates of Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable*

4.3.2.2 Zero-Inflated Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable

In the previous section, Poisson MCMCglmm with interaction effects and piecewise indicator variable was depicted and its results were compared with that of the first implementation of Poisson MCMCglmm. However, it was realized that problems which were seen in the last sections were still continuing even in updated Poisson

MCMCglmm. In this section, ZIP with interaction and piecewise indicator variable model and its efficiency will be discussed.

Significance of covariates of the model can be seen below (Table 4.5). The model converge diagnostic plots can be seen at Appendix (see Appendix-B section 2).

Table 4.5 *Summary of Zero-Inflated Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable*

	Post Mean	**l-95% CI**	**u-95% CI**	**Efficient Samples**	**p-MCMC**	
Intercept_Poisson	-7.09	-8.65	-5.03	1.95	$1x\ 10^{-3}$	***
Intercept_ZeroInf	-0.15	-0.51	2.77	31.89	$4.28x\ 10^{-4}$	***
Xlv_mv	4.29	2.67	5.21	6.11	$1x\ 10^{-3}$	***
Xin_out	-2.88	-3.85	-2.06	5.63	$1x\ 10^{-3}$	***
newMonthx	-0.61	-0.82	-0.39	11.23	$1x\ 10^{-3}$	***
Xreason1	5.35	3.55	6.61	2.26	$1x\ 10^{-3}$	***
Xreason2	0.91	-0.27	2.22	2.98	$1.38x\ 10^{-4}$	***
Month	0.52	0.35	0.69	5.79	$1x\ 10^{-3}$	***
Xlv_mv: Month	0.24	0.14	0.34	6.16	$1x\ 10^{-3}$	***
Xlv_mv:Xreason1	-4.95	-6.38	-3.43	3.66	$1x\ 10^{-3}$	***
Xlv_mv:Xreason2	-6.14	-7.60	-4.08	3.54	$1x\ 10^{-3}$	***
Xin_out: Month	0.15	0.06	0.26	7.04	$1x\ 10^{-3}$	***

The model's count generating open form for both processes (52) and (53) are:

$$\log(\lambda_i) = -7.09 + 4.29Xlv\_mv - 2.88Xin\_out + 0.52Month + 5.35Xreason1 + 0.91Xreason2 - 0.61newMonthx + 0.24Xlv\_mvxMonth - 4.95Xlv\_mvxXreason1 - 6.14Xlv\_mvxXreason2 + 0.15Xin\_outxMonth + \widehat{b_{Location}}. \quad (52)$$

$$logit(w_0) = \log(\frac{w_0}{1 - w_0}) = -0.15 \quad (53)$$

According to Poisson MCMCglmm with interaction effects and slope, only the covariate of Xreason2 behavior has changed from negative to positive.

Unfortunately, autocorrelation problem exists in all covariates of the model (see Appendix-B section 2). This problem might cause poor estimations & fits from the model. Actually, trace plot gives a hint for this poor mixing.

It is generally known that autocorrelation problem might cause faulty generating counts, however, this case needs to be proven for the situation at hand. For this aim, below observed vs. fitted values are presented for analysis (Figure 4.11).

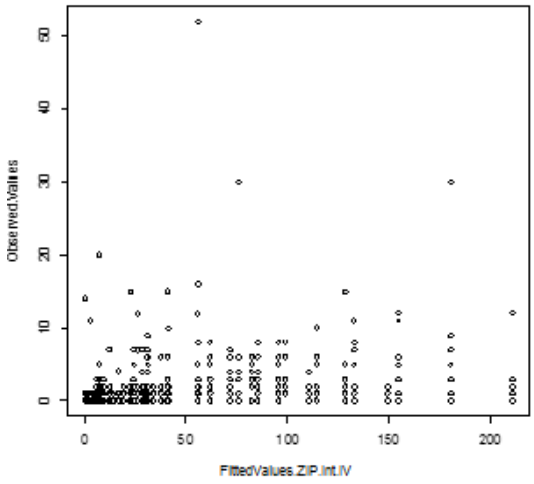

Figure 4.11 *Observed vs Fitted Values of ZIP MCMCglmm with Interaction Effects and Piecewise Indicator Variable*

From the figures, it is to be understood that adding interaction effects and adding piecewise indicator variable to the model remained insufficient in fully developing the model. In fact, it was discovered that zero-inflated model with added interaction effects and piecewise indicator variable mixes more poorly than the first implementation of zero-inflated model does. Residual plots could be consulted to observe this result clearly (Figure 4.12).

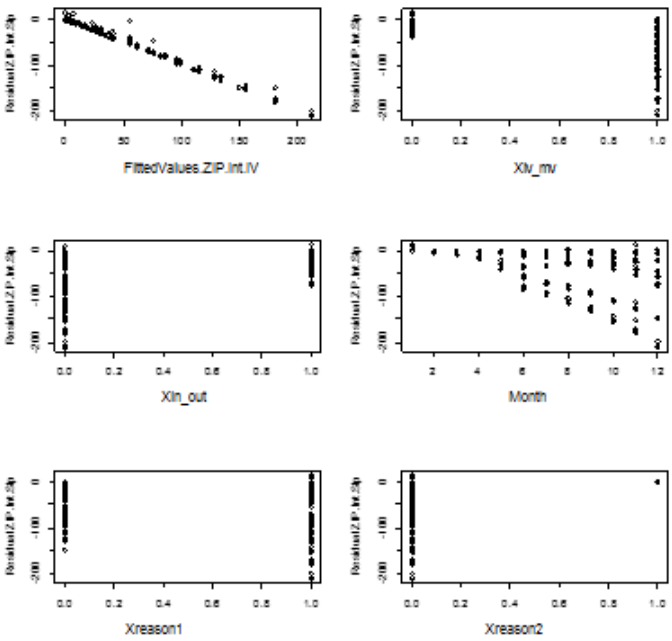

Figure 4.12 *Residual vs Fitted Values and Covariates of ZIP MCMCglmm with Interaction Effects and Slope Parameter*

Residual plots of Zero-Inflated Poisson with interaction effects and piecewise indicator variable show that model can estimate as many observed values as 200. The cause behind this poor mixing and poor estimation in models needs to be investigated. Three possible problems can arise here: Autocorrelation, over-dispersion and multi-collinearity. In order to fix these problems, prior function, variance-covariance matrix of fixed effects, number of iteration and thinning interval can be changed. Also, different methods such as centering method etc. was tried before, but unfortunately these methods cannot solve these problems (see Appendix-D).

4.3.3 Results of the Final Models on Poisson, Zero-Inflated Poisson and Hurdle Poisson MCMCglmm

In the last two sections the implementations of Poisson MCMCglmm and zero-inflated Poisson MCMCglmm were presented to show that neither of them yielded problem free solutions for modelling electricity interruption data. The diagnostic checks of these four models were not good enough to explain the specifications of the data (see Appendix-A and Appendix-B).

These results signal towards multi-collinearity problem (see Section 3.1.2.1). With the aim of eliminating multi-collinearity problem, standardized covariates will be used in this section. Besides, to understand the relationship of the covariates, it is beneficial to consider Spearman's correlation matrix of significant fixed and interaction effects (see Figure 4.13).

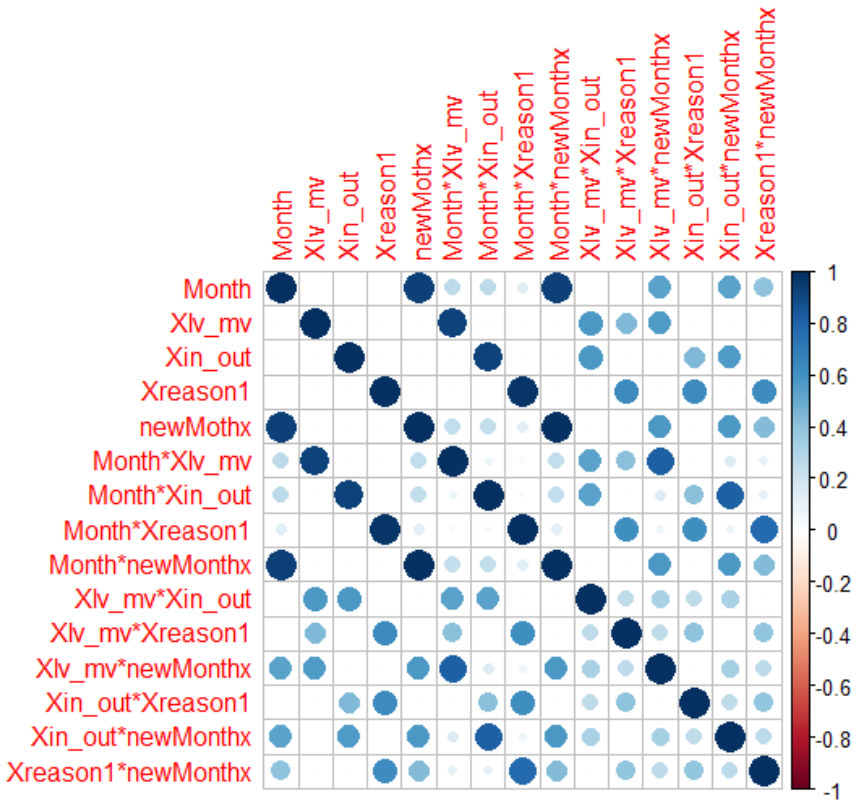

Figure 4.13 *Correlation Matrix of the Significant Covariates*

The correlation matrix shows that all of the significant covariates of the data are somehow correlated with each other. It is clear that especially the covariates of month and newMonthx are the most correlated covariates. Due to correlated covariates, variance-covariance matrix of the prior function needs to be regulated different from the default prior function (see section 4.3.2). In view of this, a new prior function needs to be formed according to the correlated covariates. From now, the new prior will have a new variance-covariance matrix with large variances (10^8) and also the large covariances ($5x10^7$) for the fixed effects in B structure.

Regarding R Structure, residual variance and also residual covariance cannot be observed in zero-inflation or Hurdle or Poisson processes. To cope with this problem, fixed residual variance whose value is 1 has been used and degree of belief parameter has been taken as 0.002. This prior is called as inverse-gamma prior with its shape 0.001. Inverse-gamma prior also captures over-dispersion problem for Poisson process (Hadfield, 2010). Considering these advantages, inverse-gamma prior has been used for all models in this section.

The conditions and the model inputs are changed as follows:

- Each model is run for varying number of iterations, because of convergence issues.
- Location is still used as random effect component.
- Burn-in period is still 3,000.
- Standardized ($\frac{X-mean(X)}{sd(X)}$) variables are used.
- The variable of Xreason2 is eliminated due to serious convergence & autocorrelation problems and Xreason1 is redefined. When it takes 1.41, it means that the interruption stems from the operator. Otherwise the reasons are either external or they are related to security.

In this case, the frequency table of the variables are almost the same (Table 4.6). The only difference is that Xreason2 eliminated. Hence, the reasons of security and external are evaluated together.

Table 4.6 *Frequency Tables of Standardized Electricity Interruption Data*

Xlv_mv.std	*Xin_out.std*	
	un-zoned (-0.99)	zoned (0.99)
meidum (-0.99)	53	23
Low (0.99)	131	80

Reasons	
Security & External	Operator
102	185

4.3.3.1 Final Implementation of Poisson MCMCglmm

According to Poisson MCMCglmm which was examined in the previous sections, autocorrelation problem of each lag was very serious. To eliminate this problem, number of iteration and thin value have been increased. The number of iteration is taken 1,000,000 and the thinning interval is taken 500. Model is run with significant covariates and the model results have been given below (see Table 4.7).

Table 4.7 *Summary Table for the Final Implementation of Poisson MCMCglmm*

	Post. Mean	l-95% CI	u-95% CI	Efficient Samples	p-MCMC	
(Intercept)	-4.66	-5.47	-3.97	235.7	$< 5x\ 10^{-4}$	***
Xlv_mv.std	1.26	0.68	1.87	416.2	$< 5x\ 10^{-4}$	***
Xin_out.std	-0.41	-0.71	-0.12	1870.3	$< 2x10^{-3}$	**
newMonthx.std	-2.04	-3.07	-0.97	278.2	$< 5x\ 10^{-4}$	***
Month.std	3.91	2.41	5.60	260.1	$< 5x\ 10^{-4}$	***
Xreason1.std	2.05	1.49	2.62	253.7	$< 5x\ 10^{-4}$	***
Xlv_mv.std:newMonthx.std	-1.15	-1.66	-0.58	1318.2	$< 5x\ 10^{-4}$	***
Xlv_mv.std:Xreason1.std	-0.90	-1.36	-0.52	447.4	$< 5x\ 10^{-4}$	***
Xin_out.std:newMonthx.std	1.26	0.87	1.67	1752.6	$< 5x\ 10^{-4}$	***
Xreason1.std:newMonthx.std	0.99	0.13	1.82	270.9	$<1{,}6x10^{-2}$	*
Xlv_mv.std:Month.std	1.73	0.98	2.46	1216.9	$< 5x\ 10^{-4}$	***
Xin_out.std:Month.std	-1.23	-1.72	-0.67	1748.1	$< 5x\ 10^{-4}$	***
Xreason1.std Month.std	-1.30	-2.66	-0.08	249.9	$<3{,}9x10^{-2}$	*

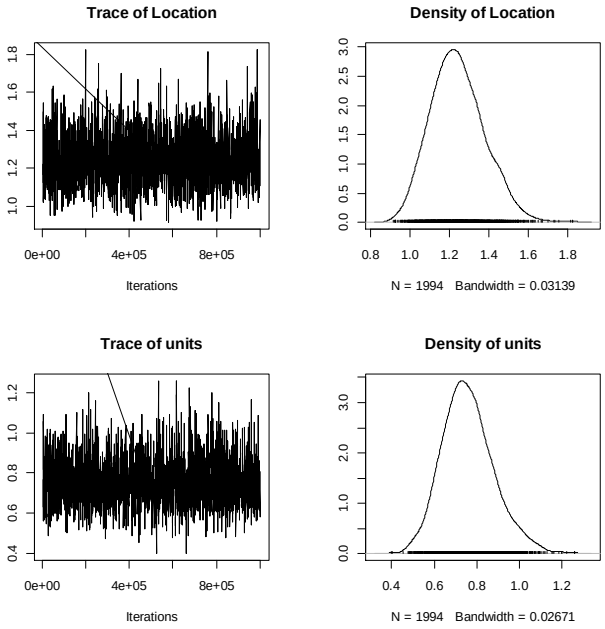

Figure 4.14 *Trace & Density Plots of Variance Component (top) and Residual Variance Component (below) of the Final Implementation of Poisson MCMCglmm*

The trace and density plots for the all effects of fixed and interaction can be seen at Appendix-C. According to Figure 4.14, there are not any trends at the trace plot for variance component. It is an evidence for converging the model. Similarly, the residual variance component seems to be converged, too. In the plot, units implicates that residuals. that MCMCglmm function always handle with over-dispersion problem (Hadfield, 2010). To check over-dispersion, it needs to be looked at the density plot of residuals. The mean of the density of residuals is not equal to 0. It is close to 0.8. It means that the model can deal with over-dispersion problem.

In addition, the results of Geweke test (see section 3.2.1.3) have been offered below (Table 4.8) to provide a check of the convergence status of the model. The test results also support the output of the trace and density plots. According to Geweke's test, none of the Geweke statistics take z-score greater than upper bound (1.96) or less than lower bound (-1.96), which indicates that posterior distribution converged for all the covariates.

Table 4.8 *Geweke Diagnostic Test Results for Final Implementation of Poisson MCMCglmm.*

Covariates	**Geweke Diagnostic z-score**
(Intercept)	0.122
Xlv_mv.std	0.551
Xin_out.std	0.668
newMonthx.std	0.398
Xreason1.std	0.096
Month.std	-0.413
Xlv_mv.std:newMonthx.std	0.335
Xlv_mv.std:Xreason1.std	-0.815
Xin_out.std:newMonthx.std	-0.230
Xreason1.std:newMonthx.std	-0.363
Xlv_mv.std:Month.std	-0.370
Xin_out.std:Month.std	0.052
Xreason1.std:Month.std	0.388

Also, in order to check the stationary and accuracy status of the model, the results of Heidelberger-Welch and Halfwidth Diagnostic tests (see section 3.2.1.3) have also been presented below (Tables 4.9 and 4.10). The results of the test verify the Geweke diagnostic and the trace-density plots. Hence, the convergences of all the covariates of the final implementation of Poisson MCMCglmm is verified by all the diagnostics.

Table 4.9 *Heidelberger-Welch Diagnostic Test's Results for Final Implementation of Poisson MCMCglmm.*

Covariates	**Stationarity Test**	**p-value**
(Intercept)	passed	0.534
Xlv_mv.std	passed	0.853
Xin_out.std	passed	0.240
newMonthx.std	passed	0.200
Month.std	passed	0.165
Xreason1.std	passed	0.499
Xlv_mv.std:newMonthx.std	passed	0.443
Xlv_mv.std:Xreason1.std	passed	0.734
Xin_out.std:newMonthx.std	passed	0.218
newMonthx.std:Xreason1.std	passed	0.173
Xlv_mv.std:Month.std	passed	0.419
Xin_out.std:Month.std	passed	0.644
Month.std:Xreason1.std	passed	0.142

Table 4.10 *Halfwidth Diagnostic Test Results for Final Implementation of Poisson MCMCglmm.*

Covariates	Halfwidth Test	Mean	Halfwidth
(Intercept)	passed	-4.647	0.064
Xlv_mv.std	passed	1.248	0.033
Xin_out.std	passed	-0.416	0.007
newMonthx.std	passed	-2.023	0.071
Month.std	passed	3.892	0.112
Xreason1.std	passed	2.036	0.047
Xlv_mv.std:newMonthx.std	passed	-1.165	0.015
Xlv_mv.std:Xreason1.std	passed	-0.899	0.024
Xin_out.std:newMonthx.std	passed	1.276	0.009
newMonthx.std:Xreason1.std	passed	0.978	0.052
Xlv_mv.std:Month.std	passed	1.756	0.022
Xin_out.std:Month.std	passed	-1.239	0.012
Month.std:Xreason1.std	passed	-1.283	0.085

The previous versions of Poisson MCMCglmm had significant autocorrelation problems. If the autocorrelation status according to the chain of the posterior distribution is examined, it can be easily seen that the final implementation of Poisson MCMCglmm does not have the autocorrelation problem (see the autocorrelation plots at Appendix-C).

At the beginning of the analysis of the final implementation, the results of variance inflation (VIF) was very high. Therefore, to cope with this problem, standardized variables were used throughout the analysis. This solution has decreased the VIF results (Table 4.11), (see Section 3.1.2.1). Thus, all of VIF results for each covariate is less than 10 which can be taken as an evidence for the elimination of multi-collinearity problem.

Table 4.11 *VIF Results of the Final Implementation of Poisson MCMCglmm*

Covariates	**VIF**
(Intercept)	1.000
Xlv_mv.std	1.000
newMonthx.std	6.072
Xreason1.std	1.000
Xlv_mv.std:Xreason1.std	1.000
newMonthx.std:Xreason1.std	6.072
Xin_out.std:Month.std	6.072
Xin_out.std	1.000
Month.std	6.072
Xlv_mv.std:newMonthx.std	6.072
Xin_out.std:newMonthx.std	6.072
Xlv_mv.std:Month.std	6.072
Month.std:Xreason1.std	6.072

The diagnostics expose that the model's assumptions are verified. Therefore, regression model of Poisson MCMCglmm (54) can be determined now.

$$\log\{E(Y_{ij}|Location)\} \quad (54)$$
$$= -4.66 + 1.26Xlv_{mv}.std - 0.41Xin_{out}.std$$
$$- 2.04newMonthx.std + 3.91Month.std$$
$$+ 2.05Xreason1.std$$
$$- 1.15Xlv_{mv}.std\ x\ newMonthx.std$$
$$- 0.90Xlv_{mv}.std\ x\ Xreason1.std$$
$$+ 1.26Xin_{out}.std\ x\ newMonthx.std$$
$$+ 0.99Xreason1.std\ x\ newMonthx.std$$
$$+ 1.73Xlv_{mv}.std\ x\ Month.std$$
$$- 1.23Xin_{out}.std\ x\ Month.std$$
$$- 1.30Xreason1.std\ x\ Month.std + \widehat{b_{Location}}.$$

In the model, all of covariates are used in the standardized version in order to prevent multi-collinearity problem. The model's main effects of the intercept, Xin_out.std, newMonthx.std and interaction effects of Xlv_mv.std: newMonthx.std, Xlv_mv.std: Xreason1.std, Xin_out.std:Month.std, Xreason1.std: Month.std have negative coefficients, whereas other main effects, interaction effects have positive coefficients. While the covariates which have negative coefficients cause a decrease in the fitted value of the model, the covariates which have positive coefficients cause an increase in the fitted value of the model. In addition, variance of the random effect, which is $b_{location}$, is 1.24. The variance is not too small not to affect the model's estimation result. This implicates that the choice of model is appropriate for the electricity interruption dataset.

With the aim of comparing the fitted values of the first implementation and the final implementation of the Poisson MCMCglmm using the same scenario which is given in section 4.3.2.1, the model standardized covariates will take these values:

- Standardized Xlv_mv will take 0.99 instead of 1
- Standardized Xin_out will take -0.99 instead of 0
- Standardized Month will take -1.59 instead of 1
- Standardized newMonthx will take -0.82 instead of 0
- Standardized Xreason1 will take 1.41 instead of 1

The scenario had defined that an interruption value which occurred due to a fault of operator in a lower voltage in an un-zoned area in January. The solution of the regression model (54) according to this scenario is when $\widehat{b_{Location}} = 0$:

$$\log\{E(Y_{i1}|Location)\} \cong -4.41$$

The estimated value of the final implementation of Poisson MCMCglmm with standardized variables equals to ≈0.010. It is closer to the observed value of 0 than the

estimated value of the first implementation of Poisson MCMCglmm (see section 4.3.2.1). When other variables are the same, and only the covariate of month is changed from January (-1.59) to July (-0.35), this situation affects the covariates elated to the newMonthx. The value of the covariate of newMonthx hence changes from -0.82 to -0.35. Then, $\log\{E(Y_{i7}|Location)\} \cong$ -2.15 which is $\approx$ 0.11. The estimated value increases from January to July. Remember that observed values increase from January to December (see Figure 4.1). This upwards movement can be observed in the fitted values in the final implementation of the model, and it is expected that the number of electricity interruption counts should be increased in the second half of the year. This increase can be seen in detail at the fitted vs observed values (see Figure 4.15).

On the other hand, when all other covariates remained same the interruption count occurred in the medium voltage area instead of the lower voltage area: the standardized Xlv_mv is taken -0.99 instead of 0. Then, the solution of the Poisson MCMCglmm regression gives a smaller value ($\log\{E(Y_{ij}|Location)\} \cong$ -17.15). The estimated value converges closely to 0. The result indicates that the probability of an interruption in January due to the operator fault, in un-zoned area, in the medium voltage network, is so low, whereas in low voltage network, this probability is more likely to occur. On the other hand, the scenario might occur in a zoned area instead of un-zoned, and then the estimated value is increased. Except for the reason of operator, the estimated value always increases. It means that the probability of the interruption counts may be higher due to reasons related to security and external factors.

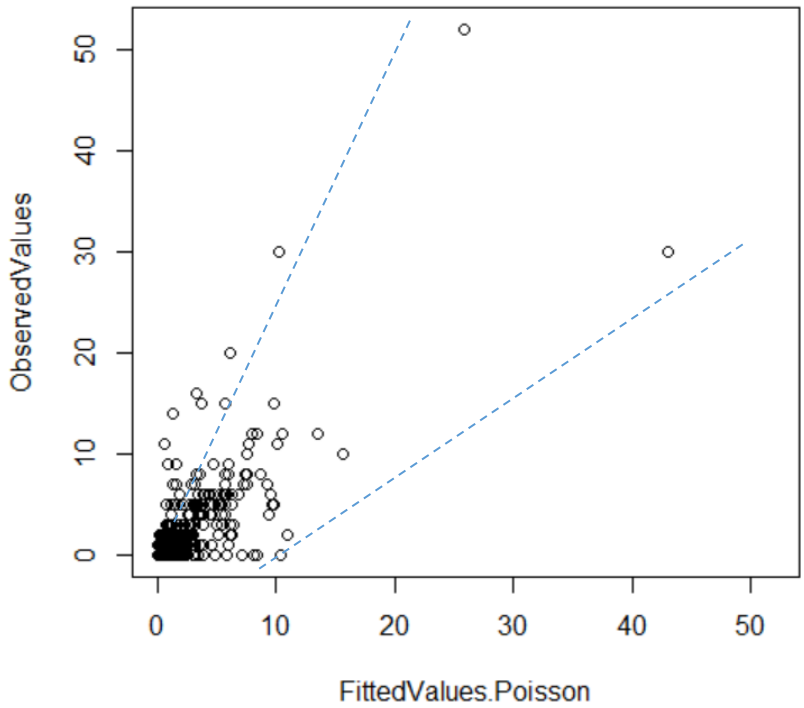

Figure 4.15 *Plot for Fitted vs Observed Plot of the Final Implementation of Poisson MCMCglmm*

Figure 4.15 shows that fitted values around zero and it implicates that the fitted values are increasing parallel with observed values. However, it can be clearly seen that the residuals spread from -10 up to +10 at the residual plots. It can be originated from the number of observed values. The number of observations are really insufficient for large interruptions. In the dataset, the number of observed values which are equal to 10 and bigger than 10 is only 18.

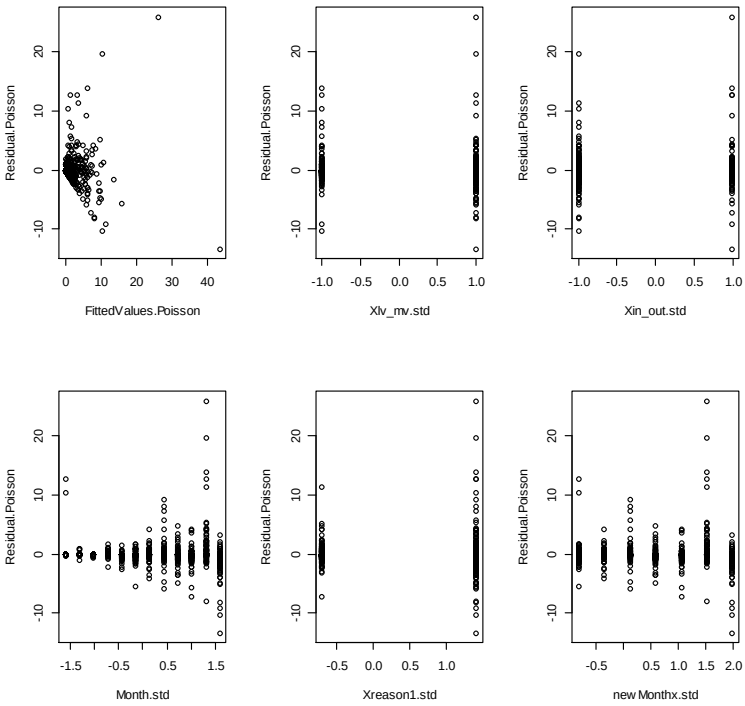

Figure 4.16 *Residual Plots vs Fitted Values and Covariates for the Final Implementation of Poisson MCMCglmm*

The residual plots above demonstrate that the residuals are increasing between June and December. To eliminate these differences, piecewise indicator variable which is newMonthx are added. The piecewise variable causes decreases in the residuals after May. Consequently, even though it cannot be completely eliminated, it should be noted that it makes significant changes to the first implemented models.

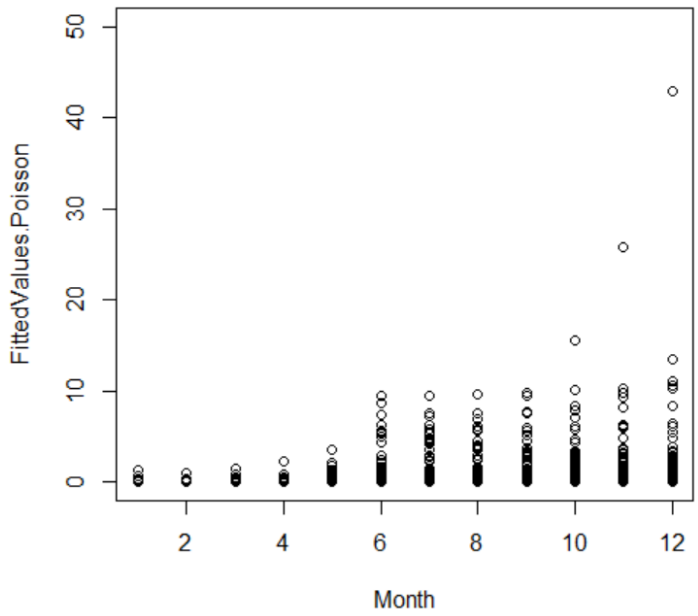

Figure 4.17 *Plot of the Fitted Values of Final Poisson MCMCglmm vs. Month*

In order to observe success of the final Poisson model, it is beneficial to compare Figure 4.1 and Figure 4.17. Figure 4.17 shows the fitted values according to Month variable. Figure 4.17 and Figure 4.1 seem close to each other. The final Poisson MCMCglmm estimate the biggest observation of the dataset as about 45. In addition, the fitted values is getting higher after June mostly. It means that piecewise indicator variable provides increasing of fitted values properly.

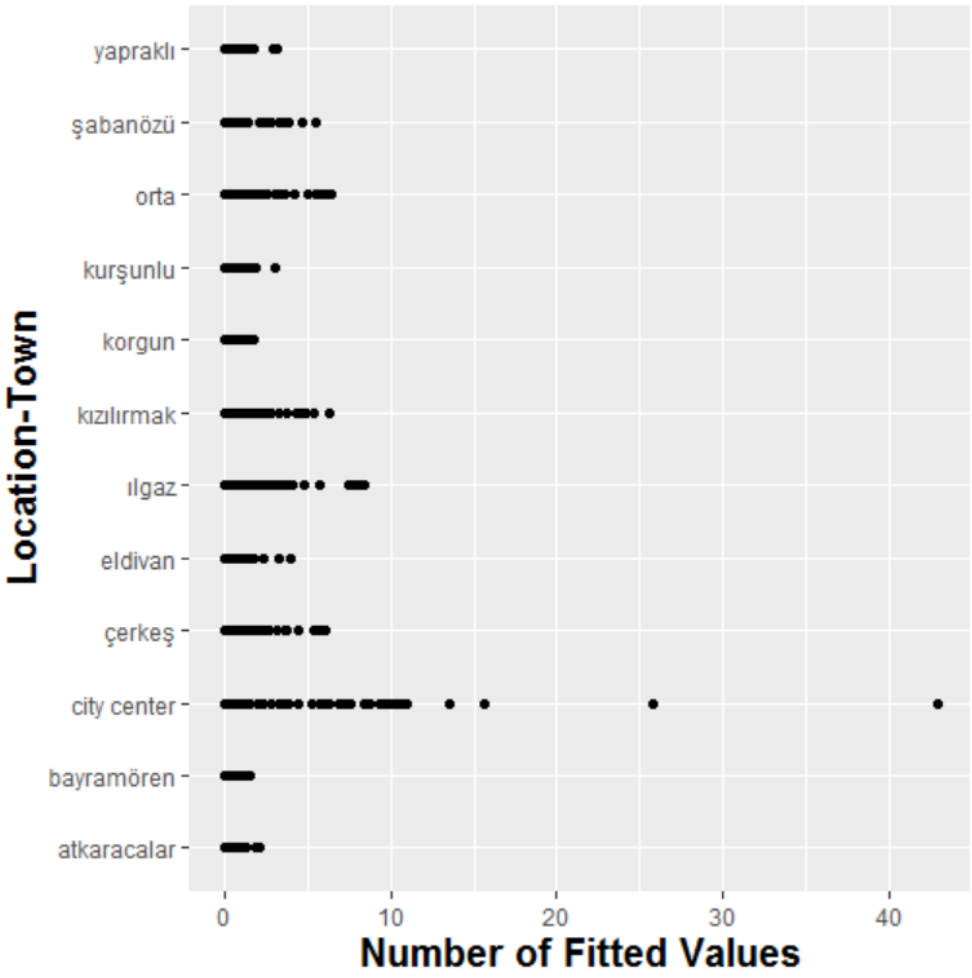

Figure 4.18 *Plot of Fitted Values for Final Poisson MCMCglmm vs Location*

The success of the final Poisson MCMCglmm can be seen at Figure 4.18. In order to understand the differences between observed vs fitted values, Figures 4.18 and 4.2 can be evaluated. The plots seems close to each other. To illustrate, we expect to exist the biggest fitted value in city center. According to Figure 4.18, fitted value of the model exist in the city center. The towns of Orta, Ilgaz, Kızılırmak and Çerkeş have other bigger fitted values like being in the observed dataset. Under this results, we may say that final Poisson MCMCglmm can explain the dataset correspondingly.

4.3.3.2 Final Implementation of Zero-Inflated Poisson MCMCglmm

Although the final implementation of Poisson MCMCglmm has satisfactory convergence status, the residuals of the model are still a bit high. Zero-inflated Poisson MCMCglmm is attempted to fix this problem. These ZIP models have two processes: named distribution (here Poisson) and zero-inflated (see section 3.4.1). Therefore, firstly all of covariates were run for each of the processes. However, the covariates for zero-inflated process did not give any significant results except for the intercept (Table 4.12).

Table 4.12 *Summary Table for the Final Implementation of Zero-Inflated Poisson MCMCglmm*

Coefficients	**Post Mean**	**l-95% CI**	**u-95% CI**	**Efficient Samples**	**p-MCMC**	
Intercept_Poisson Process	-4.63	-5.44	-3.87	2000	$< 5x10^{-4}$	***
Intercept_ZeroInflated	-3.80	-7.90	-5.14	2.689	$< 5x10^{-4}$	***
Xlv_mv.std[1]	1.23	0.65	1.85	2000	$< 5x10^{-4}$	***
Xin_out.std[1]	-0.41	-0.72	-0.12	2000	$1.4x10^{-2}$	*
newMonthx.std[1]	-2.03	-3.14	-0.99	1851.88	$< 5x10^{-4}$	***
Month.std[1]	3.91	2.39	5.68	1831.41	$< 5x10^{-4}$	***
Xreason1.std[1]	2.02	1.45	2.63	2000	$< 5x10^{-4}$	***
Xlv_mv.std:newMonthx.std[1]	-1.15	-1.71	-0.65	2000	$< 5x10^{-4}$	***
Xlv_mv.std:Xreason1.std[1]	-0.89	-1.33	-0.48	1856.51	$< 5x10^{-4}$	***
Xin_out.std:newMonthx.std[1]	1.27	0.89	1.70	2000	$< 5x10^{-4}$	***
Xreason1.std:newMonthx.std[1]	0.99	0.12	1.84	1846.39	$8x10^{-3}$	**
Xlv_mv.std:Month.std[1]	1.74	1.05	2.50	2000	$< 5x10^{-4}$	***
Xin_out.std:Month.std[1]	-1.23	-1.80	-0.72	2297.97	$< 5x10^{-4}$	***
Xreason1.std:Month.std[1]	-1.30	-2.65	-0.05	1846.42	$2.9x10^{-2}$	*

*The superscript [1] and [2] define the level of the process that [1] implicates as Poisson process and [2] implicates zero-inflation process.

Even though intercept of zero-inflated process is significant at the summary table, this does not imply that the model has good convergence. Trace and density plots along with convergence test showed that the second process of the model did not converge

very well. For this reason, the number of iteration and thinning interval were increased step by step. The summary table given as Table 4.12 shows the results when number of iterations was 50,000,000 and thinning interval was 25,000. In addition, it should be noted that this number of iterations and thinning interval include the last and the highest values of the analysis over a period that continue throughout four days. After this, it was realized that from that point onwards it is not possible to have appropriate convergence anymore, even if the number of iterations and thinning interval are increased. However, the intercept of zero-inflated process still does not converge completely. Trace and density plots, Geweke, Heilderberger-Welch tests show this result in detail (see Figure 4.19, Table 4.13, Table 4.14 and Table 4.15).

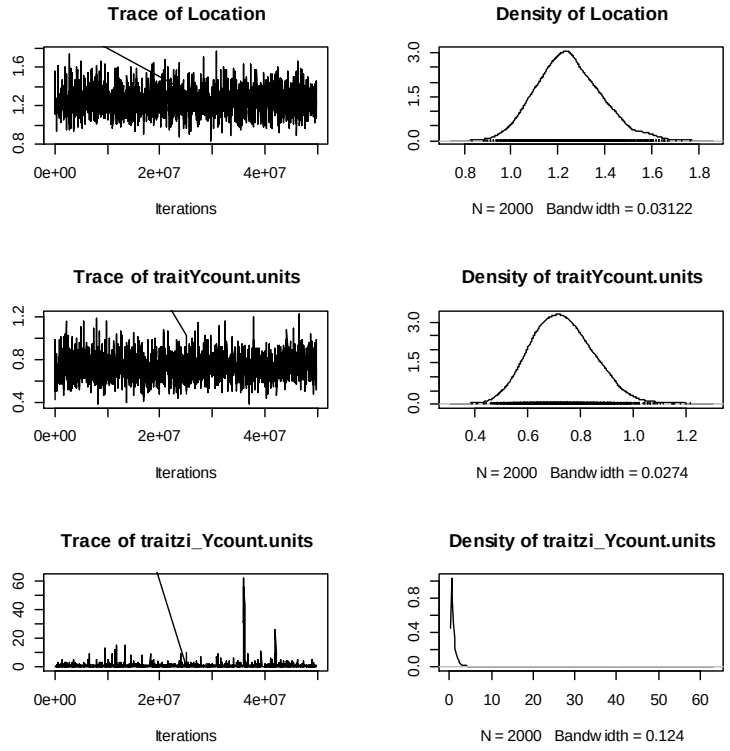

Figure 4.19 *Trace & Density Plots of Variance Component (top) and Residual Variance Component (below) for the Final Implementation of ZIP MCMCglmm*

The results of Geweke Diagnostic show that all of the covariates except for the intercept of zero-inflated process are in the interval of the diagnostic which is between -1.96 and 1.96. Although number of iterations and thinning interval were increased, the result of Geweke diagnostic did not change. Eventually, this situation suggests the possibility that the data may not be suitable for the zero-inflated distribution.

Table 4.13 *Geweke Diagnostic Test Results for Final Implementation of Zero-Inflated Poisson MCMCglmm.*

Covariates	**Geweke Diagnostic z-score**
Intercept_Poisson Process	0.568
Intercept_ZeroInflated	3.514
Xlv_mv.std[1]	0.304
Xin_out.std[1]	1.324
newMonthx.std[1]	0.978
Month.std[1]	-0.687
Xreason1.std[1]	-0.485
Xlv_mv.std:newMonthx.std[1]	0.490
Xlv_mv.std:Xreason1.std[1]	-0.006
Xin_out.std:newMonthx.std[1]	-0.742
newMonthx.std:Xreason1.std[1]	-0.308
Xlv_mv.std:Month.std[1]	-0.695
Xin_out.std:Month.std[1]	0.478
Month.std:Xreason1.std[1]	0.246

According to Geweke diagnostic, intercept of zero-inflated cannot converge very well. In parallel with Geweke diagnostic, stationary and accuracy diagnostic which is Heilderberger-Welch test gives the same results. In zero-inflated models, this kind of converge problems are frequently observed (Hadfield, 2010).

Table 4.14 *Heidelberger-Welch Diagnostic Test's Results for Final Implementation of Zero-Inflated Poisson MCMCglmm.*

Covariates	Stationarity Test	p-value
Intercept_Poisson Process	passed	0.340
Intercept_ZeroInflated	failed	0.048
Xlv_mv.std[1]	passed	0.408
Xin_out.std[1]	passed	0.413
newMonthx.std[1]	passed	0.163
Month.std[1]	passed	0.339
Xreason1.std[1]	passed	0.300
Xlv_mv.std:newMonthx.std[1]	passed	0.124
Xlv_mv.std:Xreason1.std[1]	passed	0.675
Xin_out.std:newMonthx.std[1]	passed	0.294
newMonthx.std:Xreason1.std[1]	passed	0.799
Xlv_mv.std:Month.std[1]	passed	0.070
Xin_out.std:Month.std[1]	passed	0.295
Month.std:Xreason1.std[1]	passed	0.786

Table 4.15 *Halfwith Diagnostic Test Results for Final Implementation of Zero Inflated Poisson MCMCglmm.*

Covariates	Halfwidth Test	Mean	Halfwidth
Intercept_Poisson Process	passed	-4.633	0.018
Intercept_ZeroInflated	<NA>	NA	NA
Xlv_mv.std[1]	passed	1.233	0.013
Xin_out.std[1]	passed	-0.413	0.006
newMonthx.std[1]	passed	-2.039	0.025
Month.std[1]	passed	3.916	0.039
Xreason1.std[1]	passed	2.028	0.013
Xlv_mv.std:newMonthx.std[1]	passed	-1.158	0.012
Xlv_mv.std:Xreason1.std[1]	passed	-0.891	0.010
Xin_out.std:newMonthx.std[1]	passed	1.270	0.009
newMonthx.std:Xreason1.std[1]	passed	0.991	0.020
Xlv_mv.std:Month.std[1]	passed	1.747	0.016
Xin_out.std:Month.std[1]	passed	-1.233	0.011
Month.std:Xreason1.std[1]	passed	-1.301	0.030

To check the multi-collinearity problem, VIF results must be evaluated (Table 4.16). All the VIF results of the covariates are less than 10 (see section 3.1.2.1). Multi-collinearity problem is not observed in the model.

Table 4.16 *VIF Results of the Final Implementation of Poisson MCMCglmm*

Covariates	**VIF**
Intercept_Poisson Process	1.000
Intercept_ZeroInflated	1.000
Xlv_mv.std[1]	1.000
Xin_out.std[1]	1.000
newMonthx.std[1]	6.072
Month.std[1]	6.072
Xreason1.std[1]	1.000
Xlv_mv.std:newMonthx.std[1]	6.072
Xlv_mv.std:Xreason1.std[1]	1.000
Xin_out.std:newMonthx.std[1]	6.072
newMonthx.std:Xreason1.std[1]	6.072
Xlv_mv.std:Month.std[1]	6.072
Xin_out.std:Month.std[1]	6.072
Month.std:Xreason1.std[1]	6.072

As can be observed from the values covariates have in the summary table, while the covariates of Xlv_mv.std, Month.std, Xreason1.std as well as Xin_out.std:newMonthx.std, Xreason1.std:newMonthx.std and Xlv_mv.std:Month.std have positive coefficients, other covariates do not have positive coefficients. Since the standardized variables are used, the covariates can take negative values, too. For example, the covariate of Xlv_mv.std can take -0.99 instead of 0, which is the interruption count in medium voltage area, then the covariate has decreasing effect on the model response $\log(\lambda_i)$. Otherwise, in medium voltage area, it would take -0.99, and would have positive increasing effect on the response. Therefore, to evaluate the changes on the response variable, it would be beneficial to go over the estimation of the regression model of ZIP(55) and (56) :

For the Poisson process (55):

$$\log(\lambda_i|Y_{ij}) = -4.63 + 1.23Xlv\_mv.std - 0.41Xin\_out.std - 2.03newMonthx.std + 3.91Month.std + 2.02Xreason1.std - 1.15Xlv\_mv.std\ x\ newMonthx.std - 0.89Xlv\_mv.std\ x\ Xreason1.std + 1.27Xin\_out.stdx\ newMonthx.std + 0.99\text{newMonthx. stdxXreason1. std} + 1.74\text{Xlv\_mv. std x Month. std} - 1.23\text{Xin\_out. std x Month. std} - 1.30\text{Month. std x Xreason1. std} + \widehat{b_{Location}}. \quad (55)$$

For the zero-inflated process (56) :

$$logit(w_0) = \log\left(\frac{w_0}{1-w_0}\right) = -3.80 \quad (56)$$

The scenario which was implemented to the first implementation of Poisson MCMCglmm (55) was the electricity interruption, occurring due to a fault of the operator in a low voltage and non-urban area in January. To evaluate the final implementation of ZIP MCMCglmm, this scenario is calculated with the new model. The observed value equals to zero. For this reason, zero-inflated process (56) estimates w_0 gives the $P(Y_i=0)$ as ≅0.021. On the other hand, Poisson process (55) calculates the $\log(\lambda_i|Y_{ij}) \cong -6.89$. Therefore, the model estimates the expected number of interruptions as ≅1,01.

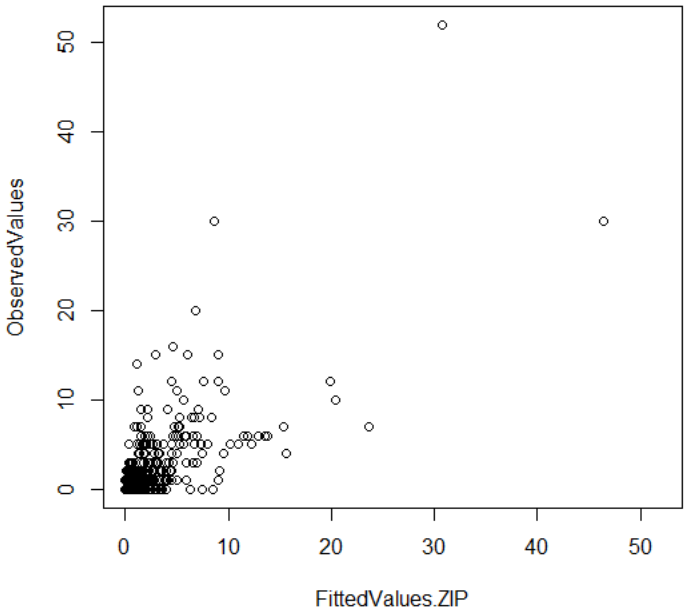

Figure 4.20 *Plot for Fitted vs Observed Plot of the Final Implementation of Zero-Inflated Poisson MCMCglmm*

Fitted vs observed plot (Figure 4.20) of the final implementation of ZIP MCMCglmm seems so close to the plot of the final implementation of Poisson MCMCglmm's plots. In this situation zero-inflated process is seen not to converge very well. Nevertheless, to overcome this convergence problem and improve the results, hurdle models which usually indicate better convergence status should be tried here.

Consequently, the only problem that occurs in the final implementation of Poisson MCMCglmm is the high residuals. It was supposed that the final implementation of ZIP MCMCglmm might solve this problem; however, as can be seen on the residual plots of the ZIP MCMCglmm, the model is not able to solve this problem (Figure 4.21).

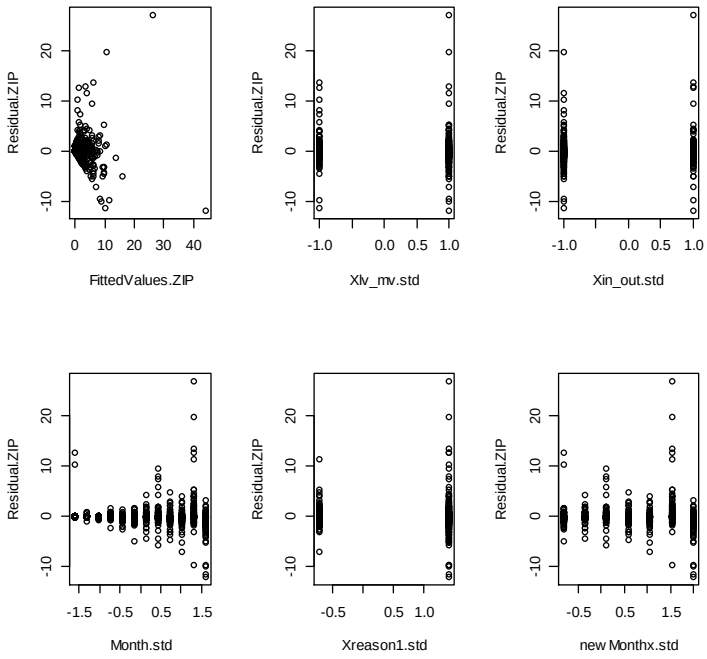

Figure 4.21 *Residual Plots of the Final Implementation of ZIP MCMCglmm*

4.3.3.3 Hurdle Poisson MCMCglmm

According to the literature, hurdle models have better convergence than the zero-inflated models since hurdle model have the ability to handle not only zero-inflation but also zero-deflation. (see section 3.4.2) In previous section, the final implementation of zero-inflated Poisson MCMCglmm had convergence problem on zero-inflation process. To fix this problem, hurdle Poisson MCMCglmm is implemented to our Bayesian data analysis.

At first, the model's diagnostics showed convergence problem in which number of iteration is 1,000,000 and thinning interval is 500. Then, the values of them has been increased progressively like zero-inflation Poisson MCMCglmm. After that, the model converge the posterior distribution when the number of iterations is 10,000,000 and thinning interval is 7,500. These values are small than the number of iterations

and thinning interval of the zero-inflated model given in previous section (see section 4.3.3.2). The summary table of significant covariates can be seen below (Table 4.17).

Table 4.17 *Summary Table of Hurdle Poisson MCMCglmm*

	Post Mean	l-95% CI	u-95% CI	Efficient Samples	p-MCMC	
Intercept_Poisson Process	0.27	-0.48	1.11	73.87	5.06×10^{-1}	
Intercept_Hurdle Process	2.85	2.53	3.16	124.11	$< 5 \times 10^{-4}$	***
Xlv_mv.std[1]	-1.29	-2.08	-0.60	94.40	$< 5 \times 10^{-4}$	***
Xlv_mv.std[2]	-0.45	-0.80	-0.12	132.14	5×10^{-3}	**
Xin_out.std[1]	-0.30	-0.79	0.22	351.39	2.19×10^{-1}	
Xin_out.std[2]	0.40	0.12	0.71	172.51	9×10^{-3}	**
newMonthx.std[1]	-0.97	-2.24	0.39	84.41	1.38×10^{-1}	
newMonthx.std[2]	1.30	0.68	1.83	113.86	$< 5 \times 10^{-4}$	***
Month.std[1]	2.01	-0.09	3.81	69.31	2.2×10^{-2}	*
Month.std[2]	-2.73	-3.48	-1.95	86.72	$< 5 \times 10^{-4}$	***
Xlv_mv.std:newMonthx.std[1]	-1.93	-3.22	-0.71	76.51	$< 5 \times 10^{-4}$	***
Xlv_mv.std:newMonthx.std[2]	1.18	0.61	1.80	96.10	$< 5 \times 10^{-4}$	***
Xlv_mv.std:Month.std[1]	3.17	1.35	5.22	78.01	$< 5 \times 10^{-4}$	***
Xlv_mv.std:Month.std[2]	-1.68	-2.42	-0.82	72.78	$< 5 \times 10^{-4}$	***
Xin_out.std:newMonthx.std[1]	1.20	0.41	1.96	495.51	3×10^{-3}	**
Xin_out.std:newMonthx.std[2]	-0.85	-1.35	-0.31	121.94	$< 5 \times 10^{-4}$	***
Xin_out.std:Month.std[1]	-1.36	-2.45	-0.29	380.41	1.3×10^{-2}	*
Xin_out.std:Month.std[2]	0.78	0.12	1.46	91.79	1.7×10^{-2}	*

*The superscript [1] and [2] define the level of the process that [1] implicates as Poisson process and [2] implicates Hurdle process.

Hurdle models have two process like the zero-inflation models as well. For this reason, the variables of the model were run for the each process. Except of the covariates of the intercept, Xin_out.std and newMonthx.std for the Poisson process, all the covariates are significant. Although both Xin_out.std and newMonthx.std are not significant, they cannot be eliminated from the analysis because of the interaction effects.

When looking at the differences of the same covariates in hurdle and Poisson process, intercepts of the processes take positive coefficients. However, the covariate of Xlv_mv.std of each process takes negative coefficients. It is understood that intercepts always increase the fitted value since using the log link for Poisson process and logit link for the hurdle process. However, while the covariate of Xlv_mv.std increases the fitted value in medium voltage area (-0.99), it decreases the fitted value in low voltage (0.99) area.

Next, it is understood that the model converge very well. In addition, it is not observed any autocorrelation between lags (see Appendix-C). Geweke diagnostic and Heilderberger-Welch diagnostics show this converge in detail for each of covariate (see Table 4.18 and Table 4.19).

Table 4.18 *Geweke Diagnostic Test Results for Hurdle Poisson MCMCglmm.*

Covariates	**Geweke Diagnostic z-score**
Intercept_Poisson Process	-0.735
Intercept_Hurdle Process	-1.403
Xlv_mv.std[1]	-0.248
Xlv_mv.std[2]	-1.007
Xin_out.std[1]	0.476
Xin_out.std[2]	-0.933
newMonthx.std[1]	-0.610
newMonthx.std[2]	-0.956
Month.std[1]	0.540
Month.std[2]	1.114
Xlv_mv.std:newMonthx.std[1]	-0.425
Xlv_mv.std:newMonthx.std[2]	-1.675
Xlv_mv.std:Month.std[1]	0.391
Xlv_mv.std:Month.std[2]	1.829
Xin_out.std:newMonthx.std[1]	0.279
Xin_out.std:newMonthx.std[2]	0.109
Xin_out.std:Month.std[1]	-0.449
Xin_out.std:Month.std[2]	0.104

According to Geweke diagnostic, it can be seen that z-scores for all of covariates are between the interval (-1.96 and 1.96).

Table 4.19 *Heidelberger-Welch Diagnostic Test's Results for Hurdle Poisson MCMCglmm.*

Covariates	**Stationarity Test**	**p-value**
Intercept_Poisson Process	passed	0.717
Intercept_Hurdle Process	passed	0.603
Xlv_mv.std[1]	passed	0.664
Xlv_mv.std[2]	passed	0.866
Xin_out.std[1]	passed	0.228
Xin_out.std[2]	passed	0.652
newMonthx.std[1]	passed	0.632
newMonthx.std[2]	passed	0.501
Month.std[1]	passed	0.671
Month.std[2]	passed	0.433
Xlv_mv.std:newMonthx.std[1]	passed	0.778
Xlv_mv.std:newMonthx.std[2]	passed	0.734
Xlv_mv.std:Month.std[1]	passed	0.719
Xlv_mv.std:Month.std[2]	passed	0.670
Xin_out.std:newMonthx.std[1]	passed	0.086
Xin_out.std:newMonthx.std[2]	passed	0.887
Xin_out.std:Month.std[1]	passed	0.052
Xin_out.std:Month.std[2]	passed	0.962

All covariates for two processes in hurdle model passed from the stationarity diagnostic. However, according to the Halfwidth diagnostic, there is an accuracy problem for some covariates of the model (see Table 4.20).

Table 4.20 *Halfwidth Diagnostic Test Results for Hurdle Poisson MCMCglmm*

Covariates	Halfwidth Test	Mean	Halfwidth
Intercept_Poisson Process	failed	0.271	0.093
Intercept_Hurdle Process	passed	2.853	0.028
Xlv_mv.std[1]	passed	-1.295	0.077
Xlv_mv.std[2]	passed	-0.451	0.029
Xin_out.std[1]	passed	-0.303	0.026
Xin_out.std[2]	passed	0.404	0.023
newMonthx.std[1]	failed	-0.978	0.148
newMonthx.std[2]	passed	1.303	0.056
Month.std[1]	failed	2.011	0.243
Month.std[2]	passed	-2.731	0.085
Xlv_mv.std:newMonthx.std[1]	passed	-1.937	0.147
Xlv_mv.std:newMonthx.std[2]	passed	1.182	0.063
Xlv_mv.std:Month.std[1]	passed	3.179	0.223
Xlv_mv.std:Month.std[2]	passed	-1.680	0.096
Xin_out.std:newMonthx.std[1]	passed	1.205	0.035
Xin_out.std:newMonthx.std[2]	passed	-0.852	0.047
Xin_out.std:Month.std[1]	passed	-1.369	0.055
Xin_out.std:Month.std[2]	passed	0.787	0.070

In order to check the multi-collinearity status, values of the VIF is figured up. According to VIF, there is no any covariate which causes the multi-collinearity problem, since all the VIF values of each covariate is less than 10 (see Table 4.21).

Table 4.21 *VIF Results of Final Implementation of Poisson MCMCglmm*

Covariates	**VIF**
Intercept_Poisson Process	1.000
Intercept_Hurdle Process	1.000
Xlv_mv.std[1]	1.000
Xlv_mv.std[2]	1.000
Xin_out.std[1]	1.000
Xin_out.std[2]	1.000
newMonthx.std[1]	6.072
newMonthx.std[2]	6.072
Month.std[1]	6.072
Month.std[2]	6.072
Xlv_mv.std:newMonthx.std[1]	6.072
Xlv_mv.std:newMonthx.std[2]	6.072
Xlv_mv.std:Month.std[1]	6.072
Xlv_mv.std:Month.std[2]	6.072
Xin_out.std:newMonthx.std[1]	6.072
Xin_out.std:newMonthx.std[2]	6.072
Xin_out.std:Month.std[1]	6.072
Xin_out.std:Month.std[2]	6.072

The diagnostic checks have shown that except the accuracy problem for three covariates, all of the covariates converge to the posterior distribution in hurdle Poisson MCMCglmm. Next, it needs to be looked at the regression model for the estimation.

Hurdle regression model has two process like ZIP models (see section 3.4.2). One of them is named distribution (here is Poisson) and the other is hurdle process. According to the summary table, the open form of the hurdle regression model such as:

For the hurdle process (57):

$$\begin{aligned} logit(w_0|Y_{ij}) = {} & 2.85 - 0.45Xlv\_mv.std + 0.40Xin\_out.std \\ & + 1.30newMonthx.std - 2.73Month.std \\ & + 1.18Xlv\_mv.std\ x\ newMonthx.std \\ & - 1.68Xlv\_mv.std\ x\ Month.std \\ & - 0.85Xin\_out.std\ xnewMonthx.std \\ & + 0.78Xin\_out.std\ x\ Month.std \end{aligned} \quad (57)$$

For the Poisson process (58):

$$\begin{aligned} \log(\lambda_{ij}|Y_{ij}) = {} & 0.27 - 1.29Xlv_{mv}.std - 0.30Xin_{out}.std \\ & - 0.97newMonthx.std + 2.01Month.std \\ & - 1.93Xlv_{mv}.std\ x\ newMonthx.std \\ & + 3.17Xlv_{mv}.std\ x\ Month.std \\ & + 1.20Xin_{out}.std\ x\ newMonthx.std \\ & - 1.36Xin_{out}.std\ x\ Month.std + \widehat{b_{Location}}. \end{aligned} \quad (58)$$

When the random effect takes 0, $\widehat{b_{Location}} = 0$, the scenario which was the electricity interruption of occurring due to a fault of the operator in a low voltage and non-urban area in January is calculated with hurdle model. The observed value equals to zero. Under these conditions, zero counts generated from the hurdle process and positive counts are generated by using truncated-Poisson process. According to the hurdle process, probability of observing 0 is (57) $w_0 \cong 0.99$. The result of the hurdle process says that the probability of existing the electricity interruption under these conditions is 0.99. The probability of observing positive counts which is 1- w_0 equals to 0.01. After all probabilities are calculated, the expected mean value is ≅0.00094 according to the model. It means that from the model fitted value is so close to 0.

The fitted vs observed plot gives a better opinion about the model (see Figure 4.22).

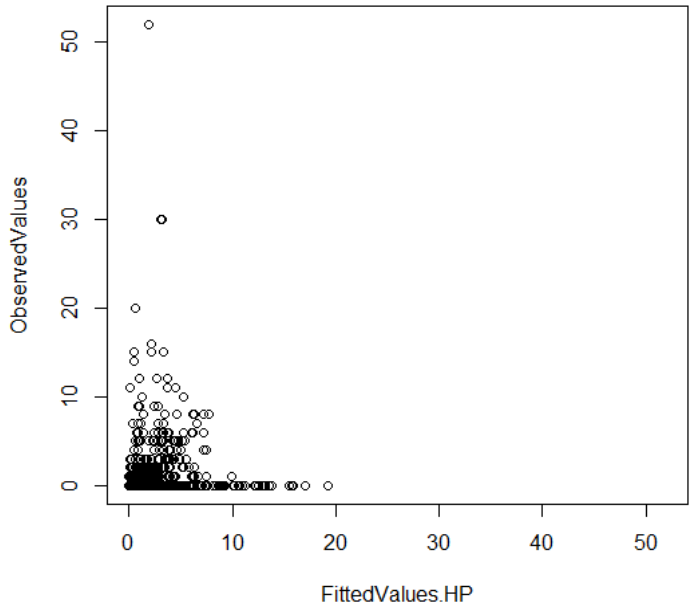

Figure 4.22 *Fitted vs Observed Values Plot of Hurdle Poisson MCMCglmm*

According to fitted vs observed values plot and residual plots, the residuals are higher than the final implementation of Poisson and ZIP MCMCglmm. The model can estimate 12 when an observed value is 0. This situation can be evaluated at residual plots in detail (Figure 4.23). Especially, in the residual plot of Month after June and newMonthx, the residuals' distributed range is higher than the previous models given in the section.

To sum up, according to the literature, hurdle model are expected to be mixing better than the ZIP models. However, this situation does not guarantee that the estimation gives better results than others.

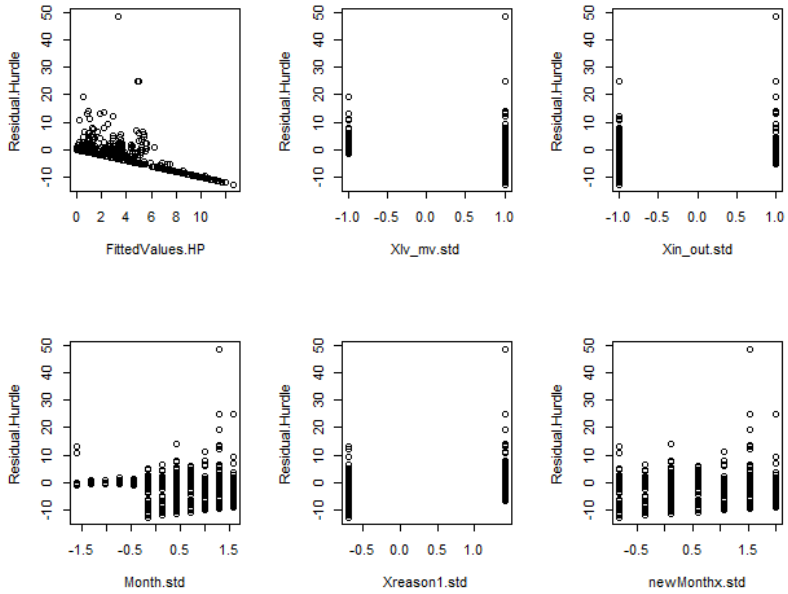

Figure 4.23 *Residual Plots vs Covariates for Hurdle Poisson MCMCglmm*

4.3.4 Posterior Predictive Checks and Comparison of the Final Models of Poisson, Zero-Inflated Poisson and Hurdle Poisson MCMCglmm

In previous sections, 7 different models and their specifications were given. In first place, in section 4.3.2, the first implementation of Poisson MCMCglmm and Zero-Inflated Poisson MCMCglmm without any interaction or other variables were explained separately. It was seen that the models were not enough to estimate some observed values and some diagnostic checks did not provide satisfactory results. Therefore, in section 4.3.3, significant interaction effects and piecewise indicator variable were added to the first implementation of Poisson and Zero-Inflated Poisson MCMCglmm. Thirdly, according to correlation relation, prior function was changed and the variables were standardized against the multi-collinearity problem. Then, the number of iterations and thinning intervals were customized against the

autocorrelation problem in the final implementation of Poisson and zero-inflated Poisson, and also the implementation of hurdle Poisson MCMCglmm was added.

Actually, the models were given in section 4.3.2 and section 4.3.3 show the development process of our data analysis. They showed the model problems and gave the possible solutions to us. Therefore, it is not necessary to compare the models given in the sections 4.3.1 and 4.3.2. However, the final implementation of Poisson and ZIP MCMCglmm and hurdle Poisson MCMCglmm need to be compared in detail.

To compare the models, the first method is posterior predictive check. Method depends on simulating data from the fitted model and comparing it to observed data (Gelman et al., 2014). To achieve this, 1000 sets of replicated dataset from each of fitted model are simulated. For different ranges of observed value, the number of replicated data bigger than the observed value is calculated, and then divided by 1,000 (59).

$$p - value_{Bayesian} = {(\mathrm{P}(T(y_{rep}, \theta) > T(y))}\Big/{1000} \tag{59}$$

where θ is unknown model parameters, $T(y)$ is denoted as captured measurement from the observed data, and $T(y_{rep}, \theta)$ is denoted as replicated data from the fitted model. The p-values that are smaller than 0.05 or bigger than 0.95 suggest that the model is generating data different from observed data.

Therefore, the method of posterior predictive check is applied for different ranges of the observed data (see Table 4.22).

Table 4.22 *Posterior Predictive Checks for Final Implementation of Poisson, ZIP and Hurdle Poisson MCMCglmm*

POSTERIOR PREDICTIVE CHECKS			
Model* / Count Range	**Poisson**	**ZIP**	**Hurdle Poisson**
=0	0.791	0.768	0
>0	0.192	0.2	1
>5	0.303	0.313	0.773
>10	0.305	0.308	0.268
>15	0.347	0.386	0.075
>20	0.189	0.164	0.026

Regarding the Poisson and ZIP models, p-values of the predictive checks are very similar to each other. In this case, it can be understood that the final implementation of ZIP model is using its Poisson process mostly in their modelling stage. In order to understand this hypothesis, it is beneficial to look at the posterior predictive histogram of the final implementation of Poisson MCMCglmm (see Figure 4.24). In Figure 4.24, the number of observed values equal to 0 exist at the middle of the histogram which are generated from the fitted values from the final implementation of Poisson MCMCglmm. This situation means that the final implementation of Poisson MCMCglmm can be seen adequate for modeling the dataset.

On the other hand, when looking at the p-values of hurdle Poisson for each count range, it is realized that the hurdle cannot satisfactorily generate values equaled to zero. The dataset has excess zeros whose percentage is 83.4%, but the model cannot generate that much fitted values equal to zero. Next, the model is also not able to satisfactorily generate fitted values bigger than 0. It is understood here that the hurdle model generate the fitted values so close to zero, but it is never able to generate the fitted values which equal to zero.

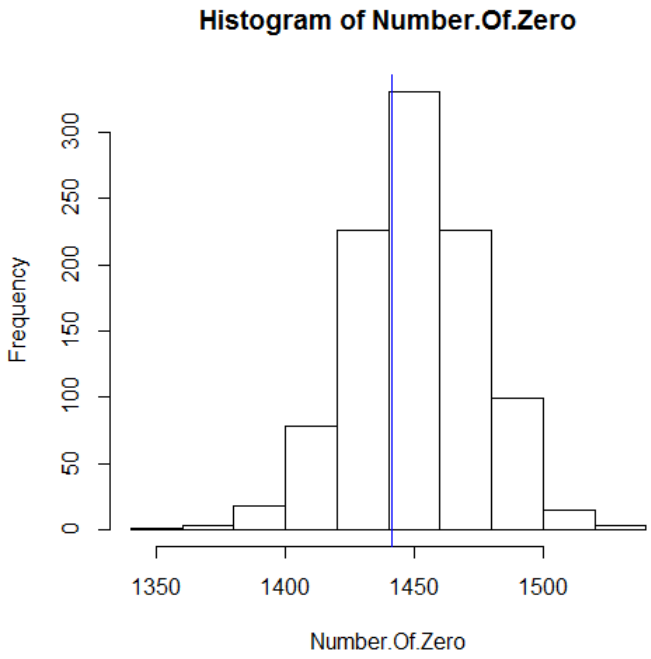

Figure 4.24 *Posterior Predictive Histogram of the Final Implementation of Poisson MCMCglmm*

To conclude that, posterior predictive checks show that the final implementation of both Poisson and ZIP MCMCglmm has satisfactory estimation of the electricity interruption dataset.

Deviance Information Criterion is another evaluation method for comparison of the Bayesian models. The deviance "D" is defined as (60):

$$D = -2\log(Prob(y|\Omega)) \tag{60}$$

where Ω is parameter set of the model. In MCMCglmm package, the mean deviance which is $\overline{D}$ is calculated over all iterations. It is the mean of the latent variables, the R-structure and the vector of fixed and random effects. The deviance is evaluated at the mean estimation of the parameters. $(D(\overline{\Omega}))$ (Hadfield, 2010). Then, the deviance information criterion (61) is that:

$$DIC = 2\overline{D} - D(\overline{\Omega}) \quad (61)$$

To compare the final implemented models, the results of DIC (61) are given (see Table 4.23). The minimum DIC gives the best model. In this case, the DIC of Poisson MCMCglmm has the minimum DIC value. Then, second preferable model is the final implementation of ZIP MCMCglmm. Unfortunately, hurdle Poisson MCMCglmm has the worst value of DIC.

Table 4.23 *Deviance Information Criterion for Final Implementation of Poisson, ZIP and Hurdle Poisson MCMCglmm*

	Poisson	**ZIP**	**Hurdle**
DIC	1662.51	1664.296	1946.685

In section 4.3.3, the number of iterations and thinning intervals of the final implementation of the models was customized for their autocorrelation status. Hence, the estimation time of the models were different as well (see Table 4.24). When looking at the computational time table, the longest time belong to the final implementation of ZIP MCMCglmm. Then, the computational time of hurdle Poisson follows the computation time of ZIP MCMCglmm secondly. Thirdly, the final implementation of Poisson MCMCglmm has the shortest computation time.

Table 4.24 *Computation Time of Final Implementation of Poisson, ZIP and Hurdle Poisson MCMCglmm*

System Time (second)	**Poisson**	**ZIP**	**Hurdle Poisson**
User	3670.08	38515.78	12225.26
System	123.46	4279.53	271.21
Elapsed	3811.55	36945.98	12523.06

To conclude that, according to posterior predictive checks, DIC, and computational time of the models, the final implementation of Poisson MCMCglmm is the best

model for this electricity interruption data. In spite of the literature, Poisson distribution gave better estimation than ZIP and hurdle models for the dataset which includes %83.4 of zero entity.

CHAPTER 5

CONCLUSION

This study aimed to give a framework of evaluation and estimation for the electricity interruption dataset which is published by the local electricity distribution companies in Turkey. Electricity interruption counts were analyzed by depending on in the frame of longitudinal data analysis and Bayesian inference. At the end of the study, it was realized that Poisson distribution was adequate for analyzing this type of longitudinal data which has 83.4% of excess zero. The data analysis were conducted in the statistical tool of R software and its package of MCMCglmm. The MCMCglmm package which was published in 2016 is developed by Jarrod Hadfield and its development process is still continued by him (Hadfield, 2016).

In the final implementation of Poisson MCMCglmm:

- Probability of electricity interruption increases from January to July.
- Probability of electricity interruption is higher in the low voltage network than medium voltage network.
- Probability of electricity interruption is bigger in a zoned area than in un-zoned area.

In the first place, the probability of electricity interruption is lower in January than July. It is reasonable since the electricity consumption is getting higher in summer season because rapid voltage changes related with higher electricity consumption cause the electricity interruptions in the network more frequently in summer season. Next, the probability of electricity interruption is bigger in low voltage network. This result can be explained with infrastructure of the electricity network. Low voltage network consist of many electricity network components, and also connection

components. Therefore, the probability of being broken down a component is higher. Also, the electricity infrastructure, and also quality of components are lower in the low voltage network than the medium voltage network. Moreover, low voltage system is more open for illegal interventions causing by residents. All of these reasons make the probability of electricity interruption is getting higher. Thirdly, the probability of electricity interruption is also higher in zoned area than in un-zoned area since voltage changes are observed more frequently in zoned area. The voltage changes made by operators in substations and distribution centers increase the probability of electricity interruptions in zoned area. However, the electricity consumption is little in un-zoned area since population density is very low. Instead of population density, electricity distribution lines exist in un-zoned area, and the components of electricity distribution lines have better quality than the components of low voltage network. These situations make the probability of electricity interruption increase in zoned area.

5.1 Limitations of the study

During the data analysis process, some limitations has been arisen:

- The package of MCMCglmm could not support the Negative Binomial distribution which is suggested by the literature as an alternative distribution to Poisson and zero-inflated Poisson. Therefore, negative binomial distribution could not be used in our data analysis (Hadfield, 2016).
- Although the number of iterations and thinning intervals were customized according to the models, autocorrelation problem of zero-inflated process in the final implementation of ZIP MCMCglmm could not be solved by us.
- Also, instead of increasing number of iteration and thinning interval, centering method was tried to solve autocorrelation problem for previous implementations of the models. However, this method could not be successful for solving the autocorrelation problem (see Appendix-D).

- Even if hurdle models are suggested by the literature for good converge status, the hurdle Poisson MCMCglmm could not be give the best estimation result in our data analysis.
- Although the data structure is constituted by EMRA with the regulation published in 2008, the regulation is not enough to standardize the electricity interruption datasets which are published by the local electricity distribution companies for data analysis. Hence, it needs to be improved.
- In zero-inflated models, zeros of data is expected to be around 30% of the total data (Hadfield, 2016). However, the dataset used in this study has 83.4% of total observation being zero. This situation can explain why the final ZIP MCMCglmm did not fit very well for zero-inflated process. However, there is not any limitation for the percentage of zero observation in the literature according to our knowledge. For this reason, this study also gives the information that ZIP models may not fit very well when zero observation is that high.

5.2 Future Studies

When looking at the Turkish electricity literature, the studies generally focused on the consumption of electricity by users. There is not so much study about the data analysis of electricity interruption to the best of our knowledge. In order to increase this type of studies, the datasets which is constituted by the local electricity distribution companies are precious resources.

The electricity interruption datasets which are published by local electricity distribution companies have many information about electricity interruptions. Some of them are:

- Duration of electricity interruptions
- Type of electricity component which causes the electricity interruption
- The number of residents who are affected from electricity interruption

If these resources might be modelled with appropriate data analysis techniques, it might give valuable information about the electricity interruptions to the decision makers and authorities. Moreover, these models might be used to predict the future electricity interruptions in the future. Second, cost of loses due to electricity interruptions can be predicted in a future study. Also, these analysis might be conducted and evaluated together for other datasets which belong to different local distribution companies if the standardized data conditions can be provided by EMRA and TEDAŞ. Consequently, we hope that this study will give a key for the researchers who study in this area to discover new resources and statistical technique.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis.* John Wiley and Sons.

Alwan, F.M. and Baharum A. and Hassan G.S. (2013). Reliability measurement for mixed mode failures of 33/11 kilovolt electric power distribution stations. *Plos One, 8(8): e69716.*

Billinton, R. and Wang, P. (1999). Teaching distribution system reliability evaluation using Monte Carlo simulation. IEEE Transactions on Power Systems, 14(2), 397-403.

Bollen, K. A. (1991). Conventional Wisdom on Measurement: A Structural Equation Perspective. *Psychological Bulletin, 110(2):305-314.*

Breslow, N. E., and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association, 88(421), 9.*

Diggle, P., Liang, K.Y. and Zeger, S.L. (2002). *Analysis of Longitudinal Data* (2nd Edition). Oxford University Press.

E.Seavy, N., Quader, S., Alexander, J. D., and Ralph, C. J. (2005). Generalized Linear Models and Point Count Data: Statisical Considerations for the Design and Analysis of Monitoring Studies. *USDA Forest Service Gen. Tech. Rep. PSW-GTR, 191, 744-753.*

EMO. (1981). Türkiye'de Elektrik Enerjisi Sektörünün Yapısı ve Tarihsel Gelişimi. *TMMOB Elektrik Mühendisliği Dergisi, 278, 81-91.*

EMRA. (2008). *Elektrik Enerjisinin Tedarik Sürekliliği Hakkında Yönetmelik. Elektrik Piyasasında Dağıtım Sisteminde Sunulan Elektrik Enerjisinin Tedarik Sürekliliği, Ticari ve Teknik Kalitesi Hakkında Yönetmelik,* Ankara.

Ertilav, M. and Aktel, M. (2015). TEDAŞ (Türkiye Elektrik Dağıtım Anonim Şirketi) Özelleştirmesi. *International Journal of Alanya Faculty of Business, 7(2), 95-108.*

Eto, J. H., and Kristina H. Lacommare, P. L. (2012). Distribution- level electricity reliability: Temporal Trends Using Statistical Analysis. *Energy Policy, 49, 243-252.*

Eto, J. H. and LaCommare, K. H. (2008). Tracking the Reliability of the U.S. Electric Power System: An Assessment of Publicly Available Information Reported to State Public Utility Commissions. *Lawrence Berkeley National Laboratory, Report No. LBNL-1092E.*

Fisher, R. A. (1925). *Statistical Methods for Research Workers.* Oliver and Boyd.

Friedman, G. D., Cutter, G. R., Donahue, R. P., Hughes, G. H., Hulley, S. B., Jacobs Jr., D. R. and Savage, P. J. (1988). CARDIA: Study design, recruitment, and some characteristics of the examined subjects. *Journal of Clinical Epidemiology.*

Fitzmaurice, G.M., Laird, N. M. and Ware, J.H. (2004). *Applied Longitudinal Analysis.* John Wiley & Sons Inc.

Gardner, W., Mulvey, E. P. and Shaw, E. C. (1995). Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin, 118(3), 392–404.*

Garwin, D. (1996). Competing on the eight dimensions of quality. *IEEE Engineering Management Review, 24: 15-23.*

Gelfand, A. E., Hills, S. E., Racine-Poon, A. and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association, 85(412), 972-985.*

Gelman, A., Carlin, J. B. B., Stern, H. S. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. B. (2014). *Bayesian Data Analysis, Third Edition.* CRC Press.

Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-6(6), 721-741.*

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments BT. *Bayesian Statistics 4.*

Geyer C. (1991). Markov chain Monte Carlo maximum likelihood. *Computing Science and Statistics: Proc. 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, VA, 156–163.*

Gibbons, R. D. and Hedeker, D. (1994). Application of Random-Effects Probit Regression Models. *Journal of Consulting and Clinical Psychology, 62(2), 285-296.*

Giles, M. B. (2015). Multilevel Monte Carlo Methods. In Springer Proceedings in Mathematics and Statistics (pp. 83–103). (https://doi.org/10.1007/978-3-642-41095-6_4)

Hadfield, J. (2010). MCMC methods for Multi–response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software, 33(2), 1-22.*

Hadfield, J. (2016). *MCMCglmm Course Notes.* Retrieved Nov 15, 2017 from https://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf.

Harrison, X. A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *Journal of PeerJ, 2, 18-37.*

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika, 57(1), 97–109.*

Heidelberger, P. and Welch, P. D. (1981). A spectral method for confidence interval generation and run length control in simulations. *Communications of the ACM, 24(4), 233-245.*

Huda, A. S. N. and Živanović, R. (2018). Efficient Estimation of Interrupted Energy with Time-Varying Load Models for Distribution Systems Planning Studies. *IFAC-PapersOnLine, 51(2), 208-213.*

Kass, R. E., Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (2006). Markov Chain Monte Carlo in Practice. *Journal of the American Statistical Association, 92(440), 1645.*

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association, 91(435), 1343–1370.*

Kingman, J. F. C. and Harrison, J. M. (1987). Brownian Motion and Stochastic Flow Systems. *The Statistician, 36(1), 66.*

Laird, N. M. and Ware, J. H. (1982). Random Effects Models for Longitudinal Data. *Biometrics, 38(4), 963-974.*

Lambert D. (1992). Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34(1), 1-14.*

Li G. and Shi J. (2012). Applications of Bayesian methods in wind energy conversion systems. *Renew Energy, (43):1-8.*

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika,73(1), 13-22.*

Lindstrom, M. J. and Bates, D. M. (1988). Newton Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association, 83(404), 1014-1022.*

Mansfield, E. R. and Helms, B. P. (1982). Detecting Multicollinearity. *The American Statistician, 36(3), 158.*

Marjoram, P., Molitor, J., Plagnol, V. and Tavare, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, 100(26), 15324–15328.*

Mccullagh, P. and Nelder, J. (1989). *Generalized linear models, Second Edition*, Chapman & Hall.

MCMCglmm-utils.R. (MoNE). (2011) *vif.MCMCglmm.* Retrieved May 25, 2019 from https://github.com/aufrank/R-hacks/blob/master/MCMCglmm-utils.R .

Meeuwsen, J.K and Kling, W. L. and Ploem, W. A.G.A. (1997). The influence of protection system failures and preventive maintenance on protection systems in distribution systems. *IEE Trans Power Delivery, 12(1): 125-131.*

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics, 21(6), 1087.*

Min, Y. and Agresti A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling: An International Journal, 5(1), 1-19.*

Montgomery, D. C. (2001). Design and analysis of experiments 5ed. *Quality and Reliability Engineering International, 3(3), 212.*

Moradkhani, A. and Haghifam M. R. and Mohammedzadeh M. (2014). Bayesian estimation of overhead lines failure rate in electrical distribution systems. *Electrical Power and Energy Systems, 56, 220-227.*

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics. 33(3), 341-365.*

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, 135, 370.*

O'Hagan, J. and McCabe, B. (1975). Tests for the Severity of Multicolinearity in Regression Analysis: A Comment. *The Review of Economics and Statistics, 57(3), 368.*

Özkıvrak, Ö. (2005). Electricity Restructuring in Turkey. *Energy Policy, 33(10), 1339-1350.*

Preisser, J. S., Stamm, J. W., Long, D. L., and Kincade, M. E. (2012). Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. *Caries Research, 46(4), 413-423.*

Prieto, F., Sarabia J. and Saez A. (2014). Modelling major failures in power grids in the whole range. *International Journal Electrical Power & Energy Systems, 54, 10-16.*

Resources, M. of N. E. and. (2017). *Info Bank Energy Electricity*. Retrieved June 20, 2017 from http://www.enerji.gov.tr/en-US/Pages/Electricity

Rowell, J. G. and Walters, D. E. (1976). Analyzing data with repeated observations on each experimental unit. *The Journal of Agricultural Science, 87(2), 423-432.*

Silverman, B. (1986). *Density estimation for statistics and data analysis.* Chapman & Hall.

Skellam, J. G. (1948). A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as Variable between the Sets of Trials. *Journal of the Royal Statistical Society: Series B (Methodological), 10(2), 257–261.*

TEIAŞ. (2018). *Hakkımızda.* Retrieved April 4, 2018 from https://www.teias.gov.tr/bolge/van/hakkimizda

Tatietse, T. T. and Villeneuve P. and Ndong E.P. N. and Kenfack F. (2002). Interruption modelling in medium voltage electrical networks. *International Journal of Electrical Power and Energy Systems, 24(10), 859-865.*

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *The American Statistician, 39(2), 95-101.*

APPENDICES

A. DIAGNOSTIC CHECKS FOR THE FIRST IMPLEMENTATION OF THE MODELS

1. Diagnostic Checks for the first implementation of Poisson MCMCglmm

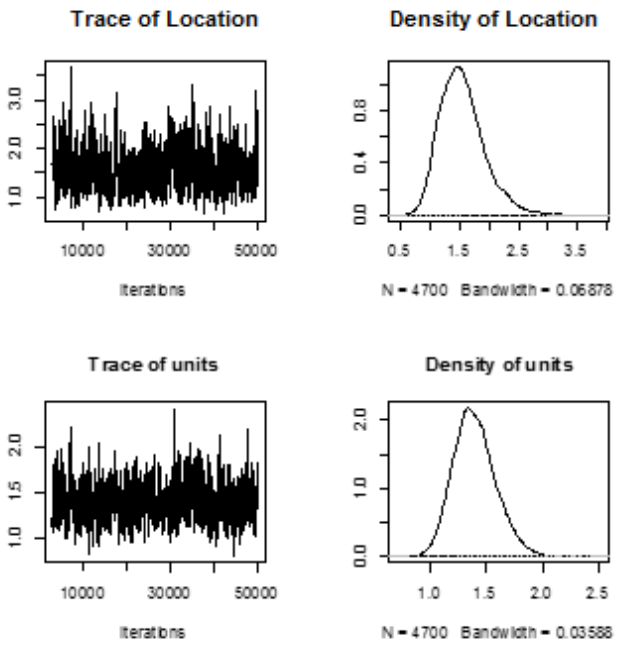

Figure A.1 *The trace and density plots of posterior distributions of (co)variances matrices of the first Poisson MCMCglmm.*

The second diagnostic check of the model will be the queue of autocorrelation (Figure A.2.).

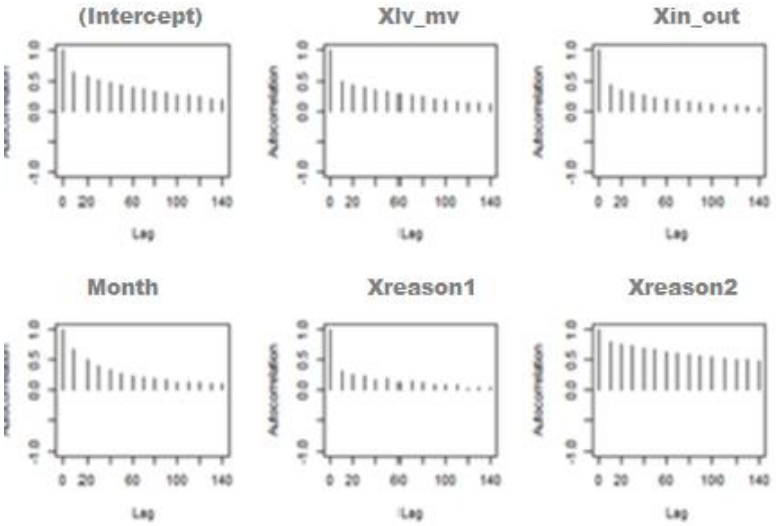

Figure A.2: *The visual diagnostic checking for autocorrelation status the first implementation of Poisson MCMCglmm.*

2. Diagnostic Checks for the first implementation of Zero-Inflated Poisson MCMCglmm

The first diagnostic check's results are given in the Figure A.3 below. Trace and Density plots of two latent variable are: Poisson and Zero-Inflated.

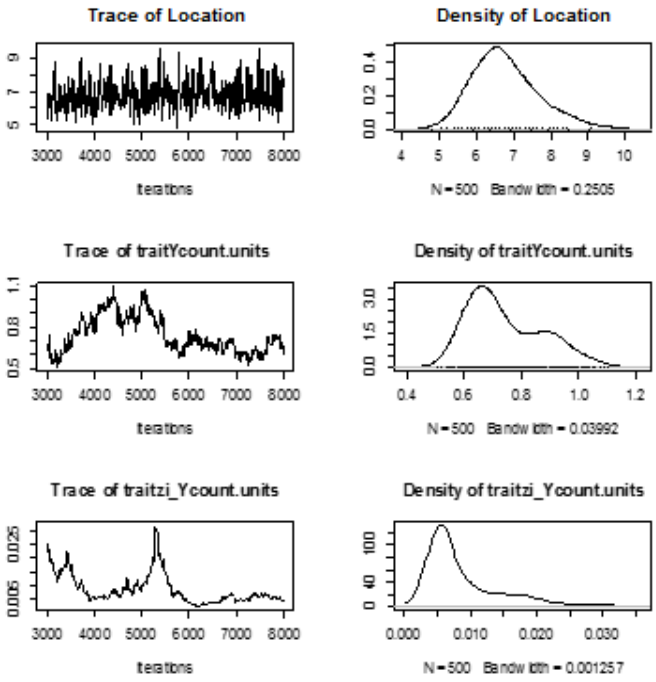

Figure A.3 *Trace and Density Plots of 2 latent variable: Poisson and Zero-Inflated.*

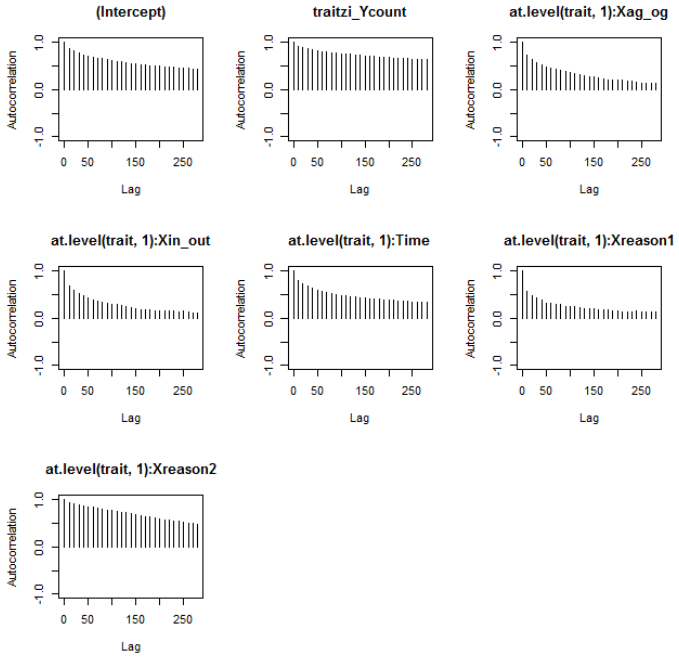

Figure A.4 *The visual diagnostic checking for autocorrelation status the first implementation of Zero-Inflated Poisson MCMCglmm.*

B. DIAGNOSTIC CHECKS FOR THE POISSON AND ZIP MODELS WITH INTERACTION EFFECTS AND PIECEWISE INDICATOR VARIABLE

1. Diagnostic Checks for the Poisson MCMCglmm with Interaction Effects and Piecewise Indicator Variable

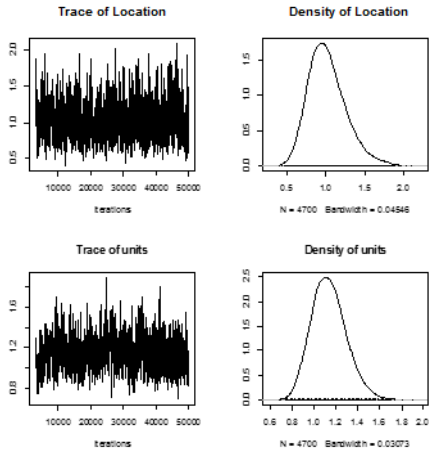

Figure B.1 *Trace and density plots of posterior distributions for Poisson MCMCglmm with interaction effects and indicator variable.*

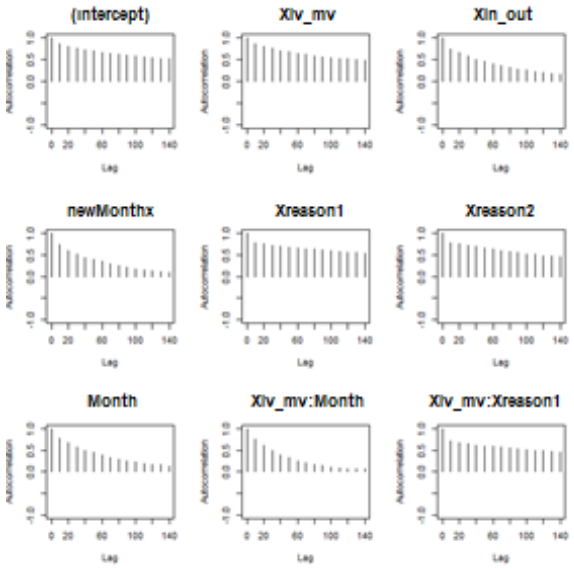

Figure B.2 *Autocorrelation plots of Poisson MCMCglmm with interaction effects and piecewise indicator variable*

2. Diagnostic Checks for the ZIP MCMCglmm with Interaction Effects and Piecewise Indicator Variable

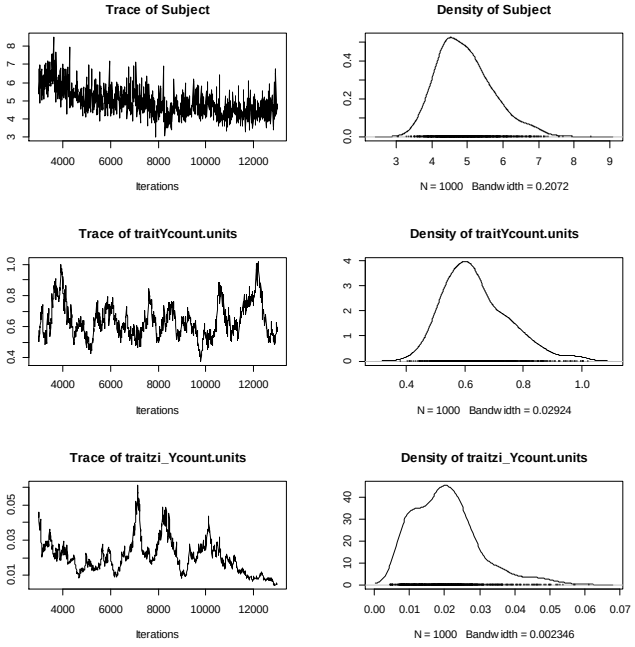

Figure B.3 *Trace and density plots of posterior distribution of Zero-Inflated Poisson MCMCglmm with interaction effects and piecewise indicator variable*

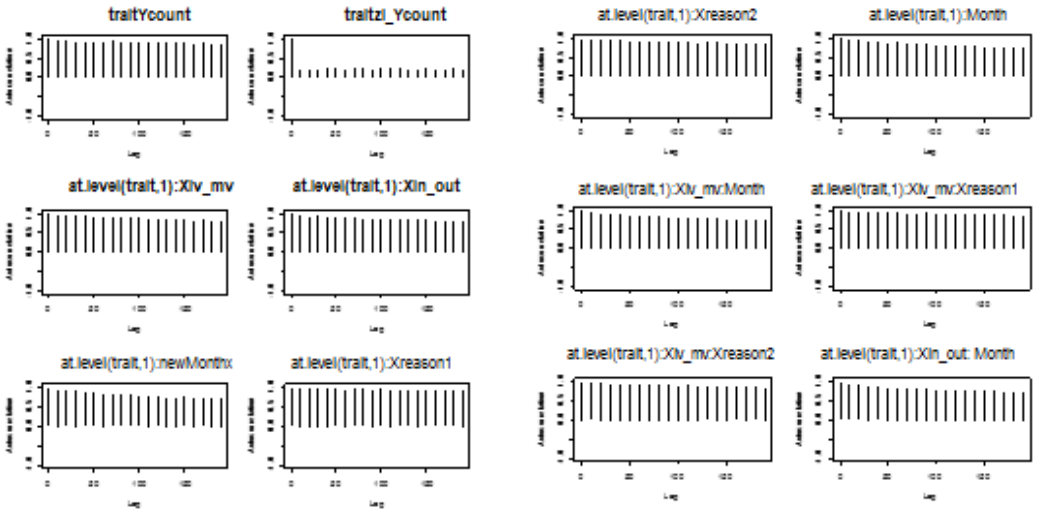

Figure B.4 *Autocorrelation Plots for Zero-Inflated Poisson MCMCglmm with interaction effects and piecewise indicator variable*

C. DIAGNOSTIC CHECKS FOR THE FINAL IMPLEMENTATION OF POISSON AND ZIP MODELS

1. The final implementation of Poisson MCMCglmm

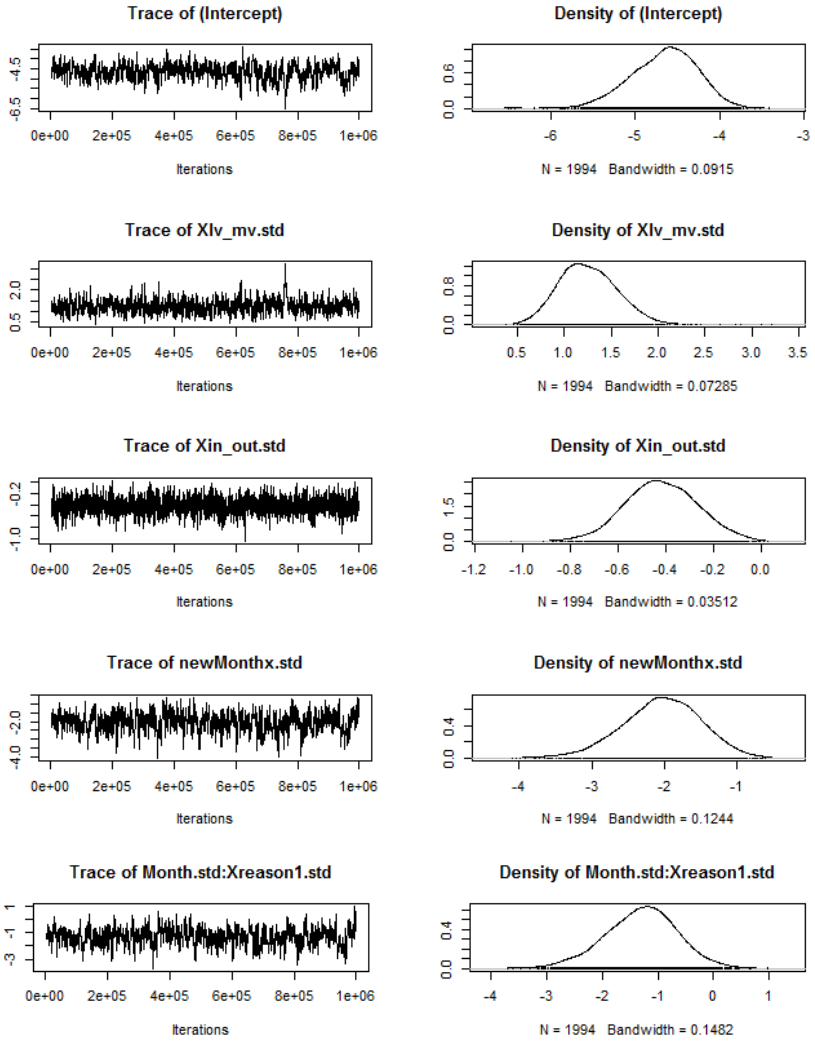

Figure C.1 *Trace and Density Plots for the final implementation of Poisson MCMCglmm with interaction effects and piecewise indicator variable*

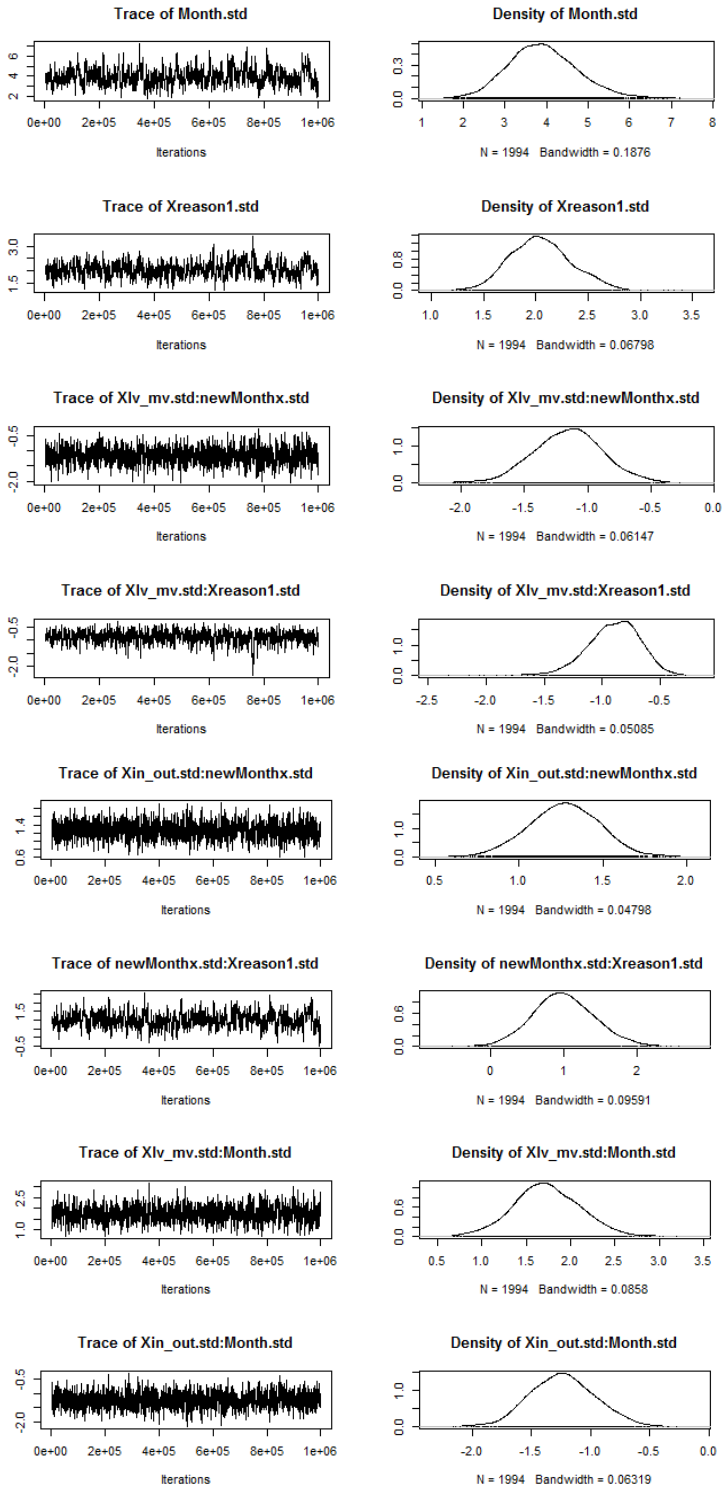

Figure C.1 *Continue Trace and Density Plots for the final implementation of Poisson MCMCglmm with interaction effects and piecewise indicator variable*

1.2 Autocorrelation Plots of the final implementation of Poisson MCMCglmm

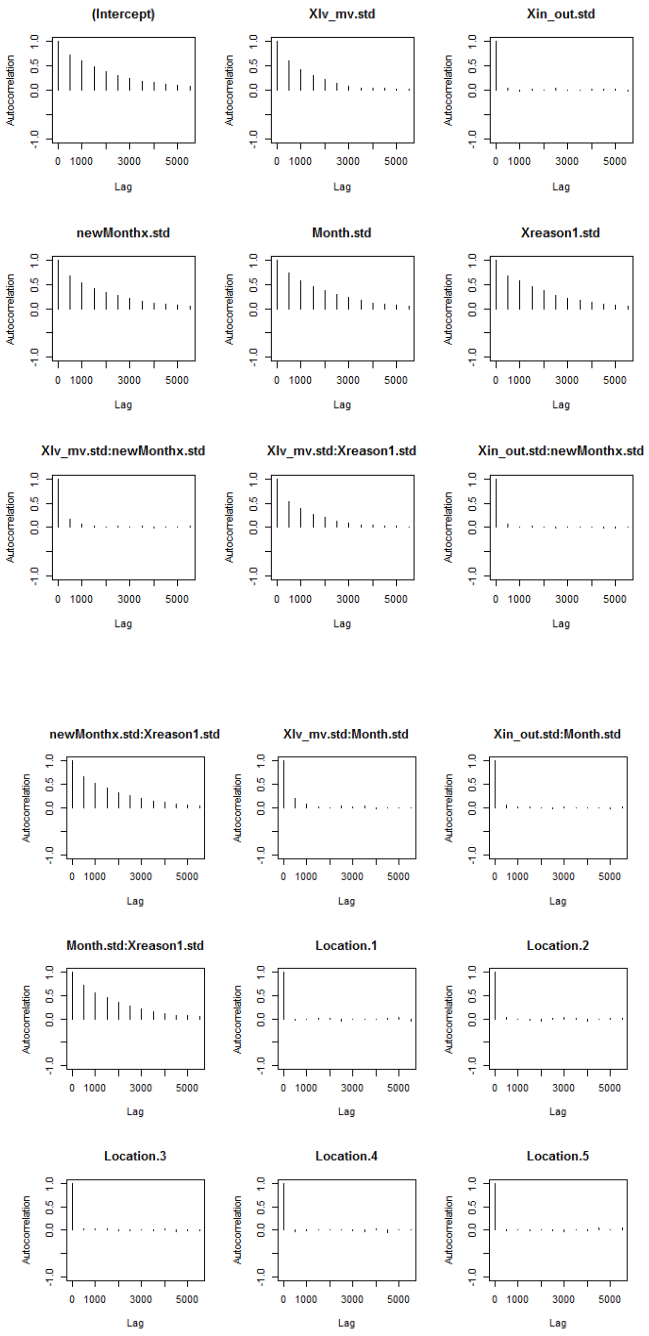

Figure C.2 *Autocorrelation Plots for the final implementation of Poisson MCMCglmm with interaction effects and piecewise indicator variable*

2. The final implementation of zero-Inflated Poisson MCMCglmm

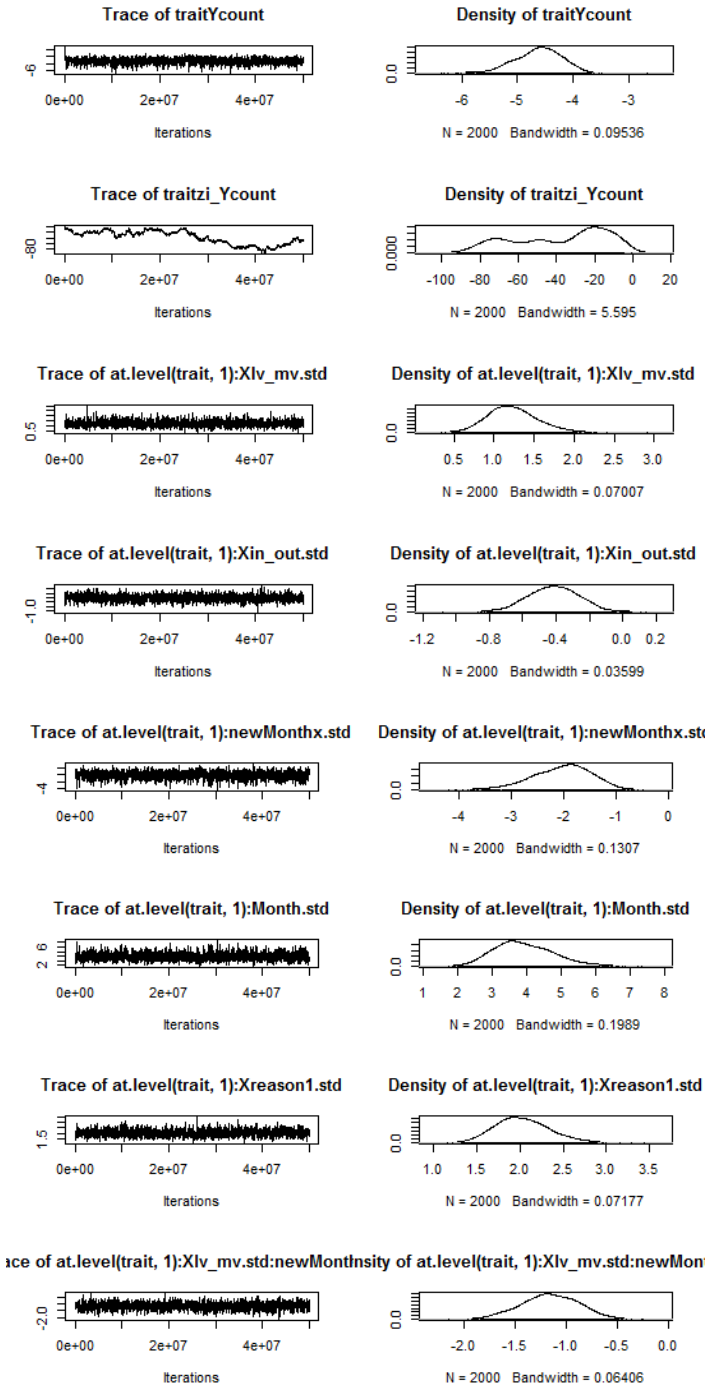

Figure C.3 *Trace and Density Plots for the final implementation of ZIP MCMCglmm with interaction effects and piecewise indicator variable*

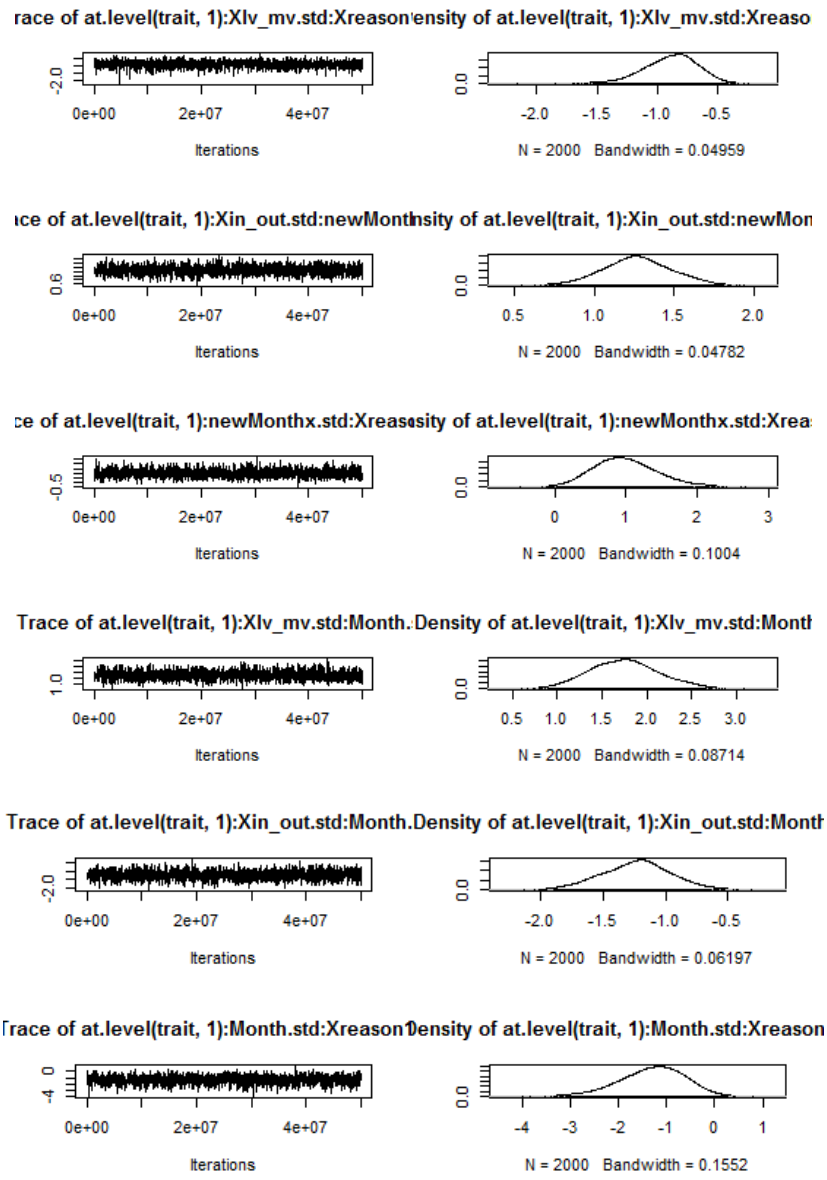

Figure C.3 *Continue Trace and Density Plots for the final implementation of ZIP MCMCglmm with interaction effects and piecewise indicator variable*

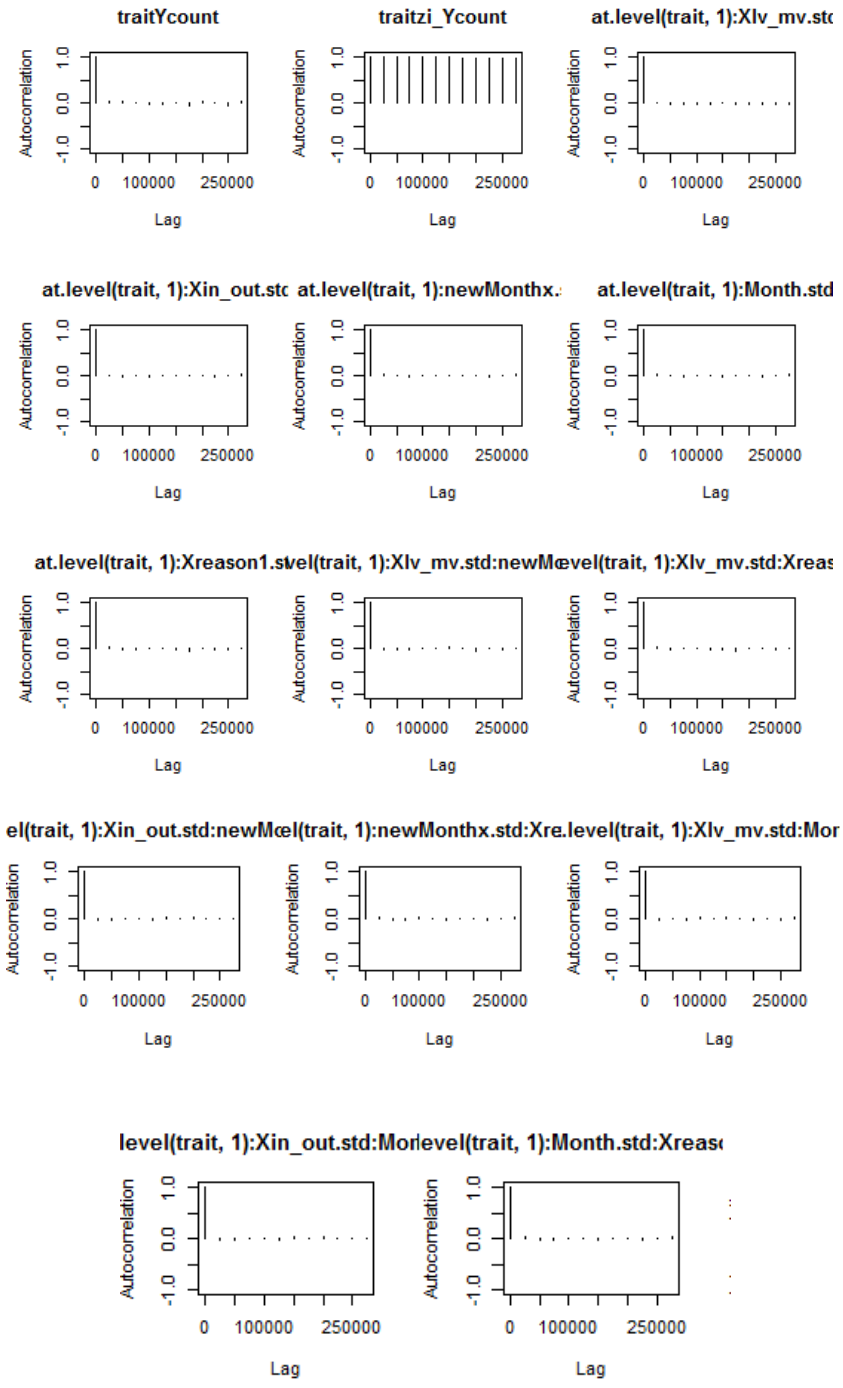

Figure C.4 *Autocorrelation Plots for the final implementation of ZIP MCMCglmm with interaction effects and piecewise indicator variable*

3. The Implementation of Hurdle Poisson MCMCglmm

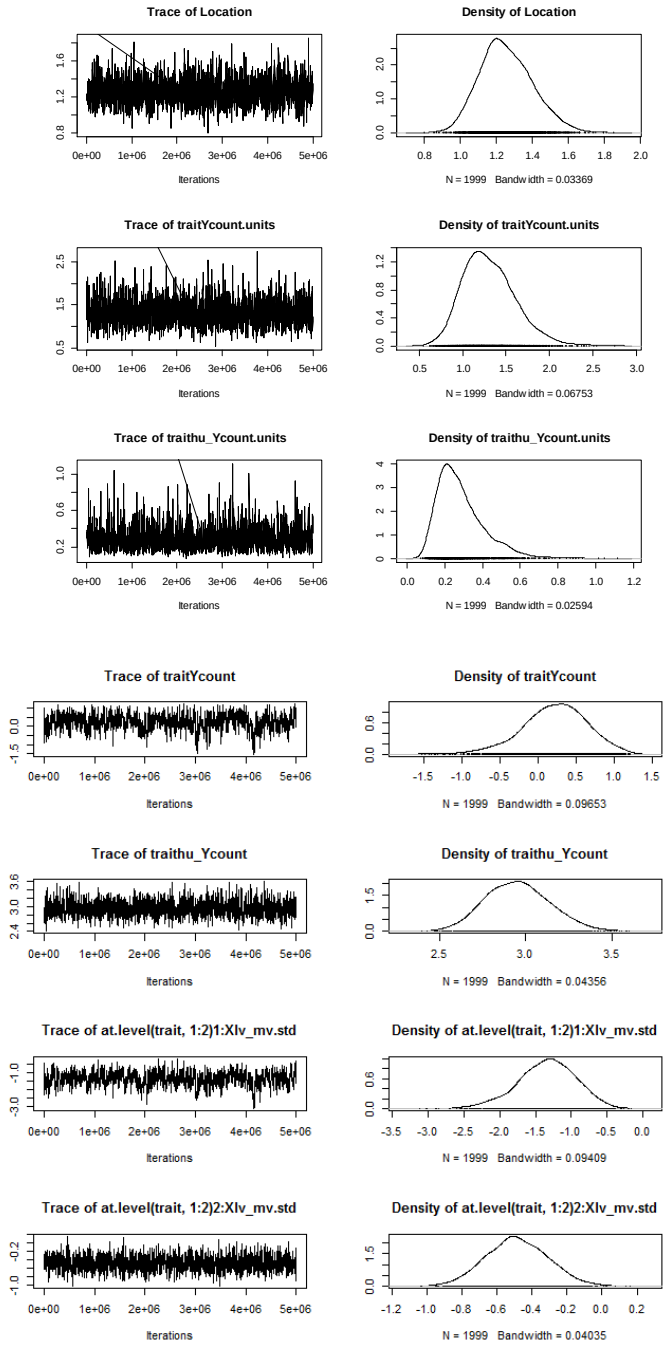

Figure C.5 *Trace and Density Plots for the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

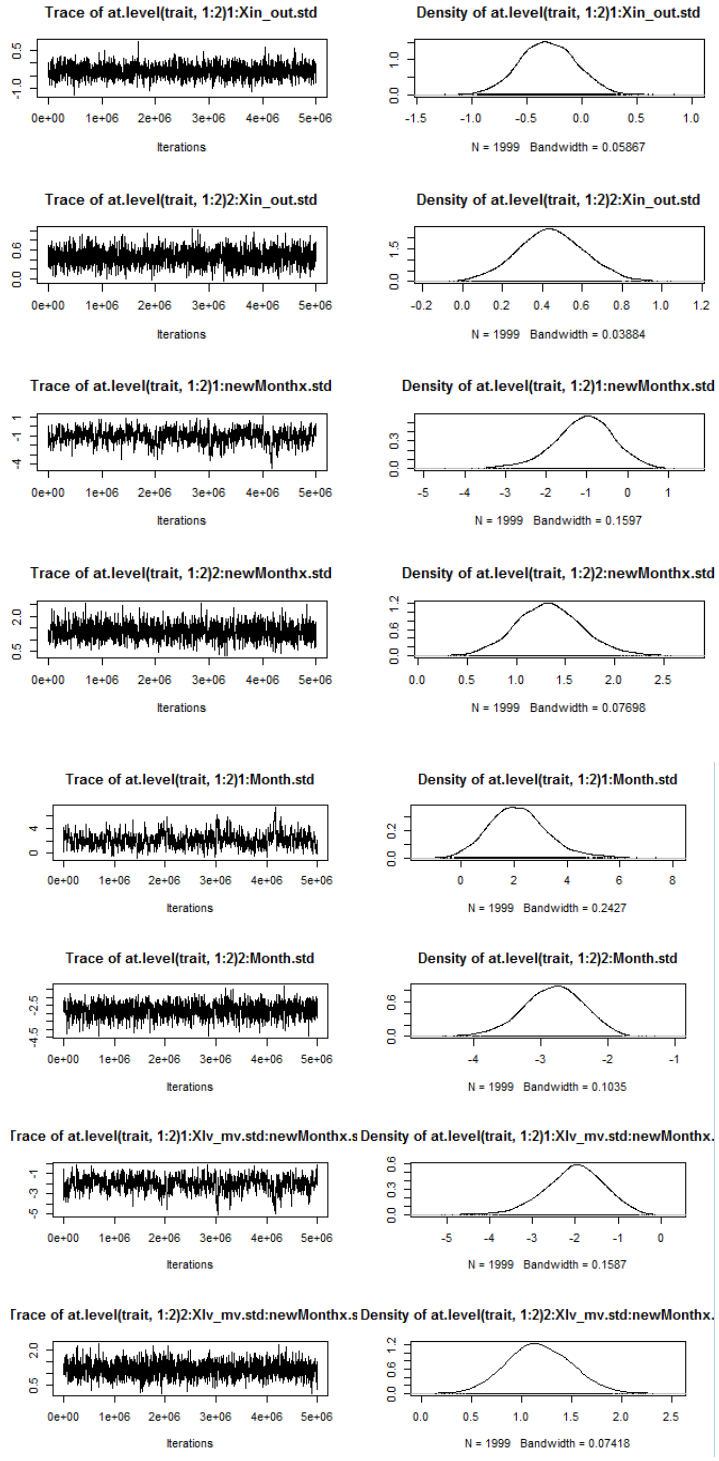

Figure C.5 *Continue Trace and Density Plots for the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

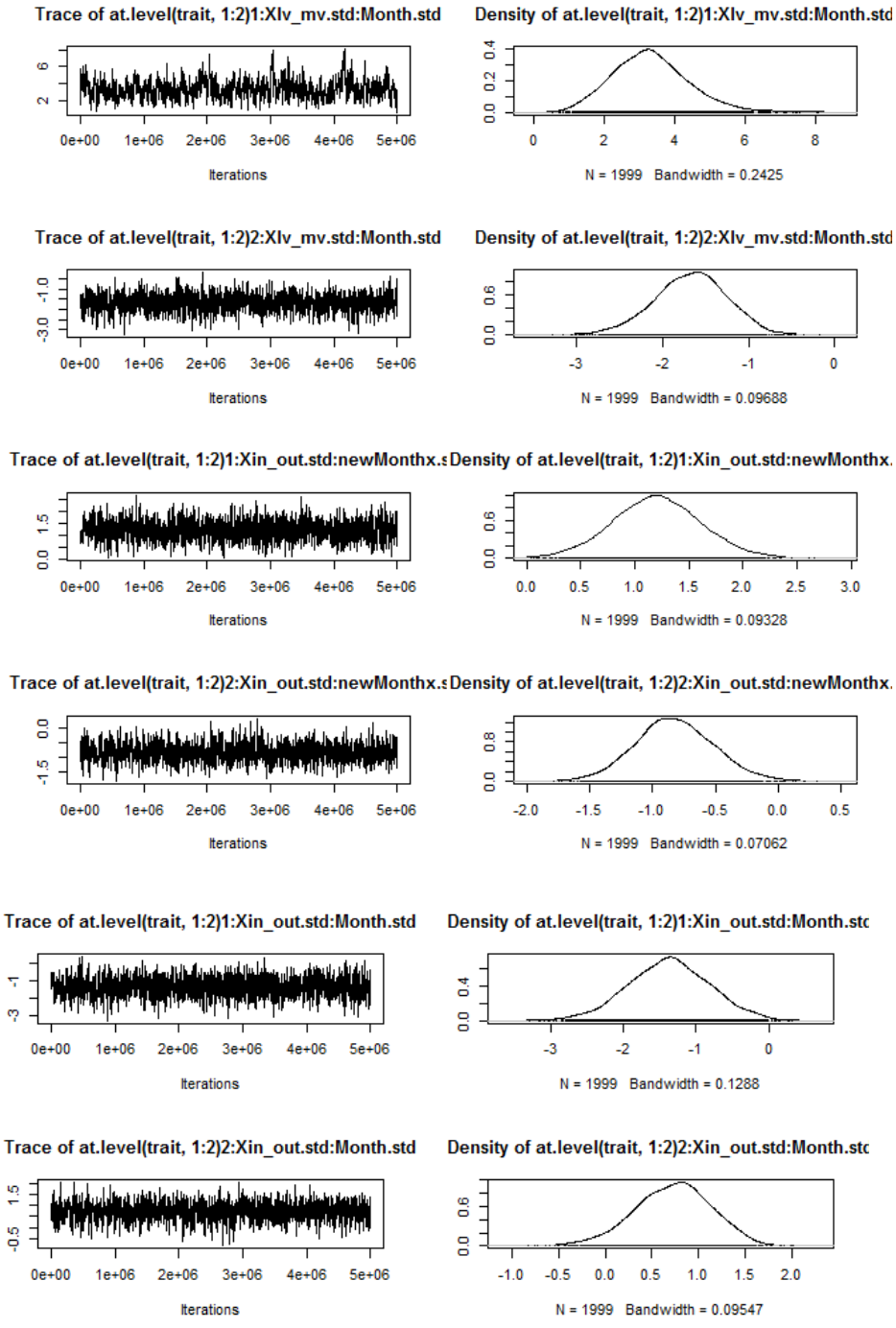

Figure C.5 *Continue Trace and Density Plots for the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

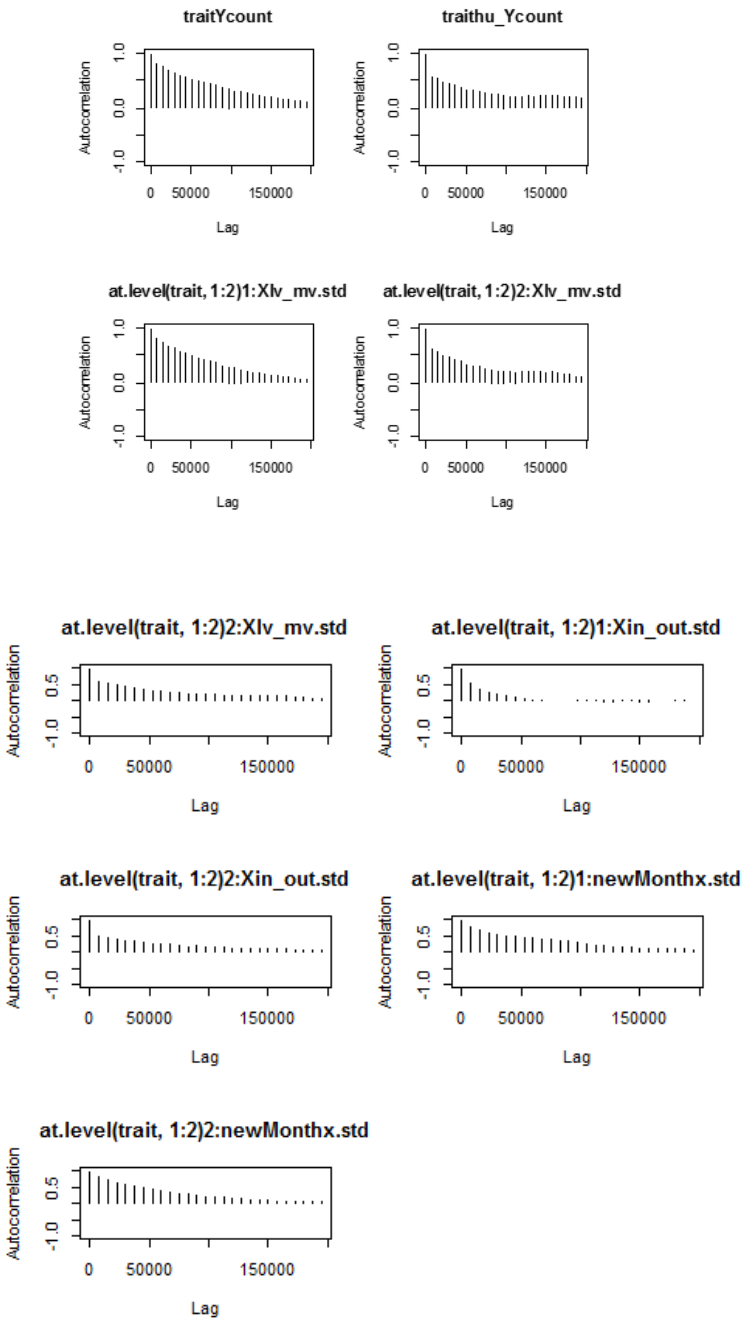

Figure C.6 *Autocorrelation Plots for the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

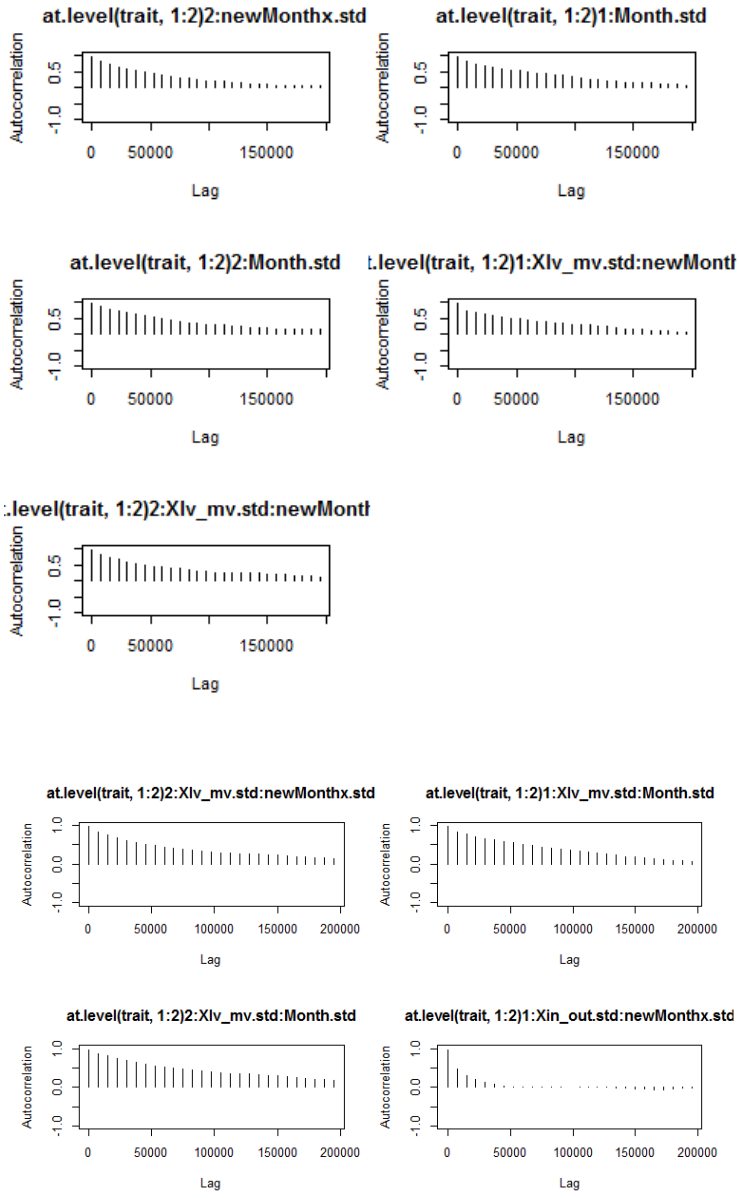

Figure C.6 *Continue Autocorrelation Plots for the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

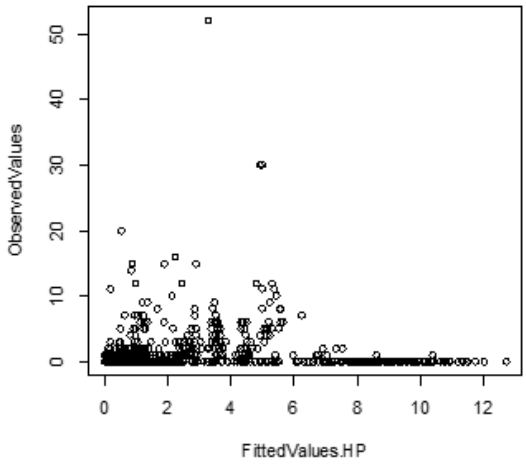

Figure C.7 *Zoomed plots for Observed vs Fitted Values of the Hurdle Poisson MCMCglmm with interaction effects and piecewise indicator variable*

D. THE RESULTS OF MODELS FOR OTHER IMPLEMENTATION

1. Poisson MCMCglmm and Zero-Inflated Poisson MCMCglmm with Method of Centering

In this section, centering method will be used to overcome autocorrelation problem in developed models, which were built by adding interaction effects and slope parameters of Poisson MCMCglmm and Zero-Inflated Poisson MCMCglmm.

Poisson MCMCglmm had autocorrelation problem (see Appendix-A). Therefore, only autocorrelation plots of covariates of developed Poisson MCMCglmm will be evaluated here (Figure D.1).

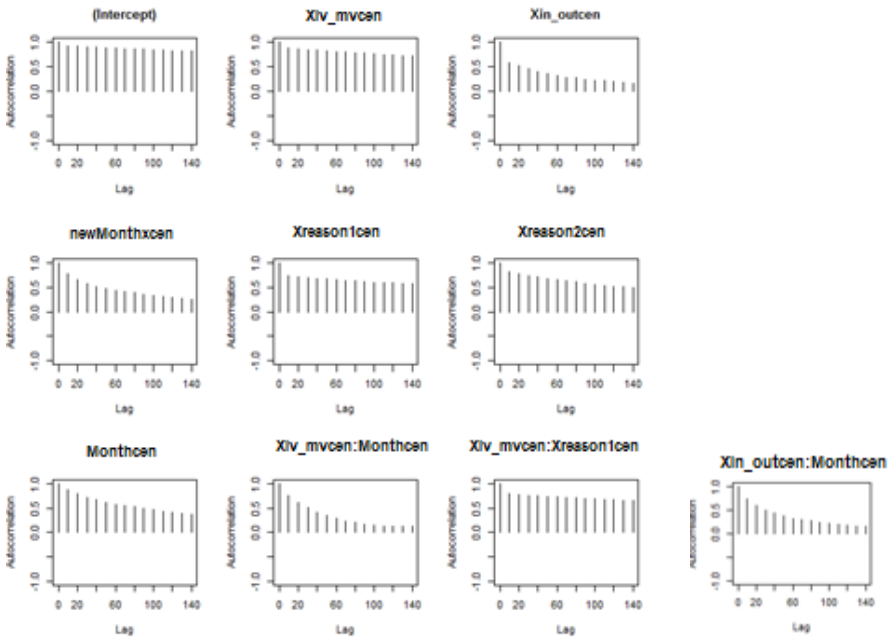

Figure D.1 *Autocorrelation Plots of Covariates for Poisson MCMCglmm with Using Centering Method.*

Centering Method could solve autocorrelation problem for covariates of developed Poisson MCMCglmm by adding interaction effects and slope parameter. Autocorrelation problem is still observed on covariates of centered Xlv_mv, Xreason1, Xreason2 and also in the interaction of Xlv_mv and Xreason1. Consequently, it is understood that autocorrelation problem in this model cannot be solved by the methods of centering or increasing number of iterations.

1.2 Zero-Inflated Poisson MCMCglmm with interaction effects and slope parameter by using Centering Method

There were not any differences in developed Poisson MCMCglmm with Centering Method. Probably, this situation will be the same at developed Zero-Inflated Poisson MCMCglmm with Centering, too. To see differences of the autocorrelation status on

the model, autocorrelation plots of covariates will be shown below (Figures D.2 and D.3).

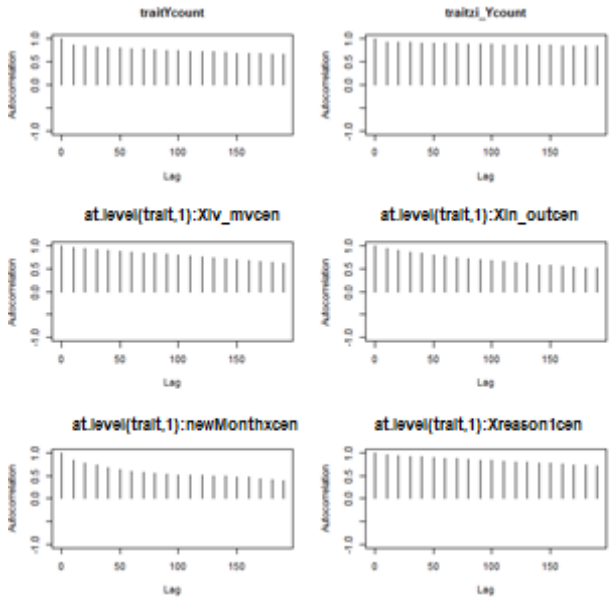

Figure D.2 Autocorrelation Plots of Covariates for Zero-Inflated Poisson MCMCglmm with using Centering Method

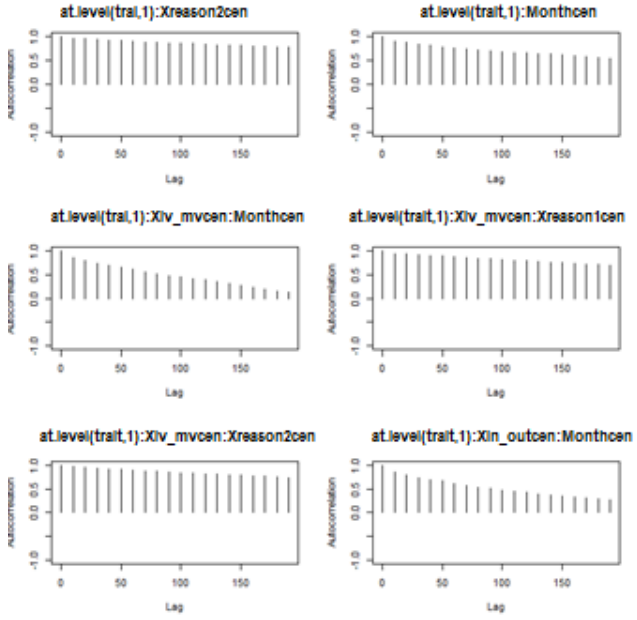

Figure D.3 *Continue Autocorrelation Plots of Covariates for Zero-Inflated Poisson MCMCglmm with using Centering Method*

According to autocorrelation plots, centering method could not solve autocorrelation problem in Zero-Inflated Poisson MCMCglmm with interaction effects and piecewise indicator variable as well. In fact, autocorrelation problem had been observed only at the covariate of Xreason2 in the first implementation of Poisson and Zero- Inflated Poisson MCMCglmm. Even so, since generating counts did not exactly fit when observed values were bigger than 10, the models were needed to be improved by adding interaction effects and piecewise indicator variable parameter. During the act of fixing and improving the strength of generating counts in models by adding interaction effects and slope parameter, it was realized that the autocorrelation problem was posing even a higher threat. While this result is understandable given the situation at hand since relations between covariates increase as a consequence of interaction effects and slope, it needs to be acknowledged that unknown relations between covariates might cause wrong generating.

E. A PART OF ELECTRICITY INTERRUPTION DATASET

No	Location	Month	Ycount	Yaverage	Xlv_mv	Xin_out	Xreason1	Xreason2
1	Çankırı_atkaracalar_LV_In_external	1	0	0	1	1	0	0
2	Çankırı_atkaracalar_LV_In_external	2	0	0	1	1	0	0
3	Çankırı_atkaracalar_LV_In_external	3	0	0	1	1	0	0
…	…	…	…	…	…	…	…	…
289	Çankırı_çerkeş_LV_In_external	1	0	0	1	1	0	0
290	Çankırı_çerkeş_LV_In_external	2	0	0	1	1	0	0
291	Çankırı_çerkeş_LV_In_external	3	0	0	1	1	0	0
292	Çankırı_çerkeş_LV_In_external	4	0	0	1	1	0	0
293	Çankırı_çerkeş_LV_In_external	5	0	0	1	1	0	0
294	Çankırı_çerkeş_LV_In_external	6	2	0.2585	1	1	0	0
295	Çankırı_çerkeş_LV_In_external	7	1	0.617	1	1	0	0
296	Çankırı_çerkeş_LV_In_external	8	1	0.067	1	1	0	0
297	Çankırı_çerkeş_LV_In_external	9	0	0	1	1	0	0
298	Çankırı_çerkeş_LV_In_external	10	0	0	1	1	0	0
299	Çankırı_çerkeş_LV_In_external	11	6	0.283	1	1	0	0
…	…	…	…	…	…	…	…	…
1093	Çankırı_kurşunlu_MV_In_security	1	0	0	0	1	0	1
1094	Çankırı_kurşunlu_MV_In_security	2	0	0	0	1	0	1
1095	Çankırı_kurşunlu_MV_In_security	3	0	0	0	1	0	1
1096	Çankırı_kurşunlu_MV_In_security	4	0	0	0	1	0	1
1097	Çankırı_kurşunlu_MV_In_security	5	0	0	0	1	0	1
1098	Çankırı_kurşunlu_MV_In_security	6	0	0	0	1	0	1
1099	Çankırı_kurşunlu_MV_In_security	7	0	0	0	1	0	1
1100	Çankırı_kurşunlu_MV_In_security	8	0	0	0	1	0	1
…	…	…	…	…	…	…	…	…
1723	Çankırı_yapraklı_MV_out_Operator	7	1	2.117	0	0	1	0
1724	Çankırı_yapraklı_MV_out_Operator	8	2	27.395	0	0	1	0
1725	Çankırı_yapraklı_MV_out_Operator	9	4	13.875	0	0	1	0
1726	Çankırı_yapraklı_MV_out_Operator	10	0	0	0	0	1	0
1727	Çankırı_yapraklı_MV_out_Operator	11	1	34.820	0	0	1	0
1728	Çankırı_yapraklı_MV_out_Operator	12	1	0.617	0	0	1	0

Table E.1 *A Part Of Electricity Interruption Data*