A DATA ADAPTIVE MODEL FOR RETAIL SALES OF ELECTRICITY

by

Johanna Marcelia

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Mathematics

Boise State University

May 2021

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Johanna Marcelia

Thesis Title: A Data Adaptive Model For Retail Sales of Electricity

Date of Final Oral Examination: 02 April 2021

The following individuals read and discussed the thesis submitted by student Johanna Marcelia, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Jaechoul Lee, Ph.D.	Chair, Supervisory Committee
Donna Calhoun, Ph.D.	Member, Supervisory Committee
Hans-Peter Marshall, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Jaechoul Lee, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor Dr. Lee for his support, guidance, and encouragement through this project. I also want to thank my committee members Dr. Calhoun and Dr. Marshall for their instruction and insight. Additionally, thanks to my mathematics professors from Whitworth University whose mentorship gave me the confidence to pursue this goal.

I would also like to acknowledge the significant role my family and friends have played in helping me reach this milestone. I am grateful for the love, encouragement, and joy you each bring to my life

Finally, I thank God, the true source of all knowledge and wisdom.

ABSTRACT

When fitting a model to a data set, the goal is to create a model that captures the trends present in the data. However, data often contains regions where the underlying model changes or exhibits shifts in certain parameters due to economic events. These locations in the data are known as changepoints, and ignoring them can result in high error and incorrect forecasts. By developing a specific cost function and optimizing using the genetic algorithm, we are able to locate and account for the changepoints in a given data set. We specifically apply this process to the retail sales of electricity in the United States by examining data sets from each state's residential, commercial, and industrial sectors. We demonstrate that, when changepoints are accounted for, model trends can be computed more accurately. We specifically explore this in the case of data sets that exhibit changepoints due to the 2020 (and ongoing) pandemic.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF MAPS

CHAPTER 1

INTRODUCTION

Almost all parts of life in the 21st century rely on direct or indirect use of electricity. Electricity has become a requirement in residential life, commercial industry, government, business, healthcare, transportation, communication, etc. Because electricity is so important, when shortages or blackouts occur, almost all parts of daily life are impacted. Thus, the models that energy companies and electricity utilities rely on must provide reliable predictions so that the supply of electricity will always meet demand.

Most electricity models fall into two basic categories: models built off of previous daily, monthly, or annual electricity demand, and models built off of "end use" consumption. Because the data required for end use consumption models is proprietary, we will be examining models that are built from public historical data. Historical data sets include not only accurate demand or supply values, but also locations in the data where anomalous events may have occurred that are not consistent with the true underlying nature of the demand patterns. Additionally, these data sets contain regions where the true demand pattern shifts. Changes to the underlying demand pattern can sometimes be explained by large scale power outages, energy intensive business openings or closures, rapid increase or decrease of regional population, etc. We use the term "changepoints" to refer to a location where the data signals a change

in pattern. Mathematically, if a model is being fitted to the data, a changepoint indicates a parameter change. While the events leading to these changepoints may be recorded in local news or through other qualitative means, they are rarely noted in data documentation. Because of this, models built on historical data may unknowingly produce erroneous results if changepoints are not accounted for.

When fitting any theoretical model to real world data, it is well known that data disruptions, changes in instruments, or shifts in economic variables all can alter the underlying model and thus indicate a changepoint. This changepoint could manifest itself as a shift in the mean or slope of the data, a change in the type of underlying model, or a change in some other parameters [2]. Due to these shifting parameters, data that contains changepoints should not be modeled using the same parameters across the entire data set. Rather, once changepoints are located in the data, data points between the changepoints may then be fitted with region specific parameters that adjust part or all of the model for that section of the data set [2].

Detection of these locations is important in many industries such as finance, climatology, medical modeling, bioinformatics, speech recognition, and energy, and because of this, many changepoint detection methods exist [3]. "Offline" methods rely on entire historical data sets to determine changepoint locations, while "online" methods account for new data points as they occur using sliding windows of data [3]. While both methods have useful applications, offline methods are typically applied when large amounts of historical data are available, and when the interpretation of historical changepoints should be considered. Additionally, parametric and non-parametric changepoint techniques have been developed. Non-parametric methods do not rely on estimating an underlying distribution and are useful when little about the structure of the data can be determined. They also show promise when working

with extremely large data sets, but, for smaller data sets and situations where data already fits a given distribution, a parametric method may perform best [3].

Regardless of method, most changepoint detection algorithms rely on developing a cost function and a search method to locate a known or unknown number of changepoints. The search space is usually fully or partially divided into a certain number of regions, such as in the Shapelet method or minimum description length (MDL) method, or the function may simply rely on Bayesian inference techniques [4]. All of these can be used in combination with likelihood models to aid in parameter optimization [3].

Once a cost function has been selected, a method of optimization must be chosen. In this paper, we will be making use of the genetic algorithm (GA) to optimize the cost function. The GA was first introduced by John Holland and is based on the theory of natural selection [5]. In this theory, an initial population of some species is identified, with some members of that population having more desirable traits than other based on the environment in which they must survive. As the species breeds the next generation, members of the parent generation with the best features pass those down to their children more often than members with less desirable traits. If there is enough variety in the starting population, subsequent generations will become better and better suited to their environment. Ultimately, after many generations, the resulting species will be better adapted to it's environment than it's far removed ancestors [6]. When translated into mathematical terms, an initial population of candidate solutions is proposed, and, as the algorithm iterates, subsequent iterations of proposal solutions pass down their best changepoint options allowing the algorithm to get closer and closer to the optimal number of changepoints and their locations.

This mechanism of optimizing a cost function may be applied to any type of

data that contains changepoints. However, these methods are not typically used by utility companies when forecasting the amount of electricity that certain sectors demand. Within the utility industry, historically, bottom up end-use models have been frequently used [7]. In these models, the amount of electricity required in a given sector is estimated based on the electricity consumption profile of the utility's individual customers. These include estimates for heating, cooling, appliances, etc. [7]. These models require very detailed personal data concerning the characteristics of individual customers, and, due to the highly protected nature of such data, are thus only feasible for utilities themselves to model. A more modern approach is based on top down econometric models that incorporate elements like historical electricity sales from a shorter time frame (five years or less) and that break down historical consumption variables by regional groups of consumers, as well as other economic and qualitative variables [7]. Some utilities employ a combination of these methods to assist with forecasting the long term impact of future electricity consumption and sales. Changepoints are not typically identified as such by utility companies, rather the historical time period is shortened to minimize the inclusion of outdated trends [7] [8]. Due to laws guarding individual customer consumption information, the data required to replicate current end-use and econometric models is not available to those outside utility companies. This poses many drawbacks to those attempting analysis in the utility industry. Therefore, our methodology relies only on public data that is aggregated by sector. Additionally, our method only incorporates historical electricity data, so no other economic data needs to be collected and evaluated for correlation.

To locate and account for changepoints, we will develop a cost function which we will optimize using the GA to mathematically determine the location and number of changepoints. Once changepoint locations have been determined, we'll use a time

series regression model with varying slopes and intercepts to fit a model to the data between each set of changepoints. We expect this method to improve modeling accuracy for electricity demand. Additionally, we expect this process to allow for better understanding of the regional impact of policies and economic events across sectors in the United States.

CHAPTER 2

DATA

This research was conducted on data from the U.S. Energy Information Agency (EIA). This data set provides the monthly retail sale of electricity by state in million kilowatthours (kWh), and can be downloaded at https://www.eia.gov/electricity/data/browser/. Each state's data is further broken down into sales to residential, commercial, and industrial sectors, making for three data sets per state. Data classifications for transportation and "other" are also provided, but some states do not report any values in these categories, so, for the sake of consistency across each state, we will only focus on the main three reported categories of residential, commercial, and industrial sales. Additionally, we will only be analyzing the 48 contiguous states, as the data from Alaska and Hawaii are incomplete and are impacted by trends that are not easily grouped along with other regions in the United States. Note also that the term "retail sales" refers to sales of electricity made to the end consumer and does not account for any sales made on the wholesale energy market where electricity producers sell excess electricity to one another [9].

The data was reported to the EIA by regional utility companies through the form EIA-826, EIA-860, EIA-923, and EIA-861 [10]. Due to federal and state level requirements and controls for utility companies, true missing values in this type of data are very rare, and, in instances where these occur, the EIA fills in these values

with estimates calculated by regression based on historical annual values prior to publicly releasing the data [10]. The resulting data set has no missing values, and no pre-processing was done on the data prior to analysis.

There are a total of 144 data sets, (three from each state), that will be examined. Observations begin in January 2001 and are recorded monthly through December 2020 providing a total of 240 observations per data set. As illustrated in Figure 2.1, Texas is by far the largest state consumer of electricity, followed by California. The least amount of retail sales of electricity were made to Vermont with other states' sales falling somewhere in between [1]. The balance between residential, commercial and industrial consumption varies greatly by state depending on each state's primary industries, population size, and natural resources.

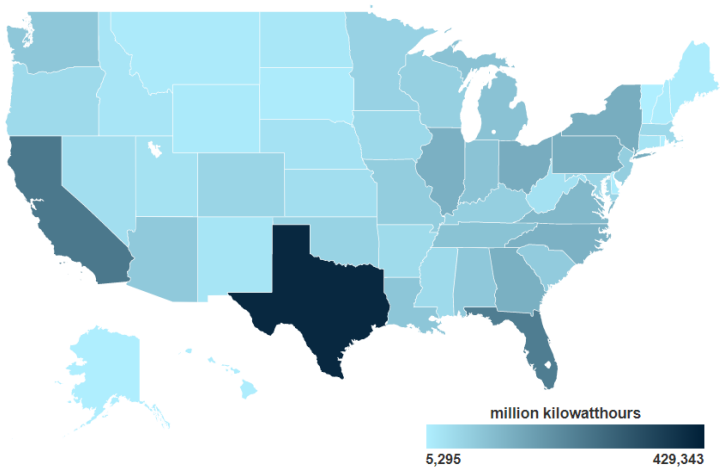


Figure 2.1: 2020 Retail Sales of Electricity by State [1]

All three of the data sets for each state are made up of time series data, meaning that the data is time dependent and exhibits seasonality and autocorrelation. Figure 2.2 shows each sector's data from Washington state, and, visually, the seasonality and time series nature of the data is apparent. For example, residential electricity

consumption peaks in the winter months and falls as temperatures become more moderate in spring, summer, and fall. This gives it a visible cyclic trend which is time dependent and can be described by quantifying the correlation of the data to itself at specific lag lengths (for example, every 12 months or every 6 months). Industrial and commercial data sets for Washington exhibit time series trends as well, but with more pronounced underlying trends.

While typically time series analysis is performed on data sets that have a large number of observations, prior to 2001, this data was not reported to a central agency and is not publicly available, limiting the time frame of this analysis. Additionally, the data does not account for all retail sales of electricity in each state, rather it only reports the electricity sales that fall into the three main categories. Non-retail sales and sales that fall into other categories are not included in this analysis but may be important in understanding a state's electricity consumption and production makeup. Potential jumps in the data may occur when a consumer newly meets or ceases to meet the criteria to fit into the residential, industrial, or commercial categories. Additionally, with the start of the 2020 pandemic, a severe drop can be seen in the trend of some of the data sets. While it can be challenging to accurately model such irregular events, they may offer excellent examples of changepoint locations.

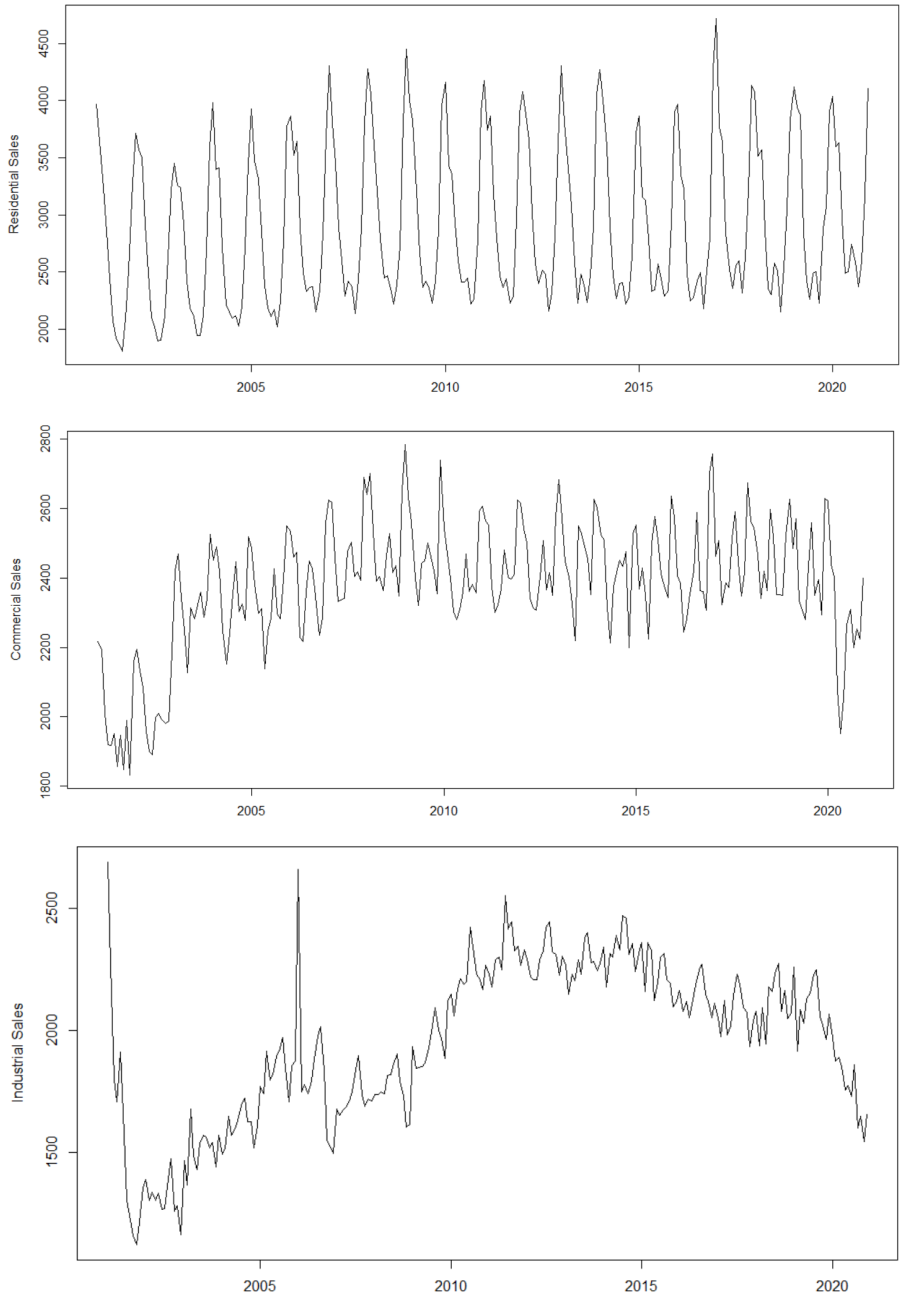


Figure 2.2: Retail Sales of Electricity in Washington in Million kWh

CHAPTER 3

METHODS

3.1 Piece-wise linear trend regression with autocorrelated errors

To arrive at a fitted model that accounts for changepoints, we must go through two main steps. Firstly, we will develop a regression model that fits the data for a known configuration of changepoints. Then, since the true configuration of changepoints is unknown, we must implement a process to locate the changepoints in the data set.

Suppose we have a data set of length n that contains time dependent monthly data as well as k known changepoints which occur in the data set at times $\tau_1, \ldots, \tau_k$. This results in a total of $k+1$ segments within the data set, each of varying lengths. The goal is to fit a linear regression model to each segment of the data between the changepoints, resulting in a partial piece-wise regression model that still accounts for autocorrelated data. In this regression model, let $\beta_0^{(i)}$ be the intercept term and $\beta_1^{(i)}$ be the slope term for the linear regression over the ith segment of the data. Due to the periodicity of the data, we also will include several harmonic terms to capture the cyclicity of the seasonal mean. Specifically, we include $\sin(2\pi t/T)$, $\cos(2\pi t/T)$, $\sin(4\pi t/T)$, and $\cos(4\pi t/T)$, where periodicity of $T = 12$ as the data is monthly. This provides offset convex and concave waves whose periodicity aligns at six month and

yearly locations in the data, modeling these periodic first moment changes. Once the coefficients $\alpha_1, \ldots, \alpha_4$ have been applied to these terms, seasonal changes to the mean that occur every year and every six months will be effectively accounted for in the model. It is worth noting that, in practice, not all sine and cosine terms are significant for each data set. However, the goal is to construct a parsimonious periodic regression model that can be applied accurately to many data sets. We therefore include all four terms to ensure the significant sine and cosine options are available to model each data set. Then our piece-wise regression model can be written as follows:

$$Y_t = \begin{cases} \beta_0^{(1)} + \beta_1^{(1)} t + s_t + \epsilon_t & \text{if } 1 \leq t < \tau_1 \\ \beta_0^{(2)} + \beta_1^{(2)} t + s_t + \epsilon_t & \text{if } \tau_1 \leq t < \tau_2 \\ \vdots & \\ \beta_0^{(k+1)} + \beta_1^{(k+1)} t + s_t + \epsilon_t & \text{if } \tau_k \leq t \leq n, \end{cases} \tag{3.1}$$

where

$$s_t = \alpha_1 \sin(2\pi t/T) + \alpha_2 \cos(2\pi t/T) + \alpha_3 \sin(4\pi t/T) + \alpha_4 \cos(4\pi t/T),$$

and ϵ_t represents the errors for the data that are not accounted for by the other terms in equation (3.1).

Because the data is time dependent, we need to determine the best model to account for the time series nature of the ϵ_t terms. We selected the seasonal autoregressive moving-average (SARMA)$(p, q) \times (P, Q)_T$ model as the mechanism to account for the time dependency. Specifically, this model allows us to calculate the month to month dependency as well as annual autocorrelation which are visually suggested

by the raw data. Here, p and q are the degree of the respective autoregressive and moving average parts of the model, while P and Q denote the degree of the seasonal autoregressive and seasonal moving average parameters. Unlike the harmonic terms which account for seasonal changes to the mean, the SARMA parameters adjust based on autocorrelation, a second moment property. The SARMA$(p,q) \times (P,Q)_T$ model has the general form

$$\epsilon_t = (\phi(B))^{-1}(\Phi(B^T))^{-1}\theta(B)\Theta(B^T)Z_t, \tag{3.2}$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p, \qquad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q,$$
$$\Phi(B) = 1 - \Phi_1 B^T - \cdots - \Phi_P B^{PT}, \qquad \Theta(B) = 1 + \Theta_1 B^T + \cdots + \Theta_Q B^{QT}.$$

Note that B is the backshift operator $X_{t-1} = BX_t$, and, once this model has been applied to ϵ_t, the final residuals will be uncorrelated with a distribution of $Z_t \sim N(0, \sigma^2)$. The parameters ϕ, Φ, θ, and Θ all will be estimated from the data once the specific SARMA model has been chosen. Data driven SARIMA model selection for this research will be discussed in further detail in Chapter 4.

3.2 Optimization via minimum distance length criterion

We now have the information needed to estimate parameters and fit this model to a data set when the location and number of changepoints is given. In actuality, the true number and location of changepoints within a given data set is almost always unknown; it must be estimated using an optimization process. Using equation (3.1),

we will develop a cost function (also referred to as a fitness function) that will be optimized when calculated using the most accurate changepoint estimates that the data can provide. While fitness functions can take on many forms, we'll be using a penalized likelihood function. This fitness function type contains a likelihood value that offers a way to measure the model's "goodness of fit", as well as a penalty term that reduces the likelihood value in a way that is proportional to the model's complexity. Common penalized likelihood fitness function options include the corrected Akaike Information Criterion (AICC), or Bayesian Information Criterion (BAC). While both of these include a likelihood term and a penalty term, these types of function place the same penalty on all model parameters; they do not tailor the penalty strength to fit the type of parameter [11]. In our model, we will be utilizing the minimum description length criterion (MDL) penalty as this penalty function often provides optimal result when applied to time dependent changepoint problems [12] [13].

The MDL penalty specifically allows different parameter types to be penalized by different amounts depending on whether they are real valued, such as the slope parameters in our model, or integer valued, such as the number of changepoints [13]. Because our model requires both types of estimates, this penalty type is ideal for creating a fitness function where certain parameters are weighted more or less heavily [11].

The MDL fitness function is made up of a log likelihood term and a penalty term, giving it the following basic form:

$$MDL(k, \tau_1, ..., \tau_k) = -2\ln(L_{opt}) + P,$$

where L_{opt} is the optimized likelihood model found using the piece-wise regression

equations in equation (3.1) for a given changepoint configuration, and P is the MDL penalty term which penalizes both the number and type of model parameters. Let ψ denote a vector of all the parameters that must be estimated in order to calculate equation (3.1) including the time series parameters in equation (3.2). Then, the expression for the likelihood value can be found using the following equation

$$L_n(\psi; Y_1, \ldots, Y_n) = f(Y_1; \psi) \prod_{t=2}^{n} f(Y_t | Y_{t-1}, ..., Y_1; \psi),$$

which accounts for the time dependent nature of the data. We further clarify this likelihood expression by writing it in terms of prediction errors where $\hat{Y}_t$ is the best one-step ahead linear predictor of Y_t. Then we can say

$$L_n(\psi; Y_1, \ldots, Y_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \left(\frac{1}{r_0 r_1 \cdots r_{n-1}}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{t=1}^{n} \frac{(Y_t - \hat{Y}_t)^2}{r_{t-1}}\right\},$$

where $r_{t-1} = \frac{E[(Y_t - \hat{Y}_t)^2]}{\sigma^2}$. This expression allows us to assess how closely the model can be fit to the data for a given configuration of changepoints. However we want to penalize solutions that contain an excessive number of parameters, unless that additional complexity allows for a better fit to the data. Since the data only contains 240 observations it can only accurately support a limited number of parameter calculations. We mathematically account for this by amending the likelihood function to include the MDL penalty.

To calculate the MDL penalty, we will adjust the penalty amount based on parameter type. Each real valued parameter being estimated from all n observations in the data set is charged a penalty of $\ln(n)/2$. These parameters include the values for the time series portion of the model, the coefficients for the harmonic terms, and

the variance estimate. Next we penalize the parameters that are estimated by the sections of data between subsequent changepoints by $\ln(\tau_j - \tau_{j-1})/2$. Each of the slope and intercept terms receive this penalty. We also want to charge a penalty of $\ln(k+1)$ for the $(k+1)$ regions that our data set is split into by the changepoints. Finally, each changepoint τ_j itself is penalized by $\ln(\tau_{j+1})$. Adding all these penalty components leads to the full MDL penalty:

$$\left(\frac{p+q+P+Q+5}{2}\right)\ln(n) + \sum_{j=1}^{k+1}\ln(\tau_j - \tau_{j-1}) + \ln(k+1) + \sum_{j=2}^{k+1}\ln(\tau_j),$$

with $\tau_{k+1} = n+1$. And, since the penalty value will only change as the number of changepoints (k) and their locations (τ_j) change, all terms that do not rely on k or τ_j may be removed. This means the penalty simplifies to:

$$P = \sum_{j=1}^{k+1}\ln(\tau_j - \tau_{j-1}) + \ln(k+1) + \sum_{j=2}^{k+1}\ln(\tau_j),$$

Which gives us our final fitness function:

$$MDL(k, \tau_1, \ldots, \tau_k) = -2\ln(L_{opt}) + \sum_{t=1}^{k+1}\ln(\tau_j - \tau_{j-1}) + \ln(k+1) + \sum_{j=2}^{k+1}\ln(\tau_j), \quad (3.3)$$

This function will be minimized when calculated using the optimal number and configuration of changepoints, so any possible estimation of changepoints can now objectively be evaluated.

3.3 Changepoint estimation via genetic algorithm

With very small data sets, it may be possible to perform an exhaustive search over the full sample space, testing all possible numbers of changepoints as well as each possible configuration. However, there are $\binom{n}{k}$ different ways to configure the changepoints, which, by the binomial theorem, requires 2^n different evaluations of the fitness function. For this data set where $n = 240$, it clearly is preferable to use an algorithm that can begin by testing different random configurations in the sample space, then iteratively converge towards an optimal solution. Otherwise the number of evaluations required may become computationally unrealistic. For analogous problems that may have many more years worth of data, or may have more frequent data points, clearly an exhaustive search is prohibitive, and an intelligent algorithm must be applied. The algorithm we have selected for this task is the GA.

The GA is not the fastest algorithm used to locate regime changes in data, however, it is the best suited to a situation where some parameters must be estimated by the full data set, while others are only estimated by a small segment of observations. More specifically, parameters such as α_i in equation (3.1), and the time series parameters ϕ, θ, Φ and Θ in equation (3.2) all are estimated by the full data set and cannot accurately be calculated by a small region of data, while the slope and intercept terms are calculated by only the data between changepoints. Common changepoint search algorithms such as Wild Binary Segmentation (WBS) method and Pruned Exact Linear Time (PELT) method are much faster, but they require that the data be fully partitioned $k + 1$ regions. They then optimize the fitness function using only the data within a specific region to compute all parameters [14][15]. They do not allow for the long-term parameters to be computed by the full data set. If our

data set was very large, it is possible that a parameter from the full data could be estimated with reasonable accuracy from a subsection of the data. However, with only 240 observations, such parameters cannot be accurately estimated from a subsection of the data. Fortunately, the GA gives us a way to compute both types of parameters while iteratively optimizing the number of changepoints, their locations, and the region specific slopes and intercept terms until an optimal solution is obtained.

The selected GA requires the number of potential changepoints and their locations to be formatted in a "chromosome" structure. Each chromosome represents a changepoint configuration as $(k; \tau_1, \ldots, \tau_k)$ where k is the number of changepoints and each τ_i is the location of a changepoint in the data set. As we apply this algorithm, we will be able to create subsequent "generations" of chromosomes that exhibit more and more of the characteristics that make them fit for their "environment." Mathematically, this means that the new iterations of chromosomes will have a better chance of optimizing the fitness function in equation (3.3) as they will carry forward the best changepoint configurations from the prior iteration at a higher rate.

The GA process is started by generating L chromosomes. For this data set, $L = 125$. In the first generation, the number of changepoints per chromosome is limited to four, as we do not want to begin the algorithm with an unrealistic number of changepoints for the given 20 year period. With enough iterations, the GA will shift that number to include a larger or smaller number of changepoint locations if the underlying data indicates that this provides a more optimal solution. We also include one chromosome with zero changepoints. Additionally, we ensure that all changepoints are at least six data points away from each other so that all segments contain at least six months worth of data. Without this stipulation, our algorithm may detect false trends that are only a few data points long and do not contain

enough information to support accurate parameter calculation.

Once the initial L chromosomes are generated, the "fitness" of each chromosome must be assessed by evaluating the fitness function in equation (3.3). The result of this calculation is referred to as a chromosomes fitness score. The next generation of chromosomes are calculated in a way that favors more "fit" chromosomes, but still allows some randomness to keep the algorithm from premature convergence. The most fit chromosome is pulled directly into the next generation of chromosomes, and the remaining chromosomes are calculated as follows: two parent chromosomes are selected from the L chromosomes in the initial generation. Let R_i be the rank (based on fitness score) of the ith chromosome where the the rank of the worst scoring chromosome is 1 and the best is rank L. Then the ith chromosome is selected to be the first parent with probability $R_i/(\sum_{j=i}^{L} R_j)$. The second parent is chosen the same way without replacement.

To create a child chromosome from the two parents, let the two parent chromosomes be as follows: $(i; \tau_1, \tau_2, \ldots, \tau_i)$ and $(j; \omega_1, \omega_2, \ldots, \omega_j)$. Then the two parent chromosomes are combined resulting in a child chromosome of $(i + j; k_1, k_2, \ldots, k_{i+j})$ where the ks are all the ordered changepoints contributed by both parents. Changepoint locations that the parents have in common are only represented once in the child. Next, each changepoint is saved or removed with probability of 0.5. This ensures that the number of changepoints in a child chromosome is close to that of the parents. Finally, each changepoint time remaining in the child chromosome shifts upwards one point with a probability of 0.3, downwards one point with a probability of 0.3, and remains unchanged with a probability of 0.4. This allows for more changepoint locations to be tested, and maintains robustness as the algorithm iterates. Additionally, we confirm that the child's changepoint locations are still spaced at least

6 data points away from each other. In GA codes, this process of assembling a member of the next generation from members of the previous generation is called crossover, and it is what allows the algorithm to "intelligently" evolve towards more and more optimal solutions.

Finally, chromosomes may occasionally mutate which helps prevent the algorithm from converging prematurely. Each changepoint for each child mutates with 0.1 probability. When mutation occurs, the selected changepoint time is replaced with a random time from the data set (excluding changepoint times already in the child chromosome, as well as times that are within 6 data points of these values). This mutation rate provided the most stable sets of solutions given this data, and, since each new generation carries over the best chromosome from the previous generation, a high mutation rate will not unnecessarily extend the iterations required for convergence.

The selection, crossover, and mutation process is repeated until L members of the new generation have been created. Members of the generation all must be unique, however, two parent chromosomes could produce multiple children in a given generation as long as the crossover and mutation processes resulted in non-identical offspring.

Once a complete new generation has been created, fitness scores are calculated and another new generation of chromosomes is formed through selection, crossover, and mutation. The process continues until the most fit members of several subsequent generations no longer improve in terms of fitness score. At this point, the GA is terminated, and the most fit member of the final generation is selected as the best changepoint configuration. In this research, the GA was terminated after 100 iterations.

Although this algorithm effectively optimizes the fitness function, it may po-

tentially to converge to multiple optimal solutions due to the randomized values chosen in the algorithm. While we cannot quantify the precise error inherent in the changepoints the GA selects, others who have applied the GA in similar time dependent contexts have found through extensive simulation study that the GA does locate true changepoint locations with a very high degree of accuracy [16]. Specific simulation studies and sensitivity analysis have been conducted using models and simulated data derived from temperature data, sea level data, and snow depth data, among others [12] [17] [18]. Because our data exhibits similar characteristics to the data sets in these studies, specifically autocorrelation and seasonality, we have not conducted a simulation study in this research. We have simply made an effort to test this algorithm with many seed values to ensure robustness.

CHAPTER 4

CASE STUDY: WASHINGTON STATE

In this section, we'll be fitting the theoretical model from Chapter 3 to the data sets for residential, commercial, and industrial retail sales of electricity for Washington state. Each data set has a different pattern to it, and the underlying trends also appear to differ. Once we have used the GA to fit our model to the optimal number of changepoints, we'll then qualitatively examine the changepoint locations to see if they can be substantiated by historical events and trends.

Before we can fit equation (3.1) to the data, we need to select a specific SARMA model that will best fit the time series nature of the residuals in equation (3.2). To determine the best fit and complexity, we tested a variety of simple SARMA options over the WA data sets along with data sets from multiple other states and calculated the corrected Akaike information criterion (AICC). The lower this value or "score," the more appropriate the SARMA model is for the data. The AICC score was specifically considered because it penalizes the number of parameters in a way that is proportional to the number of data points. This data only contains 240 observations, and we want to avoid models where a large number of parameters must be estimated by a proportionately small amount of data when possible. The AICC formula can be found below:

$$AICC = -2\ln L\left(\phi, \theta, \Phi, \Theta, \frac{S(\phi, \theta, \Phi, \Theta)}{n}\right) + \frac{2(p+q+P+Q+1)n}{n-p-q-P-Q-2},$$

where

$$S(\phi, \theta, \Phi, \Theta) = \sum_{t=1}^{n} \frac{(Y_t - \hat{Y}_t)^2}{r_{t-1}},$$
$$r_{t-1} = \frac{1}{\theta^2} diag\{v_0, v_1, ...v_{n-1}\},$$

and v_i is the diagonal entry of the $Cov(Y_t - \hat{Y}_t)$ matrix and L is the likelihood of equation (3.1). Using the AICC scores, we selected the SARMA$(1,1) \times (1,0)_{12}$ model to account for the time dependency in each data set. While certain other SARMA models may provide slightly lower AICC scores on single specific data sets, the SARMA$(1,1) \times (1,0)_{12}$ model offers greater flexibility as it fits well to a wide variety of data sets, and we want to make use of the same model for the data from the remaining 47 states. As this research does not focus solely on a single data set, a simple model that can work well with many different data configurations is an optimal choice.

The SARMA$(1,1) \times (1,0)_{12}$ model gives us the following representation for the model errors ϵ_t in equation (3.1)

$$\epsilon_t = (1-\phi B)^{-1}(1-\Phi B^{12})^{-1}(1+\theta B)Z_t,$$

where Z_t are the final uncorrelated errors that have mean zero and standard deviation σ^2. The parameters ϕ and θ are associated with the AR(1) and MA(1) parts of the model, and Φ is associated with the seasonal AR(1) part of the model. Each parameter is estimated from all n observations, but their specific values will change depending

on the changepoint configuration being tested.

Firstly, we will examine the data from Washington's residential sector. This data has been plotted in Figure 4.1, with the year on the x-axis and the amount of electricity in million kWh on the y-axis. The top graph shows the slope and intercept that would be fitted without any changepoint consideration. As we do not see any large jumps or dips. we cannot visually estimate likely changepoint locations, however they may still be present in the data.

When we run the residential data through the GA, the changepoint configuration that optimizes the fitness function is (4; 101, 160, 192, 205), which corresponds to May 2009, April 2014, December 2016, and January 2018. These locations are represented by the blue vertical lines in Figure 4.1.

Most of these changepoints are justified when qualitative and quantitative data of the time period are considered. The first changepoint in May of 2009 coincides with a period of reduced population growth in Washington state as the great recession impacted population growth. From 2001 to 2008, the growth rate of the state was between 1- 2 percent each year, but in 2009 the rate dropped to just under 1 percent, and from 2010 to 2013, the population growth rate in the state was around half of what it had been in the previous decade [19]. Visually we can see that, while the trend in residential electricity sales still increased after this time, the rate of growth has slowed and the intercept value has shifted downwards.

The next changepoint, located in April of 2014, indicates the beginning of a three year period where Washington temperatures were warmer than they had been for decades, and, in heavily populated regions such as Seattle, they were the warmest ever on record [20]. Residential electricity consumption in the more populated west side of the state is very correlated to temperature [7]. Thus it makes sense that, during

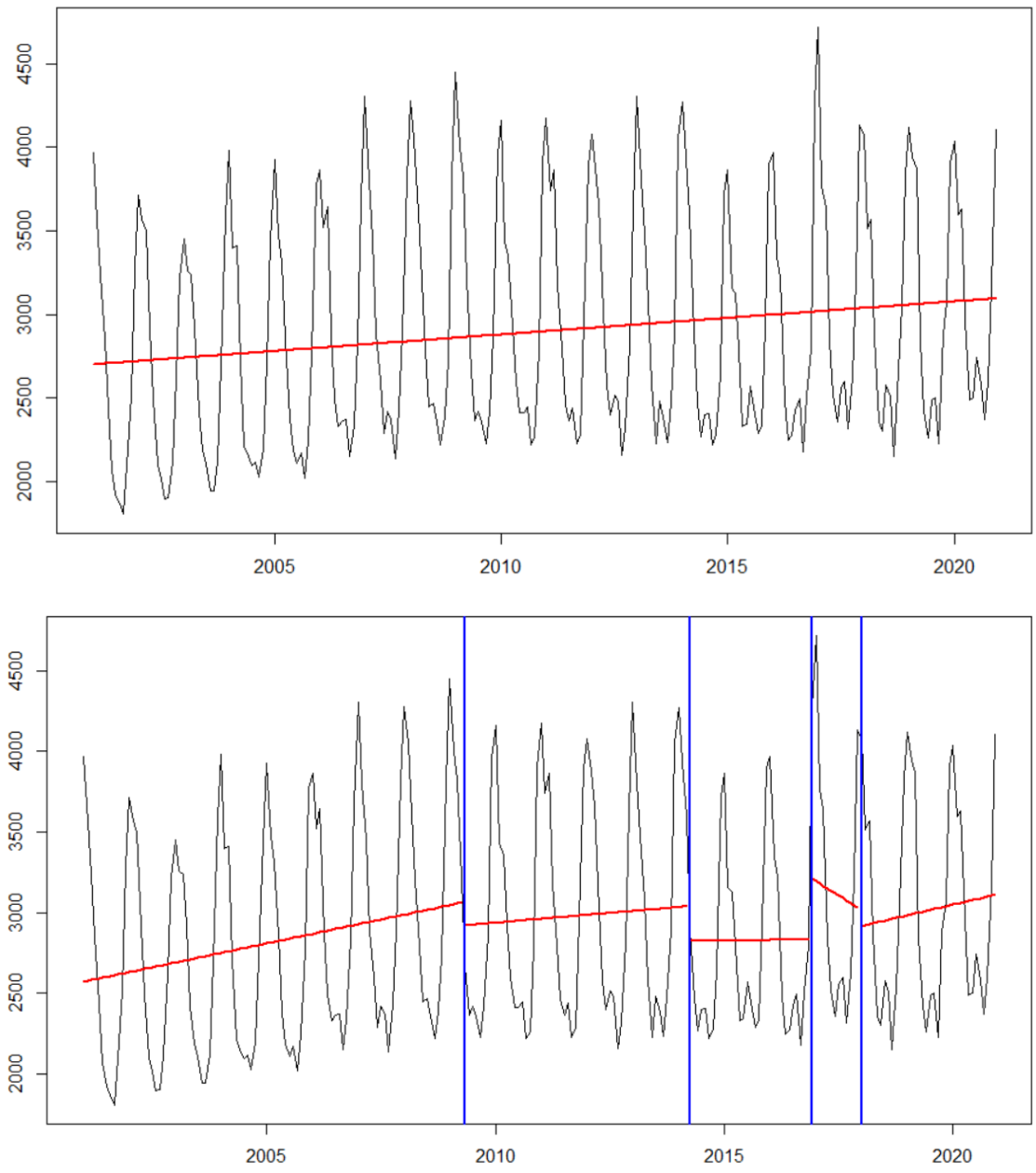


Figure 4.1: Washington residential electricity sales with and without changepoints

these warmer years of 2014, 2015, and 2016, electricity consumption fell, leading to a shift down in intercept and reduction in slope of the data across this period. This

warm period was followed by a colder than average winter across the Pacific Northwest during the winter of 2017, which coincides with the next changepoint in December of 2017 [21].[1] Since this spike, winter temperatures have remained within expected lows and highs, and the final changepoint in April of 2018 indicates the beginning several years where temperatures fell within typical averages.

While the changepoints located appear to have justification, to fully understand how well the model is working mathematically, we next want to examine some model diagnostic figures. In Figure 4.2 we can see that the final residuals almost follow a normal distribution. While the slightly heavy center and tails of the histogram and Quantile-Quantile (QQ) plot keep the the fit from being perfect, it is reasonably close. Additionally, the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots in Figure 4.3 show that almost all significant autocorrelation has been removed from the data, leaving uncorrelated residuals. Further diagnostic values for tests of correlation and normality can be found in the appendix.

The next data set we'll examine comes from Washington's commercial electricity sales. This data contains severe jumps and dips that clearly cannot be captured well by the single linear slope and intercept model in the top graph in Figure 4.4. The optimal changepoint configuration for this data set determined by the GA is

[1] Though not mentioned in any data documentation, these temperature extremes were regularly noted in local news at the time. For example, articles from KOMO news each year remarked on the unusual warmth in 2014, 2015, and 2016. Such articles can be found at the following sites:
https://komonews.com/weather/scotts-weather-blog/2014-weather-review-seattle-had-warmest-year-in-decades
https://komonews.com/weather/scotts-weather-blog/2015-weather-year-in-review-seattle-smashes-all-time-hottest-year-records
https://komonews.com/weather/scotts-weather-blog/2016-seattle-weather-another-toasty-year-but-cold-enough-to-keep-a-bug-from-being-eaten

Temperature trends during this time can be visually compared using climate data available at https://cefa.dri.edu/Westmap/Westmap_home.php?page=timeplot.php

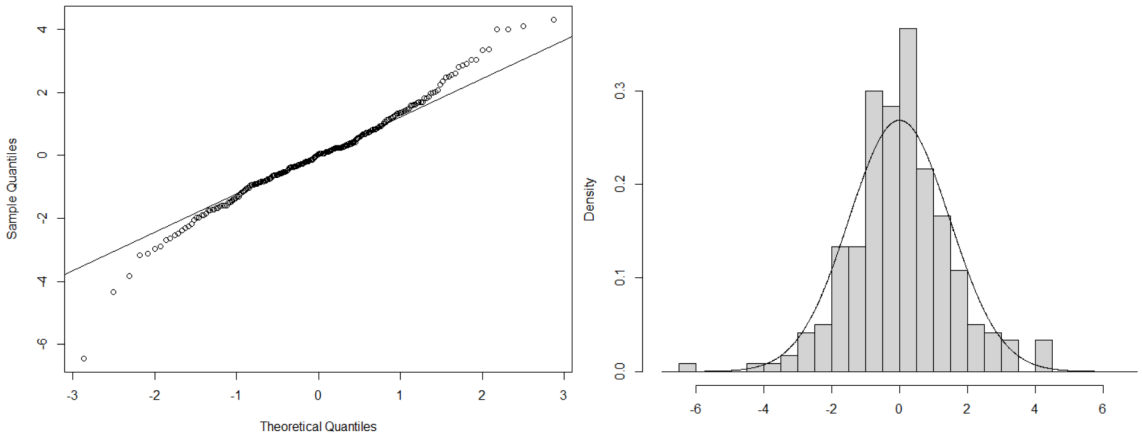


Figure 4.2: QQ plot and histogram for residential residuals

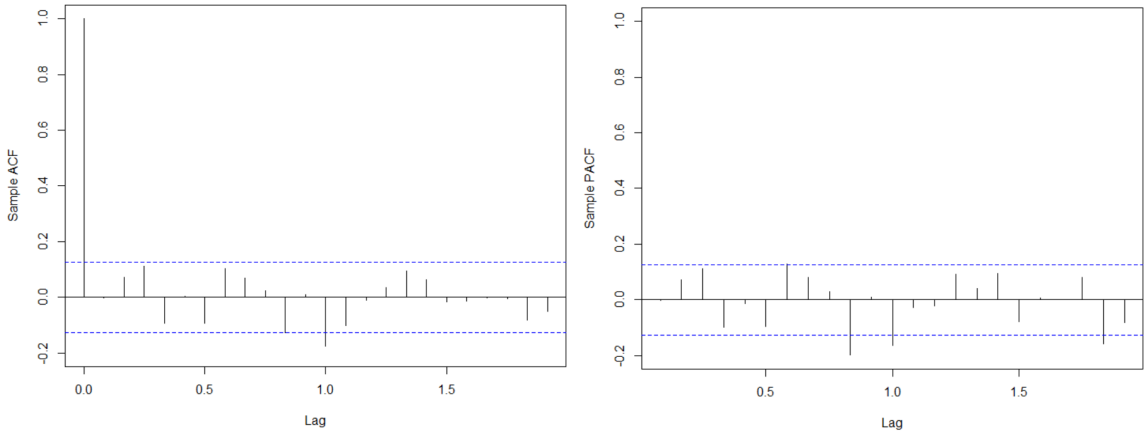


Figure 4.3: ACF and PACF for residential residuals

(4; 25, 54, 104, 230) corresponding to January of 2003, June of 2005, August 2009, and February 2020, and it is clear immediately from the graph that this changepoint configuration allows slopes and intercepts to be fitted much more closely to the data, thus greatly reducing the error of the model.

The initial changepoint in January of 2003 can be explained by a reporting change on EIA form 861M which collects this data from utilities. Previously separate categories were adjusted and aggregated into the commercial and industrial sector [10]. While many states already were reporting their sales in this way, Washington

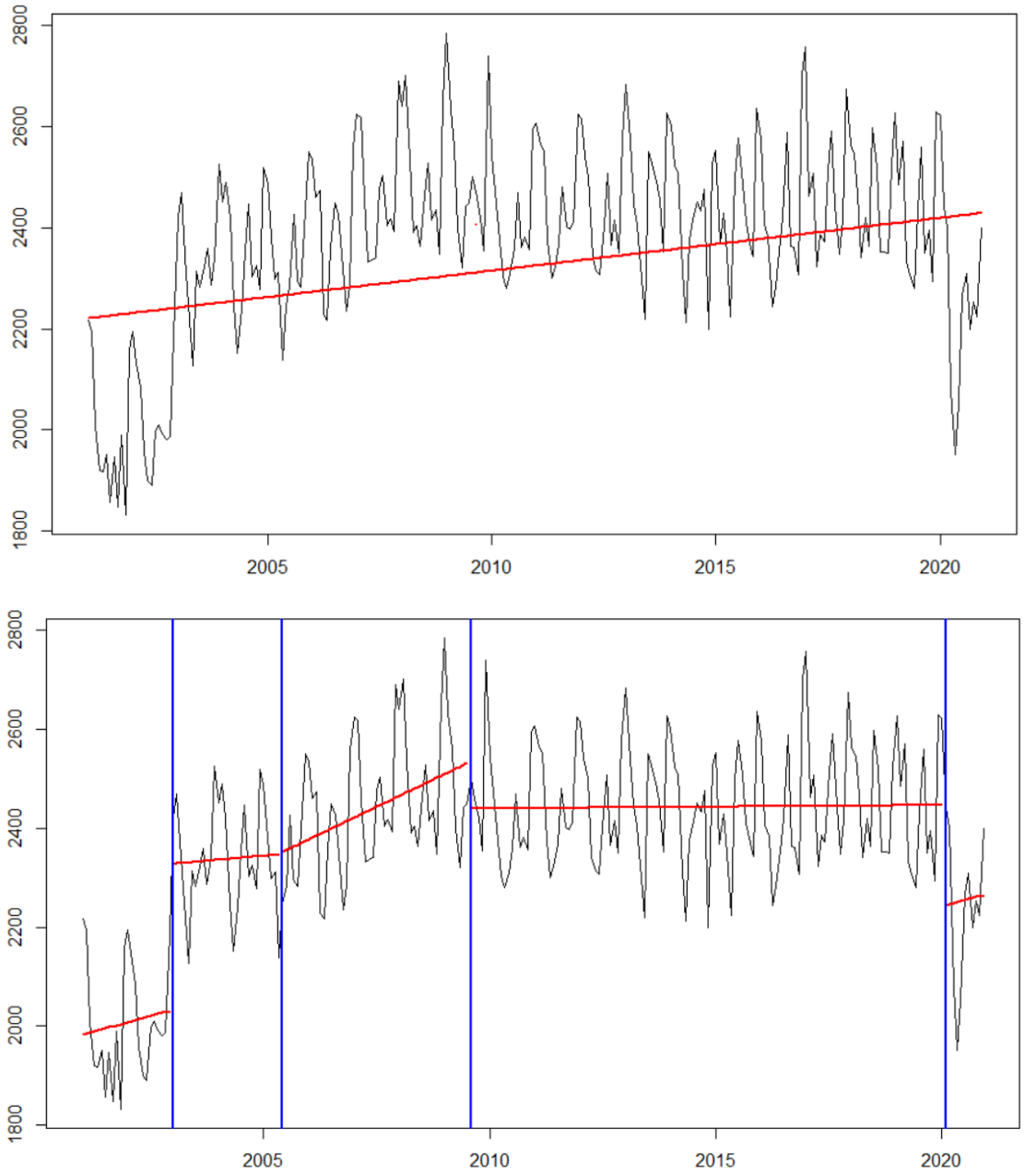


Figure 4.4: Washington commercial electricity sales with and without changepoints

was not, which lead to the jump in the data after this first changepoint. After this categorization shift, the next changepoint location is in June of 2005, and we can

see that the slope of the model after this changepoint grows steeper. Technology based businesses such as Amazon and Microsoft were expanding rapidly during this time, specifically with Amazon focusing on building and developing its corporate headquarter complex leading to large amounts of commercial growth in Seattle.[2]

The changepoint in August 2009 signals the end of this strong increase in retail electricity sales to commercial customers. Following this changepoint, electricity sales to commercial customers stagnated. Through the financial recession and subsequent economic recover during this time, commercial consumption of electricity did not grow. Notably, during the period, the Seattle region shifted more and more towards online based, technology focused businesses, and, across the US this trend seemed typical.[3]

The final changepoint location is in February of 2020, which, again closely coincides with the pandemic related shut downs which completely or partially closed commercial businesses for large portions of 2020. Following the initial drop in February, we can see that the remaining months of 2020 show a slow upward trend in electricity sales as businesses begin partial reopening. However, commercial sales of electricity did not reach their pre-pandemic level through the end of 2020.

When we examine these changepoints from a mathematical point of view, they appear to do an excellent job of fitting the model to the data. In Figure 4.5, the final residuals fit a normal distribution almost perfectly, and Figure 4.6 shows that

[2]With Amazon's massive growth came large scale developments in previously sparsely populated regions of Seattle. Such commercial growth is outlined here: https://www.seattletimes.com/business/amazon/ten-years-ago-amazon-changed-seattle-announcing-its-move-to-south-lake-union/

[3]Across the united states, the rate of growth for commerce based business grew much more than traditional brick and motor stores: https://www.digitalcommerce360.com/article/e-commerce-sales-retail-sales-ten-year-review/

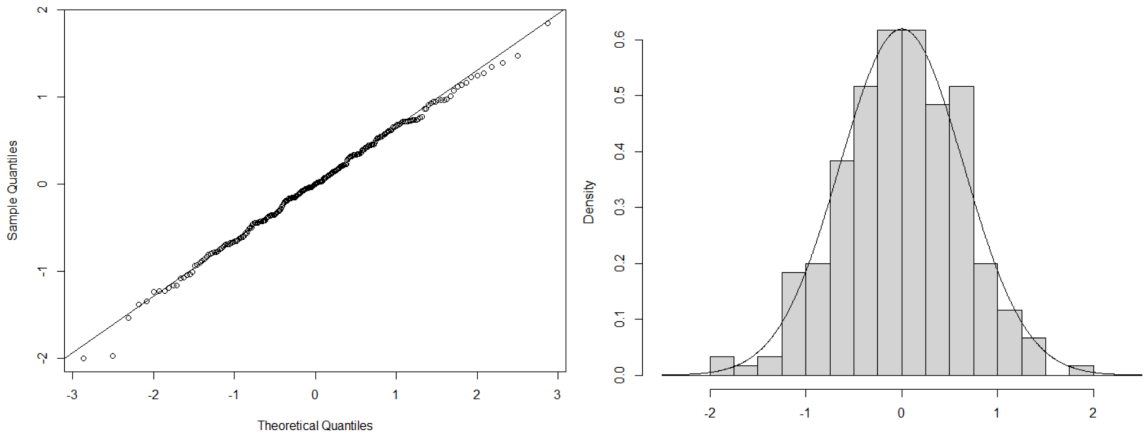


Figure 4.5: QQ plot and histogram for commercial residuals

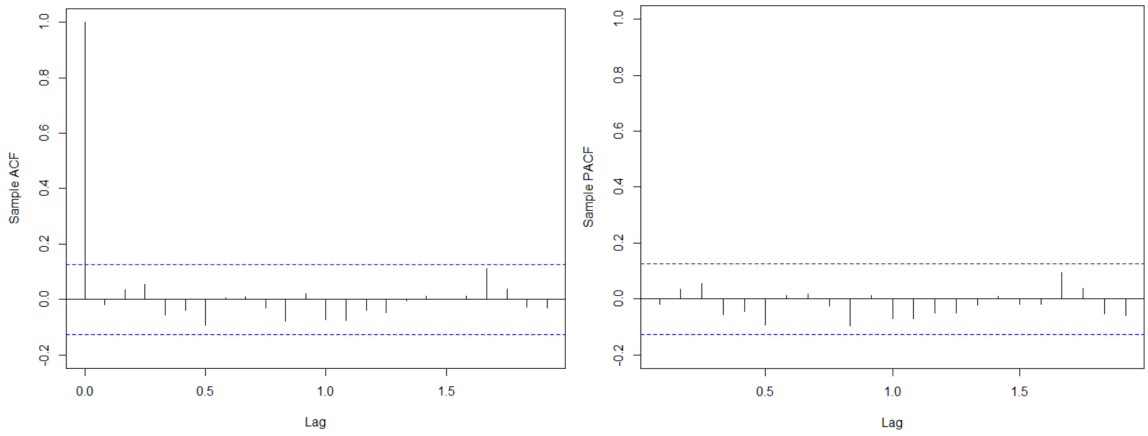


Figure 4.6: ACF and PACF for commercial residuals

the time series portion of the model has completely removed all significant correlation from the data. The residuals are normally distributed and uncorrelated, as desired.

The final case study data set is from Washington's industrial sector. This sector is dominated by manufacturing in the aerospace and transportation segments due in large part to Boeing's production plants. Other top industries include agriculture and, historically, aluminum smelting [22]. In Figure 4.7 we can see the retail electricity sales to the industrial sector 2001 through 2020. The data appears to have several spikes, plunges and regions with both positive and negative slopes. The top graph in

Figure 4.7 shows the best fit of the model slope and intercept from equation (3.1) if no changepoint values are included. Clearly this single slope does not fit well to any section of the data and underscores that changepoints must be considered in order to fit a reasonable model.

The optimal changepoint configuration for this data according to the GA was (5; 7, 56, 62, 154, 232) which corresponds to July 2001, August 2005, February 2006, October 2013, and April 2020.

The very first segment of data falls extremely steeply until it hits the first changepoint in July 2001. The main driving factor behind this drop seems to be the abrupt closure of almost all aluminum smelting plants within the state by the end of 2001. This primarily was driven by the energy crisis during the early 2000s, as the electricity required was too expensive for smelting to be profitable [23]. Coupled with the September 11 terrorist attacks impacting the airline industry (and subsequently, Boeings airplane manufacturing), and the plunge in electricity use seems well substantiated.[4]

Following this dip, an increase in trend can be seen, as economic pressures on the industrial sector began to ease. The industrial sector grew across the country, as evidenced by industrial indices such as the Industrial Production Index [24].

The next two changepoints are located in August 2005 and February 2006. These contain a region in the data that has a steep spike. Looking at the source data, this spike is caused by a single extremely large value in January of 2006. No widely noted

[4]Beoing's own annual report from the following year details some of the challenges faced during 2001 and the start of 2002: https://www.annualreports.com/HostedData/AnnualReportArchive/b/NYSE_BA_2002.pdf

Additionally, US markets dropped in the wake of the September 11 terrorist attacks indicating a challenging economic period of time: https://money.cnn.com/2001/12/17/markets/markets_review/index.htm

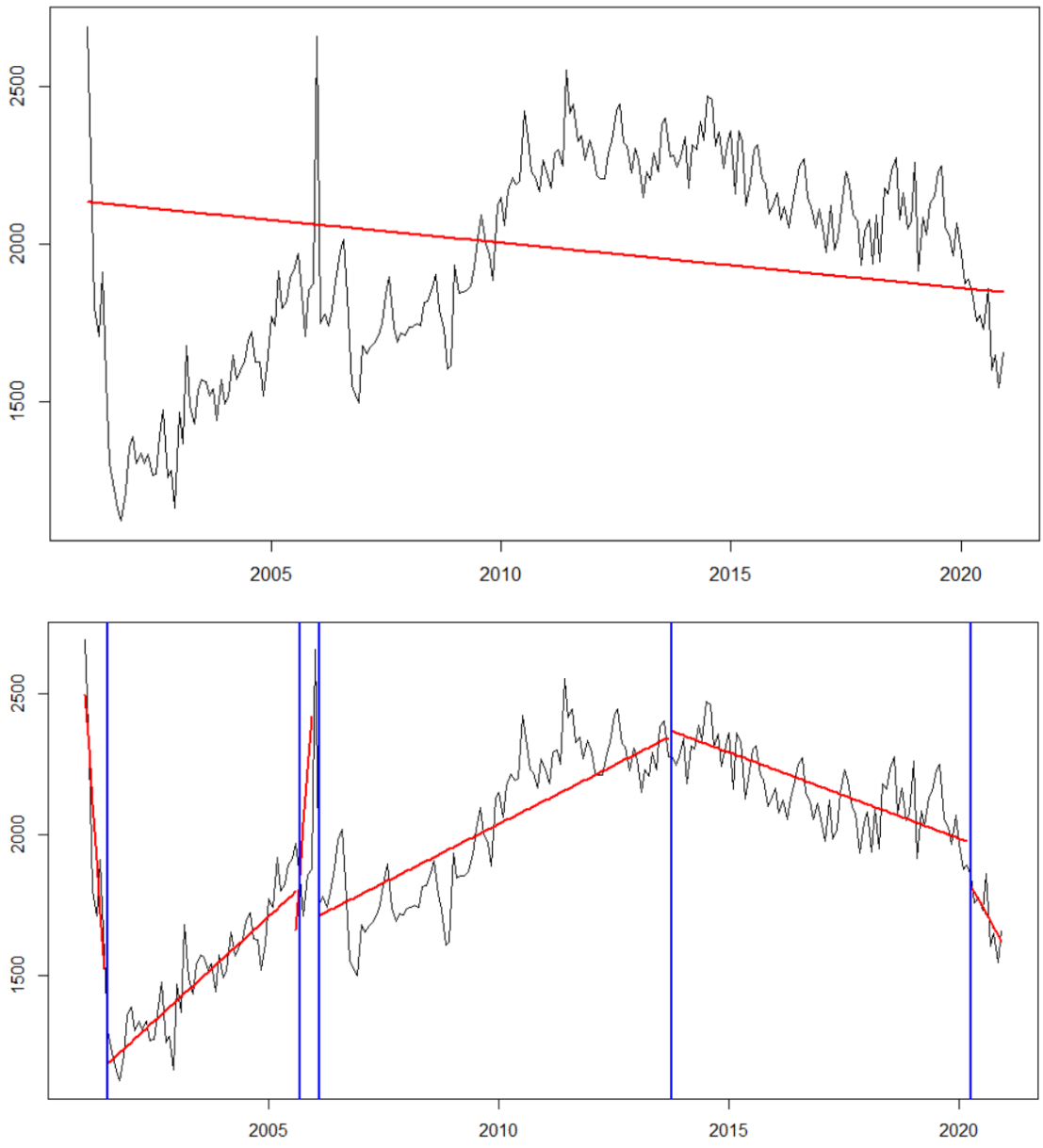


Figure 4.7: Washington industrial electricity sales with and without changepoints

regional event seems to account for this outlier, so it likely is an erroneously entered value which the GA was able to detect and contain to its own small region.

The slope after 2006 continues to rise at a rate minimally impacted by the financial

recession. Partially this could be due to the energy intensive server farms that were built and brought online during this time, increasing electricity consumption even as other industrial businesses closed.[5]

By October of 2013 (the location of the next changepoint), manufacturing as a whole in Washington markedly decreased which appears to be mirrored in the decreasing trend of electricity consumption in this segment [22]. Generally speaking, industrial production across the United States was in a slow decline through out this period as well.[6] Additionally, the aluminum smelting industry, which had only been operating at a minuscule percentage of its capacity from previous years, fully closed one of its two remaining plants in Wenatchee Washington as low aluminium prices in China continued to reduce profitability. [7]

Finally, the last changepoint is located in April of 2020, a month after full lockdown measures were imposed. This final slope shows steep decline through the end of the year as many industrial businesses were still heavily impacted by lockdowns and a struggling economy.

Mathematically, the model satisfactorily fits the data by use of these changepoints. The diagnostic figures in Figure 4.8 satisfy our expectation that residuals are normally distributed, as desired. In Figure 4.9 we can also see from the ACF and PACF that all significant correlation in the data has been accounted for by the time series model, leaving uncorrelated residuals.

---

[5]Server farm's vast electricity requirements were noted in articles such as this: https://www.nytimes.com/2012/09/24/technology/data-centers-in-rural-washington-state-gobble-power.html

[6]US based manufacturing information is outlined here: https://tradingeconomics.com/united-states/industrial-production

[7]For details regarding aluminum pricing during this time, see: https://www.indexmundi.com/commodities/?commodity=aluminum&months=240

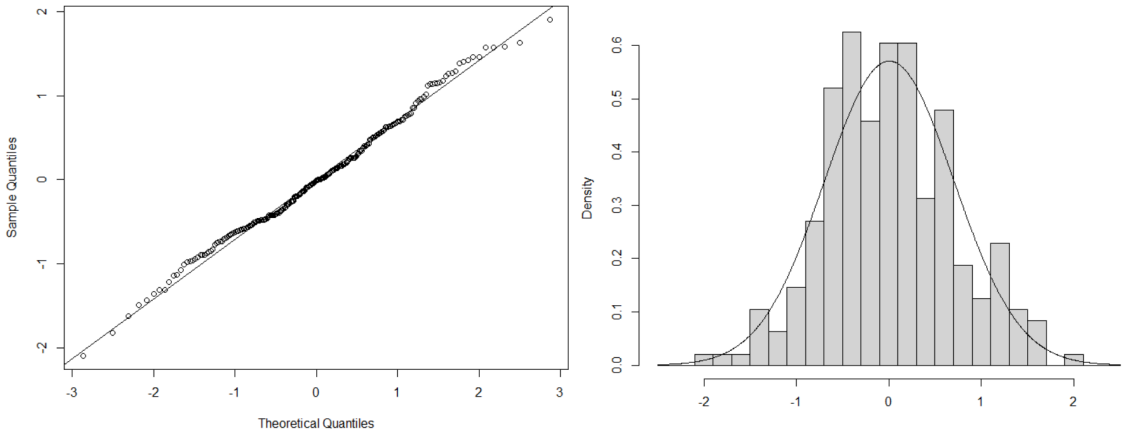


Figure 4.8: QQ plot and histogram of industrial residuals

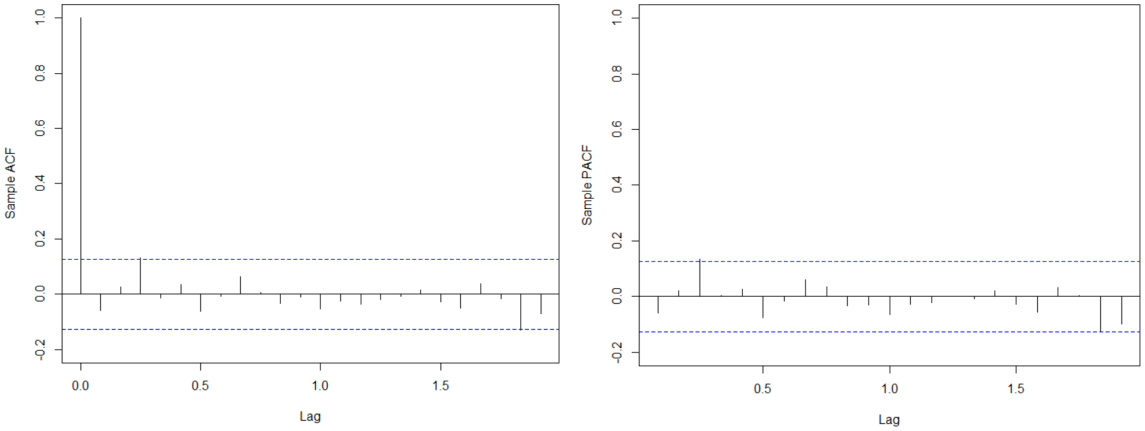


Figure 4.9: ACF and PACF for industrial residuals

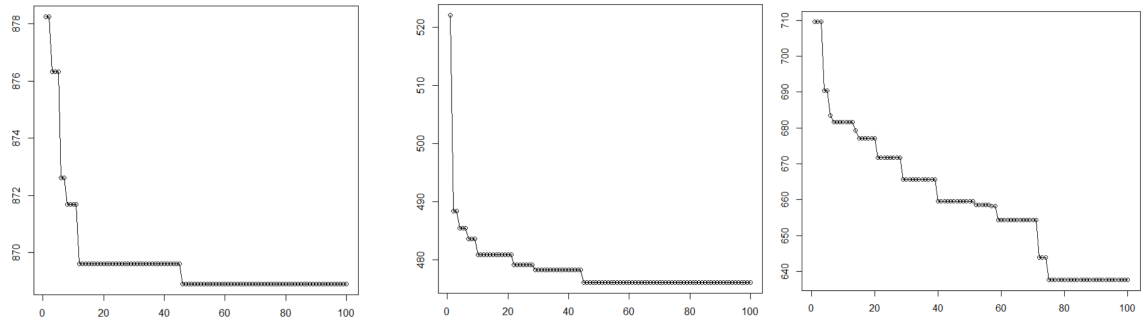


Figure 4.10: Rate of convergence for residential (left), commercial (center), and industrial (right) data sets

Overall, both by qualitative evidence as well as mathematical diagnostics, the algorithm appears to locate viable changepoints for each data set, and the final check is to determine if the algorithm runs for enough iterations to reach convergence. As stated in the methods section, the GA is terminated after 100 iteration. At this point, the best changepoint configuration is selected as the optimal solution. To visually confirm that the algorithm converged within this iteration limit, we plotted best fitness score vs. iteration, see Figure 4.10. Multiple iterations transpire without an improvement in fitness score for each data set, thus the GA appears to easily obtain a minimizing solution within the 100 iterations.

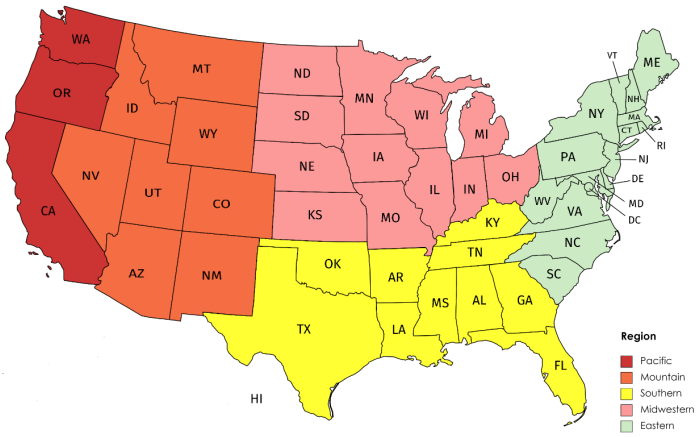
From these graphs and diagnostics, we conclude that, for the data sets from Washington state, the GA converged to changepoints that could be substantiated by local news and economic events, as well as by mathematical diagnostics. Because the model appears to be performing satisfactorily, we now will proceed to apply it to the remaining data sets from other states.

CHAPTER 5

RESULTS

When all the data sets were run through the GA, we found that, on average, the residential data sets had 2.8 changepoints, the commercial data sets had 4.8 changepoints, and the industrial data sets had 5.6 changepoints for the 20 year period. In order to more clearly understand if changepoints were more common in certain regions or during certain years, we divided the state data sets into five categories: Pacific, Mountain, Southern, Midwestern, and Eastern states (see Figure 5.1). The states contained in each of these divisions are listed in the appendix, and were chosen by combining geographically adjacent regions that were originally categorized by the EIA. These categories should allow us to better understand if trends impact the electricity sales in specific regions and sectors as a whole.

As noted, changepoint occurrence was most infrequent in the residential sector, as can clearly be seen in Figure 5.1. Interestingly, there are few regions that contain a notable number of changepoints during the 2020 pandemic outside the East Coast states. Rather, the financial crisis during 2008 and 2009 corresponds to many more shifts in the residential consumption trends in all regions, as indicated by the number of changepoints that occur during these years. It also appears that the Midwestern and Southern states have relatively fewer changepoints overall, and more years where no changepoint occurred, leading to the conclusion that the residential electricity



MAP 5.1: United States Regions

consumption trend is the most stable in these regions. The Eastern states only have three years where no changepoint was detected in any state in the region, so, additionally, the overall trends in the East Coast exhibit the most frequent changes in consumption behavior.

In Figure 5.2, the number of commercial sector changepoints is shown by regional histograms. From these histograms it appears that there are more changepoints in the early 2000s, the late 20-teens and 2020. However, different regions' commercial electricity consumption trends do have slightly different changepoint distributions. For example, in the Mountain and Eastern states, we see many changepoints during the years of the great recession as well as the 2020 pandemic. In Southern, Midwestern, and Pacific states, while we see fewer changepoints during the recession years, the pandemic coincides with more changepoints than any other historical events in our given data. Additionally, in the Eastern states, the GA determined that at least one state had a changepoint each year, while the Pacific states had five years where no changepoint was located for any state in the region. While there are many more

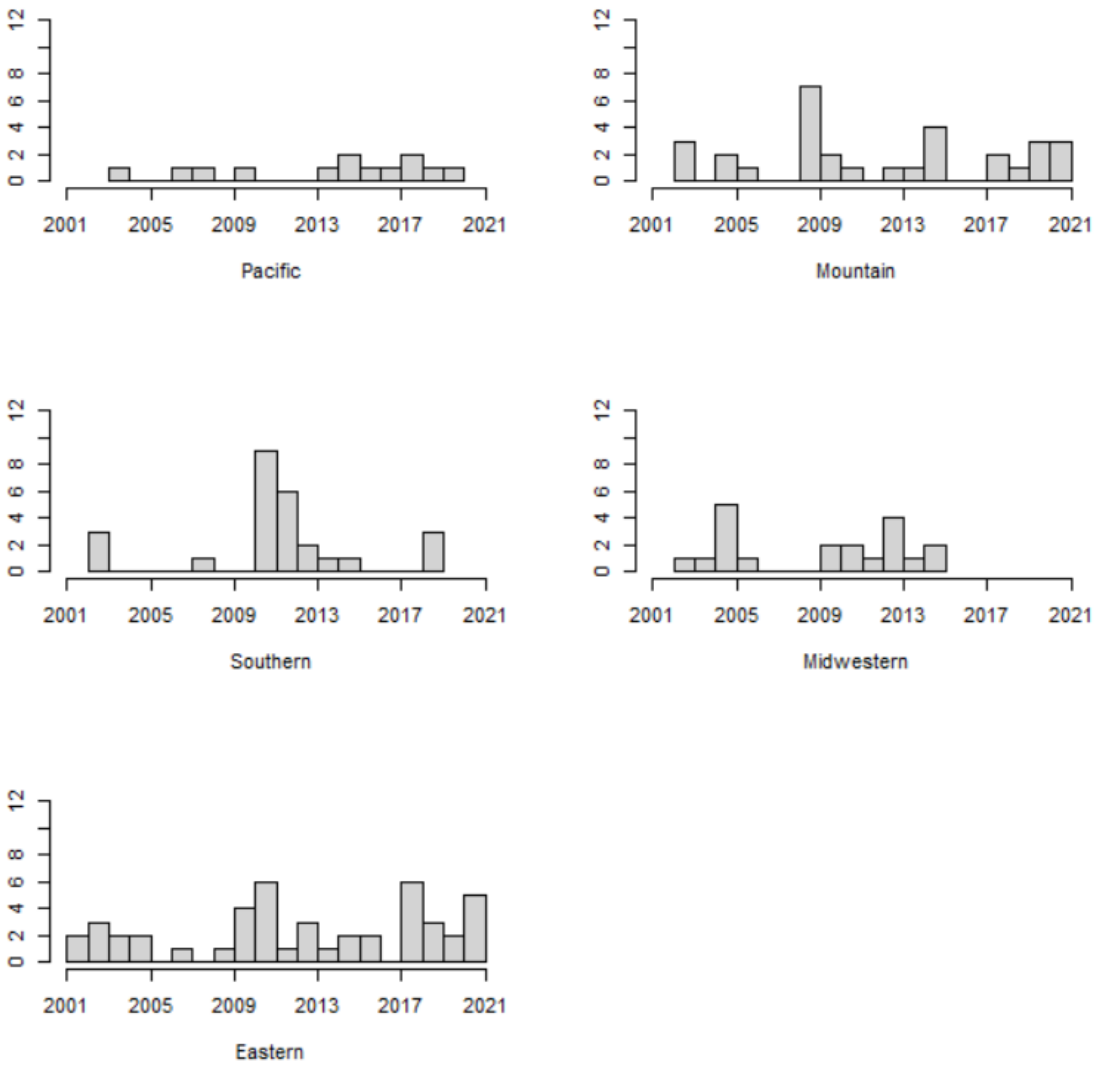


Figure 5.1: Histograms of residential changepoints by region

states in the Eastern region than the Pacific region, it still appears that the models that best fit the Pacific states' electricity consumption require less changepoints and therefore have fewer adjustment in their consumption trends.

Finally, we also can see histograms of the industrial data sets' changepoints separated by region in Figure 5.3. This sector contains the most changepoints per data set, and this increase appears to be equally spread out across all regions. Also,

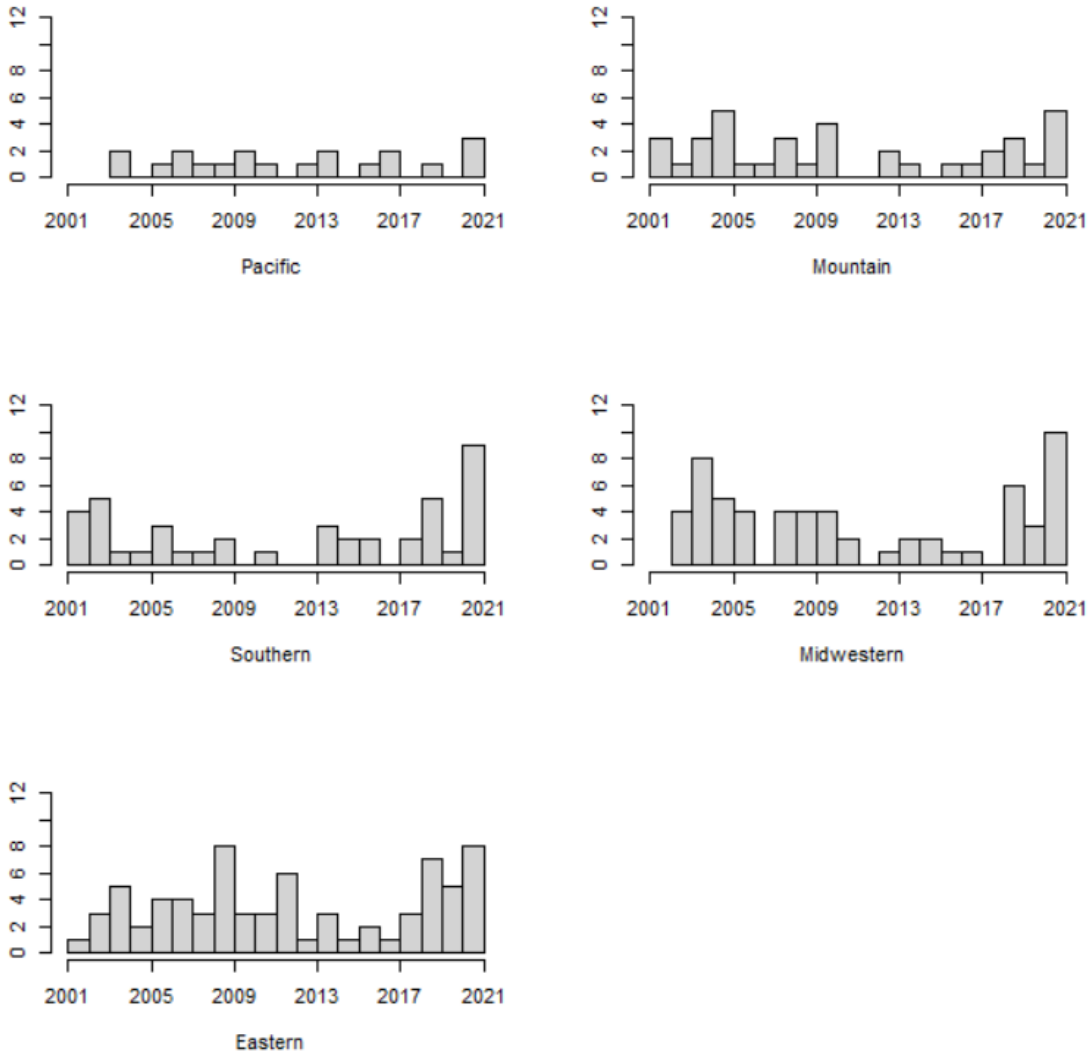


Figure 5.2: Histograms of commercial changepoints by region

there are more changepoints occurring during the years of the great recession than during the 2020 pandemic. With the exception of the Pacific states, each region has at least seven changepoints during the recession, and at least four during the pandemic. Regions such as the Mountain, Western, and Eastern states see a large proportion of their changepoints clustered around 2008, 2009, and 2010–years when recession impacts were the most heavily felt. We also see that, once again, the Eastern states

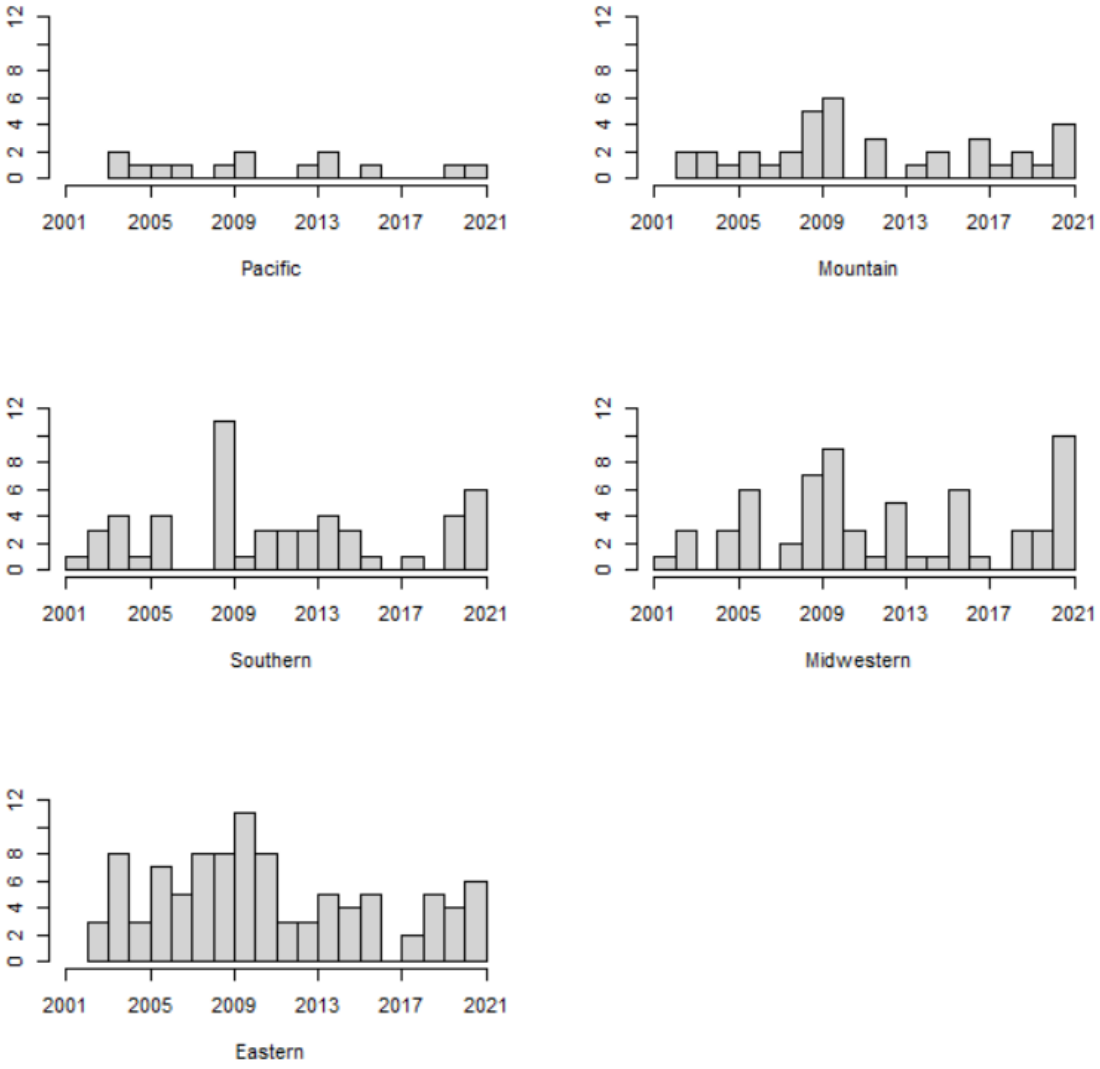


Figure 5.3: Histograms of industrial changepoints by region

have a proportionally large number of changepoints.

From all three histograms we conclude that the GA located changepoints in the data where we expected to see many shifts to the model (specifically in the years of the recession and the pandemic). However, we also have seen through case study, that the GA process easily identified changes in trends at the state level when factors specific to that state alone may cause additional changepoints. Although we do not see any

extremely surprising trends resulting from aggregating the changepoints by region, this analysis suggests that the GA is doing an effective job selecting changepoints that fit an accurate model consistent with economic events.

Changepoints allow us to fit separate slopes and intercepts to the region of data before and after the changepoint. If changepoints across the country are occurring very frequently at certain times, the change in parameter value may offer insight into the situations behind the data. As of writing, the Covid-19 pandemic is ongoing, and has had international impact on day to day life as well as electrify consumption patterns. Since we are working with electricity data and the utility industry is typically concerned with estimates of rates of growth, we will next analyze the the slopes of the data from states where electricity consumption changed with the onset of the pandemic. Because both the slope and intercept parameters are allowed to vary for each segment between changepoints, a region where the models slope changes from positive to negative does not mean that less electricity is necessarily being consumed. If the region with a negative slope is accompanied by an upward shift intercept, the actual amount of electricity consumed may be similar the amount in the prior section of the model. The trajectory of the trend may have adjusted but the real amount of consumption may not be correspondingly larger or smaller.

We began by selecting all residential, commercial, and industrial data sets that had a changepoint near or during the pandemic. Specifically, we define a changepoint occurring within six months of March 2020 as a "pandemic" changepoint. Within each sector, there are 48 data sets. In the residential sector, only nine states had a pandemic changepoint. However, in the commercial data sets, 36 states had a changepoint in the pandemic range, and, in the industrial sector, 31 states had a 2020 pandemic changepoint. For maps of the states impacted in each sector, see the

appendix.

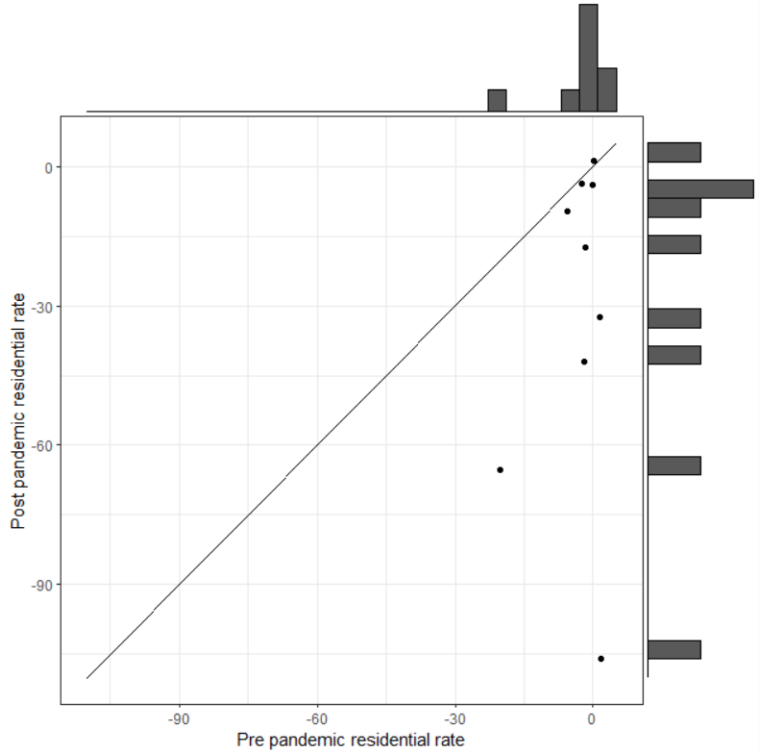


Figure 5.4: Residential pandemic related rates of change

In Figure 5.4 we have graphed the slope of the data prior to the last pandemic related changepoint on the x-axis and the slope after that changepoint on the y-axis. Points that fall on the diagonal line indicate no change in rate before and after the pandemic changepoint, so points along this line signal a shift in the slope parameter only. We can see from the x-axis that the pre-pandemic rate of change in most residential states was close to zero for all states impacted. From the y-axis it is apparent that the slope in the final segment of data had a much greater spread and demonstrate a decreasing rate of change overall. A small number of data points fall very close to the diagonal line, so some states consumption trend shifted during the pandemic in the intercept term of the model only. This does not necessarily mean

that less electricity is being consumed, but the change in slope does signal an altered trajectory for the future.

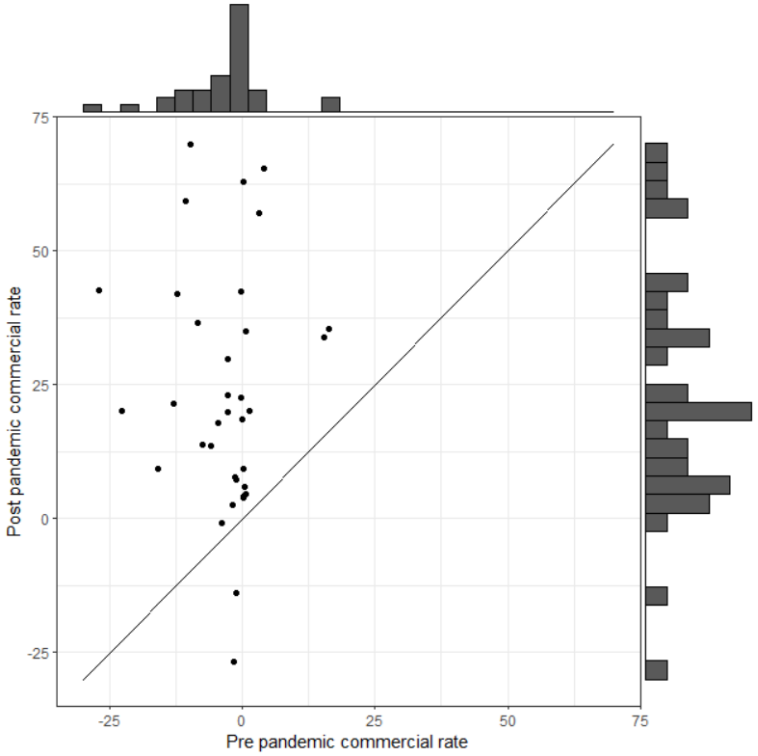


Figure 5.5: Commercial pandemic related rates of change

In Figure 5.5 we see that the commercial data sets had many more changepoints during the pandemic (likely due to the partial or complete closure of commercial businesses while strict lockdowns were in place). Once again, from the x-axis we see that pre-pandemic rates of change are mainly clustered around zero, and, from the histogram at the top of the figure, we determine these slopes are slightly skewed toward a negative rate of change. The slopes following the pandemic changepoint have a greater spread in their rates of increase or decrease. The majority of the rates fall above the diagonal line, indicating that overall, rates are increasing more quickly than they were in the pre-pandemic segment of data. Once again, increases in slope

do not indicate that more electricity is being consumed by the commercial sector than before the onset of the pandemic. Rather, it is likely that the model places the intercept term quite low in the final segment of data while the overall positive slopes for this last segment indicate recovery and upward trend in commercial electricity consumption.

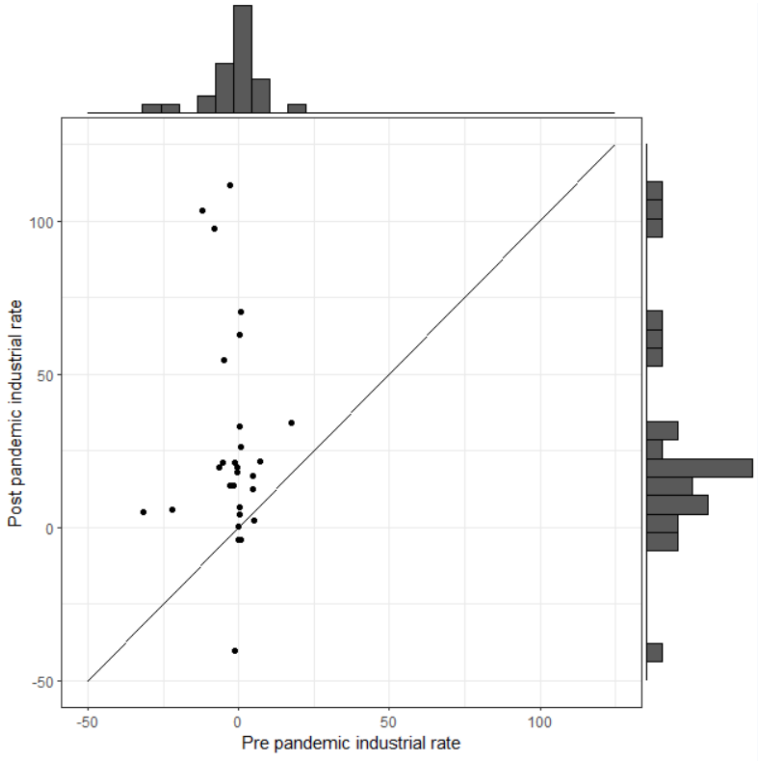


Figure 5.6: Industrial pandemic related rates of change

Finally, we have a similar plot for the industrial sector in Figure 5.6. At first glance, this figure appears quite similar to the corresponding figure for the commercial sector. The initial slopes on the x-axis are clustered around zero and slightly negatively skewed. Additionally, the majority of post pandemic slopes fall above the diagonal line on the y-axis. In the industrial sector, however, we see that the slopes in the final segment of the model are sill near zero, primarily falling between a rate

of zero and 25, and have far less spread than the post pandemic slopes in figure Figure 5.5. This indicates that, while the rate of change in impacted states is overall positive, there is less of a difference state to state in the industrial slopes compared to the commercial sector slopes in the final segment of the model.

From these graphs, we conclude that consumption in the industrial and commercial sector has a positive rate of change in most states, however, the actual rate of recovery is slow for most states' industrial sectors and varies widely for the commercial sector. Finally, the typically stable residential sector was minorly impacted, and in places where these impacts were notable, they signal that consumption is decreasing in trajectory.

CHAPTER 6

CONCLUSION

Clearly, changepoint detection is a valuable tool that allows many different trends in the data to be accurately incorporated into a model. We have shown this specifically in the case of electricity consumption data from Washington state, and additionally have analyzed data from the continental United States for changepoints. We have demonstrated that firstly, across the United States as a whole, the onset of the 2020 pandemic has coincided with changes in electricity consumption trend in the industrial and commercial sectors for most states, but has not impacted many states' residential sector. The GA's detection of the pandemic created changepoints in the US shows that the algorithm is capable of detecting shifts to consumption that are caused by both wide sweeping economic events, as well as events that may only impact a specific state or small region of the country as shown in analysis of its performance on Washington's data sets. Changepoints also allow regression models to be fitted to data much more appropriately, and they greatly decrease the error of models making them more useful and accurate.

Such results and analysis can be utilized in the electricity industry in many ways. Changepoint detection and slope categorization techniques can offer insight into the effectiveness of policies designed to create new consumption patterns. Typically, these policies are paired with incentives to assist in reaching these goals. It is challenging to

determine the effectiveness of financial incentives such as tax credits for energy rated buildings, or social campaigns to encourage residential consumers to change electricity habits. These models and the GA can detect if these policies have successfully shifted consumer behavior by fundamentally changing the underlying consumption trend, leading to a changepoint. If an incentive does not correspond to a changepoint in a related data set, further action is likely required to achieve the desired economic outcome and shift consumption.

Similarly, this process can be applied to other energy scenarios. For example, many countries plan reduce dependency on certain fuel sources (oil and coal for example) over future decades. Tax incentives are already in place to encourage utilities to sell more electricity from renewable resources. As incentives are offered to aid in reaching these goals, changepoint detection techniques can assess if incentives and policies are having the desired impact on production behavior, and offers an objective way to categorise states, cities, or regions into groups that have made progress towards these goals by shifting their supply or demand patterns and groups that have not. This can better aid understanding the incentive types and levels that are required for specific populations to adjust their supply and demand patterns towards more desirable sources.

Finally, utilities themselves can use this model and apply it their service region. If these regions are divided by consumer ZIP code or county, this model can help determine areas of more rapid growth and areas that are undergoing economic contraction. This model type allows long term time series trend calculation for certain parameters that are not typically incorporated into current models due to concerns about including outdated linear trends. In turn, this leads to lower error estimates which allows for more accurate planning.

Overall changepoint analysis and detection has uses in many areas, and is an important tool for creating accurate models across many industries and data sets. When model accuracy is paramount, these techniques provide ways to closely fit a model, even when underlying trends may be complex.

REFERENCES

[1] US Energy Information Agency. Electricity data browser, 2021.

[2] N. McClean Kahn, S. S. Zhang, and C. Nugent. Optimal parameter exploration for online change-point detection in activity monitoring using genetic algorithms. *Naval Research Logistics Quarterly*, 16:e1784, 2016.

[3] S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51:339–367, 2017.

[4] C. Truong, L. Oudre, and N. Vayatis. A selective review of offline change point detection methods. *Signal Processing*, 167:e107299, 2020.

[5] J. Holland. Genetic algorithms and adaptation. *NATO Conference Series (II Systems Science)*, 16:317–333, 1984.

[6] G. Givens and J.A. Hoeting. *Computational Statistics, 2rd edn.* New Jersey: John Wiley and Sons, 2013.

[7] L. Molander. Load forecasting & analysis manager, puget sound energy. Personal Interview. 2020-10-15.

[8] M. Marcelia. Controller & tax director, puget sound energy. Personal Interview. 2021-03-03.

[9] US Energy Information Agency. Frequently asked questions: Us energy information agency independent statistics and analysis, 2018.

[10] US Energy Information Agency. Notes on data sources, appendix c, 2021.

[11] R. Davis and C.Y. Yau. Consistency of minimum description length model selection for piecewise stationary time series models. *Electronic Journal of Statistics*, 7:381–411, 2013.

[12] Q. Lu, R. Lund, and Lee T. An mdl approach to the climate segmentation problem. *The Annals of Applied Statistics*, 4:299–319, 2010.

[13] R.A. Davis, T.C.M. Lee, and G.A. Rodriguez-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101:223–239, 2006.

[14] P. Fryzlewicz. Wild binary segmentation for multiple changepoint detection. *Annals of Statistics*, 42:2243–2281, 2014.

[15] R. Killick, P. Fearnhead, and I.A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107:1590–1598, 2012.

[16] S. Li and R. Lund. Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25:674–686, 2021.

[17] M. Lee and J. Lee. Long-term trend analysis of extreme coastal sea levels with changepoint detection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70:434–458, 2021.

[18] J. Lee, R. Lund, J. Woody, and Y. Xu. Trend assessment for daily snow depths with changepoint considerations. *Environmetrics*, 31:e2580, 2019.

[19] Washington Office of Financial Management. Washington data and research: Total population and percentage change, 2020.

[20] National Center for Enviromental Information: National Ocean and US Department of Commerce Atmospheric Administration. National climate report, 2021.

[21] National Weather Service. The month in review: January 2017, 2017.

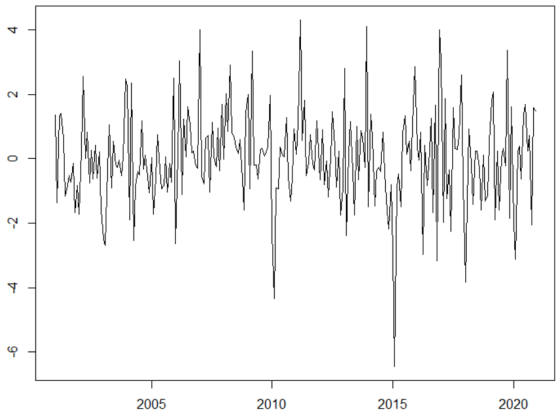
[22] National Association of Manufacturers. 2019 washington manufactuing facts, 2019.

[23] Northwest Power and Conservation Council. Aluminum, 2021.

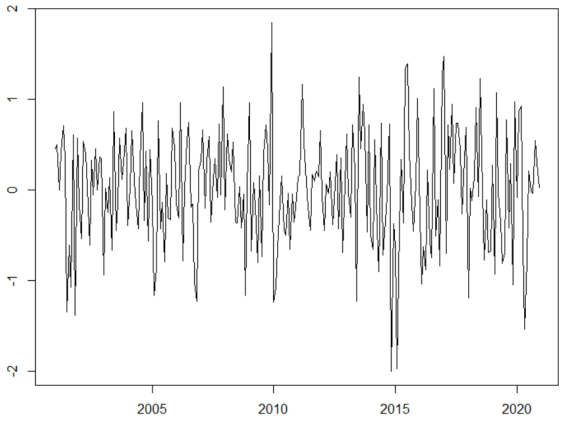
[24] Federal Reserve Economic Data. Industrial production: Total index (indpro), 2021.

APPENDIX A

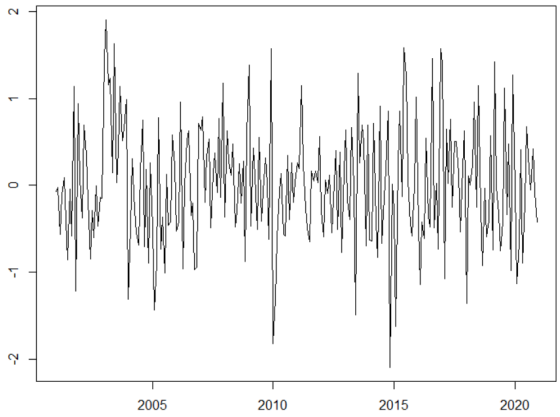
FURTHER DIAGNOSTICS AND MAPS



(a) Residential



(b) Commercial



(c) Industrial

Figure A.1: Washington Residuals

In Figure A.1, we can see the final residuals for each sector in Washington. While

it can be challenging to visually assess stationarity, these residuals no longer appear to contain any linear trends or seasonal patterns. Thus we deem them satisfactory for this analysis.

Table A.1: Residual Diagnostic Test p-values for Washington Data

Sector	Kolmogorov-Smirnov Normality Test	Shapiro-Wilk Normaility Test	Ljung-Box Correlation Test
WA Residential	0.2697	0.001618	0.0001154404
WA Commercial	0.974	0.9146	0.0172447
WA Industrial	0.803	0.329	0.0008653737

In Table A.1, the p-values that pertain to each test are listed. all three data sets pass the Kolmogorov-Smirnov normality test, and the commercial and industrial data sets also easily pass the Shaprio-Wilk normality test at any significance. The residential data does not pass this test, but we can see from the visuals that the residuals are not heavily skewed, so we will accept that the model is getting us quite close to a normal distribution, but it is not perfect.

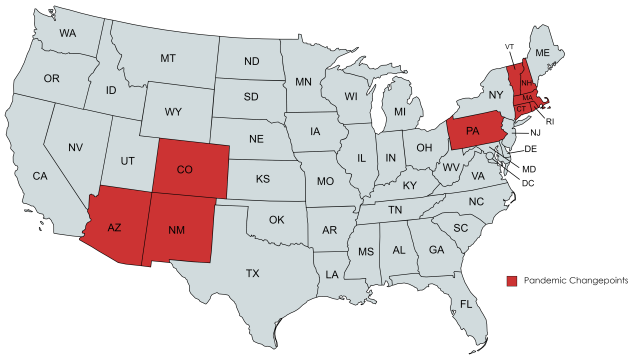
The Ljung-Box test checks for correlation remaining in the data, and it appears that, with the exception of the commercial data set, this test is failed. While we do visually see that there is a small amount of potential correlation in the data in Figure 4.3, Figure 4.6, and Figure 4.9, this test should not be considered to be fully accurate. In the Ljung-Box test, the degrees of freedom is equivalent to the lags we are testing minus the number of parameters that are being estimated. However, the number of parameters that we are estimating in some cases could be around 20, which means that, unless we compute these tests for a very large lag value, the degrees of freedom will be negative, or artificially very small. These values were found using lag = 50, and we have included them as they are a standard test for correlation,

but because of the high parameter to data ratio, we do not find them to be accurate estimates on their own. Instead we prefer to visually inspect the ACF and PACF graphs which indicate that a very small amount of correlation may be present, but in most cases, this correlation is not significant.

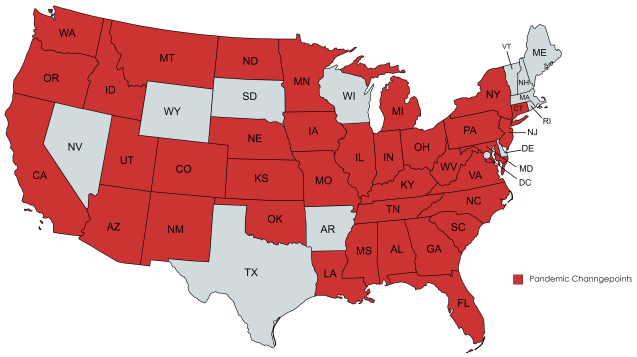
Table A.2: States by Region

Pacific	Mountain	Southern	Midwestern	Eastern
WA	AZ	TX	ND	ME
OR	ID	OK	SD	VT
CA	NM	AR	NE	NH
	UT	LA	KS	MA
	MT	MS	MN	RI
	CO	AL	IA	NY
	WY	GA	MO	CT
	NV	FL	WI	NJ
		KY	IL	PA
		TN	MI	VA
			IN	WV
			OH	NC
				SC
				DE
				MD

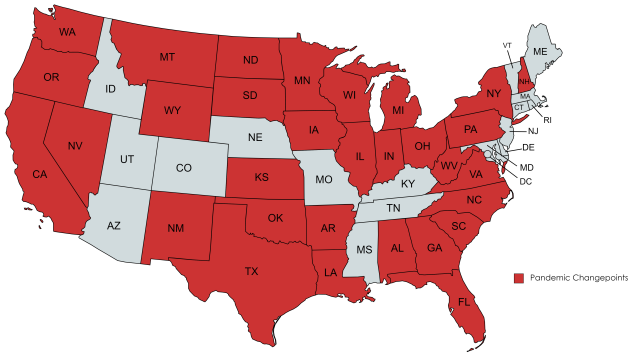
The main code for this research can be found at https://github.com/johannamarcelia/Genetic-Algorithm.



MAP A.1: Residential 2020 Changepoints



MAP A.2: Commercial 2020 Changepoints



MAP A.3: Industrial 2020 Changepoints