6-2022

Data-Driven Analysis of Drug and Substance Abuse Rates Across the Varying Regions in the United States of America

Reem Saleh
Portland State University

Data-driven analysis of drug and substance abuse rates across the varying regions in the United States of America

by

Reem Saleh

An undergraduate honors thesis submitted in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

University Honors

and

Data Science

Thesis Advisor

Dorcas Ofori-Boateng

Portland State University

2022

Abstract

Drugs and substance abuse is one of the leading causes of death for adolescents in the United States. The consequences of using these drugs are profound and can cause both damage to one's physical and psychological health. The rates of drug abuse in the United States continue to increase over the years. This paper analyzes the trends in rates of drug abuse in the four regions in the United States. It looks at the rates in cocaine, cigarettes, marijuana, and tobacco. A preliminary analysis was done to look at the trend in rates followed by an ARIMA time series model for each region. A deeper look was done at the Northeast region which had varying trends in rates. Through this, it was found that the Northeast region had marijuana and cocaine levels that require more attention from regulatory bodies and policy makers.

Contents

1 Introduction

The use of drugs is misleading to many and causes our bodies to be both physically and mentally affected. The Increasing rates of opioid-related emergency department visits and deaths among adolescents in the United States are a public health concern(Lee, Juhan, and Johannes Thrul). The mortality rate attributed to opioid abuse among children and adolescents increased by 268% from 1999 to 2016 (Gaither, Julie R., et al). The abuse of drugs such as heroin, morphine, and tobacco is a serious global problem that affects the health, social, and economic welfare of all societies, not only the United States. It is estimated that between 26.4 million and 36 million people abuse opioids worldwide (UNODC, IDS). This increase in emergencies and drug use over the years affects our generations and the way we survive and it has been devastating to see the continuous rise of these numbers. If this issue becomes normalized before anything is done the number will continue to rise and it will become an issue that is too big to handle. In order to effectively address the obstacles that come with using drugs around the United States, it is important to understand the effects that come with using and administering drugs, and not only confront the negative aspects of using drugs but also acknowledge the healing and aid role that drugs play in helping humans that suffer and are in pain.

The three main categories of drugs, split by the effect that they have on our bodies are, depressants, hallucinogens, and stimulants. Depressants slow down the function of our central nervous system, hallucinogens affect our five senses, and stimulants speed the function of the central nervous system (Australian Government Department of Health). Some of the drugs available can be put into more than one category, such as cannabis which is under all three of the categories listed. The data that will be looked at has drugs from all three categories, more specifically we will be looking at cocaine, tobacco, and marijuana, which fall within the stated groups.

Drug abuse in the United States, including the abuse of tobacco and marijuana, and illicit substances such as cocaine, is implicated in one third to half of the lung cancers and coronary heart disease in adults and in the majority of violent deaths (homicides, suicides, and accidents) in youths (Pentz, Mary Ann, et al.) To help reduce the deaths and rates of the increasing trends in drug abuse it is important to understand and create a goal where the United States brings down these increasing numbers. Recognizing the role that drug abuse plays in chronic diseases and premature mortality and continuing to provide reports and summaries on the trends of drug use and the abuse when using drugs. There have been different research and programs that aim to teach kids at a young age about the effects of drugs and the benefits of resisting them, however, these programs and their effectiveness can be questioned (Pentz, Mary Ann, et al.) As mentioned earlier in order to address the complex problem of drug abuse, we must recognize and consider the special character of this phenomenon for we are asked not only to confront the negative and growing impact of opioid abuse on health and mortality but also to preserve the fundamental role played by prescription opioid pain relievers in healing and reducing human suffering (Volkow, Nora D.). This said we must focus our scientific insight on the right balance of providing maximum relief from suffering while minimizing associated risks and adverse effects. Not focusing only on maximum relief or solely on the effects and risks that come with drugs.

In the United States by broadening the focus area to help understand at a bigger scale the trends of this abuse and finding programs that have been effective to implement them in these specific regions that require more attention. This can help reduce and close the gap that is associated with drug abuse and all the research that comes behind it and human lives that are lost.

2 Background

2.1 Data Collection

Our data was collected from the corgis data sets project, an open-source of data sets compiled together by researchers (Bart, Austin Cory, et al.) To ensure our data is reliable and in fact, accurate, the true source of the data sets are collected from the Substance Abuse and Mental Health Services Administration(SAMHSA) annual summary tables. The SAMHSA conducts yearly nationwide surveys, called National Survey on Drug Use and Health (NSDUH), to create some estimates and help with retention programs. The survey tracks trends in specific substance use and mental illness measures and assesses the consequences of these conditions by examining mental and/or substance use disorders and treatment for these disorders. The administration is a major source of statistical information on the use of illicit drugs, alcohol, and tobacco and on mental health issues among members of the U.S. civilian, non-institutional population aged 12 or older ("National Survey on Drug Use and Health (NSDUH-2015)".) The data was initially in the format of rates and population numbers for each state, separated by age groups and type of drug use. The sole focus of this research is on only the regional data, and no distinction between age groups. So, the data was reduced to be the sum of rates in the entire population, where the age groups were summed into one column for each drug. Then the states were categorized into their corresponding region. From here processing and cleaning of the data were done to get a final csv file which included the region, year, and drug rates for all age groups combined. The final csv file that was worked with was a csv file with the year and rates of each drug type in that region, from the years 2002 to 2019. With this done the next steps of data analysis conducted were creating some visuals and making observations from what was found. These preprocessing and cleaning methods were done primarily in python with some work done in R.

2.2 Analysis

A quantitative analysis approach was the focus throughout this paper since we do not contain any qualitative data such as classifiers and groups. The analysis was done through multiple visualizations created and summaries gained from the cleaned and processed data. The goal of the analysis was to be able to visualize and understand the trends in the rates of the different drugs over the years. By doing this we ensure a stronger understanding of the behavior of the data and the trends from region to region. Analysis was done using primarily R using the ggplot2 library. Analysis of the time series data was specifically done to determine the trends, seasonality, or other cyclic components that are usually present in a time series data set.

2.3 Time Series Modelling

The purpose of a time series model is to forecast and predict the future rates in the region. The reason for using a time series model rather than a regression model is because this is what made the most sense provided the data that was being used. In a typical regression model, we have predictors that influence and predict the response, which in this case is the rate. However, this data does not have any predictors to predict the response. In this case, there is simply a rate and the corresponding year of that rate value. Along with this our data is in the format of a time-series data it “is a time-oriented or chronological sequence of observations on a variable of interest”, essentially our data frame (Montgomery, Douglas, et al.) Because of this, we are able to forecast the future points, a time series model does this by using the past data and estimating the unknown parameters(aka future points) of the model using a least-squares usually.

There are three main components of time series data that need to be looked at: trend, seasonality, and cyclicity. A trend is when there is an upward or downward movement of the data over a period of time. Seasonality is where there are regular peaks or shifts of the data that occur over a specific time throughout the data, considered seasonal. Cyclicity is similar to seasonality except there is no dependence on the calendar or season, it is random spikes and changes throughout ("The Complete Guide to Time Series Analysis".)

Two different models were used and the optimal one was chosen based on the correlation values, distribution of the residuals, and the Ljung Box-test. The formulas and summaries shown below were collected and summarized from *Forecasting: Principles and Practices* by Rob J Hyndman and George Athanasopoulos. The first model was a simple method using the naive method, which is done by simply setting all future predicted values, the forecast, to be based on the last observation value. There is an adjusted method if needed for seasonal data. The formula for the naive method is provided below.

$$\hat{y}_{T+h|T} = yT$$

Although an auto.arima function will be used to help automatically choose the best one, it is important to understand the foundations of the other models.

An Auto regressive Model: In an auto regression model we predict the variable of interest using a linear combination of past values of the variable. Formula for an auto regressive model of order p is below, E_t is white noise. This model is flexible at handling a wide range of different time series patterns.

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

Moving Average Model: In this model instead of using past values of the forecast variable in a regression, this uses past forecast errors in a regression-like model. However, we do not observe the values of E_t so it is not really a regression in the usual sense. We see that the different values of Y_t can be thought of as a weighted moving average of the past few forecast errors. This is different than moving average smoothing, which is used for estimating the trend cycle of past values. The moving average model is typically used for forecasting future values and is possible to write any autoregressive model as a moving average one.

$$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q}$$

Auto regressive Integrated Moving Average Model: If we combine differencing with autoregression and a moving average model we get a non-seasonal ARIMA model. In our formula below Y_t' is the differenced series. Differencing is a method that can be used to transform a non-stationary time series into a stationary one. The first differencing value is the difference between the current time period and the previous time period. The auto.arima function combines unit root tests, and minimization of the AICc and MLE to obtain an ARIMA model that is the best fit for the data. An ARIMA model can also be used for seasonal data but additional terms needed to be added to tailor it to the seasonality that is present. The formula below is for a non-seasonal model.

$$y_t^{'} = c + \phi_1 y_{t-1}^{'} + ... + \phi_p y_{t-p}^{'} + \theta_1 \varepsilon_{t-1} + ... + \theta_q \varepsilon_{t-q}$$

3 Methodology

As mentioned the data was gathered from the SAMHSA which conducts services on a yearly basis to help guide different policy directions in the United States such as problem substances, the prevalence of mental illness, intersection of substance abuse and mental illness and provide insights in the context of data from other agencies (McCance-Katz, Elinore). The analysis will include years from 2002 until 2018 from all 50 states; the year 2019 will be used to help choose which model should be used, based on the forecast values and other factors. The data includes information from all 50 states that will be organized by region as defined by the United States Census; there are four regions, West, Midwest, South, and Northeast.

The method chosen for the time series data was done by determining whether a time series model would be effective for this purpose of forecasting. Since the data contains only two variables of interest, the year and drug rate it is the most logical decision to go with a time series model. To choose the best model, two different time series models were used and as mentioned the residuals were looked at and some tests were done to determine the best specific time series model for forecasting and viewing trends.

4 Results

4.1 Exploratory Data Analysis

Cocaine

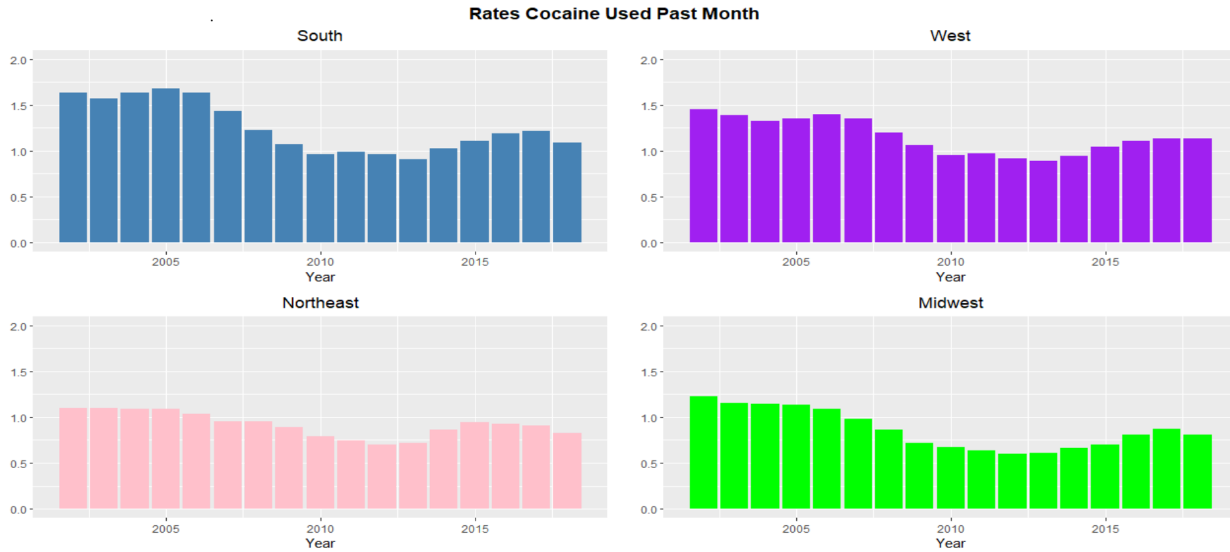


Figure 1: A downward and then upward trend is shown for all regions when looking at the Cocaine rates.

Marijuana

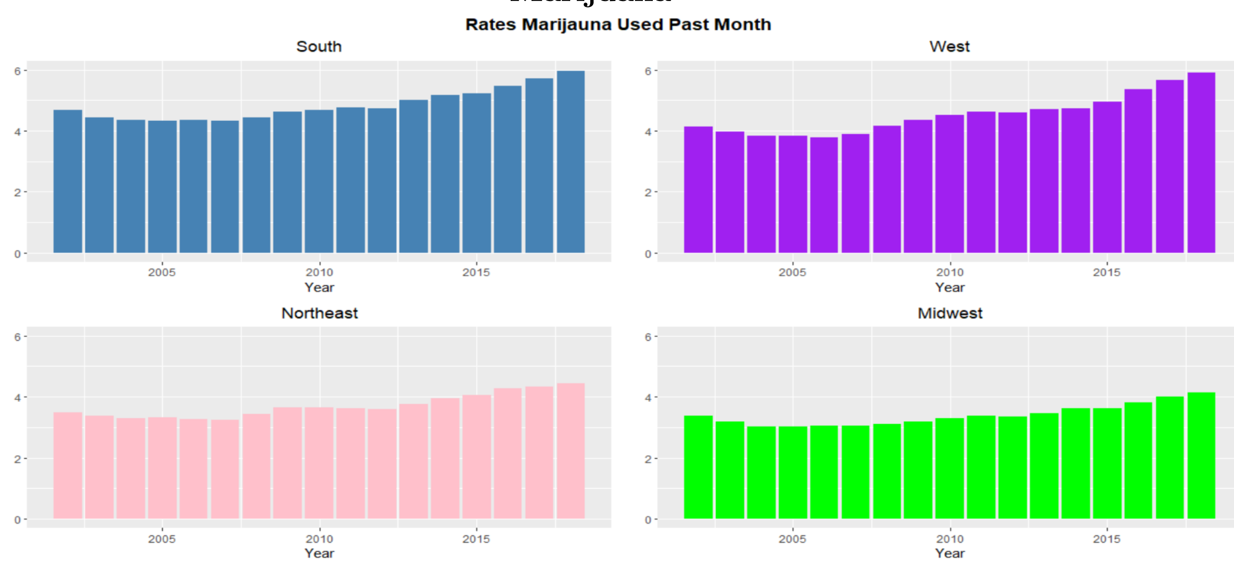


Figure 2: A downward and then upward trend is shown for all regions when looking at the Marijuana rates.

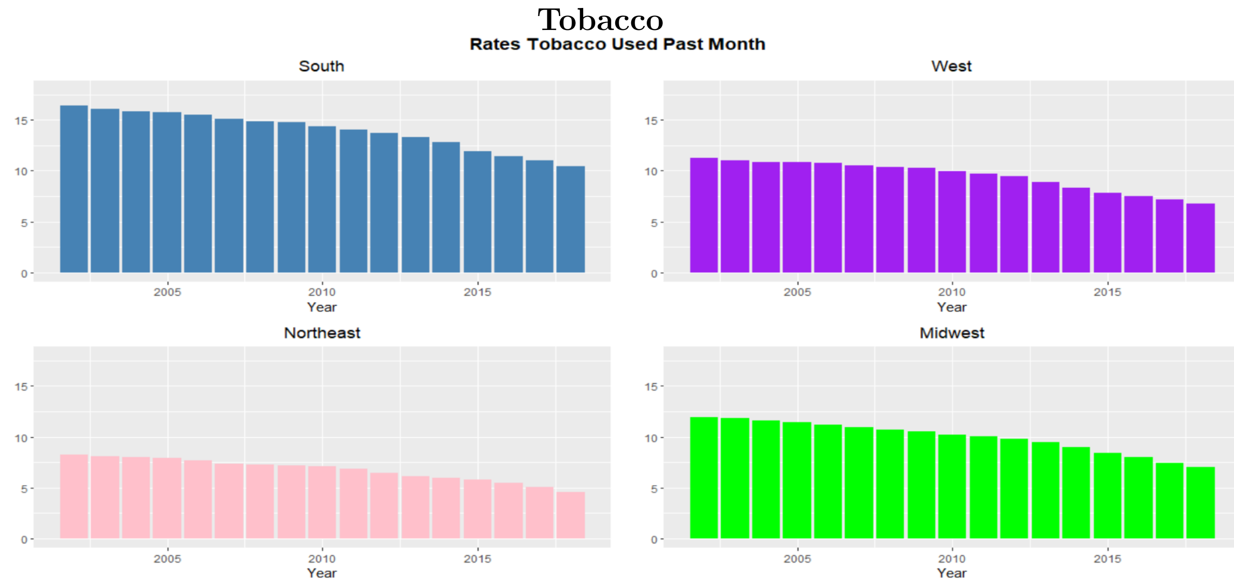


Figure 3: An overall decreasing trend in Cigarette rates is present in all regions.

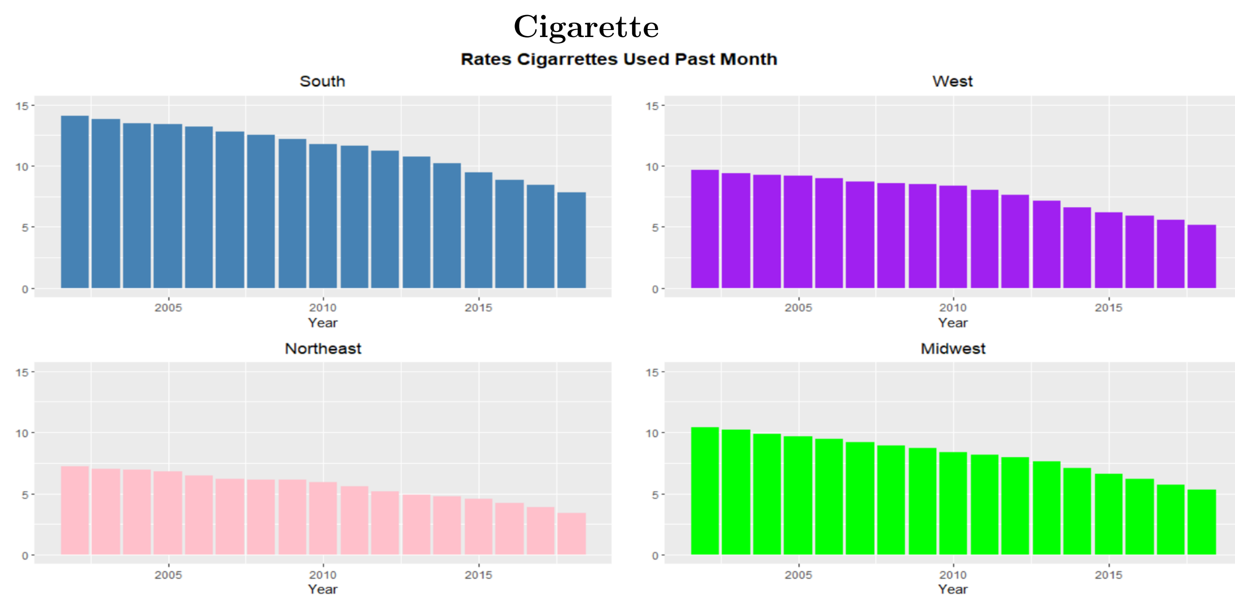


Figure 4: An overall decreasing trend in Tobacco rates is present in all regions.

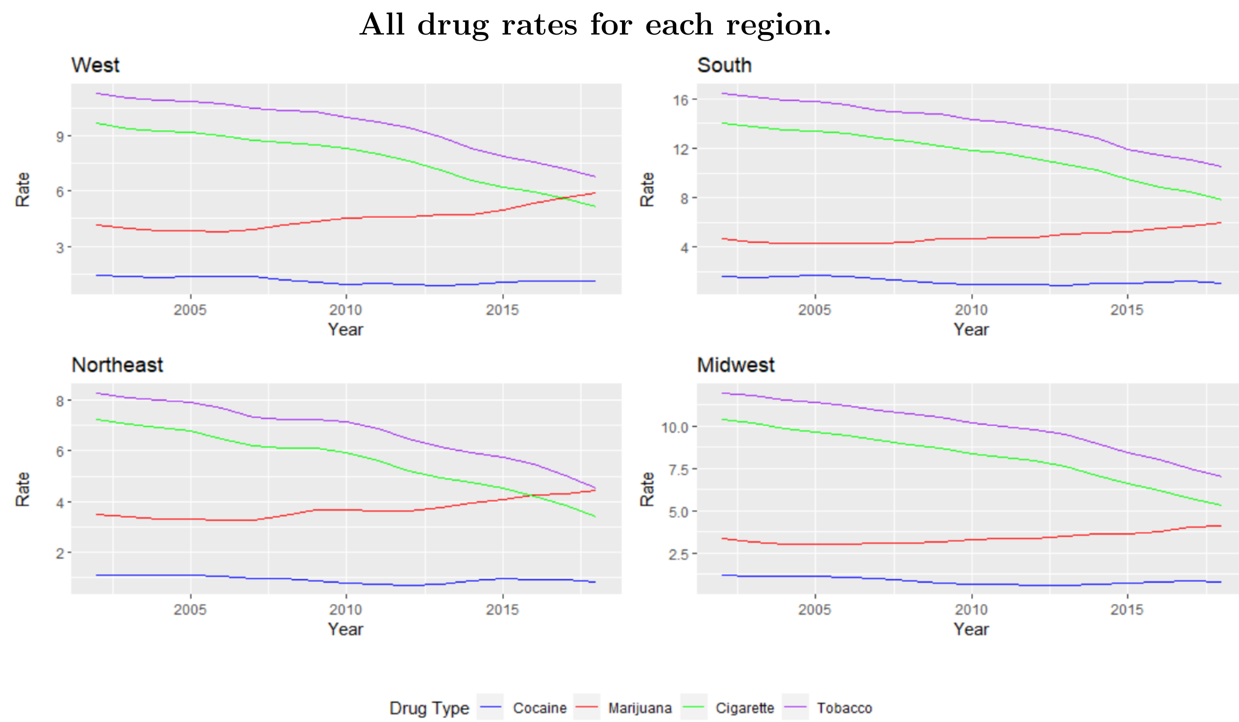


Figure 5: The same trends from earlier are present, but there is an intersection between Marijuana and Cocaine that can be seen in the West and Northeast region that was not clear in earlier figures.

4.2 Time Series

To simplify the initial steps of time series modeling. The data was split into four data sets. Each one corresponds to a region with a year and the sum of the rates of all drugs total. So, this resulted in a total of 4 time series for each region, and this was done for both the Naive Bayes and ARIMA models. The first step was to note that in a time series data it is also important to make sure that the data is stationary, with no trend and seasonality. To ensure the data was stationary the adf.test was used from the aTSA library. From the results, a p-value greater than the 0.01 significance level, is observed and our result is that the data is stationary, which is what is needed to proceed.

Checking to ensure that data is stationary

This was repeated for each region's data

```
library(aTSA)
adf.test(Sfinal[,2])
```

Output

```
Augmented Dickey-Fuller Test
alternative: stationary

Type 1: no drift no trend
     lag   ADF p.value
[1,]   0 -5.69   0.010
[2,]   1 -1.19   0.244
[3,]   2 -1.23   0.229
Type 2: with drift no trend
     lag  ADF p.value
[1,]   0 2.43    0.99
[2,]   1 2.41    0.99
[3,]   2 3.86    0.99
Type 3: with drift and trend
     lag   ADF p.value
[1,]   0 0.478    0.99
[2,]   1 0.178    0.99
[3,]   2 1.070    0.99
----
Note: in fact, p.value = 0.01 means p.value <= 0.01
```

Once this was done and the data that is being used is known to be stationary the ts() function from the stats package was used to convert the data into an R time series object that can be then put into our models.

```r
library(stats)
Southts<-ts(Sfinal[,2],start=c(2002), end = c(2018),frequency = 1)
```

Output

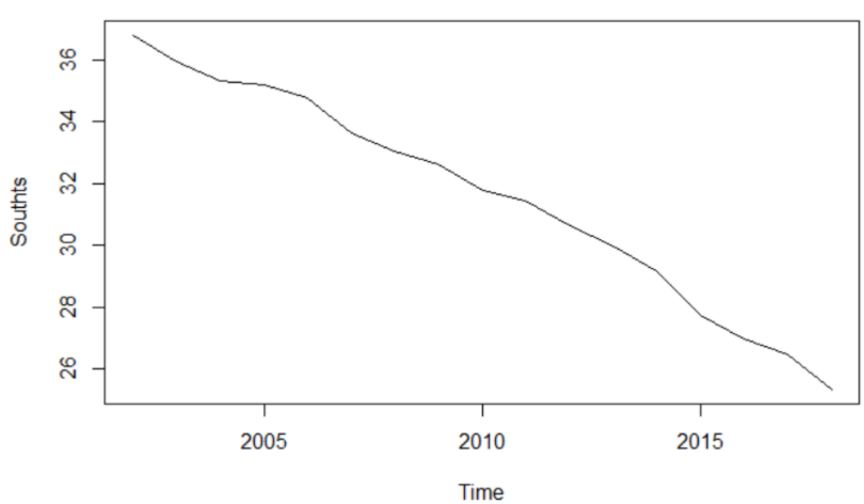


Figure 6: Time series output for South region of non-differenced data, a downward trend is present.

From the above graph, figure 6, a downward trend can be seen and this is an issue that will need to be resolved. To remove this the diff() function was applied, which provides differenced data. It is the difference between each pair of years and can see that the trend has been removed. This is shown in figure 7 below, there is no trend that can be seen in our differenced data.

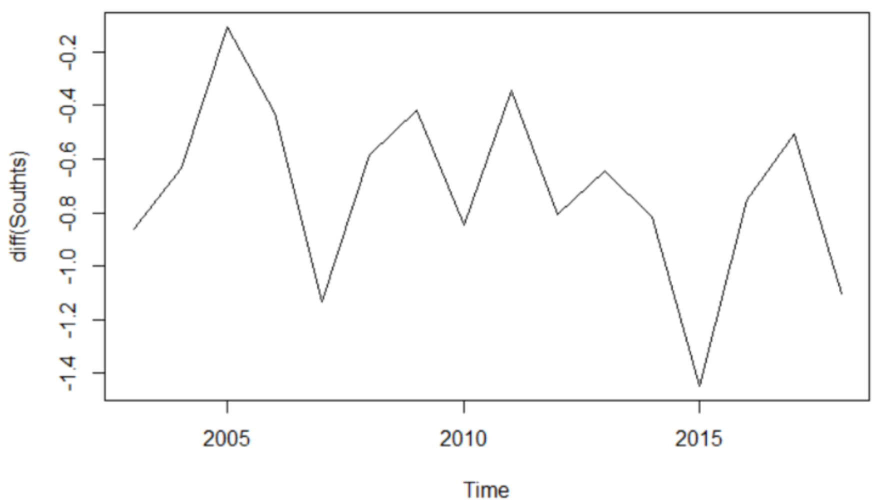


Figure 7: Time series output for South region with differenced data, no trend is found here.

Before the model was used the final item to check for was seasonality. Without having to run any functions or code it is clear from the above differenced data that no seasonality is present. There is no sign of periodic fluctuations occurring often during certain times. The data is now ready to be used in a model since there is no trend, it is stationary, and no seasonality is present.

Two models were used: the Naive Bayes and ARIMA method. To determine the best one that will be used to forecast future years, the Ljung-Box test was done and the distribution of the residuals was also looked at. This was again repeated for each region.

The two models

```
#the Naive Bayes model
nsd<-naive(diff(Southts))
#the ARIMA model
as<-auto.arima(Southts, d=1,approximation = FALSE, trace=TRUE)
```

```
        Ljung-Box test

data:  Residuals from ARIMA(0,1,0) with drift
Q* = 3.5016, df = 3, p-value = 0.3205

Model df: 1.   Total lags used: 4
```

ARIMA Model

```
        Ljung-Box test

data:  Residuals from Naive method
Q* = 7.7404, df = 3, p-value = 0.05169

Model df: 0.   Total lags used: 3
```

Naive Model

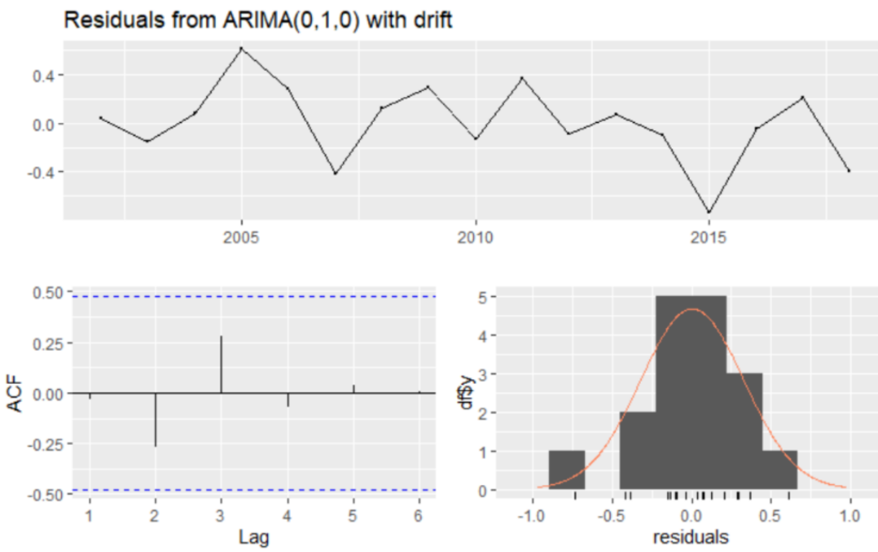


Figure 8: Residuals output example for the South for ARIMA model, the residuals are normally distributed.

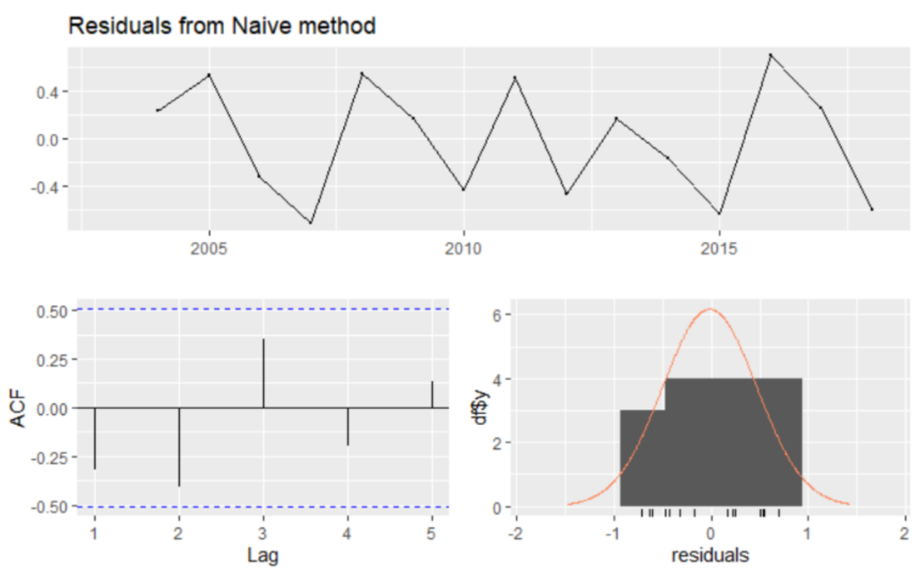


Figure 9: Residuals output example for the South for Naive Bayes model, the residuals are not normally distributed with a left-skewed distribution present.

From the Ljung-Box test, the p-value is greater than 0.05 for both models but the naive method is on the threshold with a p-value of 0.05169, whereas the auto.arima model has a p-value of 0.3205. Along with this, looking at the residuals plots, figures 8 and 9, it is clear that the ARIMA model has a more normal distribution than the naive method which is what we want to see to show that our residuals are independent of each other. From this, the ARIMA method was chosen for all regions as there were similar results for each one when looking at the results from the Ljung-Box test and the residuals plot. Thus forecasting and modeling for all regions were done using the ARIMA method.

4.3 Forecasting

Now that the model has been chosen, forecasting is done using the ARIMA model. The next step is to forecast 5 years ahead, with 2019 being one of these years that is forecasted. This step is done to view any unusual fluctuations in a certain region or if there forecasts a significant change over the upcoming years.

The forecast function was used from the forecast library. An example of our results for only the south region is shown below. This provides both 80 and 95% confidence intervals along with the specific point forecast.

```r
library(forecast)
fs<-forecast::forecast(as, h=5)
```

South Forecasts

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
2019	24.63301	24.20454	25.06149	23.97772	25.28831
2020	23.91781	23.31186	24.52376	22.99109	24.84453
2021	23.20260	22.46047	23.94474	22.06761	24.33760
2022	22.48740	21.63046	23.34434	21.17682	23.79798
2023	21.77219	20.81410	22.73029	20.30692	23.23747

The forecasting process was repeated for each region and the plots for all four are shown below.

South

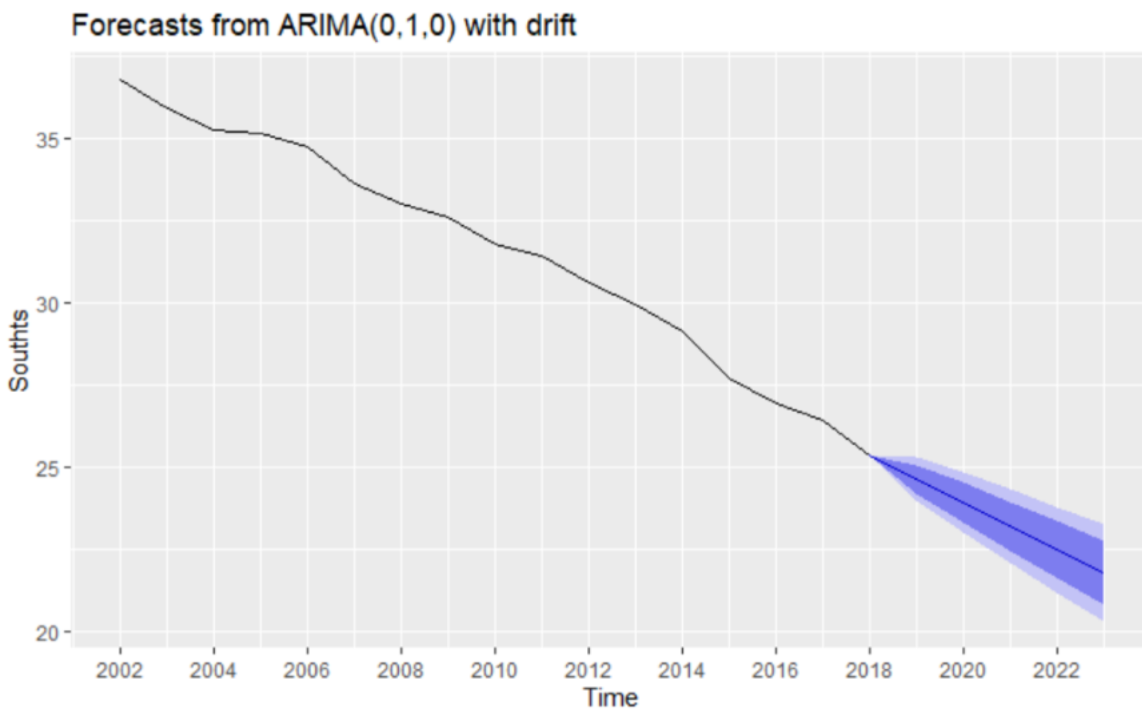


Figure 10: Forecast confidence interval for the South for overall drug rates.

Northeast

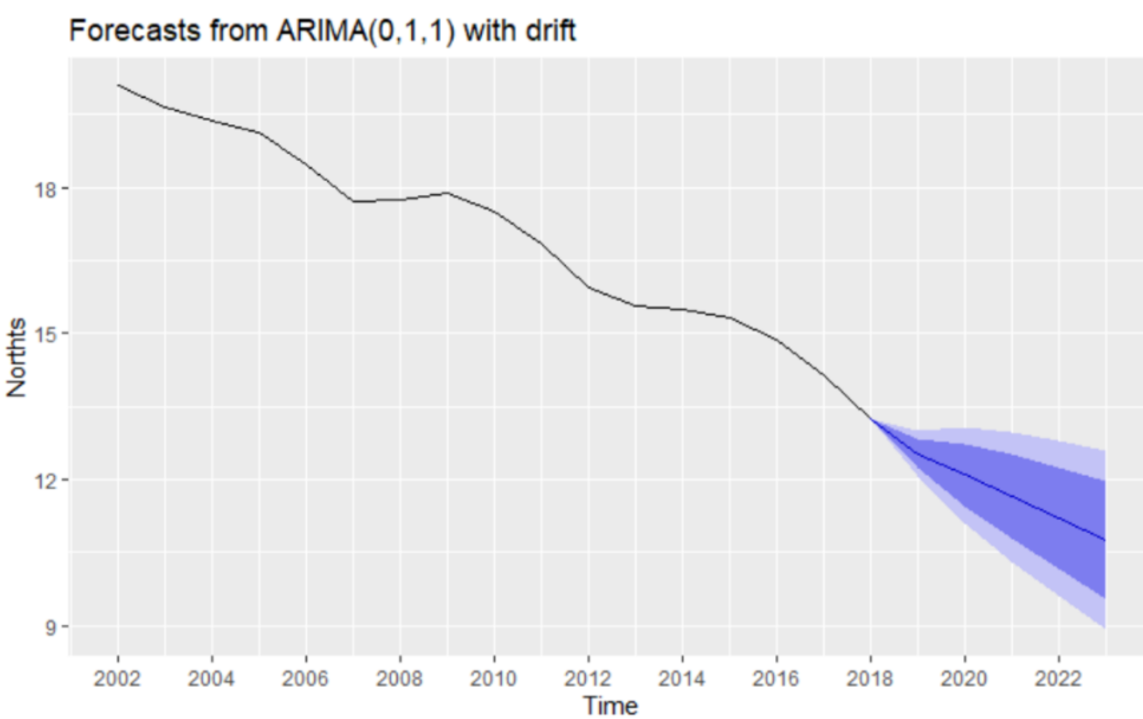


Figure 11: Forecast confidence interval for the Northeast for overall drug rates.

West

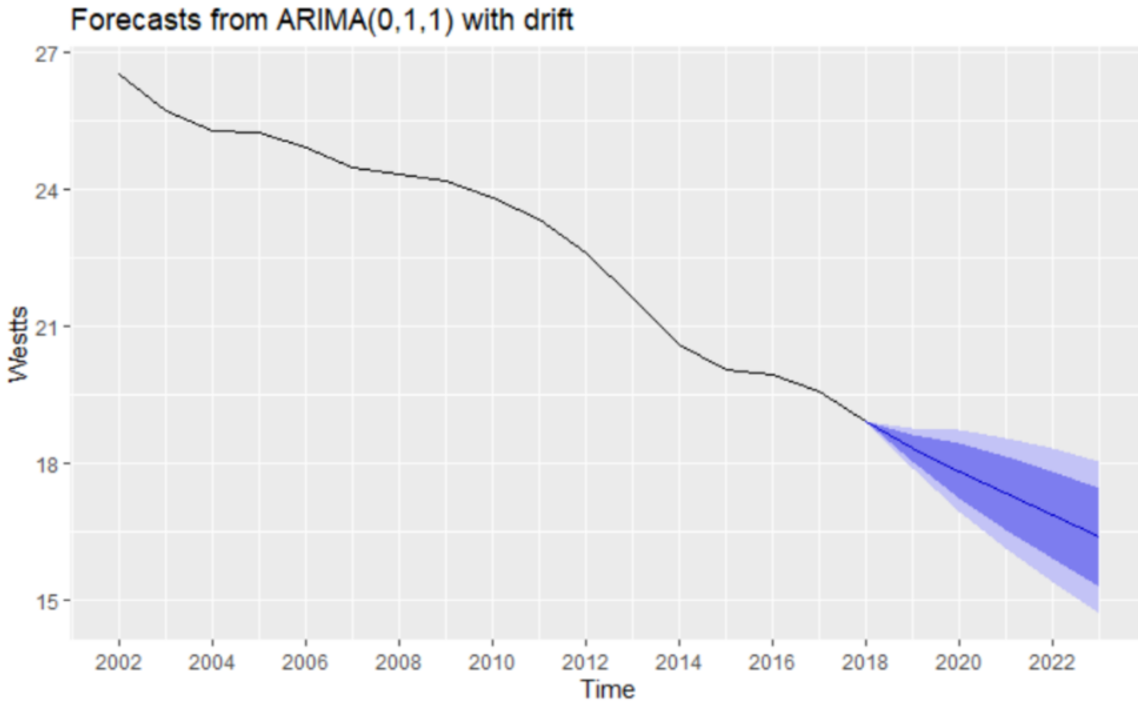


Figure 12: Forecast confidence interval for the West for overall drug rates.

Midwest

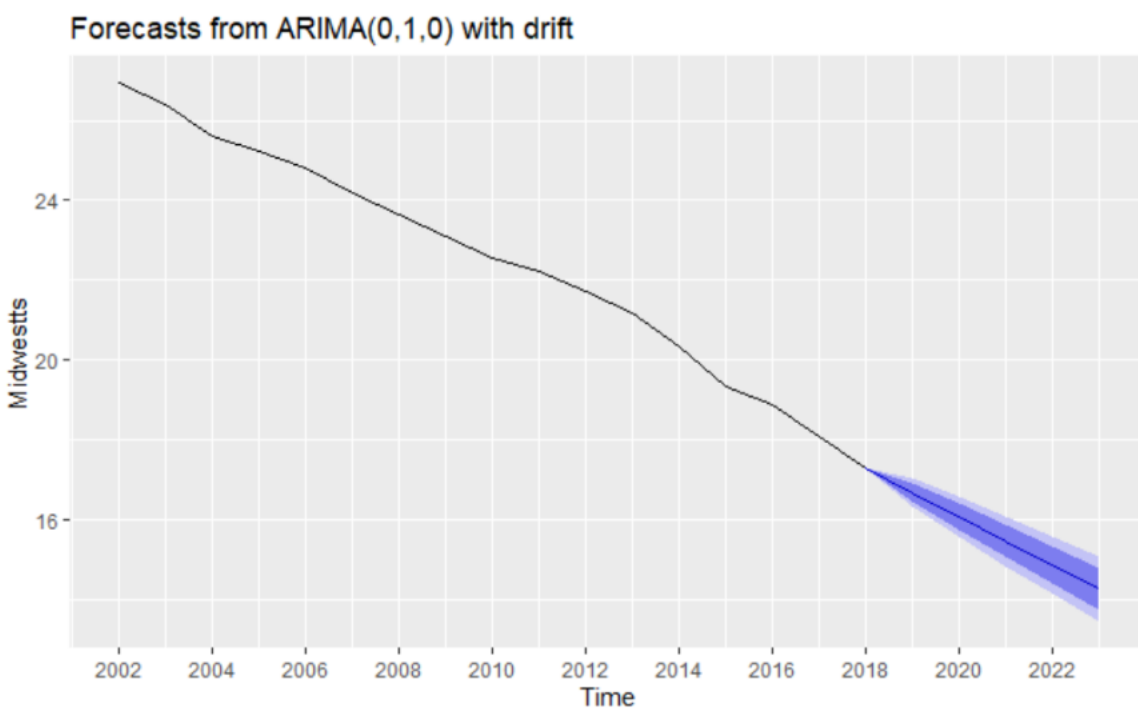


Figure 13: Forecast confidence interval for the Midwest for overall drug rates.

From the above forecasts, it is clear that all regions have a downward trend present for all drugs and this will likely continue to decrease over the years. Although all regions have a downward trend, one region stood out because of the small fluctuations that occurred. The Northeast region, figure 11, compared to all the other regions had an increase for a couple of years before it went back down. Whereas the other regions had an almost consistent decrease over the years. Because of this, four new time series were created on the Northeast region only for each drug there was.

4.4 Northeast Focus

4.4.1 Time Series

Continued with the auto.arima function because it had the best results from the previous analysis. The entire process was repeated the same for each drug type in the Northeast region. The residuals were looked at, a Ljung-Box test was done, and forecasting was also done.

```
Ncigarrette <- N[1:2]
NCocaine <- N[,c(1,3)]
NMarijuana <- N[,c(1,4)]
NTobacco <- N[,c(1,5)]

NorthCig<-ts(Ncigarrette[,2],start=c(2002), end = c(2018),frequency =
    1)
NorthCoc<-ts(NCocaine[,2],start=c(2002), end = c(2018),frequency = 1)
NorthMar<-ts(NMarijuana[,2],start=c(2002), end = c(2018),frequency =
    1)
NorthTob<-ts(NTobacco[,2],start=c(2002), end = c(2018),frequency = 1)
```

Creating Model and checking fit

```
ancig<-auto.arima(NorthCig, d=1,approximation = FALSE, trace=TRUE)
ancoc<-auto.arima(NorthCoc, d=1,approximation = FALSE, trace=TRUE)
anmar<-auto.arima(NorthMar, d=1,approximation = FALSE, trace=TRUE)
antob<-auto.arima(NorthTob, d=1,approximation = FALSE, trace=TRUE)
```

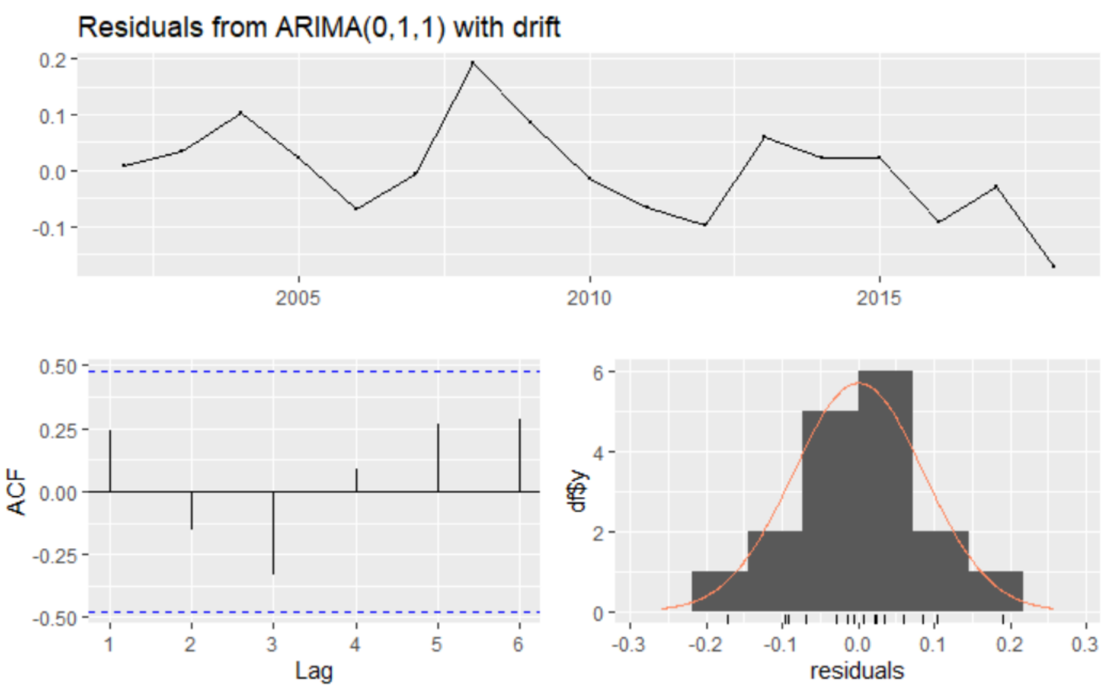
```
checkresiduals(ancig)
checkresiduals(ancoc)
checkresiduals(anmar)
checkresiduals(antob)
```

Ljung Box-test and residuals Results for cigarette model

```
        Ljung-Box test

data:  Residuals from ARIMA(0,1,1) with drift
Q* = 6.38, df = 3, p-value = 0.09452

Model df: 2.   Total lags used: 5
```



It is clear that once again the residuals are normally distributed, and there is a p-value greater than 0.05. From this, it is safe to continue with the ARIMA model as had been done earlier for all the regions. This was done for all four drugs in the Northeast region.

4.4.2 Forecasting

Here are the forecasting results for the four drugs in the Northeast region. There are very significant differences for each drug.

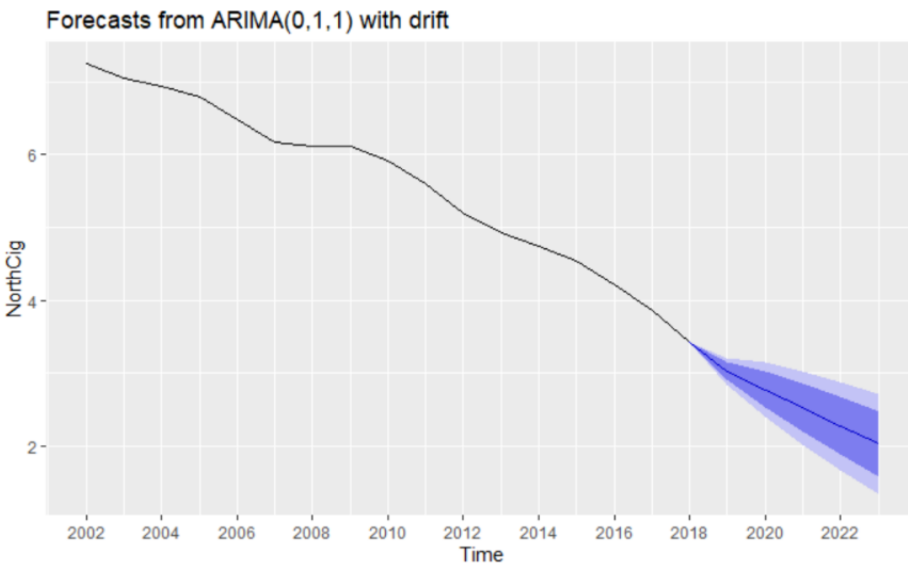


Figure 14: Forecast confidence interval for Cigarettes in the Northeast, a continuous decrease is predicted.

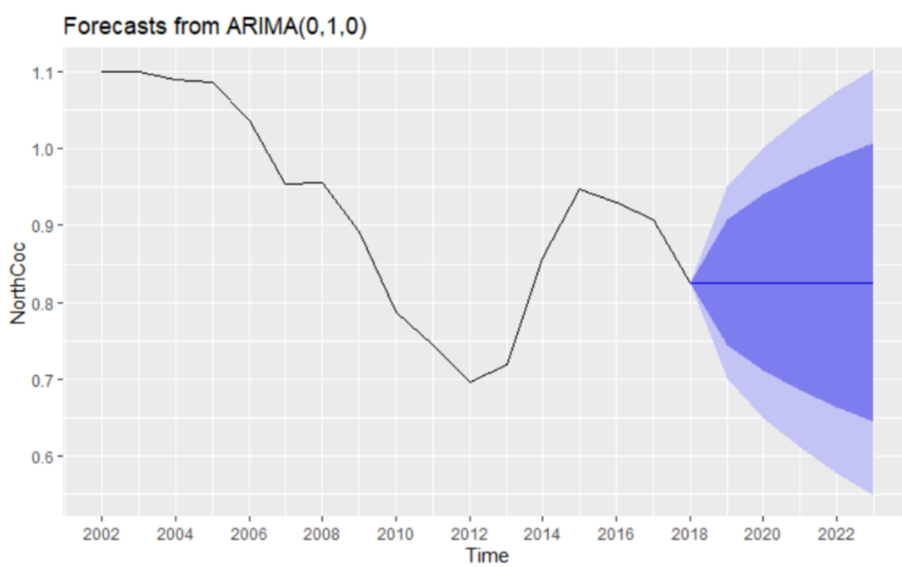


Figure 15: Forecast confidence interval for Cocaine in the Northeast, model has a difficult time predicting future rates because of irregular increases and decreases present.

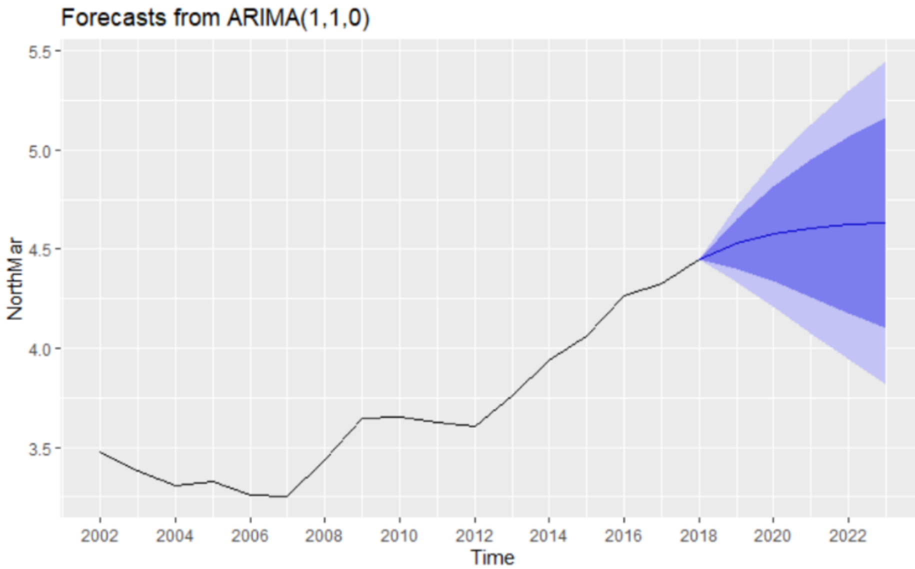


Figure 16: Forecast confidence interval for Marijuana in the Northeast, a continuous increase is predicted.

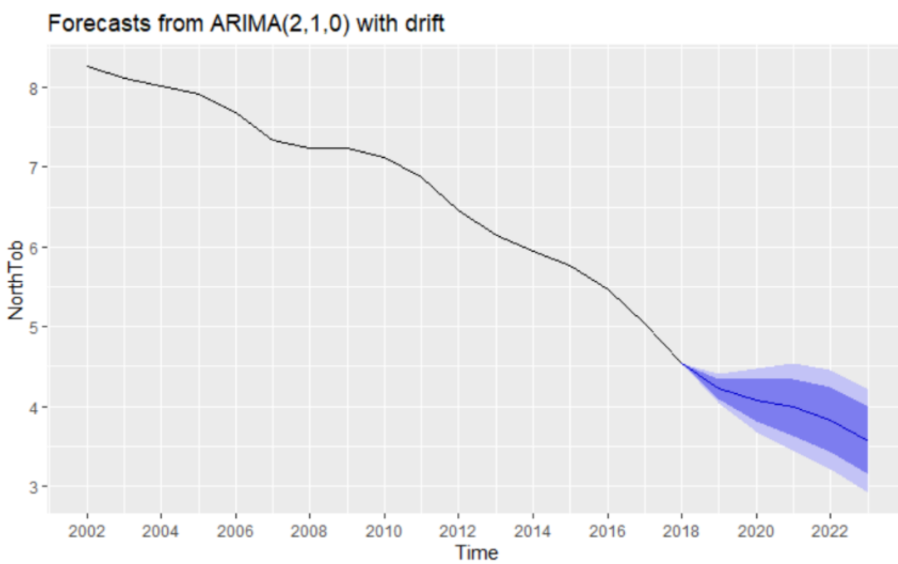


Figure 17: Forecast confidence interval for Tobacco in the Northeast, a continuous decrease is predicted.

5 Discussion

From the results, it is clear that the overall rates of drug use have been decreasing over the years and the trend will continue to be a decreasing trend for all regions. In the initial analysis with the plots done in ggplot2, section 4.1, where we compared the different drug rates in each region we saw the pattern of a continuously decreasing trend in tobacco and cigarettes. However, the pattern we saw in marijuana and cocaine, was a decreasing and then increasing trend, figures 1 and 2. This is the same for all regions. Although it seems that tobacco products are not an issue since they have a continuously decreasing trend, an important thing to note is the y-axis. The two drugs, tobacco, and cigarettes have the highest rates overall compared to the other two drugs. We are able to see from figures 1-4 that tobacco and cigarette rates range from approximately 5% to 15%, whereas cocaine and marijuana range from a high of 0.5% to around 6%. These high rates for tobacco and cigarettes are important to keep a close eye on as adolescent cigarette use may in fact be a risk factor for opioid misuse (Lee, Juhan, and Johannes Thrul.) If an adolescent begins with cigarettes it increases the chances of misusing other types of drugs. With this in mind, a time series was done for each region separately but not separated by a drug. This allowed us to find any patterns in overall drug use from region to region, and forecast some future years. We are able to see that all of the time series by region are similar, but there is one region that had a few ups and downs, which was broken down to determine what drug may be causing these jumps.

The small fluctuations and differences in trends can be seen in the Northeast region, in figure 11, between the years 2007 and 2009 there was a small increase in the drug rates before they went back down. There was also an almost flat line where the rates stayed very similar between 2013 and 2015. Due to these observations and irregularities present in the Northeast but no other region, another time series and forecasts was done for each drug type in the Northeast region only. Once this was done there were some interesting results shown in figures 14-17. The first thing to note is tobacco and cigarettes have the same downward trend that was present in the overall region from earlier forecasting. We can infer that retention programs and efforts targeted towards these drugs have been successful as the trend continues to decrease and is forecasted to continue also. However, we see very different results from the drugs marijuana and cocaine. With marijuana, figure 16, we see that in the early years 2002-2007 the rates had been decreasing, but after those years it increased drastically. The forecasted years also continue to have an increasing prediction, with wide ranges for confidence intervals. Looking at cocaine, figure 15, there is no clear pattern present and our forecast has such a wide range it shows that the model is not able to predict future years accurately because of the variance in the rates over the years. From the years 2002-2012, there is a decreasing trend that is present, but after this, there are different changes. The line goes from increasing at a fast rate to decreasing. The forecasted line is a plato, flat, with no trend present. From these observations and trends found in marijuana and cocaine specifically, it is clear that these are the two that are uncertain of what will happen and what these rates will become in future years.

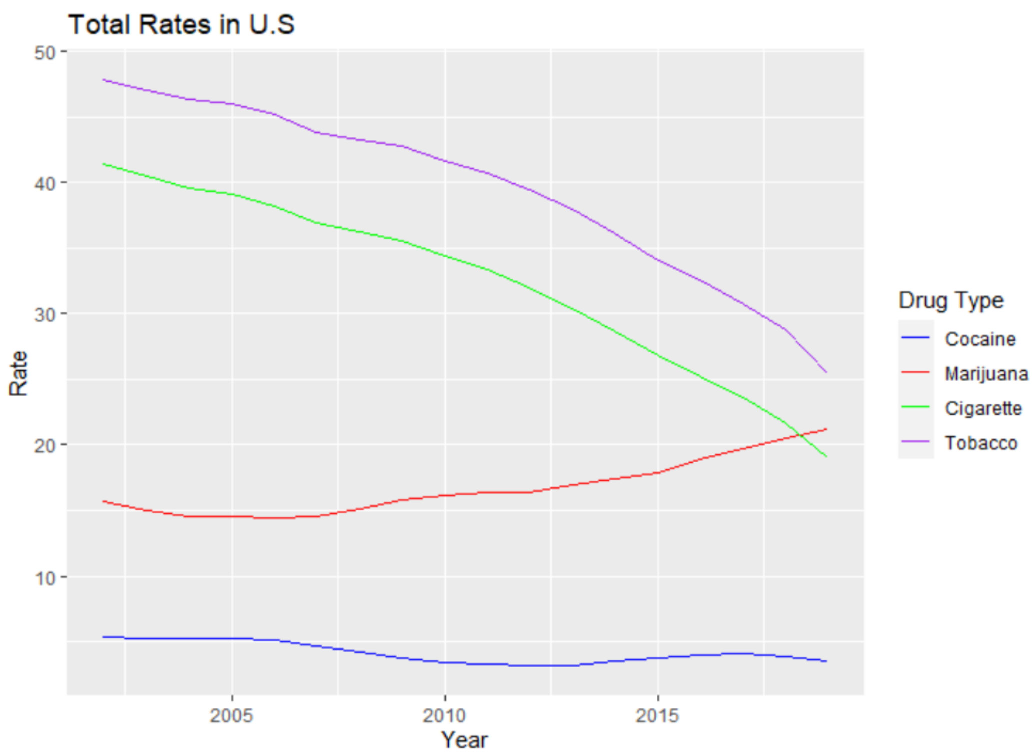
6 Conclusions

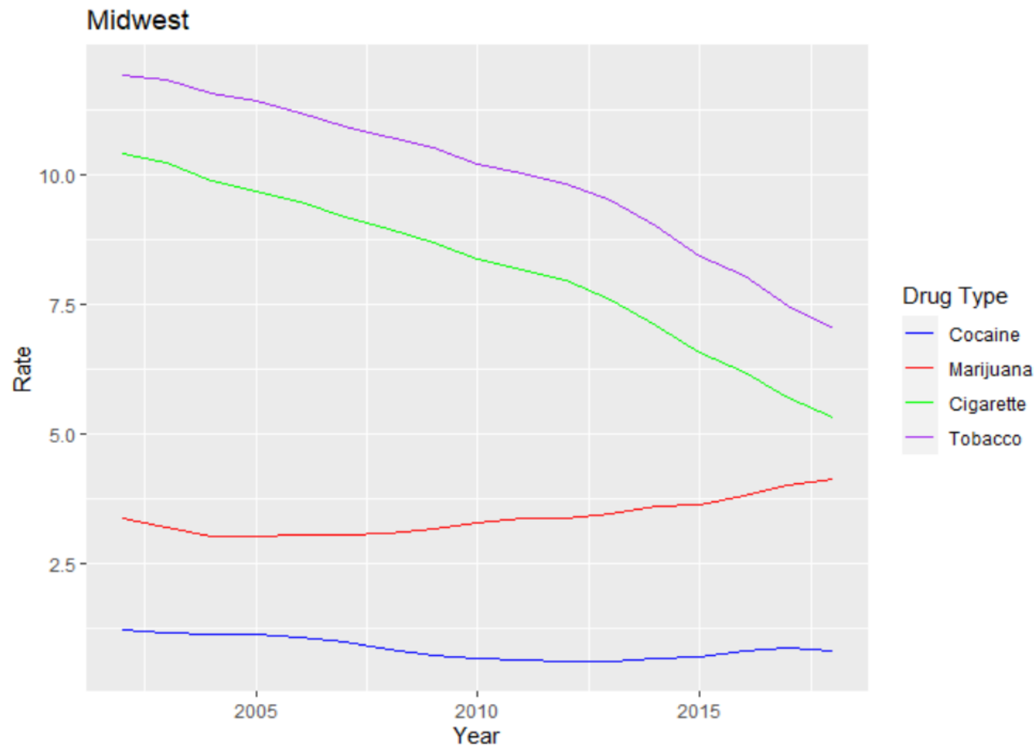
Throughout this thesis, it is clear that focus and efforts should be pushed towards the Northeast region, more specifically work towards creating retention programs and stricter measures towards cocaine and marijuana. The varying accessibility to drugs from region to region plays a large impact on these rates and who is more affected by it.

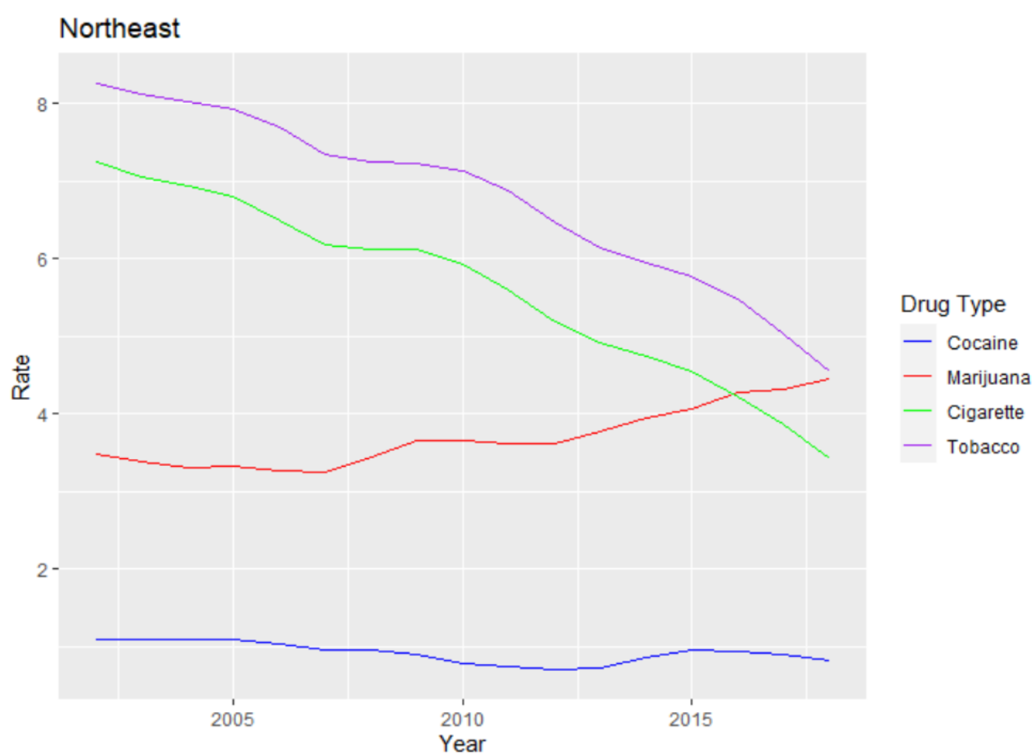
An interesting continuation to this study could be to break this down further by age groups and look at patterns between the groups and drugs from region to region. This would be interesting to take on mainly because of the different restrictions there are on marijuana from state to state. For example, some states have an age limit that ranges from 18 to 21 or greater, but some don't have any age restrictions("Alcohol and Drugs: Marijuana Laws."). Although this mainly applies to medicinal use, lower-income communities are often able to easily access the drugs easier than expected, with more reports of drug use in minority communities(Moore, Lisa D., and Amy Elkavich.) This can play an impact from region to region and the different trends in the rates that might appear. Overall, it is important to be aware of the risk that comes with drug use and the impact it can have not only on oneself but on the communities around us. Continuing to analyze these trends in rates and pushing resources towards focus areas is a small step to helping the fight against drug abuse.

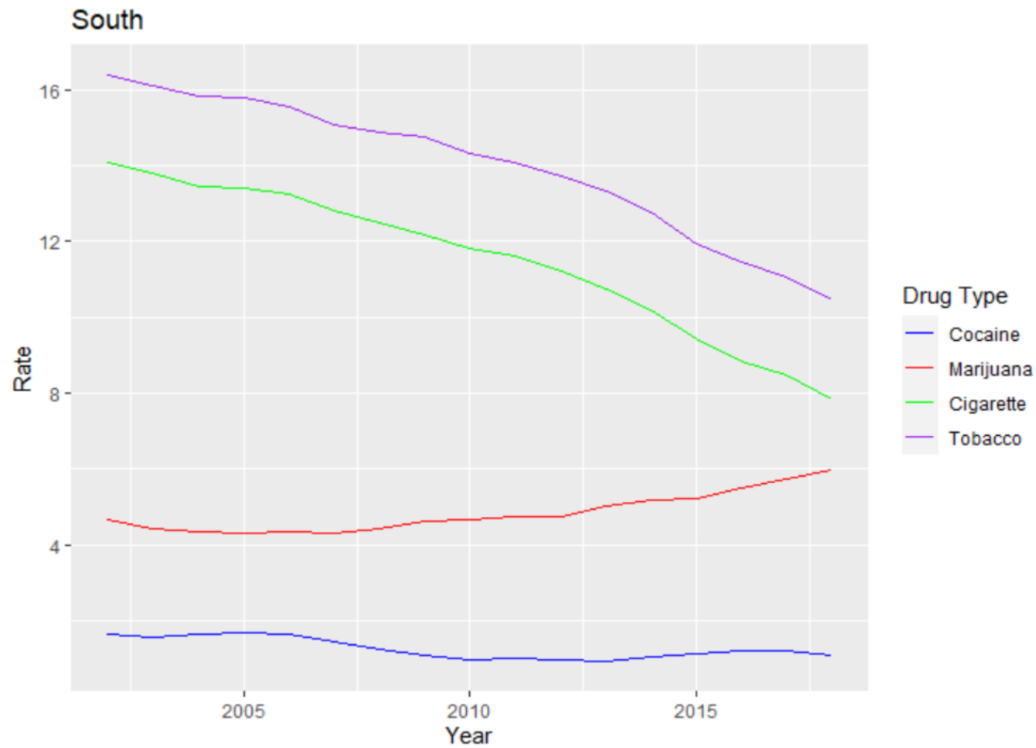
7 Appendix

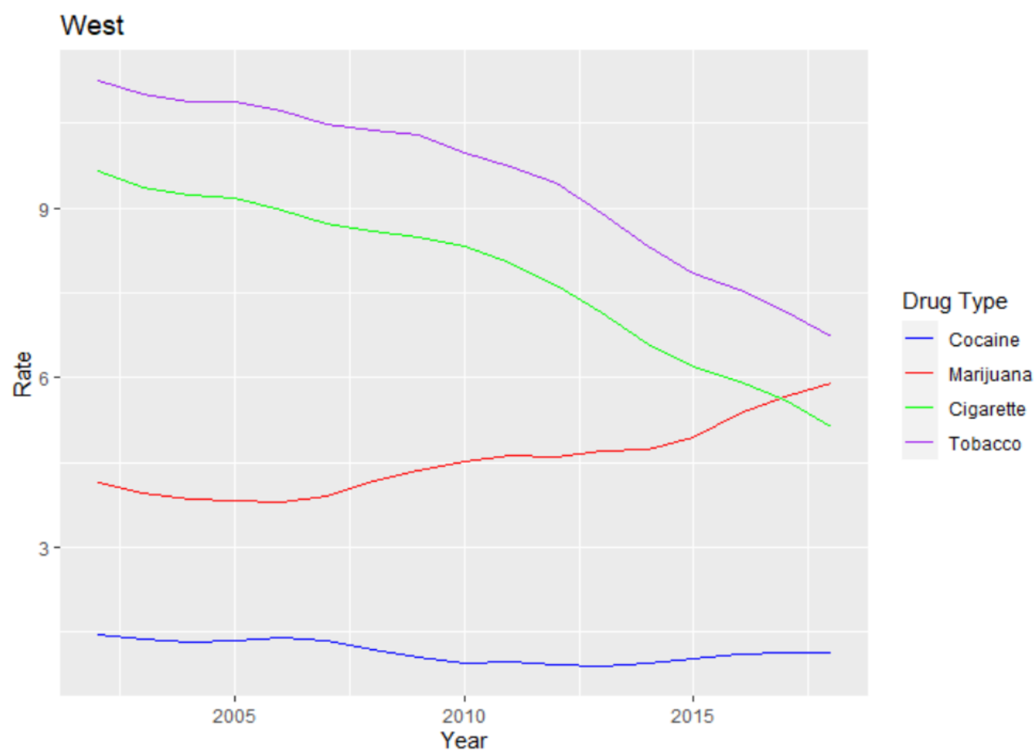
7.1 Appendix A: Initial Analysis



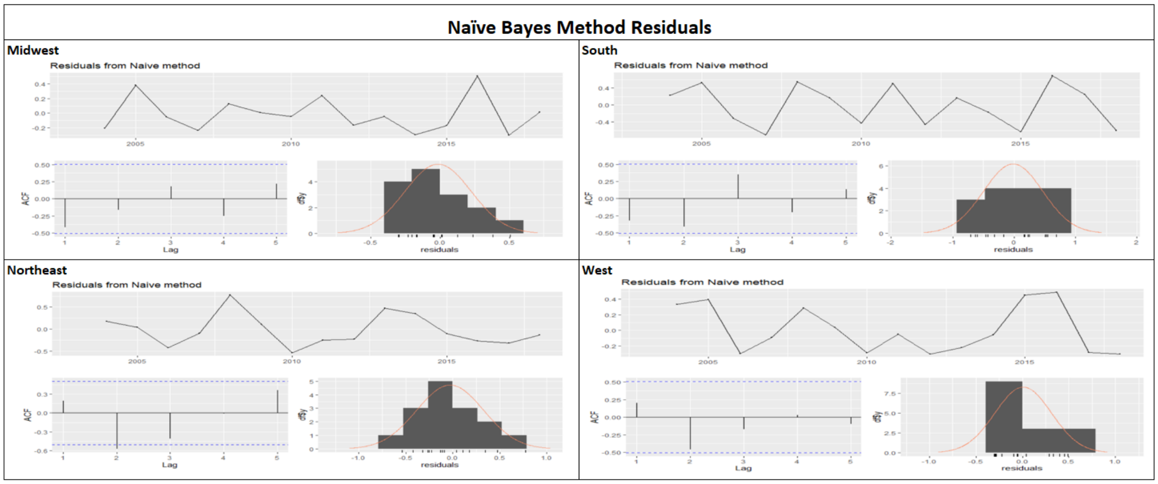

Midwest
Rate
10.0
7.5
5.0
2.5
2005
2010
2015
Year
Drug Type
Cocaine
Marijuana
Cigarette
Tobacco

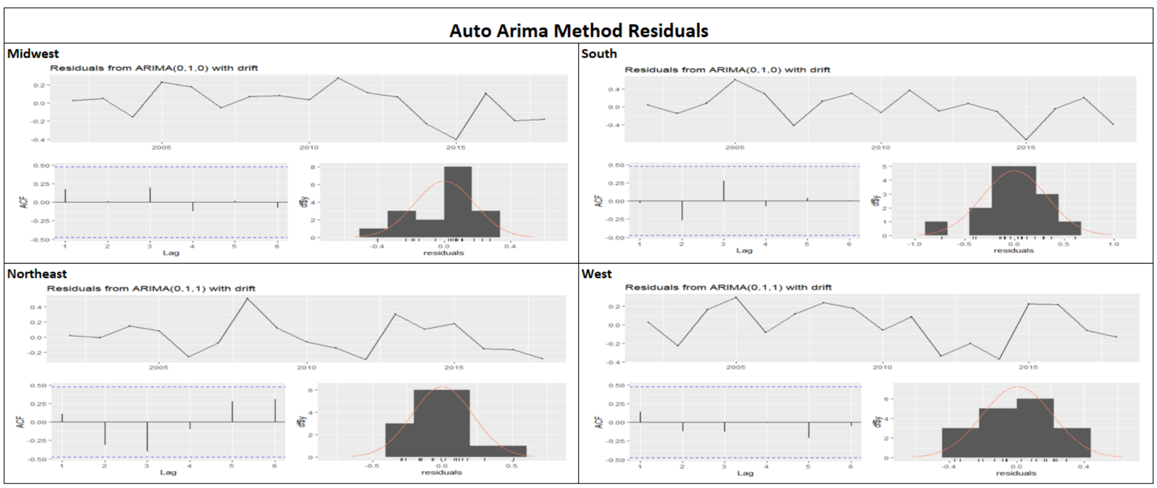

Northeast
Rate
8
6
4
2
2005
2010
2015
Year
Drug Type
Cocaine
Marijuana
Cigarette
Tobacco

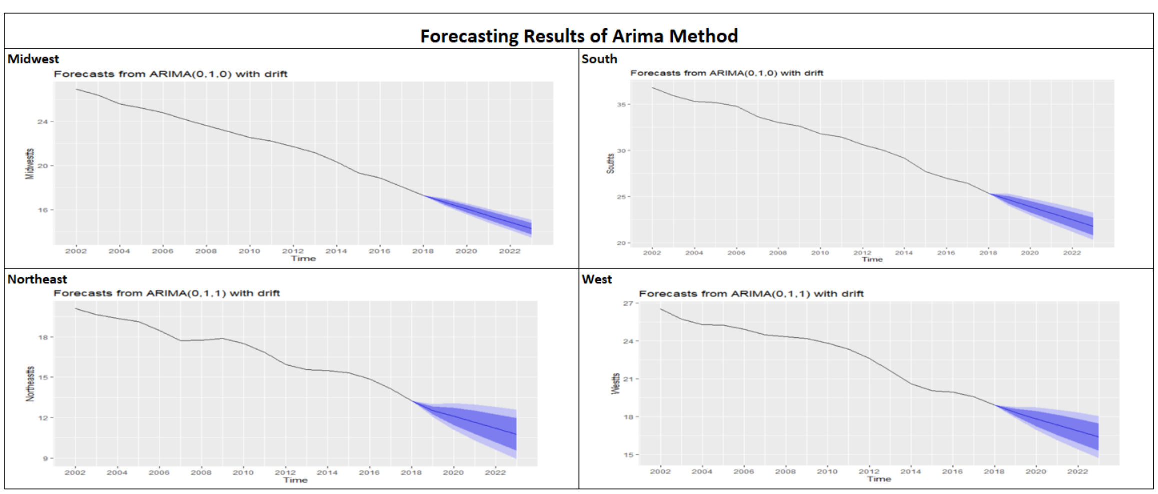

South
16
12
8
4
Rate
2005
2010
2015
Year
Drug Type
Cocaine
Marijuana
Cigarette
Tobacco


West
9
6
3
Rate
2005
2010
2015
Year
Drug Type
Cocaine
Marijuana
Cigarette
Tobacco

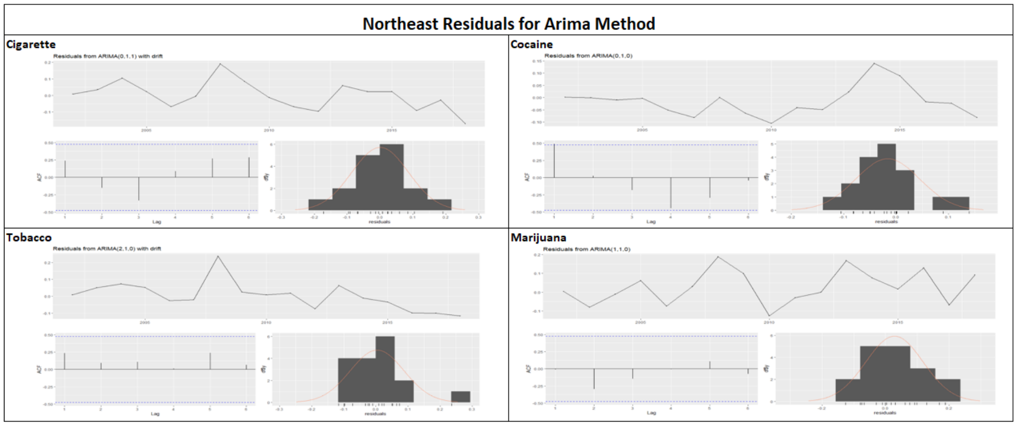
7.2 Appendix B: Time Series Model

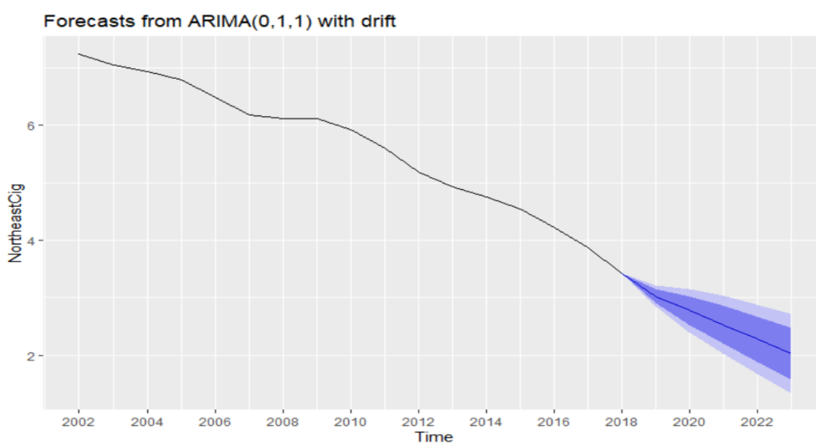


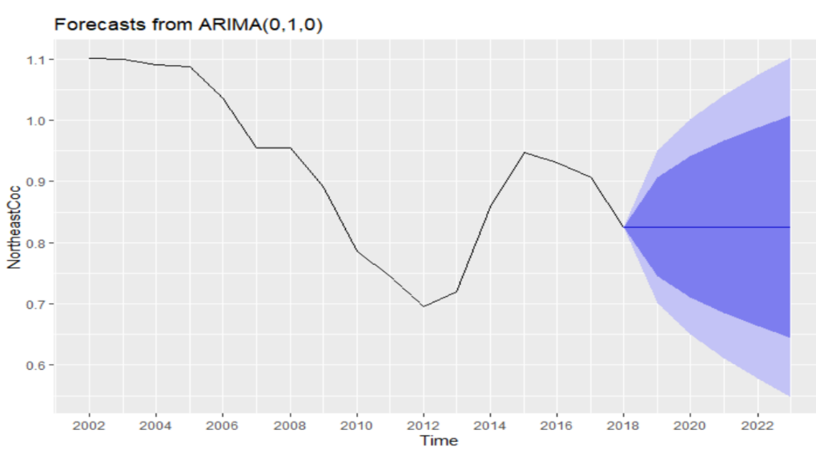


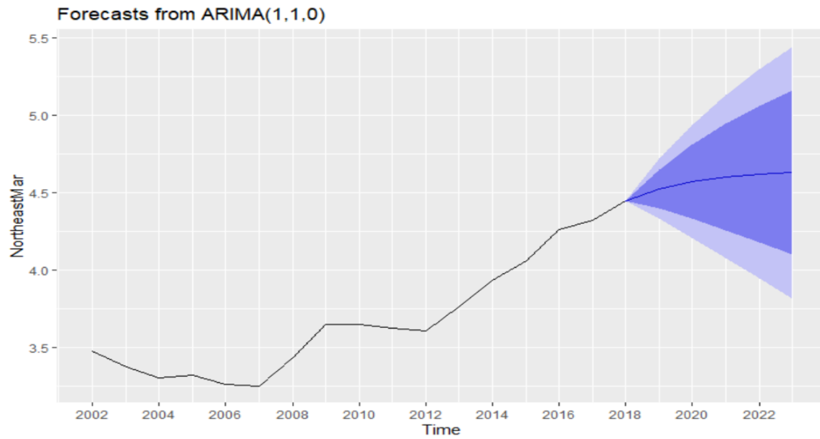


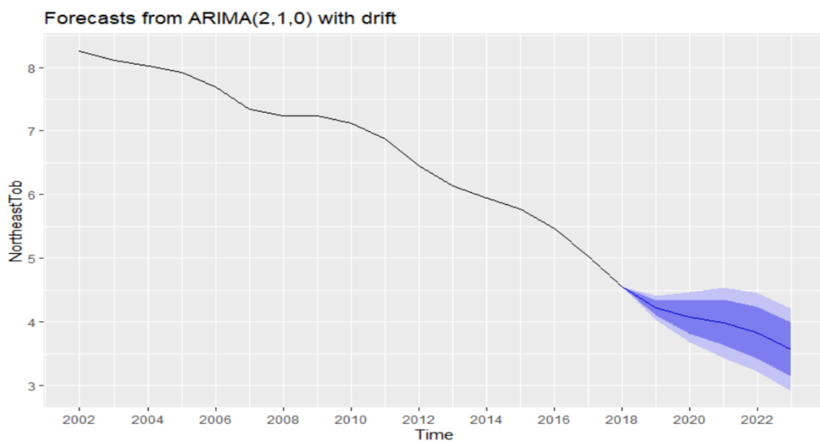
Northeast Focus








Forecasts from ARIMA(1,1,0)
NortheastMar
Time


Forecasts from ARIMA(2,1,0) with drift
NortheastTob
Time

8 Bibliography

"Alcohol and Drugs: Marijuana Laws." *IIHS-HLDI Crash Testing and Highway Safety*, 2022, www.iihs.org/topics/alcohol-and-drugs/marijuana-laws-table

Australian Government Department of Health. "Types of Drugs." *Australian Government Department of Health*, 21 June 2021, https://www.health.gov.au/health-topics/drugs/about-drugs/types-of-drugs

Bart, Austin Cory, et al. "CORGIS Datasets Project." Think.cs.vt.edu, https://think.cs.vt.edu/corgis/csv/

Gaither, Julie R., et al. "US national trends in pediatric deaths from prescription and illicit opioids, 1999-2016." *JAMA network open*, vol. 1, no. 8, 2018, pp. e186558-e186558, doi:10.1001/jamanetworkopen.2018.6558

Hyndman, Rob, and George Athanasopoulos. *Forecasting: Principles and Practice.* 2nd ed., OTexts, 2018.

Lee, Juhan, and Johannes Thrul. "Trends in Opioid Misuse by Cigarette Smoking Status among US Adolescents: Results from National Survey on Drug Use and Health 2015–2018." *Preventive Medicine*, vol. 153, Dec. 2021, p. 106829, doi:10.1016/j.ypmed.2021.106829

McCance-Katz, Elinore. "The National Survey on Drug Use and Health: 2017." *samhsa*, 2017, https://www.samhsa.gov/data/sites/default/files/nsduh-ppt-09-2018.pdf

Montgomery, Douglas, et al. *Introduction to Time Series Analysis and Forecasting.* 2015.

Moore, Lisa D., and Amy Elkavich. "Who's Using and Who's Doing Time: Incarceration, the War on Drugs, and Public Health." *American Journal of Public Health*, vol. 98, no. Supplement 1, Sept. 2008, pp. S176–S180, doi: 10.2105/AJPH.2007.126284

"National Survey on Drug Use and Health (NSDUH-2015)." *Data.gov*, Substance Abuse Mental Health Services Administration, 12 Mar. 2021, catalog.data.gov/dataset/national-survey-on-drug-use-and-health-nsduh-2015

Pentz, Mary Ann, et al. "A multicommunity trial for primary prevention of adolescent drug abuse: Effects on drug use prevalence." *Jama*, vol. 261, no. .22, 1989, pp. 3259-3266.doi:10.1001/jama.1989.03420220073030

"State Data Tables and Reports from the 2019-2020 NSDUH — CBHSQ Data.", *samhsa*, www.samhsa.gov/data/nsduh/state-reports-NSDUH-2020

Substance Abuse and Mental Health Services Administration. "National Survey on Drug Use and Health — CBHSQ." *samhsa*, 2017, www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health

"The Complete Guide to Time Series Analysis", *Data Science Central*, 1 Feb. 2021, www.datasciencecentral.com/the-complete-guide-to-time-series-analysis/

UNODC, IDS. "World drug report." *United Nations*, 2009. https://www.unodc.org/unodc/en/data-and-analysis/WDR-2012.html

Volkow, Nora D. "America's addiction to opioids: heroin and prescription drug abuse." *Senate Caucus on International Narcotics Control*, vol. 14, 2014, pp. 1-16. https://www.nih.gov/sites/default/files/institutes/olpa/20140514-senate-testimony-volkow.pdf