A Comparison of the United States' Coastal Climates Through Time

Joseph Foley
BSc Statistics
University of Reading
March 2013

Abstract

Utah State University has made available the climate data for all US monitoring stations under the National Weather Service's Cooperative Observer Program. This study makes use of the maximum daily temperatures recorded at these stations in order to compare the United States' coastal climates through time. The rationale behind this aim is to see what affect the Atlantic and Pacific Oceans have on the US climate. Temperatures are taken from the States of Washington, Oregon, California, Maine, Virginia and Florida over a period of nine years (2000 – 2008).

Two analyses were conducted. First an exploratory analysis was conducted which consisted of basic statistical tests such as t-tests and F-tests. The second analysis utilised time series methods. Seasonal autoregressive integrated moving average models were fitted to the temperature data of each State. The results were mostly inconclusive due to the small sample size used in this study. However the analyses did suggest that the East Coast had a much more varied climate than the West Coast. It is believed that this is possibly due to the Gulf Stream's changing nature as it heads northward along the East Coast.

Contents

CHAPTER 1
Introduction

The United States of America is a nation that contains many different climates. This is due to its size, location and the geographical features within. The US mainland (the USA excluding Alaska and Hawaii) is 2,680 miles wide and 1,582 miles long. It is the 3rd largest country in the world and so it is to be expected that there will be differences in the climate across the country. The country's geographical position also contributes to its varied climate. The mainland occupies the latitudes between 24° and 49° which has resulted in Mediterranean-like climates in the south and Scandinavian-like climates in the north. Even the geographical features such as the mountain ranges and forests influence the various climates in the United States and they can also be consequences of these climates (for example deserts and swamps).[1]

Not only does the climate vary across the country but it also changes with time. The ever changing Gulf Stream in the Atlantic Ocean and the weather phenomenon known as El Niño in the Pacific Ocean have been altering the weather systems in the US mainland for centuries. Mankind may also have contributed to the changing climate. The effects of Global Warming are being intensely studied by climatologists and it is possible that it has had an effect on the US climate.

Since 1890 the National Weather Service (NWS) Cooperative Observer Program (COOP) has been recording daily climate data in hundreds of locations across the United States. Utah State University provides this climate data freely to the public (http://climate.usurf.usu.edu/products/data.php). They provide daily maximum temperature, daily minimum temperature, 24 hour precipitation totals and snowfall totals at each observatory. This study makes use of this data and more specific details are given in Chapter Two.

There are many investigations one could undertake with the data recorded by COOP. For example one could study the effects each hurricane season has on the climates of the southern states. One could compare how much the climate has changed in the north of the country to the south. Another investigation could be how the relationship between temperature and precipitation has changed through time. Such ideas are not new and there have been numerous reports published on the subject of the US climate.

The main aim of this study is to compare the temperatures of the West Coast Region to the temperatures of the East Coast Region. The purpose of doing so is to see how the US mainland temperature is affected by the Atlantic and Pacific Oceans through time. The coastal waters surrounding the United States are ever changing; this is why the comparisons being made must factor in time. The Gulf Stream in the east is the northward flow of warm water along the east coast but this flow never maintains a constant rate and even the direction of the flow can change (*BBC Weather Centre, 2009)*. El Niño is the collection of warm water in the East Pacific Ocean due to trade winds; it is an event that occurs once every three to seven years (*SCRIPPS , 1997*). It mostly affects Latin America but it could even influence the West coast of the United States. These oceanic events are the rationale behind this time series investigation.

Whilst the main aim is to compare the West Coast to the East Coast there is still more that can be learned from this investigation. The Gulf Stream and El Niño may affect the north and south differently, for instance thermal imaging shows that less warmth from the Gulf Stream reaches Maine than many of the other States (see *figure 1.1*). Something similar may occur with El Niño and so a secondary aim of this study will be to compare the northern coastal temperatures to the southern coastal temperatures.

[1] USA land statistics from worldatlas.com

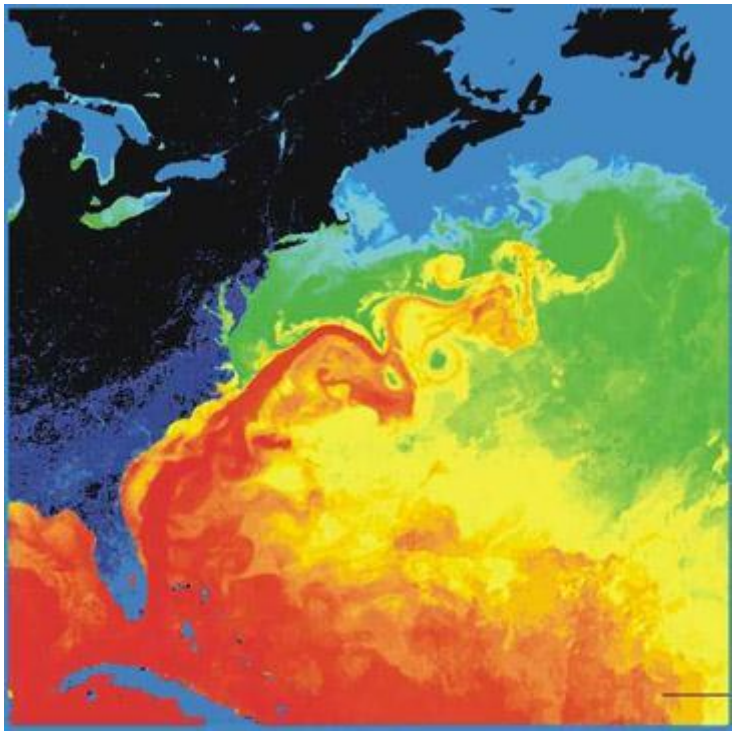


Figure 1.1 *A satellite thermal image of the Gulf Stream provided by NASA.*

Only one variable in this investigation is studied (temperature) however multiple comparisons are made. Specifically the coasts are compared by their temperature variability, extreme values, temperature distribution, long term trend and by the nature of their seasonality. Often this study will make use of the word “climate” which encompasses many weather features. However the use of the word “climate” in this study explicitly refers to the nature of the temperatures within the United States.

Since temperature data was only available on a per State basis this study investigates the temperatures of six States. The West Coast States are Washington, Oregon and California. These States are in the study because they are the only US mainland States on the West coast. The East Coast States are Maine, Virginia and Florida. These States were chosen because they are spaced evenly apart and so should give good coverage of the East coast. It is also worth mentioning where each of the observations was taken. In the West the recordings were made at the airports of Seattle, Portland and Los Angeles. In the East the recordings are from the airports of Augusta, Norfolk and Miami. All of the observatories are within 50 kilometres of the sea. *Figure 1.2* is a map showing where the observations were taken.

There were two stages to the analysis and each has its own chapter. The exploratory data analysis was conducted first and it consisted of numerous well known statistical tests and processes. It is presumed that the reader is already familiar with the tests used in the exploratory analysis, so they will not be explained in great detail. The exploratory analysis was conducted to provide some insight into the data so that an ideal method of time series analysis could be applied. The main time series analysis consisted of fitting an Auto Regressive Integrated Moving Average (ARIMA) model to the data for each State. This method of analysis will be explained in detail and its use will be justified.

Figure 1.2 *A map of the USA showing where the temperature observations were taken (dots indicate an observatory). Source: Wikimedia.com (it has been modified to include observatory locations).*

1.1 Overview

Chapter Two features the exploratory data analysis. The data is given a full description and the data distributions of each State are compared. Various tests are conducted such as t-tests, F-tests and correlation tests. The purposes of these tests are to compare mean temperature, temperature variability and show if there is any relationship between any of the States' temperatures. Box plots and scatter plots also accompany the analysis.

Chapter Three describes the time series analysis methods that can be employed. The chapter covers basic inspection of time series plots and the various models that can be applied to time series data. Much detail is given on how to select the parameters of a model, how to estimate these parameters, check if a model adequately fits the data and how the model can be used to forecast future values in the series.

Chapter Four features the time series analysis. The methods described in Chapter Three are applied to the US climate data. A simple inspection of the time series plots is given and detailed explanations are given on how the data is modelled for each State. These models are also interpreted and compared with one another. Finally forecasts are generated by each of the models and these are carefully examined.

Chapter Five consists of a discussion about the entire investigation. The results and conclusions from both analyses are summarised. The investigation is also evaluated and suggestions are made for its improvement.

At the end of this research report there is a references section and an appendix section that contains results from the analysis that were not included in the main analysis chapters.

CHAPTER 2
Exploratory Data Analysis

In this chapter the data used in this investigation is fully described. Following this are five individual sections covering distributions, Box-plots, Means, Variances and Correlations. Finally an overall conclusion of the exploratory analysis is given at the end of the chapter.

2.1 The Data

As stated in the introduction this study has used temperature data collected at the airports of Seattle (Washington), Portland (Oregon), Los Angeles (California), Augusta (Maine), Norfolk (Virginia) and Miami (Florida). The National Weather Service Cooperative Observer Program (COOP) was responsible for collecting this data and Utah State University was responsible for making it available.

The data used in this study is the maximum daily temperatures (in degrees Fahrenheit) recorded by COOP and it covers nine years (2000 to 2008). Maximum temperature was chosen instead of minimum temperature for no reason in particular. There were a few missing observations which were dealt with. If a day had a missing temperature observation then an average of the temperatures from the day before and the day after would be put in the missing observation's place.

Daily observations were not used directly in the analysis; instead monthly averages were taken and used in the analysis. There were several reasons for doing this. The first reason is that it reduces a lot of the random noise (variability) in the data; it is very difficult to determine with precision the factors that have led to a particular day's temperature. This is not so with the temperature of a month and is the reason why so many climatological studies are based on monthly averages. The reduction in random noise also meant that the models and forecasts made in this study are more reliable. The data is also much more manageable in monthly form as a full temperature cycle (which consists of one year) is made up 12 even units instead of 365 units (where each month consists of 28,30 or 31 days). Leap years also pose no problems when monthly temperature averages are examined as opposed to daily temperatures. However yearly averages were used for the scatter plots and this will be explained in that section.

For the time series analysis the data had to undergo some transformations and this will be discussed further in that chapter. Otherwise the data used in the exploratory analysis has not been transformed.

2.2 Data Distributions

It is always necessary to examine how the data is distributed in order to determine how best to proceed with the analysis. By inspecting the temperature distributions of each State, useful information can be revealed such as whether one of the coasts typically has more cold days in a year than hot days. Insight into the range of temperatures that each coast experiences could also be gained by examining the temperature distributions.

To compare the distributions a histogram was constructed in Minitab using the monthly average temperatures of each State. An outline of the normal distribution is placed over the histogram and is used as a reference to see whether the data is normally distributed. *Figure 2.1* Displays these histograms.

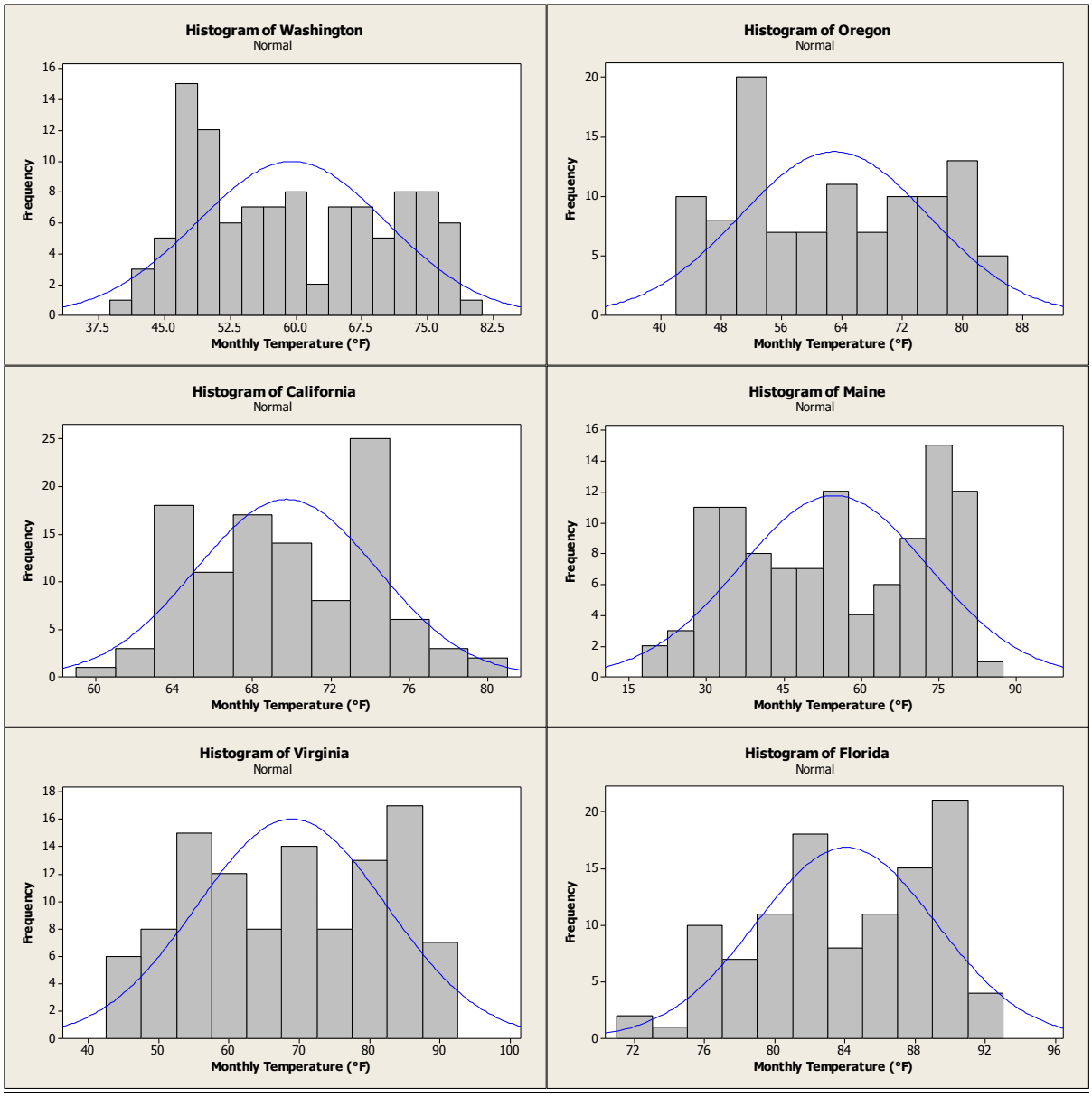


Figure 2.1 *Histograms for the six States' monthly temperature data with normal distribution outline.*

None of the data sets appear to be normally distributed although California does seem to best fit the normal distribution curve out of the six States. The West Coast States of Washington and Oregon have positively skewed distributions so these States typically have more days that are colder than the average than days that are hotter than the average. Florida on the East Coast has a negatively skewed distribution and so it typically has more days that are hotter than the average than days that are colder than the average. It is difficult to determine how the temperature distributions of Maine and Virginia are skewed. However it does appear that West Coast States and East Coast States have dissimilar temperature distributions, this suggests that there may be differences between their climates. It seems as if the West Coast has more cold days than the East Coast which has a more evenly distributed temperature throughout the year.

These skewed distributions indicate that the data will have to undergo a transformation of sorts if time series models are to be fitted to the data. This is because most time series models can only be fitted to data that is normally distributed. The transformations required may reveal more about the differences between the two coasts.

2.3 Box Plots

A box and whisker plot has been constructed for each of the states so that further interpretations on the temperature distributions can be made. The Box plot was created in Minitab and is shown by *Figure 2.2*.

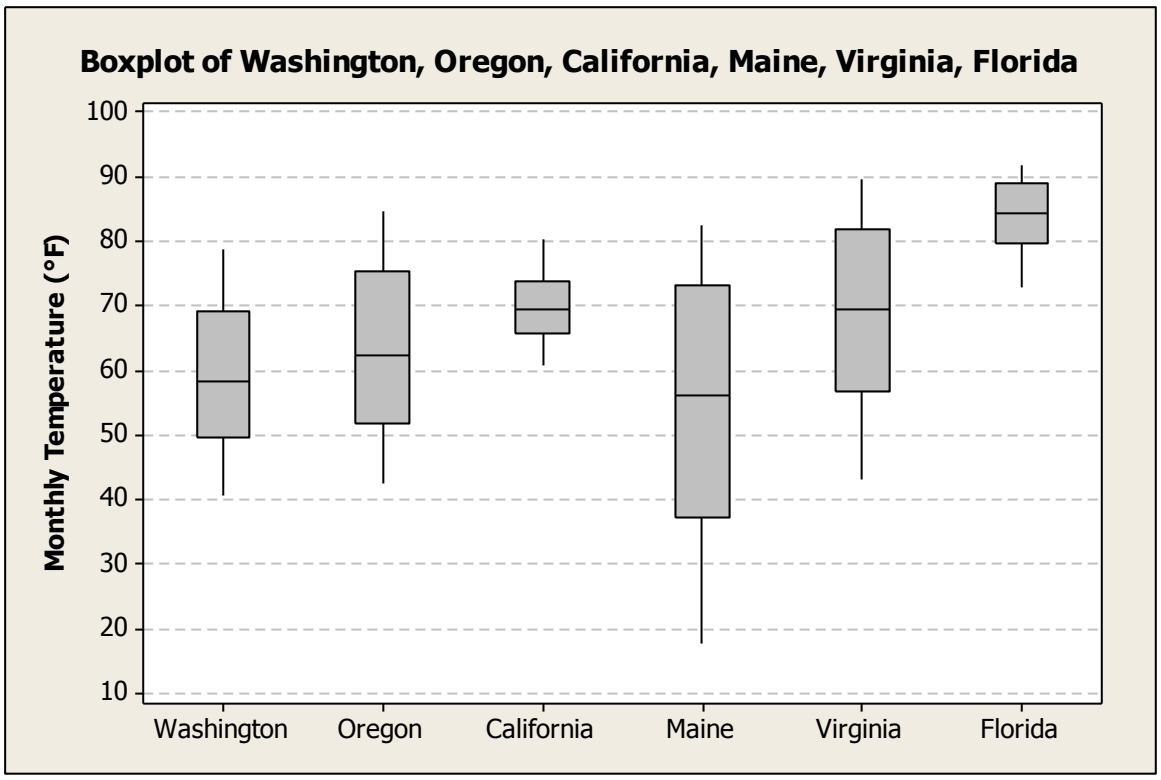


Figure 2.2 *Boxplot of the six States' monthly temperatures* (°F).

The box plot shows that there are no outliers for any of the states so no state has distinguished itself by having an unusually cold or hot month. This may be because there was no climatological event that seriously influenced the temperature of the United States during the nine years that this study is investigating. Since there are no outliers no procedure was needed to minimise their impact on the rest of the analysis.

It can be seen that Maine has the greatest temperature range and interquartile range out of the six states. So it is plausible that Maine has the most variable climate out of the states being studied; this could be due to the Gulf Stream being more variable that far north on the west coast. Augusta's latitude (Maine) is similar to Seattle's and Portland's (Washington and Oregon) yet it can get much colder there. Perhaps the Gulf Stream has almost no influence on Maine's climate.

The West Coast States appear to have smaller temperature ranges than the East Coast States. This may imply that the west coast of the United States has a more constant temperature than the east coast. Perhaps this is due to the ever changing nature of the Gulf Stream on the east coast.

The southern states of California and Florida have the smallest ranges which is to be expected as they are closer to the equator than the other states.

2.4 Comparison of Mean Temperatures (Paired t-tests)

The purpose of a t-test is to determine whether there is a difference between the means of two variables. A paired t-test is conducted when the samples are related in such a way that each observation in one sample can be paired with another observation in the other sample. In this case each monthly average temperature in a state can be paired with another monthly average temperature in another state (e.g. Jan' 06 in Maine with Jan' 06 in Oregon.)

A z-test could also be used to compare means but it cannot be used in this case as the true population variances are not known. It is also necessary for each variable to be normally distributed and each observation must be independent from all others. It has already been seen that the data is not normally distributed. It is also logical to assume that each observation is not completely independent because the temperature of one month could influence the temperature of the next. Therefore t-tests cannot be conducted on the entire data set. However it would be acceptable to conduct a t-test on a specific month of the year as each observation would be independent and follow a normal distribution (e.g. the temperature of each January does not have any influence on the temperature of any other January).

The reason for doing these tests is to see whether they can reveal anything about the eastern and western coastal climates. Every combination of the six States has been considered in this analysis and paired t-tests were conducted for each month of the year. This analysis was conducted in excel using the "T.TEST" formula.

Table 2.1 and *table 2.2* show the p-values of each t-test for the months of January and July. A p-value is the two tailed probability that the mean of one State's temperature is not significantly different from the mean of the other State's temperature. It is conventionally regarded that if a p-value is lower than 0.05 then there is significant evidence to suggest that there is a difference between the means. All of the p-value tables for each month can be seen in appendix A but only the most interesting tables are shown here. The top right box shows the p-values for the West Coast States versus the East Coast States and the highlighted cells are p-values which are greater than 0.05.

The January and June tables do not necessarily represent the tables of the other months. However they best illustrate the following interpretations.

In some months there was no significant difference in mean temperature between the Western States, yet this never occurred for the Eastern States. Between Maine, Virginia and Florida there was always a significant p-value (the mean monthly temperatures are significantly different from each other for each month). This may imply that there is a more diverse climate along the East Coast.

The t-tests revealed that for some months the Eastern States of Maine and Virginia did not have mean temperatures that were significantly more different than the Western States. However Florida always had a significantly different mean temperature. This is due to Florida being a much hotter State than the other five States.

November was the only month where every p-value was significant. This allows one to conclude that during that month each state had its own mean temperature that was significantly different to the mean temperatures of the other states.

Table 2.1 *P-values for each t-test conducted for every two way combination of the six States during the month of January.*

JANUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.577	0.000	0.000	0.114	0.000
Oregon	0.577	-	0.000	0.000	0.077	0.000
California	0.000	0.000	-	0.000	0.000	0.000
Maine	0.000	0.000	0.000	-	0.000	0.000
Virginia	0.114	0.077	0.000	0.000	-	0.000
Florida	0.000	0.000	0.000	0.000	0.000	-

Table 2.2 *P-values for each t-test conducted for every two way combination of the six States during the month of June.*

JUNE	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.000	0.056	0.015	0.000	0.000
Oregon	0.000	-	0.177	0.972	0.000	0.000
California	0.056	0.177	-	0.091	0.000	0.000
Maine	0.015	0.972	0.091	-	0.000	0.000
Virginia	0.000	0.000	0.000	0.000	-	0.001
Florida	0.000	0.000	0.000	0.000	0.001	-

Not much can be learnt by comparing the means. The tables show that for most of the months each State had a significantly different mean temperature to most of the other States which should be expected of six different locations that are separated by hundreds of miles. It is also difficult to make any firm conclusions with this data due to the sample size. There were just nine observations for each test which is generally considered to be very small in regards to the study of climatology.

2.5 Comparison of Temperature Variability (F-tests)

An F-test is used to determine if there is significant evidence to suggest that the variances of different variables are different. It can be used on multiple variables and the result of the test would determine if at least one of the variables had a unique variance. To be more specific this study has conducted an F-test between a maximum of two variables at a time. Every combination of the six States has been considered in this analysis and F-tests were conducted for each month of the year. This analysis was conducted in Excel using the "F.TEST" formula.

As in a t-test the data must be normally distributed with each observation being independent so this analysis will only focus on one specific month at a time. The purpose of this test is to reveal whether there is a difference in temperature variability between the two coasts. It is possible that one of the coasts will have a more constant temperature than the other.

Table 2.3 and *table 2.4* show the p-values of each F-test for the months of February and July. A p-value is the two tailed probability that the variance of one State's temperature is not significantly different from the variance of another State's temperature. It is conventionally regarded that if a p-value is lower than 0.05 then there is a significant evidence to suggest that there is a difference between the variances. All of the tables for each month can be seen in Appendix B but only the most interesting tables are shown here. In the top right box are the p-values for the West Coast States versus the East Coast States and the highlighted cells are p-values which are less than 0.05.

Table 2.3 *P-values for each F-test conducted for every two way combination of the six States during the month of February.*

FEBRUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.863	0.103	0.019	0.012	0.192
Oregon	0.863	-	0.141	0.027	0.018	0.253
California	0.103	0.141	-	0.418	0.325	0.727
Maine	0.019	0.027	0.418	-	0.859	0.250
Virginia	0.012	0.018	0.325	0.859	-	0.187
Florida	0.192	0.253	0.727	0.250	0.187	-

Table 2.4 *P-values for each F-test conducted for every two way combination of the six States during the month of July.*

JULY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.896	0.753	0.502	0.672	0.010
Oregon	0.896	-	0.854	0.588	0.770	0.014
California	0.753	0.854	-	0.719	0.913	0.021
Maine	0.502	0.588	0.719	-	0.802	0.046
Virginia	0.672	0.770	0.913	0.802	-	0.027
Florida	0.010	0.014	0.021	0.046	0.027	-

In most cases there was no significant evidence to suggest that there was difference in temperature variance between any of the States with the exception of Florida which often had a significantly different variance to many of the States in many of the months.

There is little else that can be derived from the results of these F-tests but it might be worth noting that in the months of February, October and December there was significant evidence to conclude that there was a difference in temperature variability between some of the West and East Coast States. It is possible that there were so many insignificant results because of the small sample size that was used. However there is at least some evidence that in some of the locations studied there is a more constant temperature than some of the others.

Despite the mixed results it is still worth directly comparing the variances themselves. *Table 2.5* displays the average monthly variance for each of the states.

Table 2.5 *Average monthly variance for each State's temperature data*

State	Average variance
Washington	5.048
Oregon	5.477
California	5.405
Maine	9.939
Virginia	9.736
Florida	2.405

The table shows that Florida has the lowest average variance (it is also known that Florida's variance is significantly different to the other State's variances for most of the months.) The Gulf Stream originates around Florida and may be the reason why this State has the most constant temperature out of the other States being studied. Further north on the east coast the Gulf Stream is constantly changing and could be part of the reason why Maine and Virginia have larger temperature variances than Florida.

Maine and Virginia have similar average variances which appear to be quite different to those of the Western States yet for only a few months there was enough evidence to suggest that their variances were different to the Western States. This may have happened because the small sample size that was used (nine observations).

The West Coast States all have similar average variances; this may be due to the Pacific Ocean having an equal effect on temperature variability across the entire coast.

2.6 Scatter Plots and Correlations

A scatter plot allows one to visually compare an observation of one variable directly with the corresponding observation of another variable. In this case the scatter plots compare the average temperature of a State for a specific year with the average temperature of another State for that same year. The nine years' worth of data are represented as nine points on each scatter plot which make it possible to determine if there is any relationship between the temperatures of each state. Yearly data was used because it eliminated the seasonality and is easier to view than monthly data (there is one plot instead of twelve).

A regression analysis was attempted but a requirement of regression analysis is that each observation is independent like in the t-tests and F-tests. The observations are not independent so the analysis had to be conducted on specific months again. The analysis produced no significant p-values due to the small sample size. Since the regression analysis was inconclusive it is not possible to make firm conclusions about the nature of the relationship between the temperatures of the States, however it still may be possible to learn something from the scatter plots of the data.

Figure 2.3 is a scatter plot of Washington vs. Oregon and a scatter plot of Virginia vs. Florida; these plots are being shown because they are the most interesting. *Figure 2.4* is a matrix plot that shows a scatter plot for every two-way combination of the six States, this is being shown for completeness.

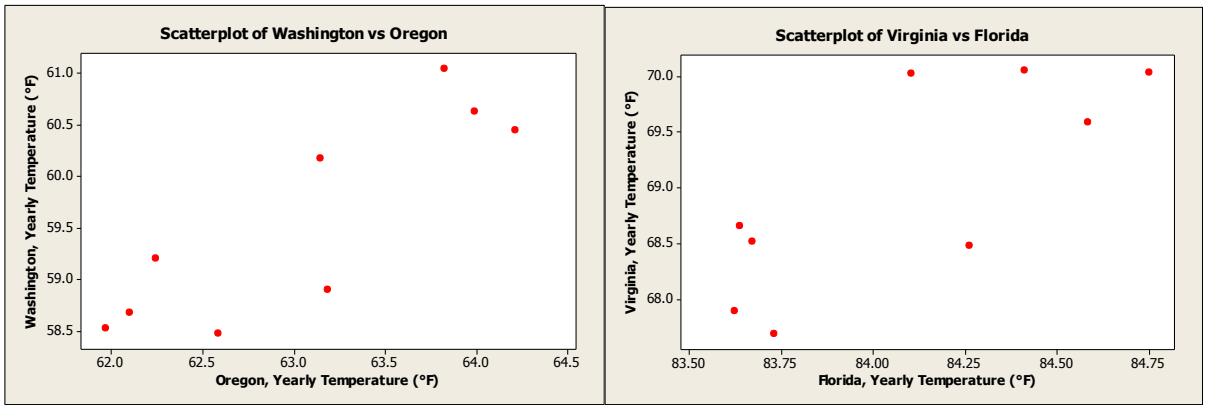


Figure 2.3 *Scatter plots showing Washington vs. Oregon and Virginia vs. Florida using average yearly temperatures.*

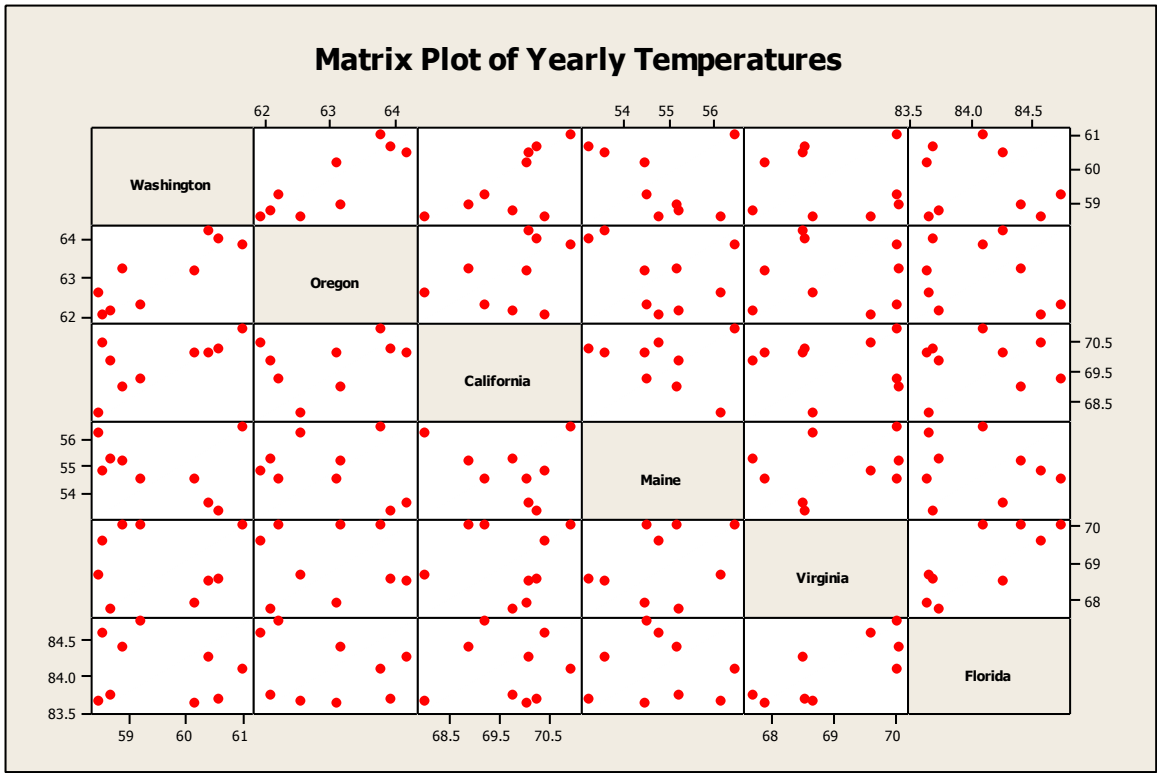


Figure 2.4 *A matrix plot which contains the scatter plots for every two way combination of the Six States.*

The matrix plot shows that there is generally no relationship between the temperatures of the six States. There are some exceptions though. There is a clear positive relationship between the yearly temperatures of Washington and Oregon. This is due to the fact that the locations of where the observations were made (Seattle and Portland) are very close to each other. Washington also appears to have a positive relationship with California but Oregon does not which is strange because Oregon is actually closer to California than Washington is. Perhaps this would not be the case if the sample size was larger.

On the East Coast it appears that Virginia has a positive relationship with Florida. Maine does not have a close relationship with the other East Coast States; this may be because this State has a large temperature range and variance. However Maine may well have a negative relationship with the West Coast States of Washington and California.

To quantify these relationships some correlation coefficients have been calculated. A correlation coefficient shows how strongly linked two variables are and whether there is a positive or negative relationship. *Table 2.6* shows the correlations for each two way combination of the six States. The top number in a cell is the correlation coefficient and the bottom number is its corresponding p-value which is the probability that the correlation coefficient is significantly different from zero.

Only two of the correlation coefficients have significant p-values (p-values less than 0.05) so no firm conclusions can be made about the relationships between the temperatures of the six States. More data is needed. However since the p-values indicate that many of the correlations could be zero and since most of the State combinations have a random looking scatter plots then it could be possible that the temperature of the West Coast does not influence the temperature of the East Coast. So if there is a hotter than average year in the west then there may not necessarily be a hotter than average year in the east.

Table 2.6 *Correlation coefficients for each two way combination of the six States with corresponding p-values.*

Correlations	Washington	Oregon	California	Maine	Virginia	Florida
Washington	_	0.861	0.656	-0.290	-0.039	-0.203
		0.003	0.055	0.450	0.922	0.601
Oregon	0.861	_	0.361	-0.321	-0.045	-0.235
	0.003		0.340	0.400	0.908	0.543
California	0.656	0.361	_	-0.252	-0.010	0.064
	0.055	0.340		0.513	0.979	0.869
Maine	-0.290	-0.321	-0.252	_	0.311	-0.063
	0.450	0.400	0.513		0.416	0.872
Virginia	-0.039	-0.045	-0.010	0.311	_	0.778
	0.922	0.908	0.979	0.416		0.013
Florida	-0.203	-0.235	0.064	-0.063	0.778	_
	0.601	0.543	0.869	0.872	0.013	

2.7 Conclusions

The major problem that this analysis faced was that the temperature observations were not independent. Only individual months could be properly tested which reduced the sample size to nine. As a result of this much of the exploratory analysis was inconclusive. However the exploratory analysis still indicates the possibility that there may be differences between the climates of the East and West coasts.

The temperature distributions of each State revealed that the West coast typically had more colder than average months than the East coast. The box plot showed that of the nine years being studied there were no months of extreme temperature, they also revealed that Maine had the largest temperature range and that it could get colder in Maine than anywhere else. This may be because the Gulf Stream has little influence on Maine's temperature.

The t-tests suggested that there may be a more diverse climate along the east coast, so it is possible that the Gulf Stream does not influence the entire coast equally. The F-tests and variances agreed with this conclusion as they showed that Florida which is at the source of the Gulf Stream had the most constant temperature, whereas Maine had the most variable temperature. Otherwise the F-tests were largely inconclusive and did not reveal much about the differences between the coasts.

The scatter plots suggested that there was no relationship between the yearly temperatures of the two coasts.

This analysis revealed some minor details about the temperatures of the two coasts. An important discovery was that the data is not normally distributed so some transformations had to be employed during the time series analysis. One of the main limitations of the exploratory analysis is that it did not consider the data as a series through time; so much more will be revealed about the climates of the two coasts in the time series analysis (Chapter Four).

CHAPTER 3
Time Series Methodology

A time series is a set of observations that have been recorded sequentially through time, so there is a natural order to the data that is governed by the flow of time. Time series analysis is a very expansive field in statistics and is only partially covered in this chapter. The methods used in this investigation are discussed in this chapter as well as some others; the source of this information primarily comes from Chatfield (2003). For more information on time series analysis the reader should consult the references page.

3.1 Purpose of Time Series Analysis

Chatfield identifies four different objectives of time series analysis. These are description, explanation, prediction and control.

By using time series analysis one can describe the time related features of a data set such as the nature of its seasonality or the presence of a trend. Describing data in this way can help aid the understanding of various time related phenomena (such as the seasons of the USA). The structure of a time series can also be described through the use of modelling. Various time series models are able to reveal the data's core components that make up its structure.

The explanation of many events can be aided by the use of time series analysis. When multiple time-related variables are being examined it is possible that one may explain the variation observed in another (e.g. temperature variation explaining ice cream sales). Sometimes an unobservable process can be indirectly detected by time series analysis (e.g. Gravity waves using a laser interferometer (*LIGO Scientific Collaboration,* (*1997*))).

When time series models are made it becomes possible to forecast future events such as the likelihood of rain in a particular month. Prediction via time series modelling is also a very important tool used by many sales based businesses.

Time series analysis can also be used to aid in the control of certain processes. Many manufactures constantly monitor time series graphs in order to maintain the best possible quality control over their products. Extensive models are also made to develop control strategies. Even governments manage their fiscal policies based on economic time series data.

3.2 Understanding the Data

Whilst the core focus of this chapter is on modelling the data via advanced time series methods it is always necessary to conduct a few minor checks on the data. Modelling techniques often depend on the nature of the data so the analyst needs to be aware of the kind of data that they have. For instance the analyst must be sure if their data is continuous or discrete.

Before going forward one must be aware of three key features of a time series. They are trends, seasonal variation and cyclic variation. Chatfield defines a trend to be "a long-term change in the mean level", an example of this could be the Earth's global temperature declining as it heads into an ice age over the course of thousands of years. What constitutes "long-term" is mostly a subjective matter though. A declaration of what is "long-term" should be based on in-depth knowledge of the subject being analysed.

Within a set period of time the mean level of a time series can change and then revert back to its original value, this is known as seasonal variation or seasonality. An example of this is the sale of air conditioning units which typically sell in higher numbers in the summer than in the winter, so in this case a full seasonal cycle is completed in a year. Cyclic variation is similar to seasonal variation but it occurs over a much longer period of time. An example of this can be seen in the economy which experiences growth for a number of years (each year being one seasonal cycle) and then it enters a recession for a few years before returning to growth.

A simple inspection of the time series plot may reveal whether a trend is present. Seasonality is also an important factor to be aware of, as are long term cyclical patterns. The presence of any of these factors will change how a time series analysis is approached; this will be discussed further in section 3.3.

If there is no trend, seasonal variation or cyclical variation and there is no systematic change in the mean then the time series is said to be stationary. Supposedly there is no such thing as a "stationary time series" but it is still a useful concept as various time series models require the data to be "stationary" before they are fitted. So it is important for the analyst to know if their data is stationary and what they can do to make it so if it isn't. Making a time series stationary will be discussed in section 3.3.

As discussed in the exploratory analysis the analyst may need to decide whether a transformation needs to occur. A time series plot can sometimes show whether the variance of a data set is increasing or decreasing. In such cases logarithmic or square root transformations must be done to stabilise the variance, this is because a stable variance is required by most time series models.

In most cases the data must be normally distributed if a time series model is to be fitted. If the data is not normally distributed then the analyst must decide on a transformation to make it so. Transformations may also deal with outliers or extreme values which can also be revealed by a time series plot.

3.3 Differencing

Before discussing the time series modelling process it is necessary to explain how non stationary data can be made stationary. In order to make a time series data set stationary a process called differencing is used. To difference a data set one must form a new data set by subtracting each observation from a previous observation. Equation (3.1) demonstrates this.

$$y_t = x_t - x_{t-1} = \nabla x_t \tag{3.1}$$

Where $t = 1, 2 \ldots N$

Sometimes the data must be differenced again to make it stationary. This is known as second order differencing and is demonstrated by equation (3.2).

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1} = x_t - 2x_{t-1} - x_{t-2} \tag{3.2}$$

If a trend is present within the data it is usually sufficient to take a difference of one (so $y_2 = x_2 - x_1$, $y_3 = x_3 - x_2 \ldots y_N = x_N - x_{N-1}$). If seasonality is present within the data then the data should be differenced by the length of time it takes to complete a full seasonal cycle, for example if the data set consists of monthly observations then the data should be differenced by twelve (so $y_{13} = x_{13} - x_1$, $y_{14} = x_{14} - x_2 \ldots y_N = x_N - x_{N-12}$).

If the data contains a trend and seasonality then second order differencing is required. The data must be differenced by one and by the amount of time units in a seasonal cycle; this can be done in any order.

3.4 Time Series Modelling

The next five subsections focus on constructing time series models. The first section discusses some of the time series models that can be used. The remaining subsections explain the four steps of time series modelling which are parameter selection, parameter estimation, model checking and forecasting.

3.4.1 Types of Models

Before proceeding it is necessary to explain what a backwards shift operator is. The backwards shift operator is denoted by B^k. If it is placed next to a variable it is indicating that we are actually considering the variable k units of time ago. Equation (3.3) shows it in use.

$$B^k x_t = x_{t-k} \tag{3.3}$$

Equation (3.4) shows the backwards shift operator being used in a polynomial of order k.

$$\boldsymbol{\vartheta_k}(B) = 1 + \vartheta_1 B + \vartheta_2 B^2 + \cdots + \vartheta_k B^k \tag{3.4}$$

The remainder of this subsection will consist of the actual time series models.

The Autoregressive model

An autoregressive model expresses the current term in the series (x_t) as a linear arrangement of the previous terms. Equation (3.5) demonstrates an autoregressive model of order one.

$$x_t = \varphi x_{t-1} + z_t \tag{3.5}$$

Where φ is a parameter and z_t is an error term.

This is known as an AR(1) model. This can be generalised to multiple terms and equation (3.6) demonstrates an AR(p) model.

$$x_t = \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + z_t \tag{3.6}$$

Equation (3.7) shows how the model can be written using a backwards shift operator.

$$\boldsymbol{\varphi_p}(B) x_t = z_t \tag{3.7}$$

Where $\boldsymbol{\varphi_p}(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p$

The Moving Average model

A moving average (MA) model involves the mean of the time series and the residuals of previous observations. It plots the current value of a series against the current and previous error terms. The error terms are assumed to be independently and identically distributed under a normal distribution with a mean of zero. Equation (3.8) shows an MA model of order one.

$$x_t = \mu + z_t + \theta z_{t-1} \tag{3.8}$$

Where μ is the mean of the time series, z_t is a residual term and θ is a parameter.

This is otherwise known as an MA(1) model. There can of course be MA models with any number of terms, equation (3.9) shows an MA(q) model.

$$x_t = \mu + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \tag{3.8}$$

Often the μ term is left out of the model; particularly if the data set has been differenced (see section 3.3) and has become stationary, the mean of such a data set is often zero anyway. Equation (3.9) shows how the model can be shown using a backwards shift operator and without a μ term.

$$\begin{aligned} x_t &= (1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q) z_t \\ x_t &= \boldsymbol{\theta_q}(B) z_t \end{aligned} \tag{3.9}$$

Where $\boldsymbol{\theta_q}(B)$is a polynomial of order q in B.

The Autoregressive Moving Average (ARMA) model

A single time series model may contain both AR and MA terms; this model is known as an ARMA model. Equation (3.10) shows an ARMA model of order (*p, q*).

$$x_t = \varphi_1 x_{t-1} + \cdots + \varphi_p x_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q} \tag{3.10}$$

Here *p* corresponds to the number of AR terms in the model and *q* corresponds to the number of MA terms in the model. Equation (3.11) shows how the model can be written using backwards shift operators:

$$\boldsymbol{\varphi_p}(B) x_t = \boldsymbol{\theta_q}(B) z_t \tag{3.11}$$

It is important to note that the data set must be stationary for this model to be applied, if the data is not stationary and differencing is required then an appropriate model would be an ARIMA model.

The Autoregressive Integrated Moving Average (ARIMA) model

This model is used when differencing is required to make the data stationary, it is the differencing process that corresponds to the "integrated" term. Essentially the ARIMA model is just an ARMA model applied to data that has been differenced. The equation is the same but it must be clear that the data has been differenced so let:

$$w_t = \nabla^d x_t = (1 - B)^d x_t$$

Then an ARIMA model of order (*p, d, q*) is simply:

$$w_t = \varphi_1 w_{t-1} + \cdots + \varphi_p w_{t-p} + z_t + \theta_1 z_{t-1} + \cdots + \theta_q z_{t-q}$$

The *d* in ARIMA(*p, d, q*) refers to the order of differencing that occurred. This model is just an ARMA(*p, q*) model for w_t. Equation (3.12) is an example of an ARIMA(1,1,1) model.

$$w_t = \varphi_1 w_{t-1} + z_t + \theta_q z_{t-1} \tag{3.12}$$

Equation (3.13) shows the same model but using the original x_t variable (note that $w_t = x_t - x_{t-1}$).

$$x_t = x_{t-1} + \varphi_1 (x_{t-1} - x_{t-2}) + z_t + \theta_q z_{t-1} \tag{3.13}$$

The model can be expressed using a backwards shift operator in exactly the same way as the ARMA model but x_t is replaced with w_t.

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model

This model takes into account the seasonality present within a dataset. Unlike the previous models this model is multiplicative so it is generally easier to demonstrate it using polynomials of backwards shift operators, see equation (3.14). The model is denoted by ARIMA (*p, d, q*) x (*P, D, Q*)$_S$.

$$\boldsymbol{\varphi_p}(B)\boldsymbol{\Phi_P}(B^s)x_t = \boldsymbol{\theta_q}(B)\boldsymbol{\Theta_Q}(B^s)z_t \tag{3.14}$$

Where $\boldsymbol{\varphi_p}$ is s non-seasonal autoregressive polynomial of *order p,* $\boldsymbol{\Phi_P}$ is the seasonal autoregressive polynomial of order *P,* $\boldsymbol{\theta_q}$is the non-seasonal moving average polynomial of order *q,* $\boldsymbol{\Theta_Q}$ is the seasonal moving average polynomial of order *Q*. The number of time units in a seasonal cycle is denoted by S. The order of seasonal differencing is denoted by D whilst *d* corresponds to the order of non-seasonal differencing. Equation (3.15) shows how w_t can be expressed as a variant of the original variable x_t.

$$w_t = \nabla^d \nabla_s^D x_t \tag{3.15}$$

So if for example *d*=1, *D*=1 and S=12 then:

$$\begin{aligned} w_t &= \nabla \nabla_{12} x_t \\ &= \nabla_{12} x_t - \nabla_{12} x_{t-1} \\ &= (x_t - x_{t-12}) - (x_{t-1} - x_{t-13}) \end{aligned}$$

To demonstrate the model an example of an ARIMA $(1,0,0)\times(0,1,1)_{12}$ will be shown. This model has one non-seasonal AR term, one seasonal MA term and seasonal differencing of one (s=12 so the data is really differenced by 12). Example:

$$\begin{aligned} \boldsymbol{\varphi}(B)w_t &= \boldsymbol{\Theta}(B^{12})z_t \\ (1-\varphi B)w_t &= (1+\Theta B^{12})z_t \\ w_t - \varphi w_{t-1} &= z_t + \Theta z_{t-12} \\ w_t &= \varphi w_{t-1} + z_t + \Theta z_{t-12} \\ x_t &= x_{t-12} + \varphi(x_{t-1} - x_{t-13}) + z_t + \Theta z_{t-12} \end{aligned}$$

(Since $w_t = \nabla_{12} x_t = x_t - x_{t-12}$*)*

3.4.2 Parameter selection

Selecting the correct parameters to appear in the initial model requires careful inspection of the autocorrelation function (ACF) and partial autocorrelation function (PACF). Before explaining what those are it will be necessary to explain what autocorrelation is.

A correlation coefficient is a measurement of the linear association between two variables (see equation (3.16)). If say there are *N* pairs of observations for variables x and y then the sample correlation coefficient is:

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} \tag{3.16}$$

The correlation coefficient is positive if large values of x are paired with large values of y. The correlation coefficient is negative if large values of x are paired with small values of y.
(Note: $-1 \leq r \leq 1$)

The autocorrelation coefficient is very similar to the correlation coefficient but it only considers one variable that is examined through time. If there are *N* observations on a time series it is possible to make *N*-1 pairs of observations like so $(x_1, x_2), (x_2, x_3), \ldots, (x_{N-1}, x_N)$ where the observations in a pair are separated by one unit of time. One can consider the first observation of each pair to be a variable and the second observation to be another variable. With these two "variables" the correlation coefficient formula can be applied to give the autocorrelation coefficient. This is shown by equation (3.17).

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x}_{(1)})(x_{t+1} - \bar{x}_{(2)})}{\sqrt{\sum_{t=1}^{N-1}(x_t - \bar{x}_{(1)})^2 \sum_{t=1}^{N-1}(x_{t+1} - \bar{x}_{(2)})}} \tag{3.17}$$

Where $\bar{x}_{(1)}$ is the mean of the first observation in each pair (so the mean of the first *N*-1 observations) and $\bar{x}_{(2)}$ is the mean of the second observation in each pair (so the mean of the last *N*-1 observations). For a large *N,* there is a negligible difference between $\bar{x}_{(1)}$ and $\bar{x}_{(2)}$ so the above formula can be approximated. This is shown by equation (3.18).

$$r_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \tag{3.18}$$

The above formula is for when each observation is separated by one unit of time but it can be generalised to when there is separation by *k* units of time (this is known as lag *k*). Equation (3.19) calculates the sample autocorrelation between observations separated by *k* units of time:

$$r_1 = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2} \tag{3.19}$$

It is possible to plot a set of autocorrelation coefficients, the resulting graph is known as a correlogram. In this graph the sample autocorrelation coefficients are plotted against the lag to which they correspond. *Figure 3.1* is an example of a correlogram.

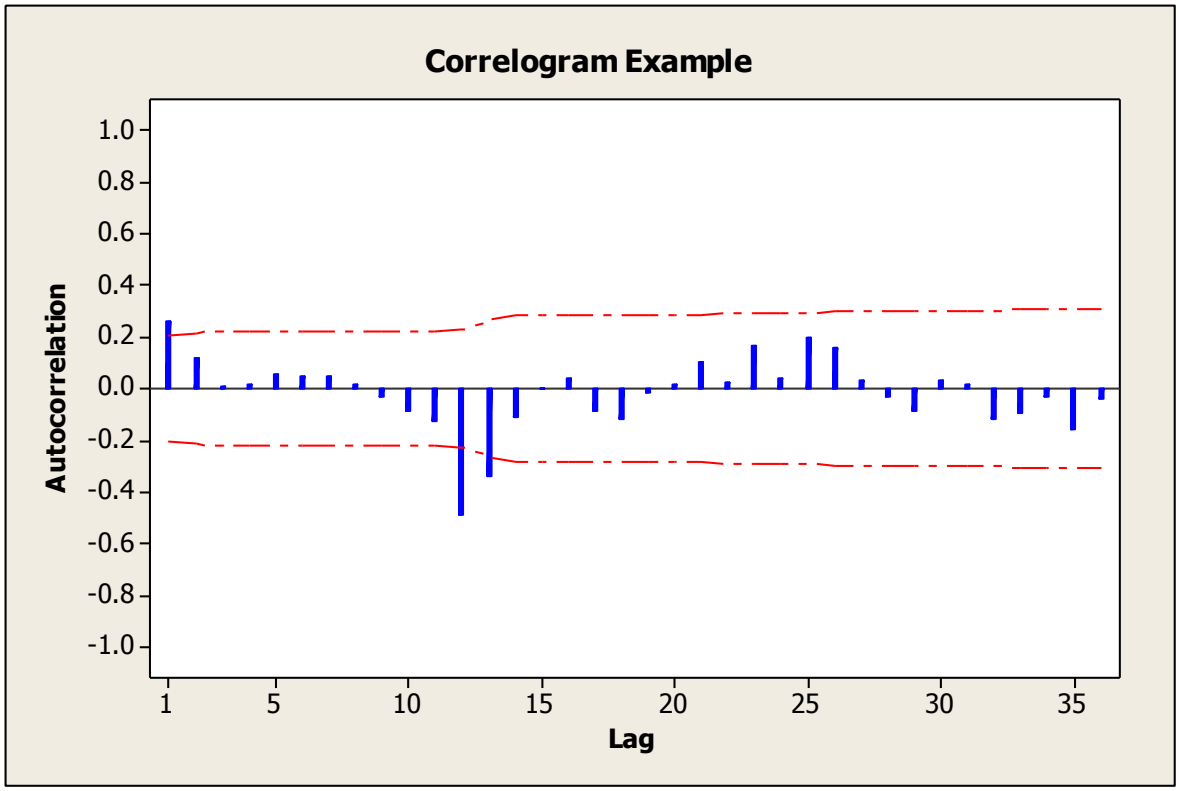


Figure 3.1 *Example of a correlogram.*

The correlogram is also known as the auto correlation function (ACF) and it is a useful tool for determining what parameters should be included in an ARIMA model. This will be discussed in detail later.

The partial autocorrelation function (PACF) is also needed when determining what parameters should be in an ARIMA model. A partial autocorrelation coefficient measures the association between two observations whilst accounting for the other intermediate observations. So if one wants to calculate the partial autocorrelation between observation one and observation eight the statistic will also take into account observations two to seven as well.

To find the partial autocorrelation coefficient between "variables" at time *t* and "variables" at time *t+k* a linear regression model must first be constructed. The response variable is the "variable" at time *t* and the explanatory variables are the "variables" from time *t+1* up to and including *t+k.* An example is given by equation (3.20).

$$x_t = \beta_0 + \beta_1 x_{t+1} + \beta_2 x_{t+2} + \cdots \beta_k x_{t+k} + \varepsilon_t \qquad (3.20)$$

This model can then be used to predict values for x_t . Residuals can be found by taking the difference between the predicted and observed values of x_t .The partial auto correlation coefficient for lag *k* can then be calculated by calculating the correlation between these residuals and the observed values of x_{t+k}. These coefficients can then be plotted in the same manner as the autocorrelation coefficients to give the PACF. *Figure 3.2* is an example of a PACF.

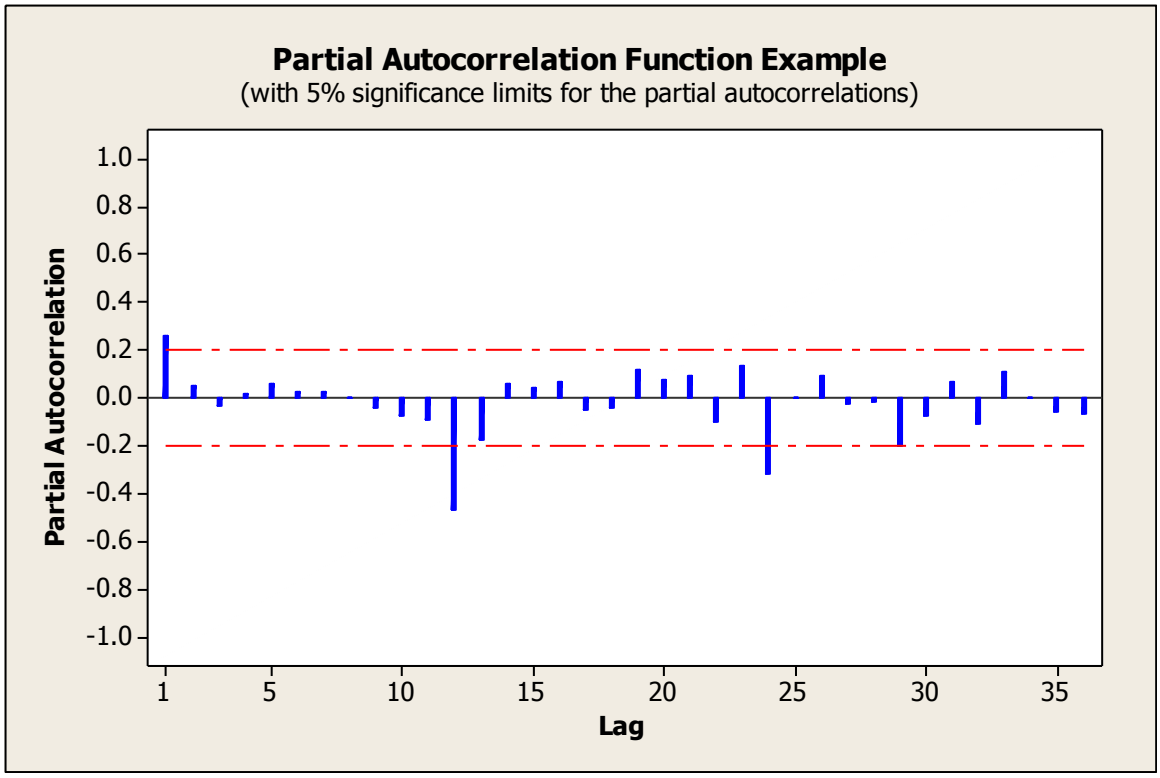


Figure 3.2 *Example of a PACF.*

Note how it looks very similar to the ACF. Both the ACF and PACF can be used to decide on what parameters should feature in an ARIMA model. However this is only true if the time series is stationary. If a trend is present then the ACF and PACF can be very misleading and may cause one to select the wrong parameters to be in a model. The same is true for a time series that has a seasonal component. *Figure 3.3* is an ACF for a time series that contains a seasonal component. It displays a typical sinusoidal pattern because observations with a small time lag between them will be positively correlated and observations with a large time lag between them (but still within the same seasonal cycle) will be negatively correlated. For example the months in summer will all be of similar temperatures but the winter months will have very different temperatures to the summer months.

Once the data is stationary and the ACF and PACF have been produced (from the stationary dataset), parameter selection for the initial model can begin. The parameters to be selected are MA and AR terms; they can also be seasonal or non-seasonal.

For a non-seasonal data set one should select MA terms for the model if significant spikes are present in the early lags of the ACF and a series of declining spikes are seen in the early lags of the PACF (a significant spike is one that exceeds the confidence interval). AR terms should be considered when significant spikes appear in the early lags of the PACF and there is a series of declining spikes in the early lags of the ACF. If both the ACF and PACF are showing a series of declining spikes then both AR and MA terms are be needed. If there are no major spikes in either graph then the series is a random walk and an ARIMA model cannot be fitted. Figure 3.4 is an example of how to select terms for a model.

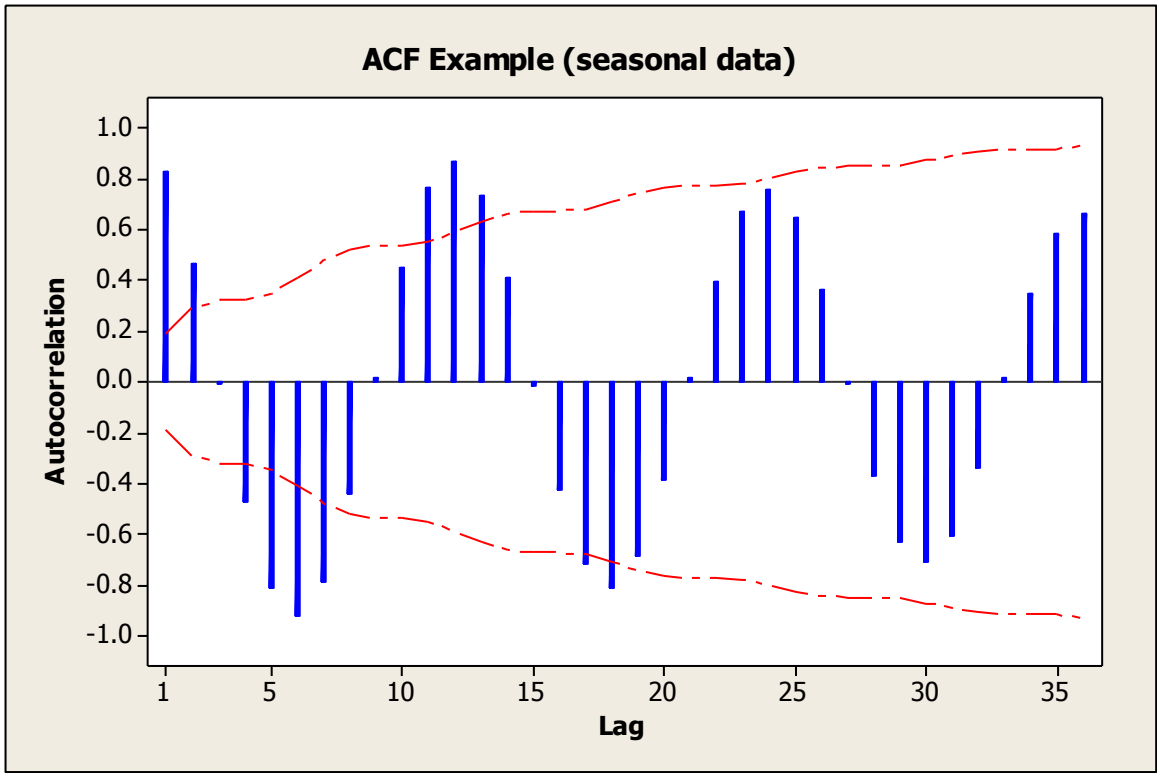


Figure 3.3 *Example of an ACF of seasonal data.*

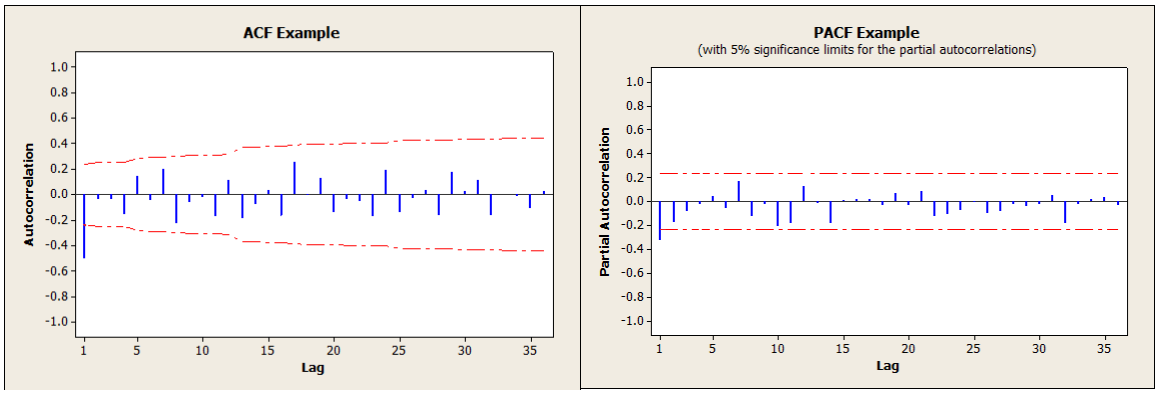


Figure 3.4 *There is a major spike at lag one in the ACF, this is accompanied by a series of declining spikes at the early lags of the PACF. This indicates at MA terms should be in the model.*

If the data contains seasonality then seasonal moving average (SMA) and seasonal autoregressive (SAR) terms will be required. If the ACF is showing significant spikes at lags which are a multiple of the seasonal cycle's time span (for months of the year that's 12, 24, 36 etc.) and the PACF has a declining pattern at those same lags then SMA terms should be considered. If it is the PACF that has the significant spikes at those lags and the ACF has the declining pattern then SAR terms should be considered. If both graphs have a declining pattern then SAR and SMA terms should be selected. Models for seasonal data can also have non-seasonal terms in them as well and they should be selected using the same criteria used for non-seasonal data sets. Figure 3.5 is an example of how to select terms for a seasonal model.

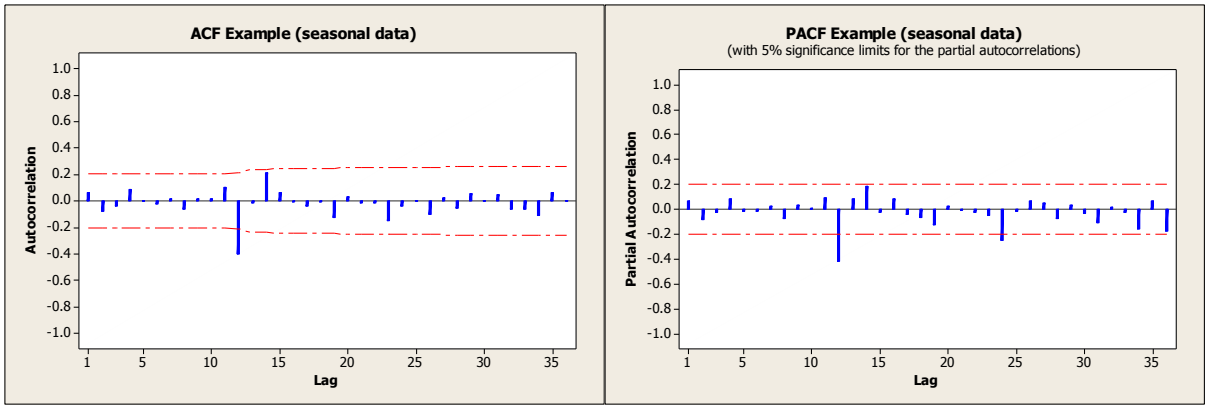


Figure 3.5 *There is a significant spike at lag 12 in the ACF and there is a series of declining spikes (each at a multiple of 12) in the PACF. This indicates that an SMA term should be in the model. There are no spikes at the early lags so the model may not need non-seasonal terms.*

It is suggested that when a specific parameter has been selected to be in the model, only one term should be used at first. After model checking has taken place more terms can be added to see if there is an improvement in the model. However it is usually of interest to keep the model as simple as possible so it is worthwhile starting with just one term. Selecting ARIMA model parameters is not an exact science and requires a lot of guess work and experience. The ACF and PACF are useful for selecting parameters for the initial model but the final model should only be decided on after careful checking has occurred (see section 3.4.4).

3.4.3 Parameter Estimation

Autoregressive Model

Estimating the parameters of an autoregressive model is no different than estimating the parameters for a linear regression model. Least squares estimates can be used; equation (3.21) shows how the slope parameter of a regression model with one explanatory variable can be estimated:

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{3.21}$$

The same formula can be applied in an autoregressive model with one predictor term (first order process) except y is now x_{t+1} and x becomes x_t. This is demonstrated by equation (3.22).

$$\hat{\varphi}_1 = \frac{\sum_{t=1}^{N-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=1}^{N-1}(x_t - \bar{x})^2} \tag{3.22}$$

Similarly for higher order AR models the same procedure used for multiple regression models can be used for estimating the parameters (estimation by least squares), it is presumed that the reader is already familiar with that process and so it will not be explained in detail.

Moving Average Model

To estimate the parameters of an MA model an iterative procedure must be used. First suitable starting values for the parameters must be used and then the residuals (z_t) can be calculated. So for an MA model of order one it must first be assumed that $z_0 = 0$ so that $z_1 = x_1 - \mu$, and then $z_2 = x_2 - \mu - \theta_1 z_1$, and so on until $z_N = x_N - \mu - \theta_1 z_{N-1}$.

The residual sum of squares can now be calculated (Σz_t^2).This procedure must then be repeated for neighbouring values of the parameters. A grid of the residual sum of squares can now be constructed. By inspecting this grid one can determine the values of the parameters that minimise the residual sum of squares. These values are the best estimates for the parameters. The same procedure can be used for MA models of higher orders.

ARMA, ARIMA and SARIMA Models

Since an ARMA model may contain MA terms its parameters must be estimated in the same manner as for an MA model. The following example shows how the residuals for an ARMA (1, 1) can be calculated:

The model is:

$$x_t - \mu = \varphi_1(x_{t-1} - \mu) + z_t + \theta_1 z_{t-1}$$

Suppose that there are N observations, make some initial estimates for μ, φ_1 and θ_1. Then assume $z_0 = 0$ and $x_0 = \mu$. The residuals are calculated like so:

$$z_1 = x_1 - \mu$$
$$z_2 = x_2 - \mu - \varphi_1(x_1 - \mu) - \theta_1 z_1$$
Up to
$$z_N = x_N - \mu - \varphi_1(x_{N-1} - \mu) - \theta_1 z_{N-1}$$

The residual sum of squares can now be calculated. Repeat this process with neighbouring parameter estimates. The best combination of estimates is the one that minimises the residual sum of squares.

The exact same process can be used for ARIMA models since they are the same model but use differenced data. The same is true for SARIMA models but they tend to have many more parameters so the estimation process is incredibly arduous. Typically an analyst will use a software package such as SAS, R or Minitab to carry out the estimation procedure.

3.4.4 Model Checking

As with all statistical model building an important step in time series analysis is checking the model to make sure that it adequately fits the data. There are various factors one should consider when determining lack of fit but perhaps the most important factor to consider is whether the estimates for the model parameters are statistically significant. The parameter estimates are assumed to be normally distributed so a t-test can be used to test their significance.

To test the significance of a parameter estimate divide it by its standard error and compare that value to a t-distribution where the degrees of freedom are equal to the number of observations in the differenced series minus the number of parameters in the model. The significance level is up to the analyst but 5% is typically used. If the test statistic exceeds the t-value then it can be concluded that the parameter estimate is significantly different from zero and should be included in the model. When using software to test for significance a p-value will be given which is the probability that parameter is not equal to zero (typically a p-value less than 0.05 suggests that the parameter estimate is significant).

A residual is the difference between an observed value and its corresponding fitted value determined by the model. For a time series model to adequately fit the data these residuals must be normally

and randomly distributed (they must be white noise). If the residuals are not normally distributed then the model does not fit the data well as there is still some variation in the data that it fails to explain. Various graphs can be used to judge whether the residuals are normally and randomly distributed, *figure 3.6* shows four of these graphs.

Ideally for a well fitted model one would expect to see a "normal" looking histogram of residuals and points that do not deviate from the line in the normal probability plot. As for the residuals vs. fits and residuals vs. order plots one should expect to see a random scatter of points that contain no discernible pattern such as curvature or increasing/decreasing spread. If there is a pattern then the analyst should consider transforming the data to remove the change in variability within the data.

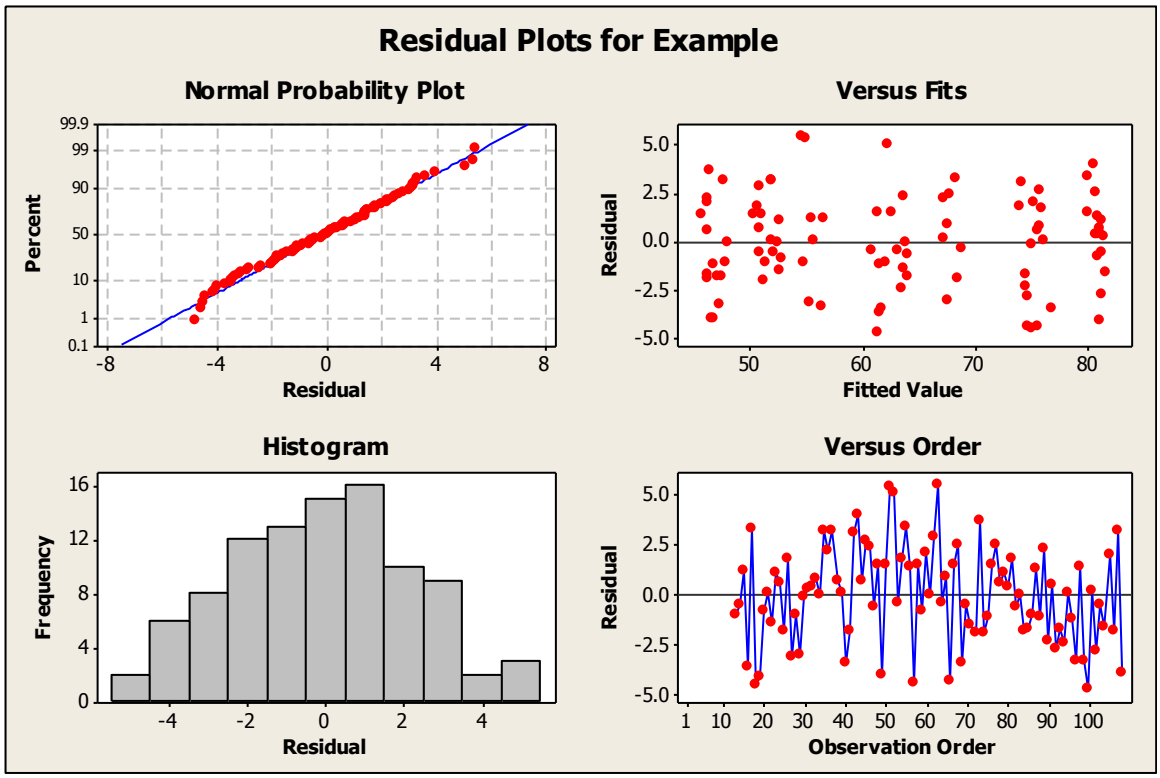


Figure 3.6 *Four in one residual plot showing a normal probability plot, versus fits plot, histogram and versus order plot.*

The residuals can also form a time series of their own which means that an ACF and PACF of the residuals can be calculated as well. The residuals are supposed to be uncorrelated if the data has been modelled well so the ACF and PACF of the residuals should show no major spikes (that is to say that the correlation and partial correlation between residuals at a specified lag is not significant). Direct tests can be used to the same effect such as the portmanteau lack-of-fit test which examines the first K values of the residual ACF; it is given by equation 3.23.

$$Q = N \sum_{k=1}^{K} r_{z,k}^2 \tag{3.23}$$

Where N is the number of terms in the differenced series and K is a value chosen by the analyst (typically between 15 and 30).

If the model is a good fit then Q approximately follows a chi-squared distribution with ($K-(p+q+P+Q)$) degrees of freedom where *p, q, P* and *Q* are the number of AR, MA, SAR and SMA terms. However this approximation can be inadequate if N<100, fortunately there is another statistic one can use called the Ljung-Box-Pierce statistic. This is given by equation 3.24:

$$N(N+2)\sum_{k=1}^{K} r_{z,k}^2 /(N-k) \tag{3.24}$$

This statistic is used by most software packages and will be employed in this study as well.

3.4.5 Forecasting

One of the main purposes of time series modelling is to make forecasts. To generate forecasts one can simply use the model equation as they would for finding the fitted values. This is best demonstrated by an example. The following example shows how future values in a time series can be predicted with the model equation. The model used is an ARIMA $(1, 0, 0)\times(0, 1, 1)_{12}$ model.

Model: $x_t = x_{t-12} + \varphi(x_{t-1} - x_{t-13}) + z_t + \Theta z_{t-1}$

Let $\hat{x}_N(h)$ be the forecasted value that is h units of time ahead of the final observed value in the series, then:

$$\hat{x}_N(1) = x_{N-11} + \varphi(x_N - x_{N-12}) + \Theta z_{N-11}$$
$$\hat{x}_N(2) = x_{N-10} + \varphi[x_N(1) - x_{N-11}] + \Theta z_{N-10}$$

The example shows how forecasts that are further into the future will need to rely on previous forecasts that have been made (for example $\hat{x}_N(2)$ in the above series requires $\hat{x}_N(1)$). This means that forecasts further into the future are less reliable than more recent forecasts.

3.5 Summary

A time series analysis typically follows five steps. First the analyst must examine the time series plot so that they can prepare the data for modelling which may involve some kind of transformation or differencing procedure. The parameters of the model are then selected by examining the ACF and PACF. These parameters are then estimated. This is followed by checking the model and reconstructing the model if necessary. Once an adequate model has been found it can be used for forecasting future values of the time series.

This entire procedure can be done using various software packages. This investigation makes use of Minitab. The application of these methods and their results are in the next chapter.

CHAPTER 4
Time Series Analysis

In this chapter the time series methods discussed in Chapter Three are applied to the US temperature data. There will be two main sections to this analysis. The first will be a simple inspection of the time series graphs for each State. The following section will consist of the SARIMA models constructed for each State and their forecasts.

A time series analysis is being conducted because the US climate is constantly changing and only a time series analysis can take into account these changes that are occurring through time. The Gulf Stream in the Atlantic Ocean is never static and is ever changing. El Niño is an event that occurs every three to seven years in the Pacific Ocean and the magnitude of its effect is always different. So time is a crucial element in this investigation therefore time series methods must be applied.

SARIMA models are used to determine the core components of the coastal climates; this makes it possible for a comparison between the two coasts to be made. SARIMA models are also used due to the complex nature of climate data. Naturally there is seasonality within the data and there could possibly be a trend too, only SARIMA models can account for these factors.

4.1 Time Series Plot Inspection

As discussed in Chapter Three it is always necessary to inspect the time series plot of a data set before a model is fitted. In this section the time series plots for each State is examined and interpreted (*figures 4.1 to 4.6*).

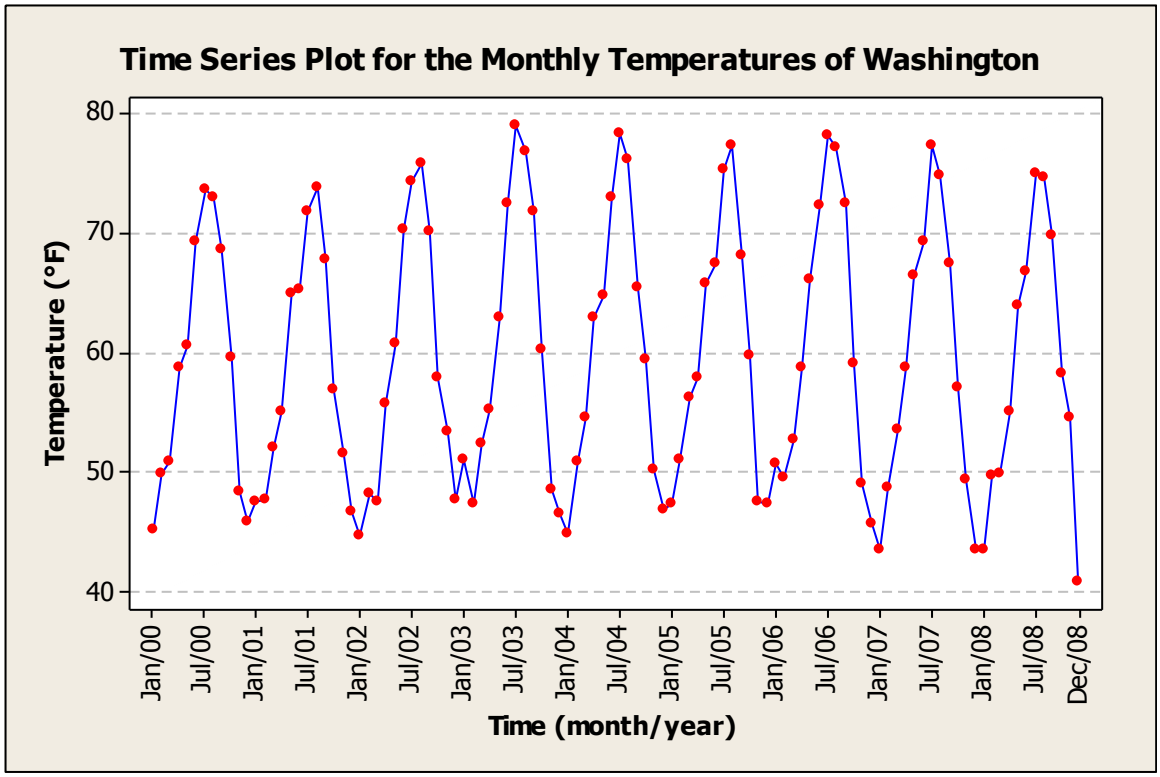


Figure 4.1 *Time series plot for the monthly temperatures of Washington.*

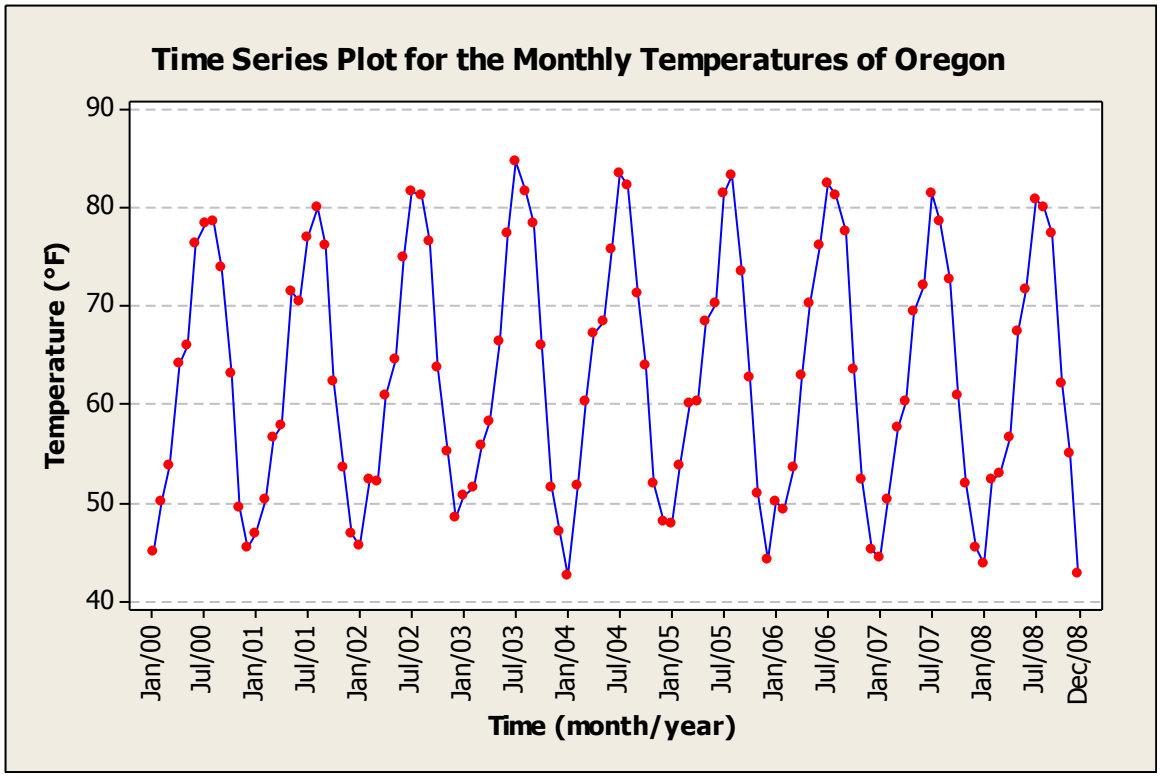


Figure 4.2 *Time series plot for the monthly temperatures of Oregon.*

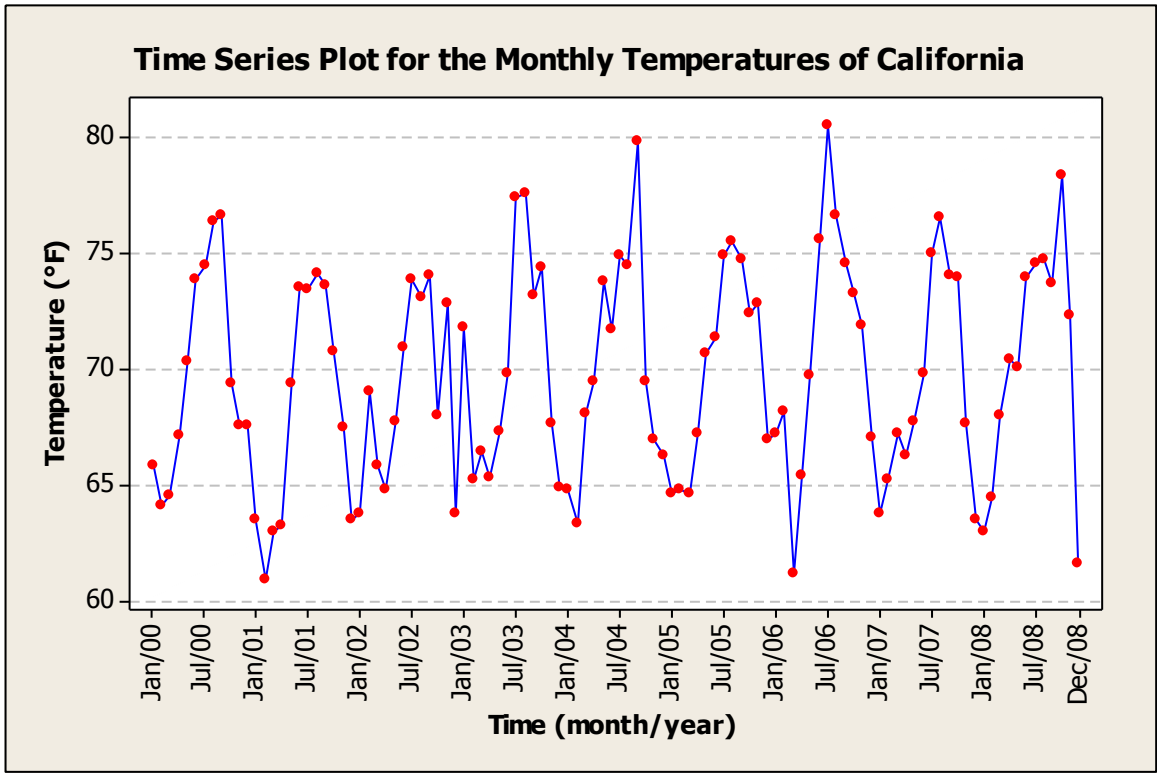


Figure 4.3 *Time series plot for the monthly temperatures of California.*

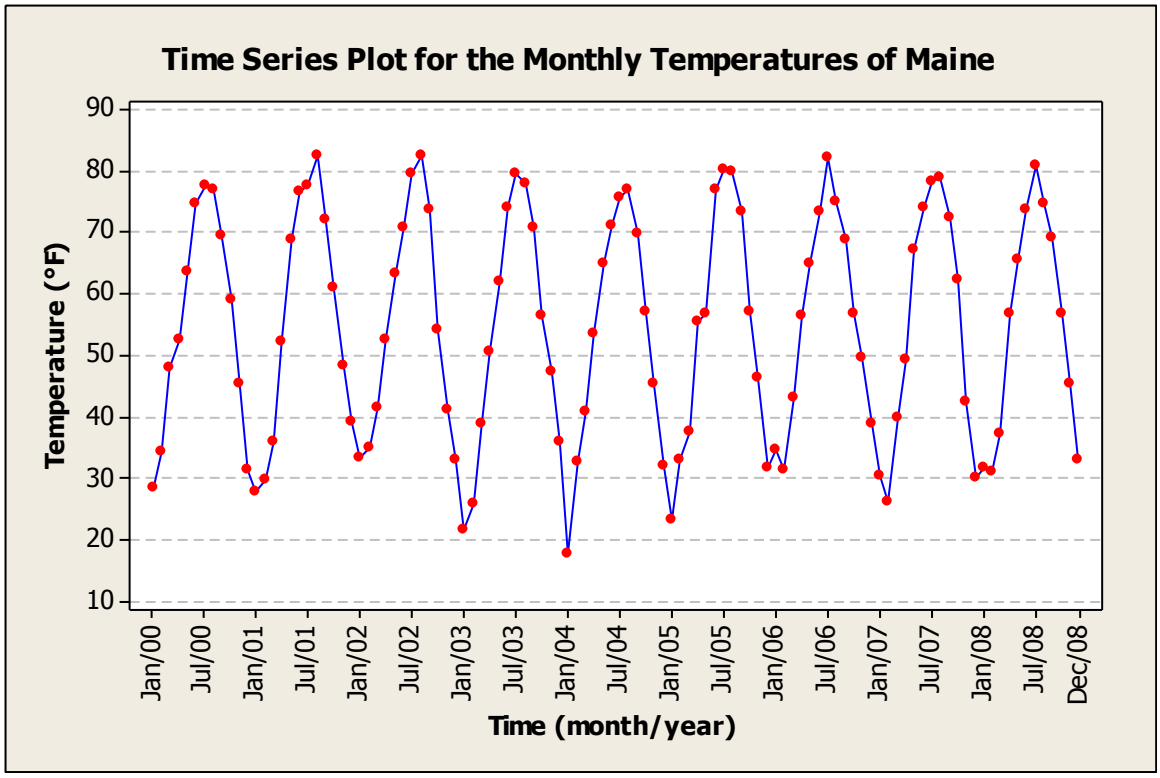


Figure 4.4 *Time series plot for the monthly temperatures of Maine.*

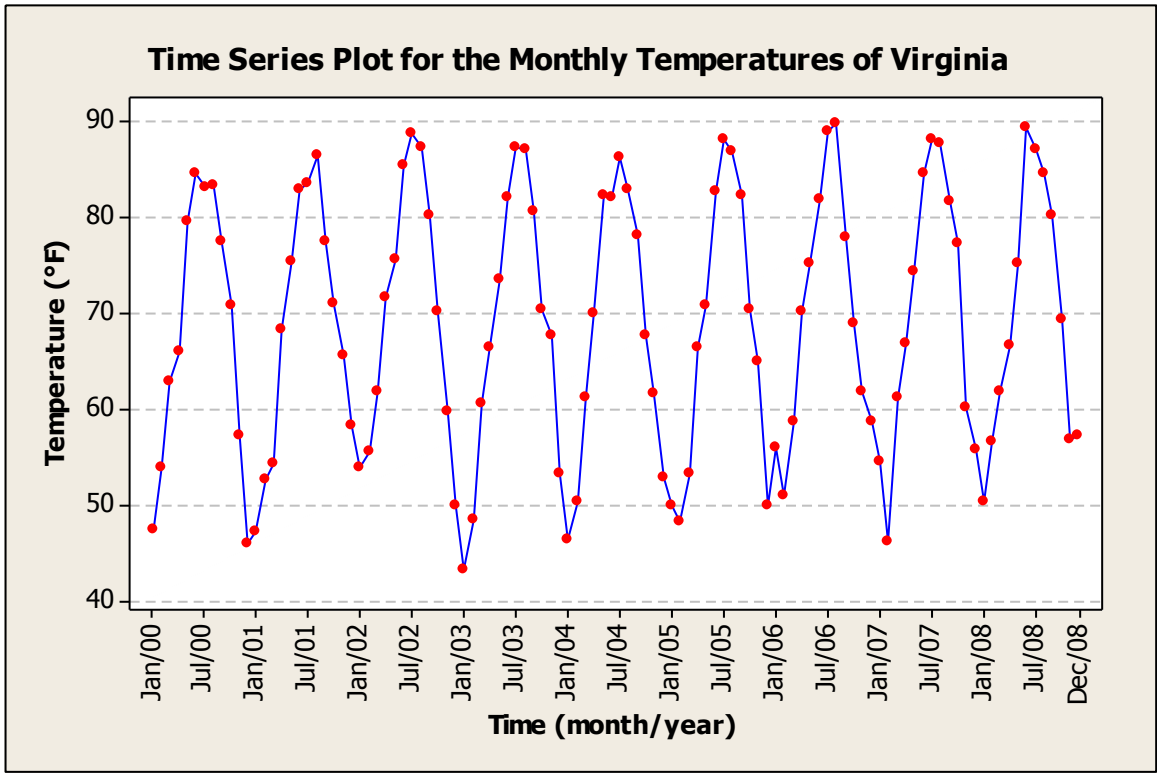


Figure 4.5 *Time series plot for the monthly temperatures of Virginia.*

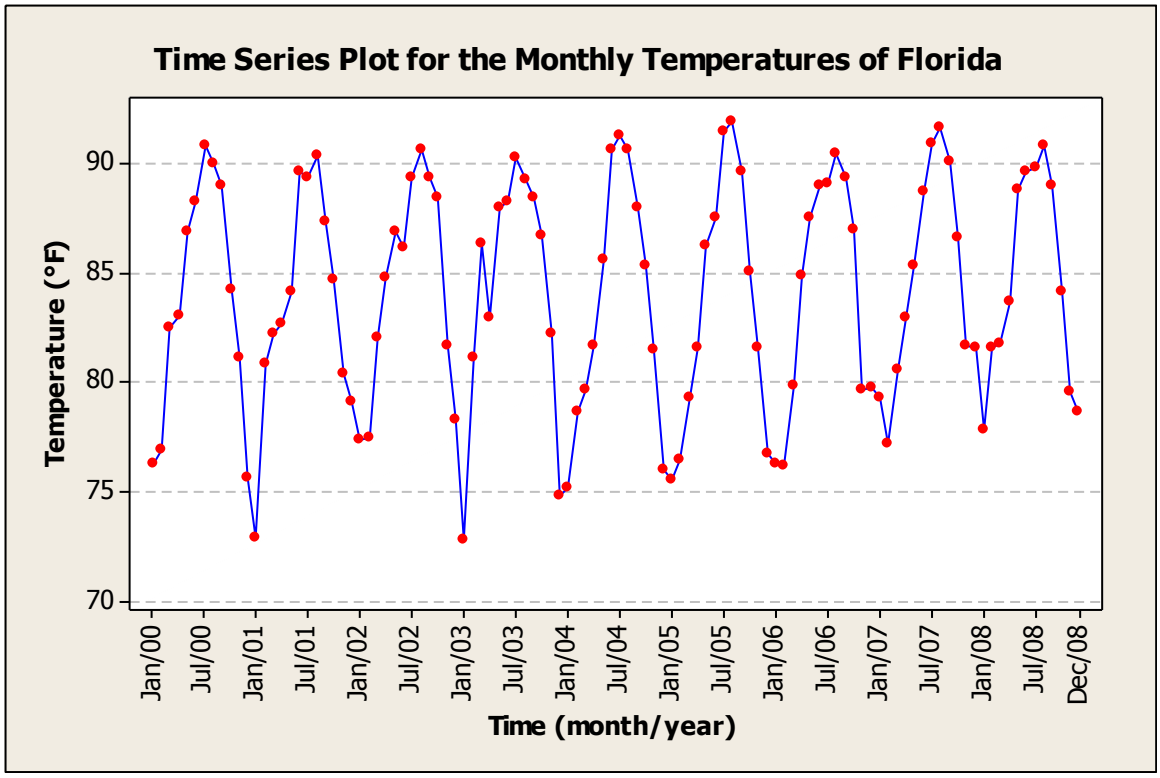


Figure 4.6 *Time series plot for the monthly temperatures of Florida.*

4.1.1 Washington

The time series plot (*figure 4.*1) appears to show no linear trend, however there may be some curvature. There seems to be a rise in temperature from the year 2000 to the year 2004 and then a steady decrease thereafter.

It is difficult to determine if there has been any change in the variance over time. There also doesn't appear to be any major anomalies in the data, although there was an unusually warm January in 2004 but such events are entirely possible.

4.1.2 Oregon

Oregon has a very similar time series plot (*figure 4.2*) to Washington because the observations made in each State were taken at locations very close to each other (Portland and Seattle.) The curvature seen in Washington is also present in Oregon. It can also be seen that the latter years in the graph have a higher temperature range than earlier years which suggests that there is an increasing variance present in Oregon's temperature data. As a result of this it was decided that the data should undergo a log transformation to stabilise the variance.

4.1.3 California

This time series plot (*figure 4.*3) appears to be far more erratic compared to the time series plots of the other West Coast States. Since the observations were taken in Los Angeles which is very far to the south, it was expected that there would have been a more constant and predictable temperature.

The most erratic year was 2002. The unusually hot months were September 2004 and July 2006. The unusually cold months were February 2001, March 2006 and December 2008. There may also be an upward trend. The nature of this data is very strange and it caused difficulties when it was being fitted by an ARIMA model.

4.1.4 Maine

Maine's time series plot (*figure 4.4*) appears to be quite a uniform time series plot with almost no irregularities. There appears to be no trend but there is a slight decline in temperature for the years of 2004 and 2005 so it is possible that there is some curvature in the data.

The variance appears to be static through time, so a transformation was not needed to fit an ARIMA model to the data. The exploratory analysis did show that the data was not normally distributed but the ARIMA model was fitted to the data after it had been differenced. The differenced series did have a normal distribution which also indicated that no transformation was needed.

4.1.5 Virginia

The time series plot (*figure 4.*5) for Virginia seems to be showing an upward trend. The variance seems stable although there is a possibility that it is decreasing. There are no major irregularities in the plot but the January of 2006 was unusually warm.

4.1.6 Florida

Like California, Florida also has an erratic time series plot (*figure 4.6*) relative to its East Coast counter parts. This was not detected in the explanatory analysis as that did not consider how the temperature behaved through time. There seems to be no clear pattern for the first four years of data and then there is possibly an upward trend for the last five years. The data also seems to have a decreasing variance through time so a transformation was needed to fit an ARIMA model to the data.

There were unusually cold Januarys in 2001 and 2003 but in the intervening year of 2002 there was a rather warm January. The May of 2003 was also unusually cold.

4.2 SARIMA model construction

In this section a SARIMA model is fitted to the temperature data for each State. The parameters of the model were first determined by the ACF and PACF plots (after appropriate differencing had occurred). Minitab was then used to estimate the model parameters, after which the model was checked to see if it adequately fitted the data. If the model was unsatisfactory a new model was constructed and was also checked. This process continued until the best possible model had been acquired. Forecasts were then made using the final model.

A step by step model construction process is given for the State of Washington. The model construction processes for the other States are summarised. However more details are provided on the construction of Virginia's model.

4.2.1 Washington

As explained in Chapter Three the parameters of an ARIMA model are determined by an examination of the ACF and PACF plots. *Figure 4.7* is the ACF plot for Washington's temperature data.

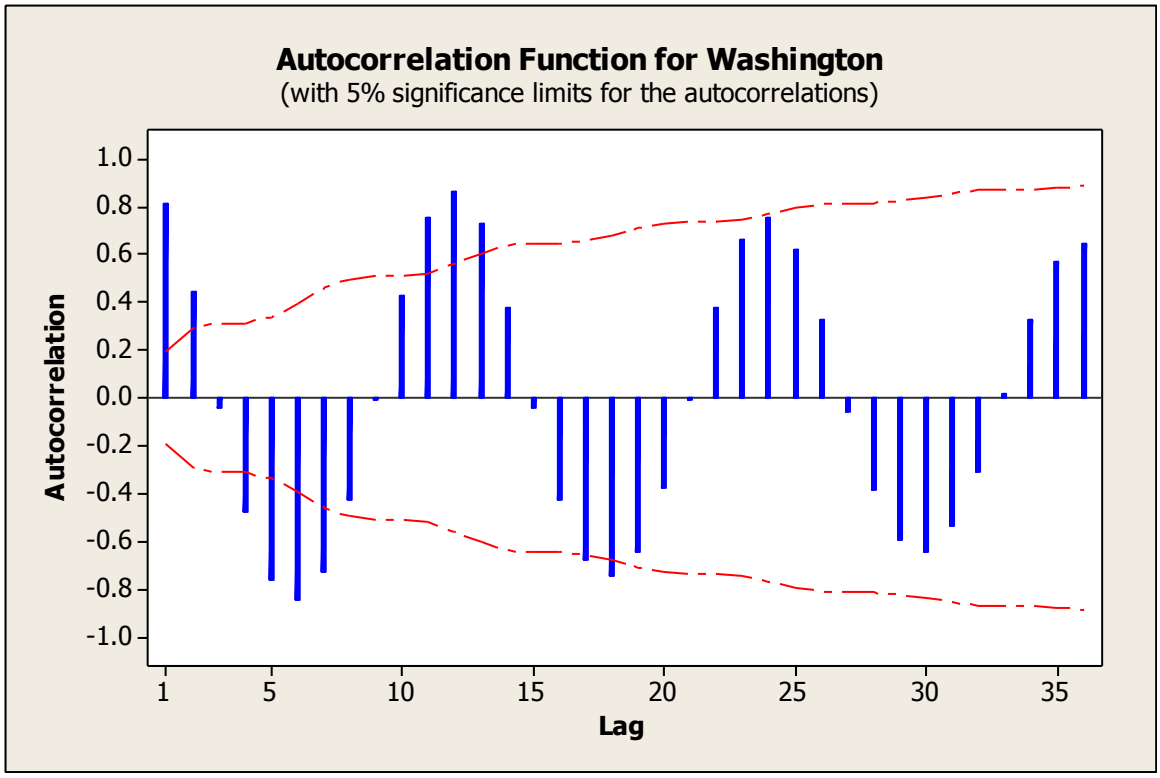


Figure 4.7 *ACF plot for Washington's data.*

In this plot one can see the standard sinusoidal pattern that accompanies seasonal data. So it was necessary to difference the data by 12 to remove the seasonality from the data (since a season cycle is completed in 12 months). *Figure 4.8* shows the ACF and PACF plots of the differenced data.

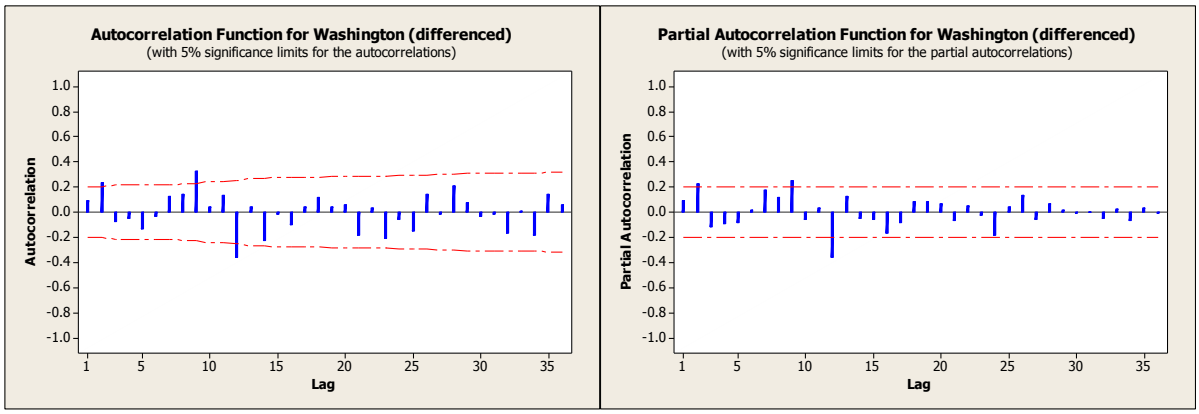
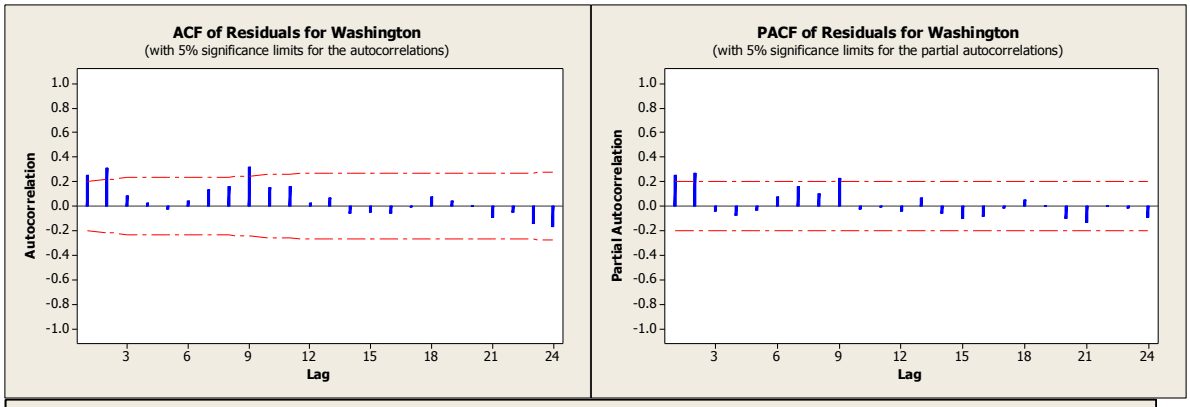
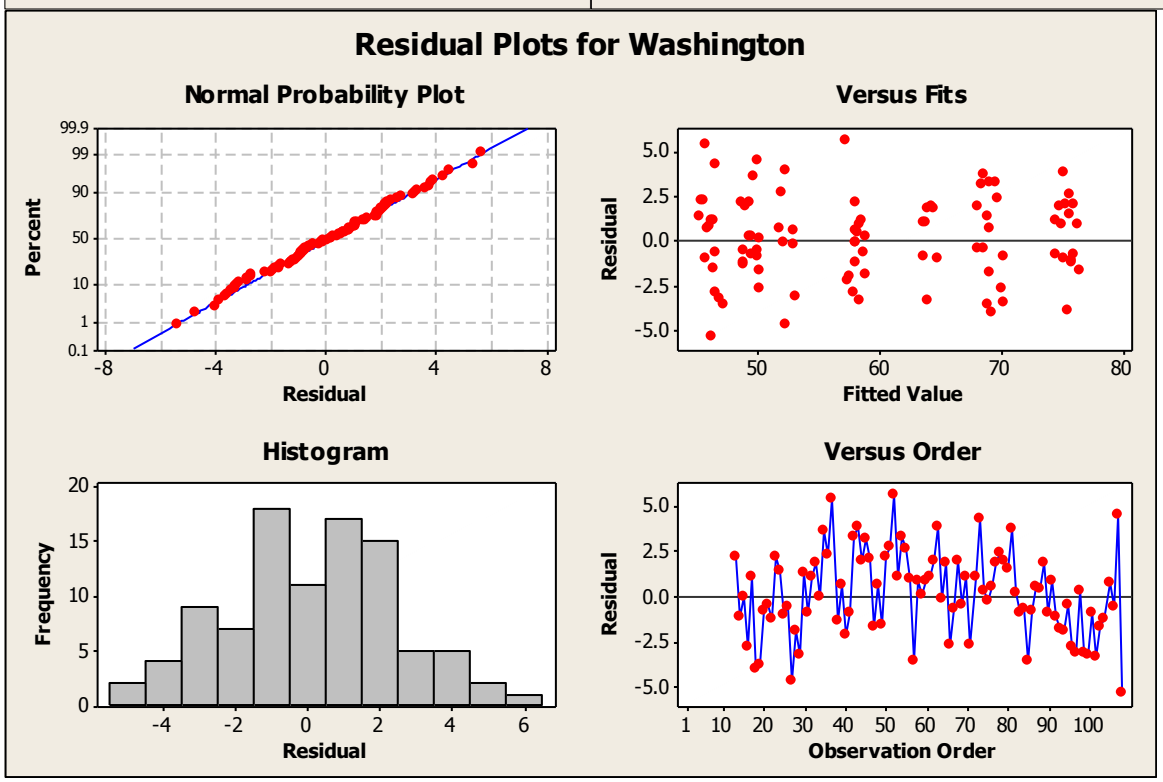


Figure 4.8 *ACF and PACF plots for Washington's differenced data.*

In both plots there are spikes at lags two and nine, since there is no significant spike at lag 1 it is unclear on what these spikes mean. There is a major spike at lag 12 on the ACF which is accompanied by spikes that tail off in multiples of 12 on the PACF. This feature indicates that there should be a seasonal moving average (SMA) term in the model.

The first model constructed was ARIMA $(0, 0, 0) \times (0, 1, 1)_{12}$. (*Figure 4.9* Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)





Modified Box-Pierce (Ljung-Box) Chi-Square statistic

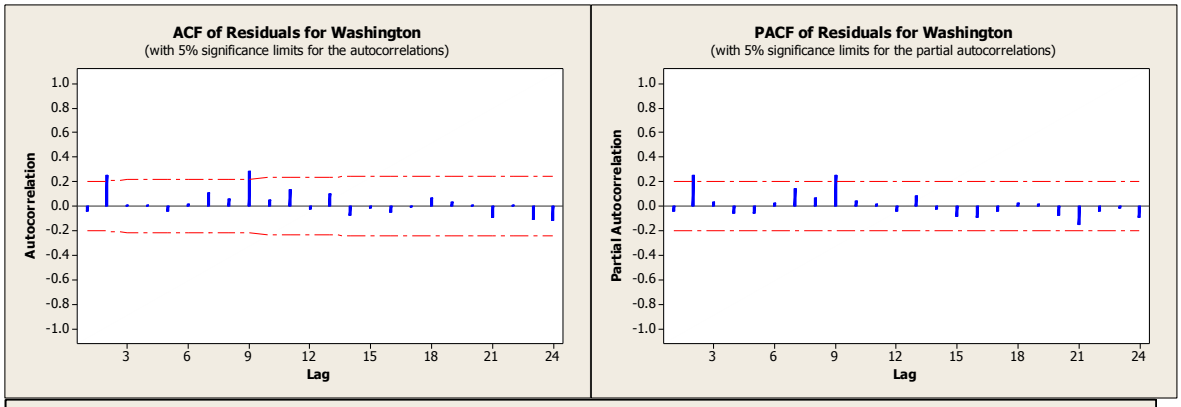
Lag	12	24	36	48
Chi-Square	37.7	47.9	63.1	76.7
DF	10	22	34	46
P-Value	0.000	0.001	0.002	0.003

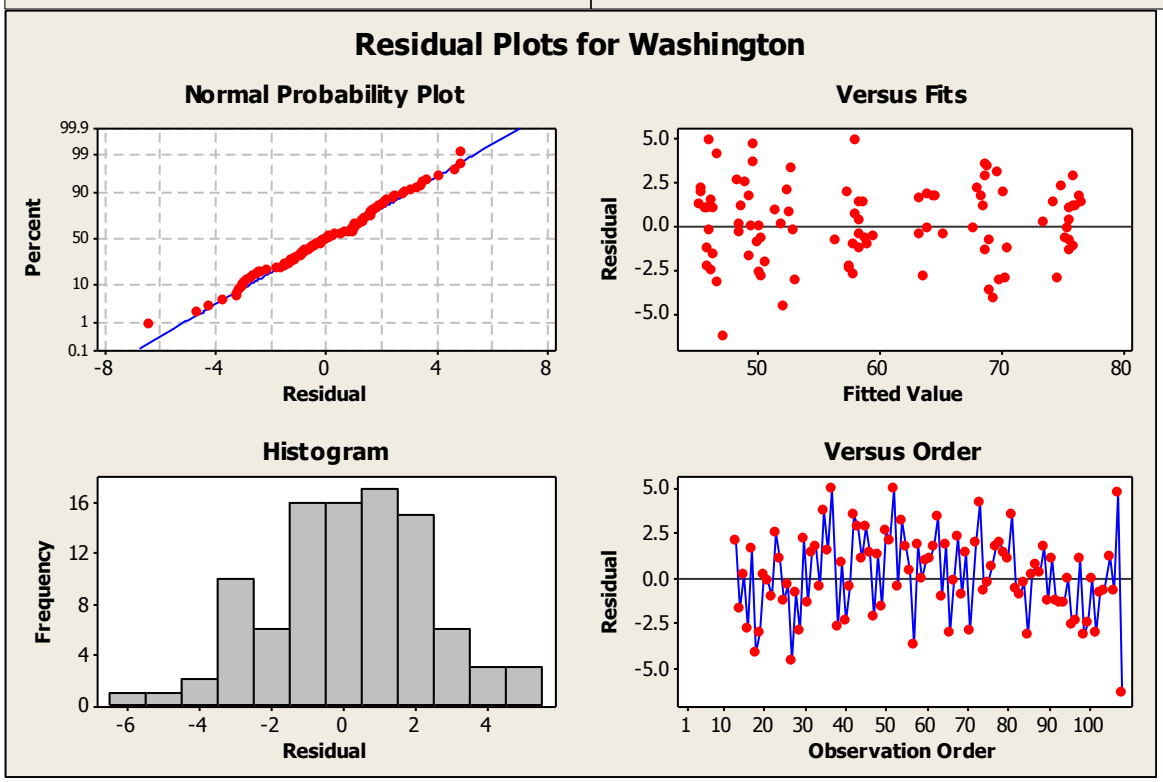
Final Estimates of Parameters

Type		Coef	SE Coef	T	P
SMA	12	0.8790	0.0863	10.19	0.000
Constant		0.08331	0.06079	1.37	0.174

Figure 4.9 *Residual plots, box-pierce statistics and parameter estimates for Washington's ARIMA (0, 0, 0) × (0, 1, 1)$_{12}$ model.*

The SMA term was significant but the model had highly significant box-pierce p-values which suggested that the residuals were correlated (this implied that the model did not adequately fit the data). The ACF of residuals was showing spikes at early lags, so it may have been wrong not to include a non-seasonal term. Another ARIMA model was constructed and it included a non-seasonal autoregressive (AR) term. The second model was ARIMA (1, 0, 0) × (0, 1, 1)$_{12}$. (*Figure 4.10* Shows the residual plots and the model fitting statistics.)





Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	19.7	26.9	45.0	55.0
DF	9	21	33	45
P-Value	0.020	0.173	0.079	0.146

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	0.2473	0.1033	2.39	0.019
SMA	12	0.8926	0.0838	10.65	0.000
Constant		0.05520	0.05429	1.02	0.312

Figure 4.10 *Residual plots, box-pierce statistics and parameter estimates for Washington's ARIMA (1, 0, 0) × (0, 1, 1)$_{12}$ model.*

The new AR term was significant but the box pierce p-values remained significant as well. The time series plot for Washington suggested that there may have been some curvature in the data. It was thought that this was preventing an adequate model from being found. Curvature is not too dissimilar from a trend so it was decided that the data should be subject to further differencing. A non-seasonal difference of one was applied to the already differenced data. *Figure 4.11* shows the ACF and PACF plots of the newly differenced data.

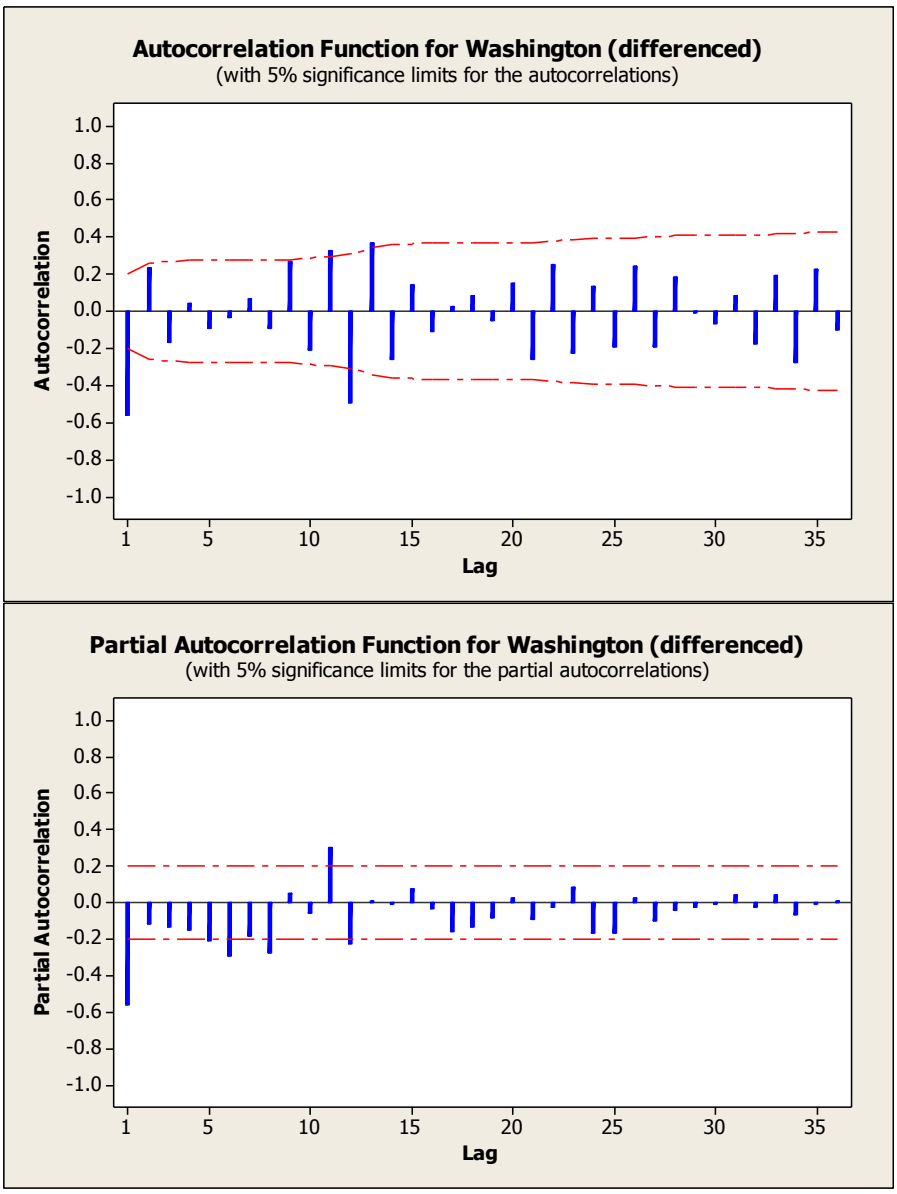
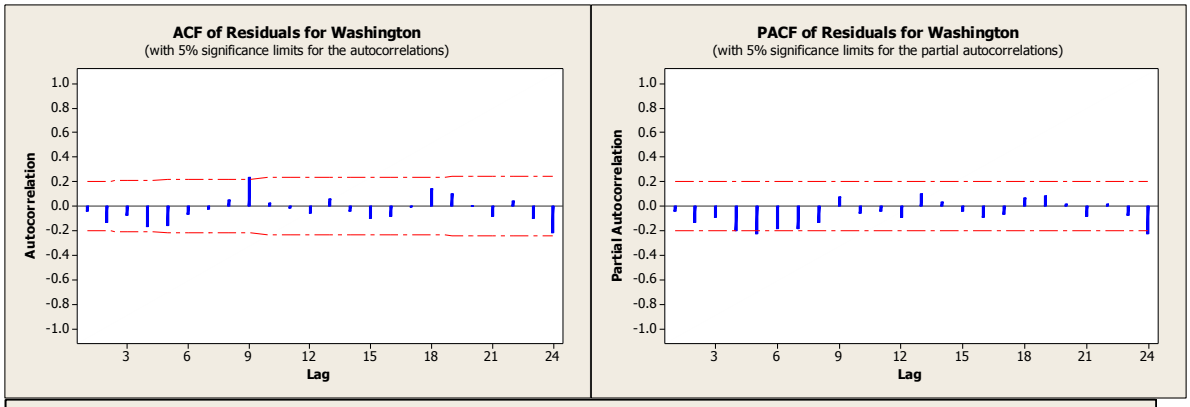
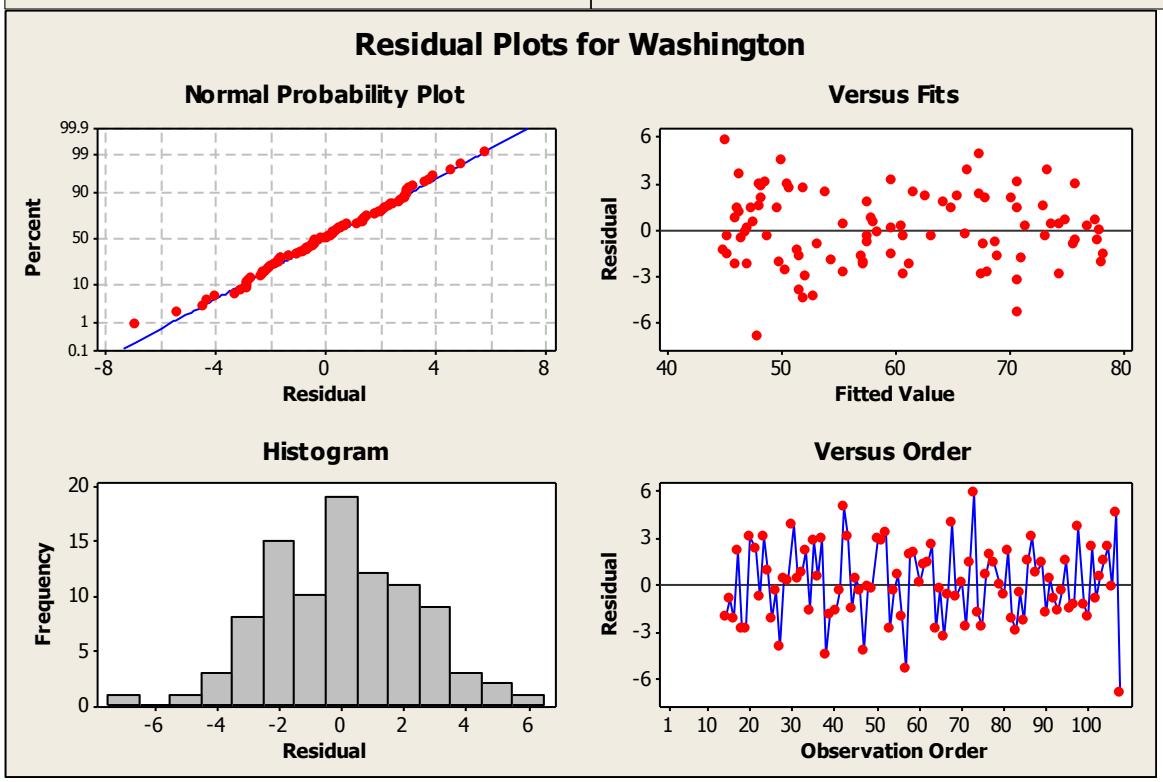


Figure 4.11 *ACF and PACF plots for Washington's differenced data.*

These plots appeared to be much neater than the previous ones and new model was constructed that factored in the non-seasonal differencing.

The third model constructed was ARIMA (1, 1, 0) × $(0, 1, 1)_{12}$. (*Figure 4.12* Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)





Modified Box-Pierce (Ljung-Box) Chi-Square statistic

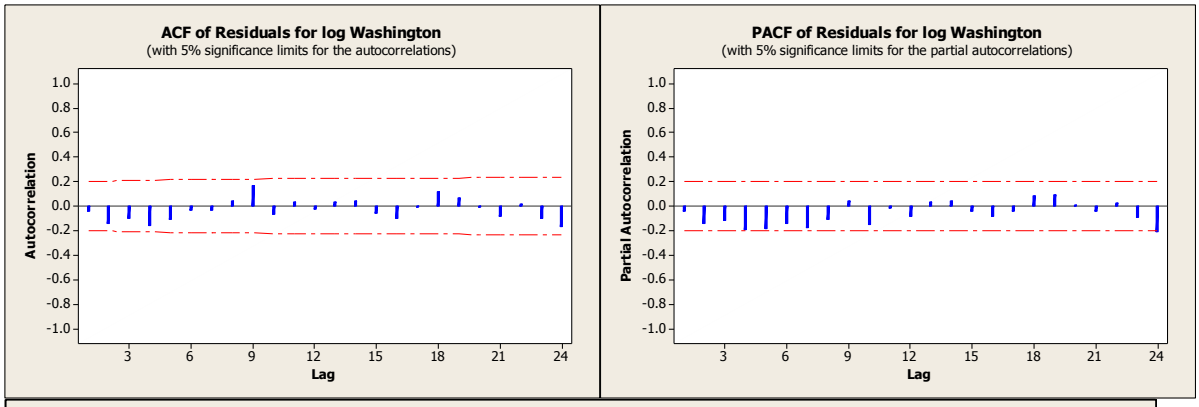
Lag	12	24	36	48
Chi-Square	15.2	29.8	41.4	50.9
DF	9	21	33	45
P-Value	0.085	0.097	0.150	0.251

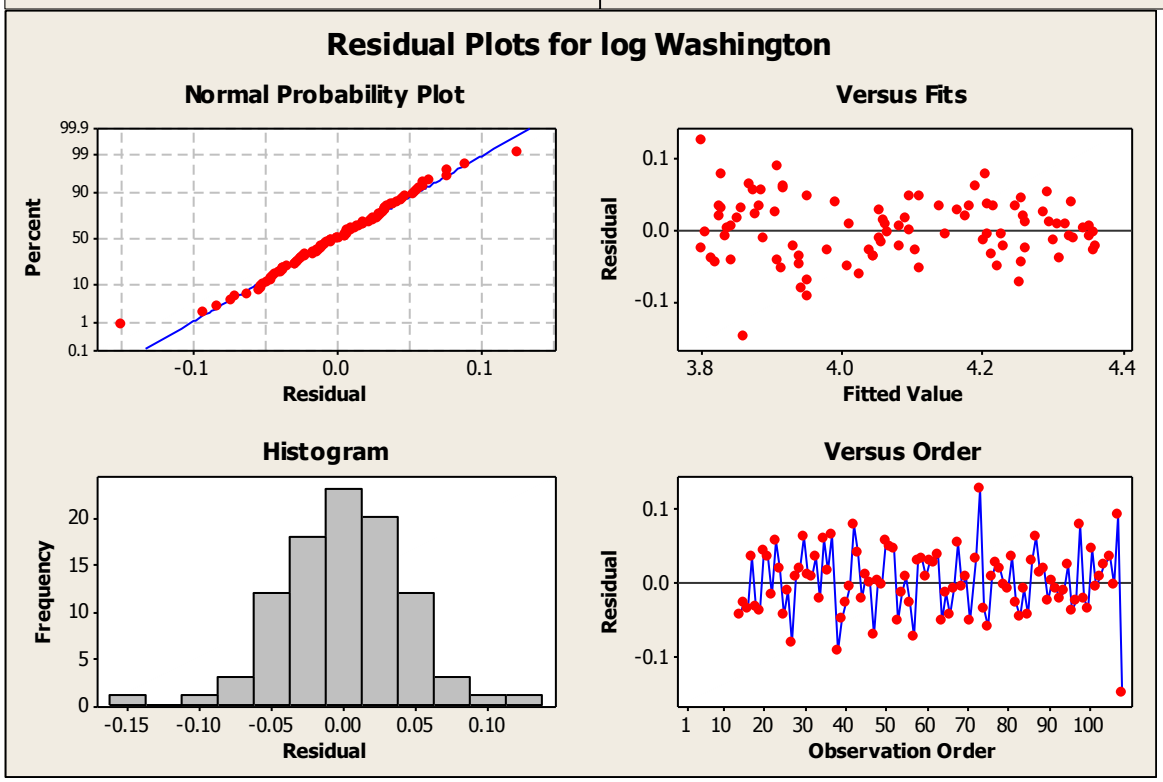
Final Estimates of Parameters

Type		**Coef**	**SE Coef**	**T**	**P**
AR	1	-0.5825	0.0947	-6.15	0.000
SMA	12	0.8831	0.0910	9.71	0.000
Constant		-0.02599	0.05167	-0.50	0.616

Figure 4.12 *Residual plots, box-pierce statistics and parameter estimates for Washington's ARIMA (1, 1, 0) × (0, 1, 1)$_{12}$ model.*

The box-pierce p-values were insignificant (so the residuals were not correlated). The residuals appeared to be normally and randomly distributed. This model was adequate but it may have been possible that the variance was increasing with time so a log transformation was applied to the data and the same model was fitted again, (*Figure 4.13* Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)





Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	10.9	20.8	31.9	41.3
DF	9	21	33	45
P-Value	0.281	0.473	0.522	0.629

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	-0.6079	0.0964	-6.30	0.000
SMA	12	0.8993	0.0784	11.47	0.000
Constant		-0.0006195	0.0009522	-0.65	0.517

Figure 4.13 *Residual plots, box-pierce statistics and parameter estimates for log Washington's ARIMA (1, 1, 0) × (0, 1, 1)$_{12}$ model.*

This model appeared to be superior. The box-pierce results were even more insignificant so it is more likely that the residuals are uncorrelated. The residuals also appeared to be more "normally" distributed. This was the final model constructed for Washington and was used for interpretative and forecasting purposes.

4.2.2 Oregon

Oregon had a very similar time series plot to Washington so it was thought that it would have the same parameters in its ARIMA model. A log transformation was applied to the data and the same model was fitted. However this particular model had highly significant box-pierce p-values. Strangely there were actually almost no ARIMA models which adequately fitted the data. The best model was judged to be ARIMA (0, 0, 0) × $(0, 1, 1)_{12}$ (using log data.) (*Figure 4.14* (*page* 39) Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)

This was the one model that did not have significant box-pierce p-values (its residuals were uncorrelated). The residual histogram appeared to be relatively normal but the residuals vs. fits plot indicated that the variance is not entirely constant through time. The ACF and PACF residual plots do have a spike at lag nine that exceeds the 95% confidence interval but this could have happened by chance (since one in twenty values do exceed a 95% confidence interval). The difficulties experienced in fitting this model may have been due to the small sample size being used and the limitations of ARIMA modelling. An ARIMA model is unable to deal with curvature. It may have been better to have created two ARIMA models for this state, one for each side of the curve.

4.2.3 California

For this State there were many models that possessed significant parameter coefficients and insignificant box-pierce p-values but the residuals of these models were always highly skewed even when a log transformation or root transformation had been applied. Eventually the best model was judged to be ARIMA (0, 0, 0) × $(0, 1, 2)_{12}$ (using log data.) (*Figure 4.15* (*page* 40) Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)

The histogram of residuals was still slightly skewed but it could pass as being normal. The same could be said about the normal probability plot. The points on the residuals vs. fits plot appeared to be randomly scattered. The ACF and PACF residual plots were also in good shape with only one spike exceeding the confidence interval at a very late lag. There are some very minor issues regarding the residuals of this model otherwise it adequately fits the data.

4.2.4 Maine

The differenced data for Maine appeared to be normally distributed and had a rather uniform time series graph with no irregularities so no transformations were needed in order to construct an ARIMA model for the data. It was relatively easy to fit an ARIMA model for Maine's data. The best model was judged to be ARIMA (1, 1, 0) × $(0, 1, 1)_{12}$ (*Figure 4.16* (*page* 41) Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)

This model had highly significant parameter estimates and insignificant box-pierce p-values. The residuals appeared to be normally and randomly distributed. There was a single significant spike in the ACF and PACF residual plots but this is quite possible even in a well fitted model.

4.2.5 Virginia

In order to inspect the ACF and PACF plots the data was differenced by 12 to remove the seasonality as usual but *figure 4.17* (*page* 42) clearly shows that the seasonality still remained.

This was very strange and previous work was checked to see if there were any miscalculations, there were none. To solve this problem further differencing of one was applied to the data. *Figure 4.18* (*page* 42) is the resulting ACF plot.

These plots then made selecting an initial model possible. Since a differencing of one had to be done in order to produce ACF and PACF plots that could be interpreted it is likely that there is a trend within the data. The time series plot for Virginia also suggested that there was a trend.

Many models were constructed for Virginia but the residual plots all indicated that the decreasing variance seen earlier was a problem. To deal with this a square root transformation was applied to the data. Even after this was done only a few models were able to adequately fit the data. The best model was judged to be ARIMA (0, 1, 1) × $(1, 1, 1)_{12}$ (using root data.) (*Figure 4.19* (*page* 43) Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)

This model did have significant coefficients and insignificant box-pierce p-values but the residuals were far from perfect. The normal probability plot and histogram suggested some normality but the residuals vs. fit plot suggested that a decreasing variance was still present (although not as much now that the square root transformation had been applied.)The ACF and PACF residual plots do not have all of their spikes within the confidence intervals.

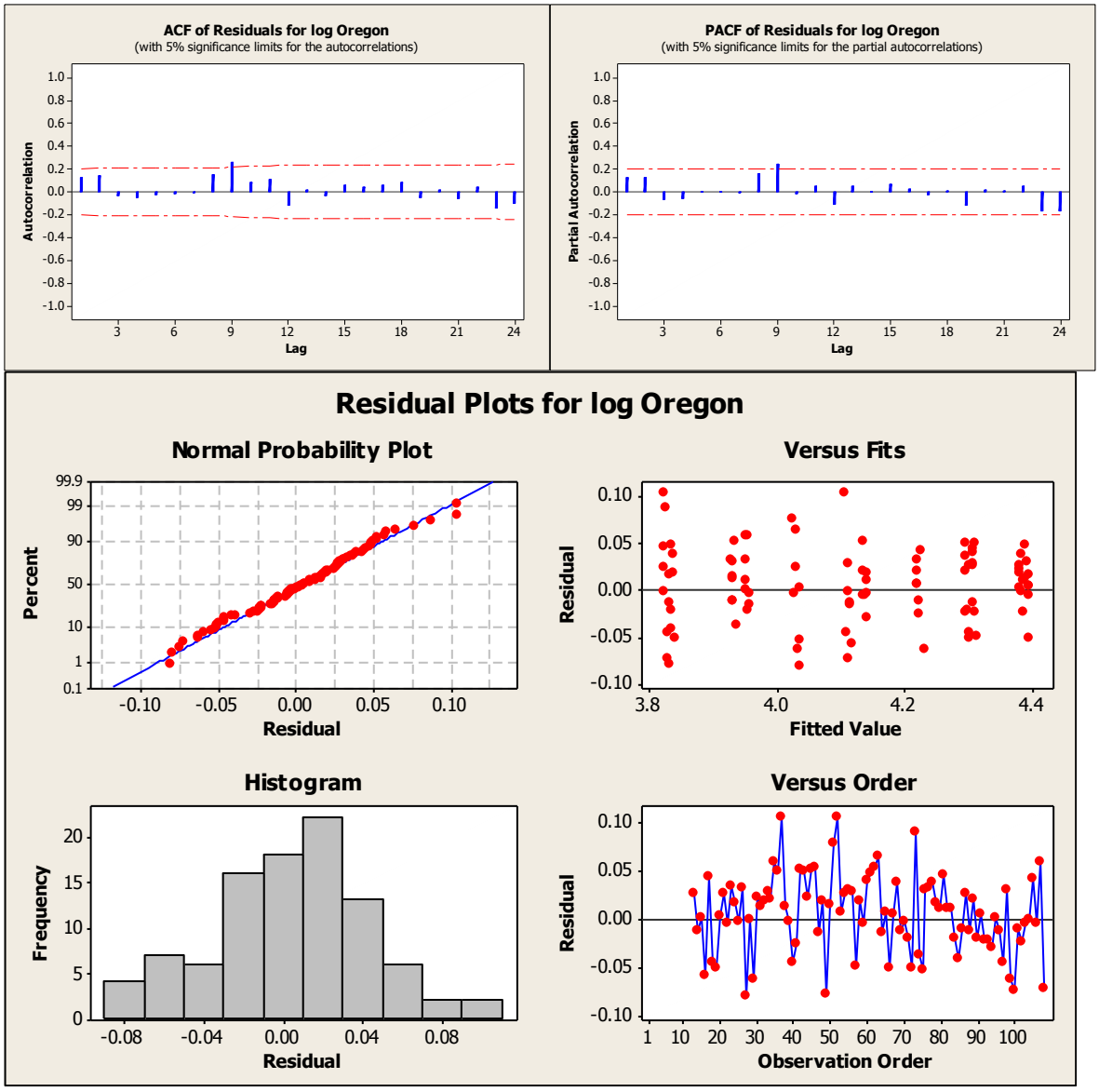
Other transformations were applied but they only yielded poorer results. Other models were also adequate but they were overly complex for such a little improvement in the residual plots and so were not selected as the final model.

4.2.6 Florida

The decreasing variance seen in the time series plot posed such a problem in the model building process that not even a square root transformation could solve. An ideal transformation was eventually found though. The data was reflected in the y-axis and a constant was added (so that the variance was now increasing instead of decreasing), then a log transformation was applied.
(Full transformation: $ln(100 - x_t)$)

The best model was judged to be ARIMA (0, 0, 1) × $(2, 1, 0)_{12}$ (with transformation) (*Figure 4.20* (*page* 44) Shows the ACF and PACF plots of residuals, four in one residual plot, box-pierce statistics and the parameter estimates with p-values for the model.)

The parameter coefficients were significant and the box-pierce p-values were insignificant. The residuals appear to be randomly distributed but the histogram suggested that they were still slightly skewed. It was difficult to find a good model for the Florida data but this final model should be adequate.



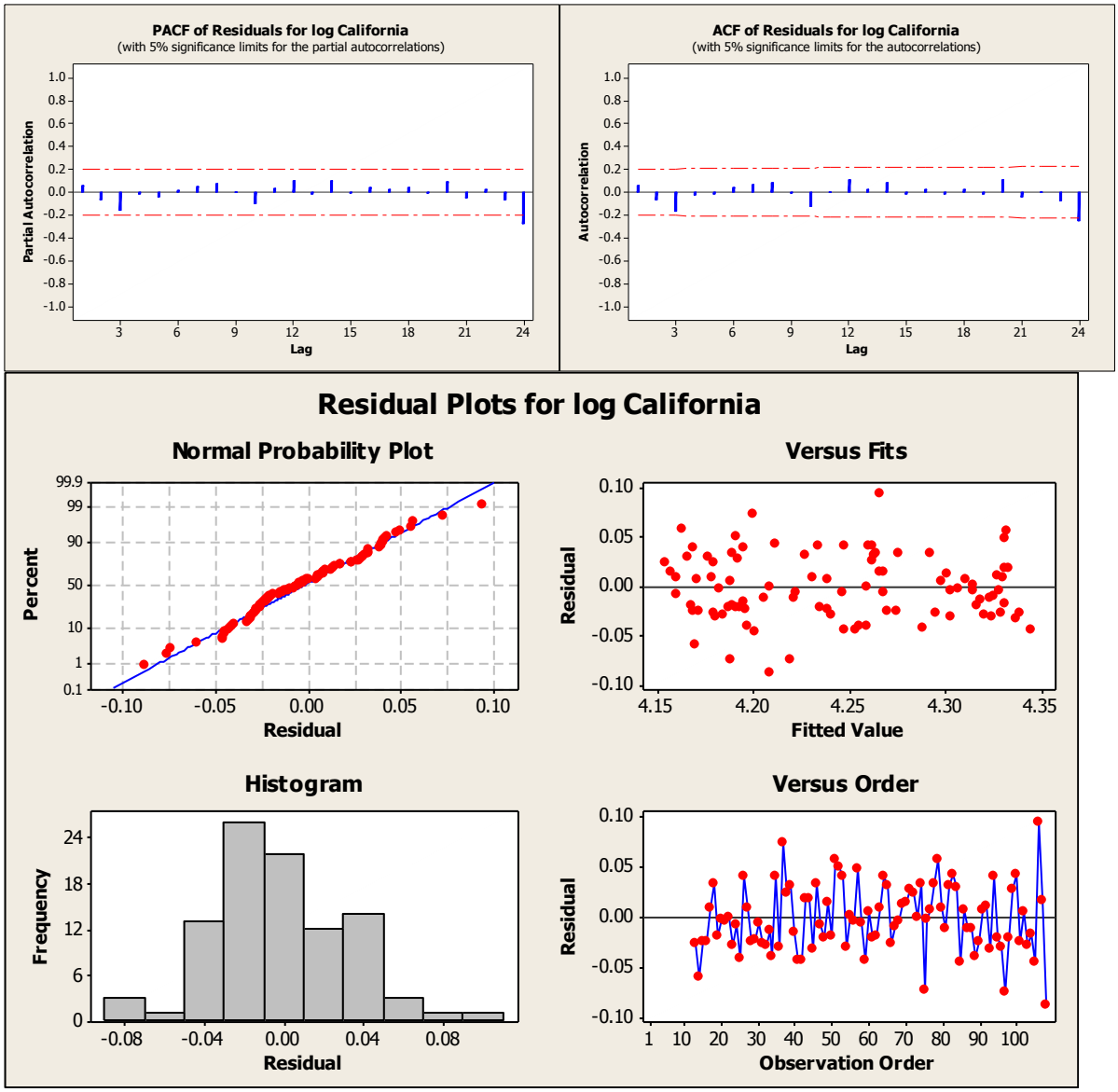

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag             12     24     36     48
Chi-Square    17.4   24.5   38.7   47.5
DF              10     22     34     46
P-Value      0.065  0.321  0.267  0.413

Final Estimates of Parameters

Type            Coef     SE Coef       T       P
SMA  12       0.9001      0.0882   10.20   0.000
Constant  -0.0008661  0.0009958   -0.87   0.387
```

Figure 4.14 *Residual plots, box-pierce statistics and parameter estimates for log Oregon's ARIMA (0, 0, 0) × (0, 1, 1)$_{12}$ model.*



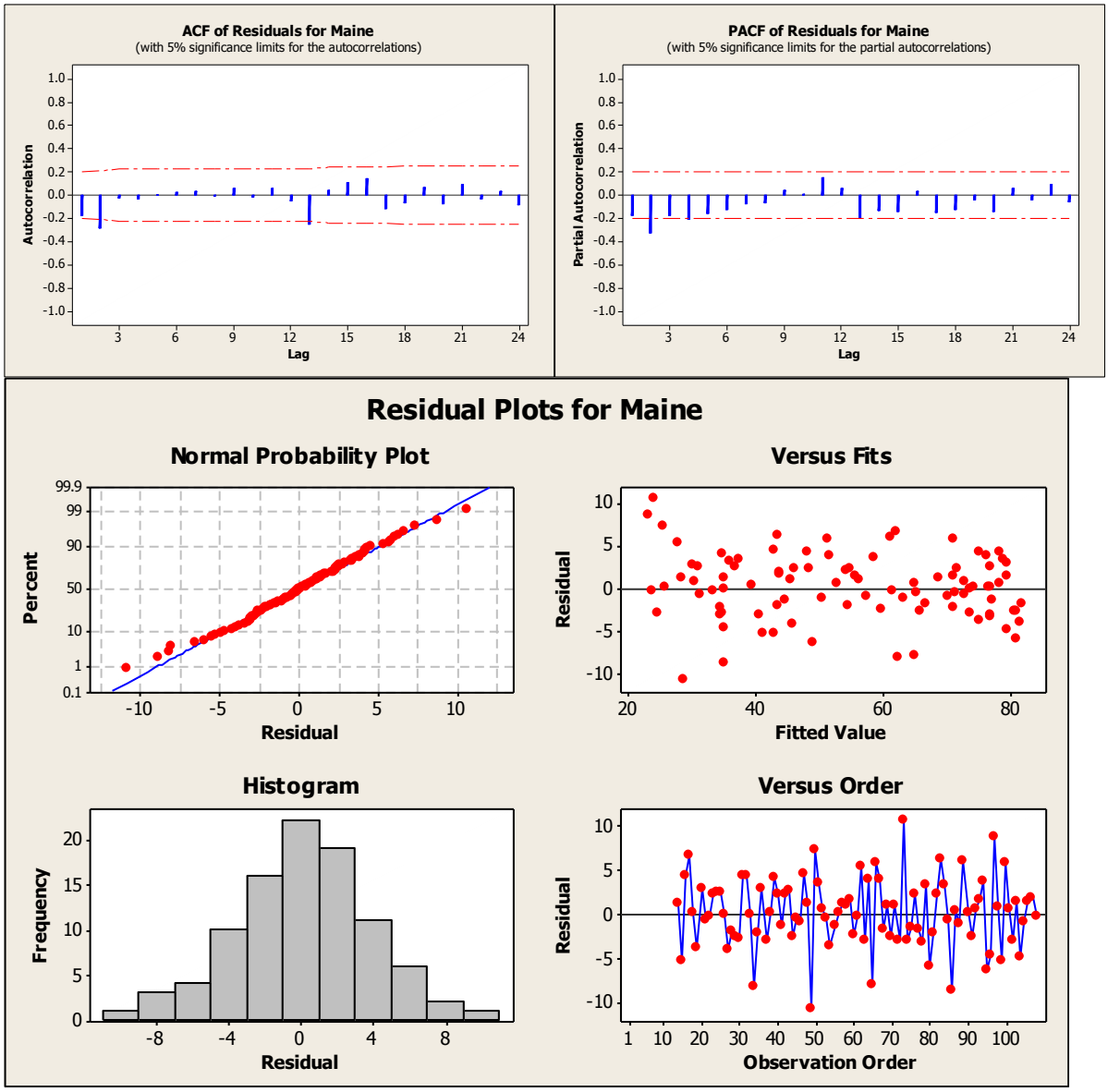
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	7.9	19.9	30.3	41.1
DF	9	21	33	45
P-Value	0.540	0.529	0.602	0.637

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
SMA	12	1.3447	0.1179	11.41	0.000
SMA	24	-0.4932	0.1305	-3.78	0.000
Constant		0.0028099	0.0005711	4.92	0.000

Figure 4.15 *Residual plots, box-pierce statistics and parameter estimates for log California's ARIMA (0, 0, 0) × (0, 1, 2)$_{12}$ model.*



Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	12.6	28.9	43.7	52.9
DF	9	21	33	45
P-Value	0.183	0.117	0.101	0.195

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	-0.4635	0.0908	-5.11	0.000
SMA	12	0.8405	0.0930	9.04	0.000
Constant		0.00487	0.09719	0.05	0.960

Figure 4.16 *Residual plots, box-pierce statistics and parameter estimates for Maine's ARIMA (1, 1, 0) × (0, 1, 1)$_{12}$ model.*

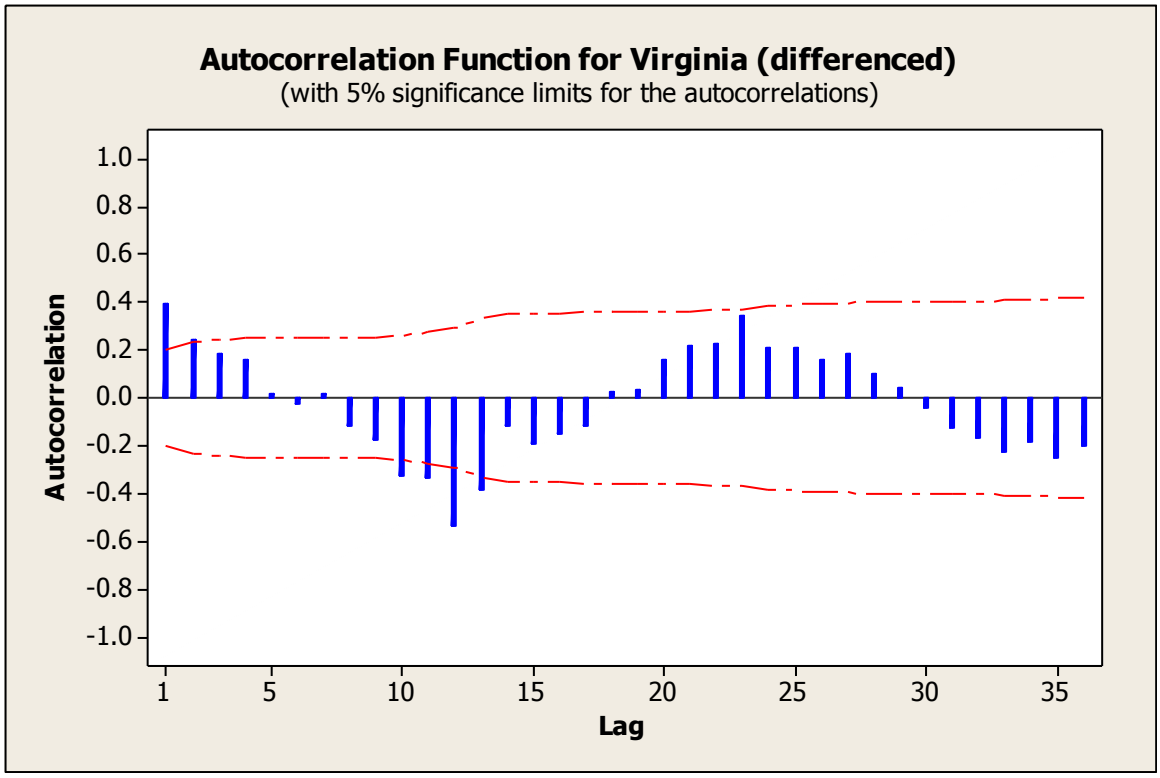


Figure 4.17 *ACF plot for Virginia's differenced data.*

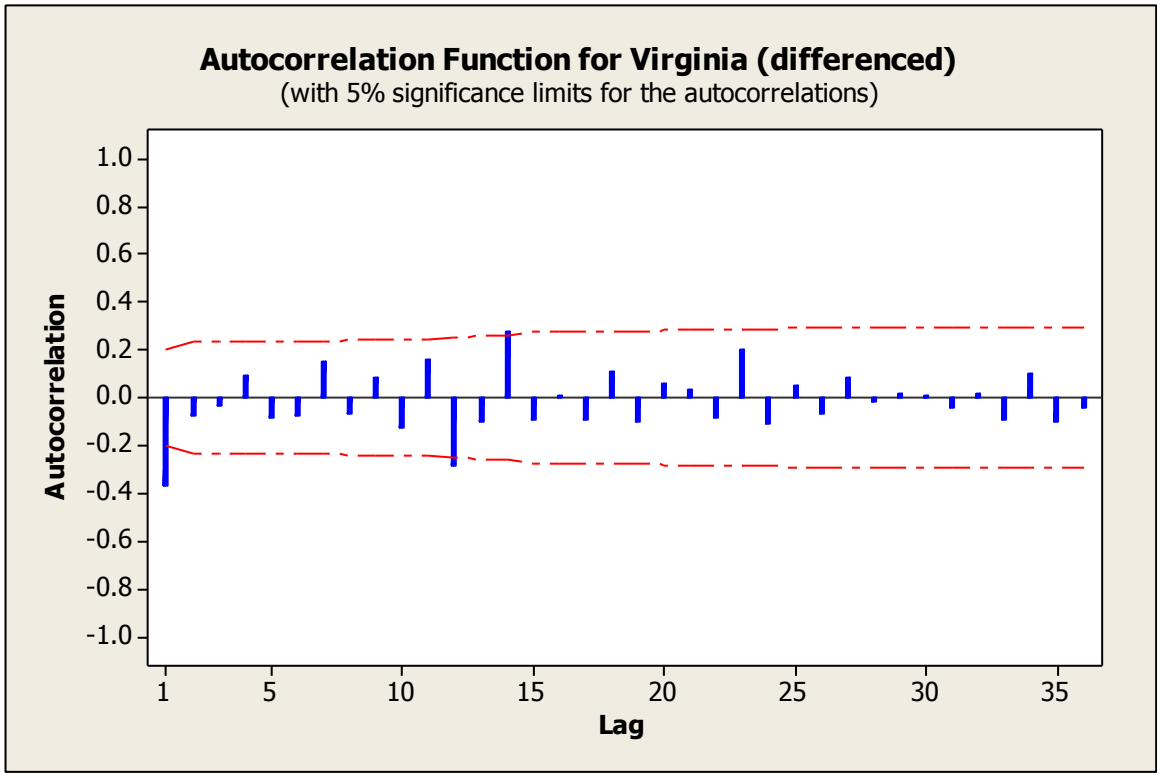
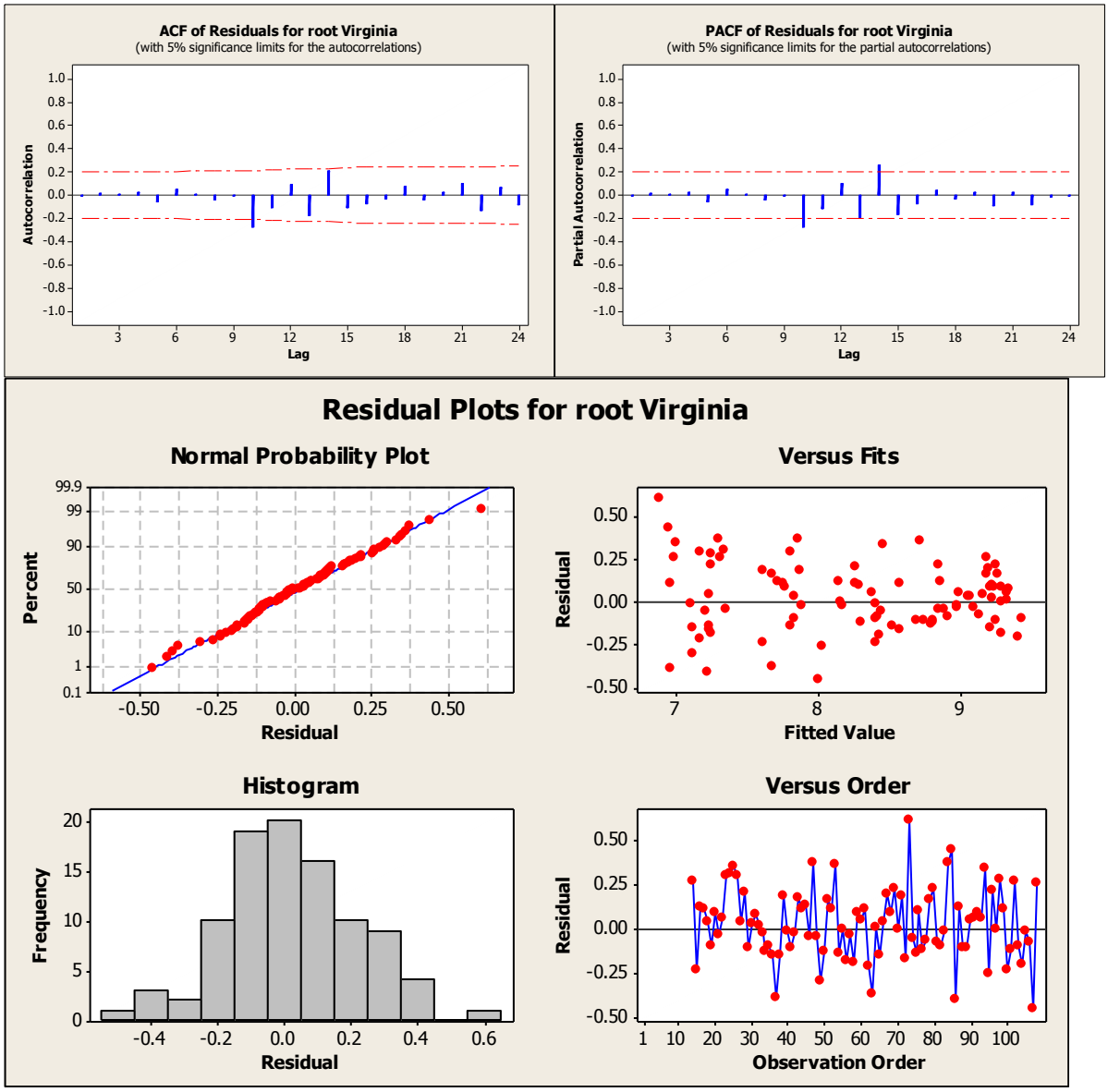


Figure 4.18 *ACF plot for Virginia's differenced data (2nd order).*



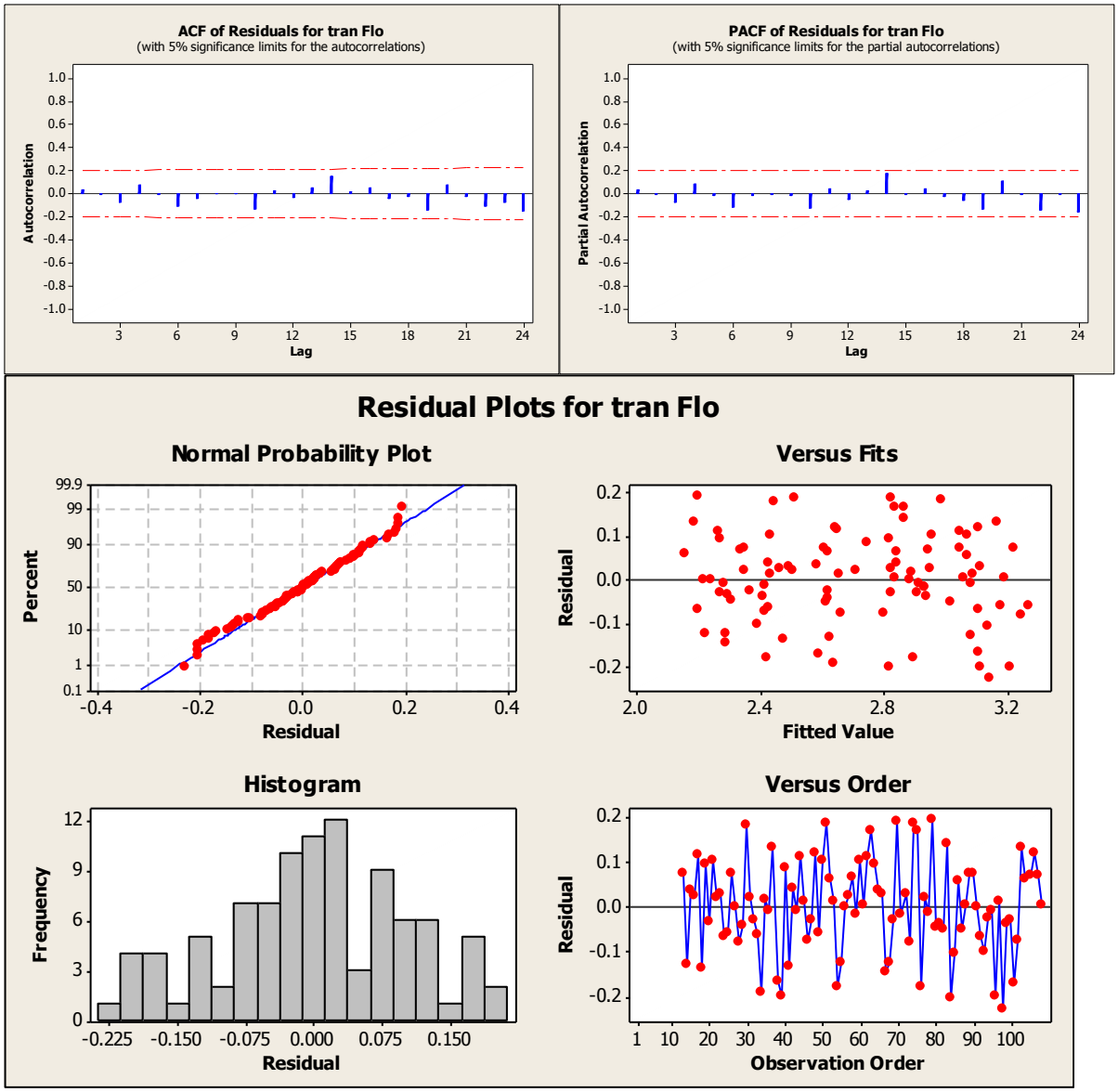
Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	11.2	28.1	44.3	52.0
DF	8	20	32	44
P-Value	0.192	0.108	0.073	0.190

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
SAR	12	-0.2802	0.1236	-2.27	0.026
MA	1	0.8875	0.0536	16.55	0.000
SMA	12	0.8219	0.1104	7.44	0.000
Constant		-0.0000241	0.0006255	-0.04	0.969

Figure 4.19 *Residual plots, box-pierce statistics and parameter estimates for root Virginia's ARIMA (0, 1, 1) × (1, 1, 1)$_{12}$ model.*



Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	4.9	16.7	27.9	34.5
DF	8	20	32	44
P-Value	0.772	0.673	0.675	0.848

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
SAR	12	-0.7637	0.0939	-8.13	0.000
SAR	24	-0.5289	0.0953	-5.55	0.000
MA	1	-0.2310	0.1025	-2.25	0.027
Constant		-0.01507	0.01304	-1.16	0.251

Figure 4.20 *Residual plots, box-pierce statistics and parameter estimates for transformed Florida ARIMA (0, 0, 1) × (2, 1, 0)$_{12}$ model.*

4.3 Model comparisons

Most of what can be judged by this process is what had to be done to find a suitable model. The model equation tells us very little about the state's temperature, it is primarily used for forecasting purposes (see section 4.4). *Table 4.1* had been prepared to compare the models directly:

Table 4.1 *Parameters and differencing procedures used by each State's model.*

	d=1	D=1	AR(1)	MA(1)	SAR(1)	SAR(2)	SMA(1)	SMA(2)
Washington	*	*	*				*	
Oregon		*					*	
California		*					*	*
Maine	*	*	*				*	
Virginia	*	*		*	*		*	
Florida		*		*	*	*		

This table reveals many similarities between the coastal states. It shows that most of the states required at least one seasonal moving average term; this implies that current temperatures in the United States are not necessarily determined by the temperatures of the previous years but rather the residuals of these temperatures.

It can be seen that Washington and Maine have almost identical models. They both required non seasonal differencing of one and they both have the same parameters (AR(1) and SMA(1)). The coefficients of these parameters are very similar as well; the 95% confidence intervals for the parameter estimates even overlap so there is no significant evidence to conclude that the models are different. This is shown by *table 4.2.*

Table 4.2 *Parameter estimates and their confidence intervals.*

State	Parameter	Estimate	Lower CI	Upper CI
Washington	AR	-0.6097	-0.801536	-0.417864
Maine	AR	-0.4635	-0.644192	-0.282808
Washington	SMA	0.8993	0.743284	1.055316
Maine	SMA	0.8405	0.65543	1.02557

It may be possible to conclude that the natures of the two States' temperatures are very similar except Maine is generally colder than Washington which is probably due to it being at a higher latitude. However it was necessary to apply a log transformation to Washington's data so one behavioural difference could be that Washington's temperature has become more varied through time.

The other West Coast States also required a log transformation to deal with the increasing variance within their data. Two of the East Coast States required transformations that dealt with a decreasing variance. This suggests that the West coast's temperature is becoming more erratic whilst the East coast's temperature is stabilising.

Oregon and California have similar models as they only required SMA terms. However California's model did need a second SMA term which suggests that California has a more erratic climate than Oregon. This was also evident in the time series graph for California. Virginia and Florida share a similar relationship as they too have similar parameters in their models but Florida required one extra SAR term because its temperature was slightly more erratic than Virginia's (which can also be seen by comparing their time series graphs).

The West Coast States have more simple models than the East Coast States since their models contain fewer parameters. The West Coast models are also more similar to each other than the East Coasts models. This suggests that the West Coast may have a more homogeneous climate than the East Coast. The exploratory analysis also showed that the East Coast had a more varied climate. This may be due to the Gulf Stream not having an equal influence upon the entire coast. Satellite thermal imaging does show that the Gulf Stream's warmth dissipates as it moves along the East Coast, it even changes direction towards Europe (see *figure 1.1*). The models and time series graphs seem to be reflecting the Gulf Stream's behaviour.

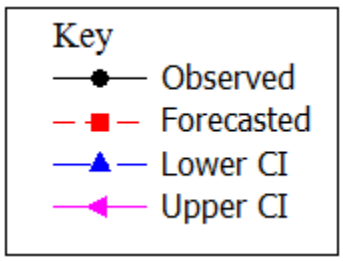
There is some evidence that a trend is present for some of the States. The best models for Washington and Maine required non seasonal differencing of order one. These models also contain non seasonal AR terms which implies that the current month's temperature is dependent on the previous month's temperature (even when the seasonal effect has been removed). Earlier inspections of the time series plots revealed no trends but there may have been some curvature in the data. This curvature may be the reason why the ARIMA models are suggesting that a trend is present. This highlights how important a visual inspection of the time series plot is and shows that ARIMA models are unable to model curvature. Perhaps another model should have been applied to the data such as a Gompertz curve (see Chatfield, 2003 chapter 2).

It was an absolute necessity for Virginia's data to undergo non-seasonal differencing (the sinusoidal pattern in the ACF remained until this was done); this was accompanied by a non-seasonal MA term. Virginia's model suggests that a trend is present and a visual inspection of the time series graph indicated an upward trend. So it is likely that between the years of 2000 and 2008 the average monthly maximum temperature of Virginia had been increasing.

Florida's model does contain a non-seasonal MA term but it did not require any non-seasonal differencing. So there could be a trend but it is less certain that there is one than in the case of Virginia. The time series plot did show some kind of upward trend in the latter years but due to the erratic nature of Florida's temperature data it would be unwise to conclude that there is a trend at this point, more data is needed to be sure that a trend is present.

4.4 Model Forecasts

For each State the monthly temperatures for the year 2009 have been forecasted using the SARIMA models. *Figures 4.21* to *4.26* show a graph of the forecasts for each State:



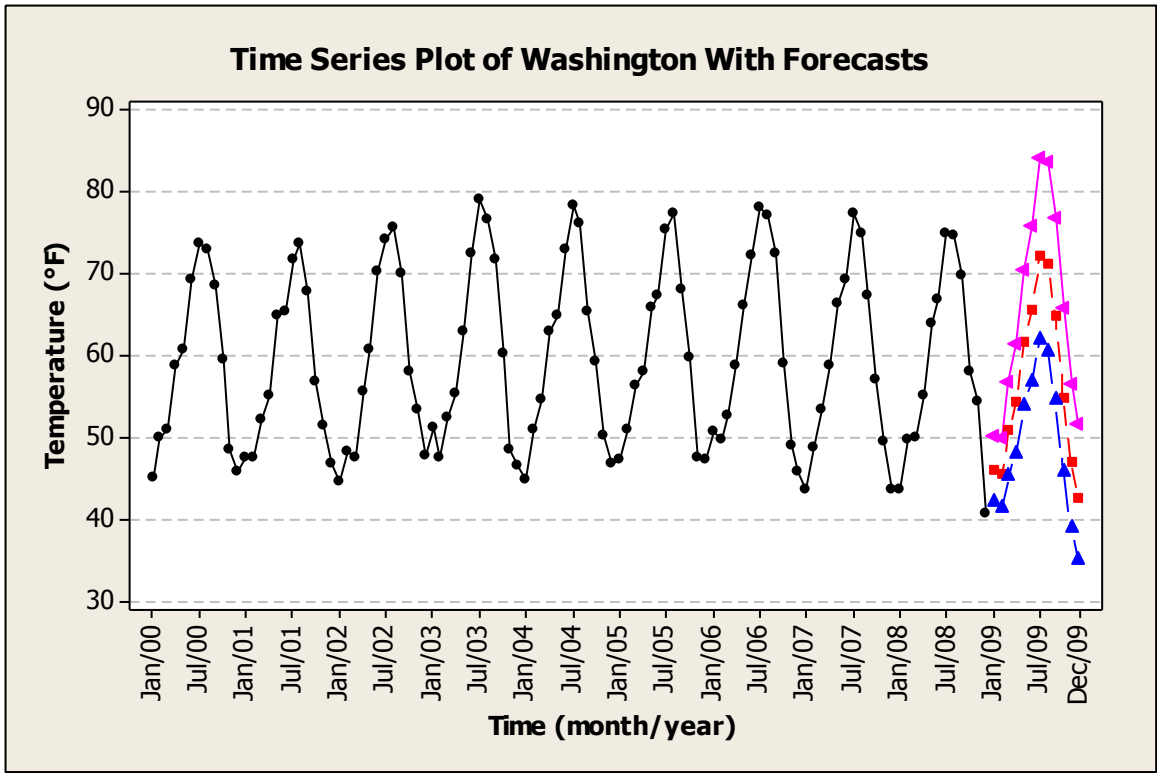


Figure 4.21 *Time series plot for Washington's monthly temperature. Forecasted values (with confidence intervals (CIs)) are plotted for the year 2009.*

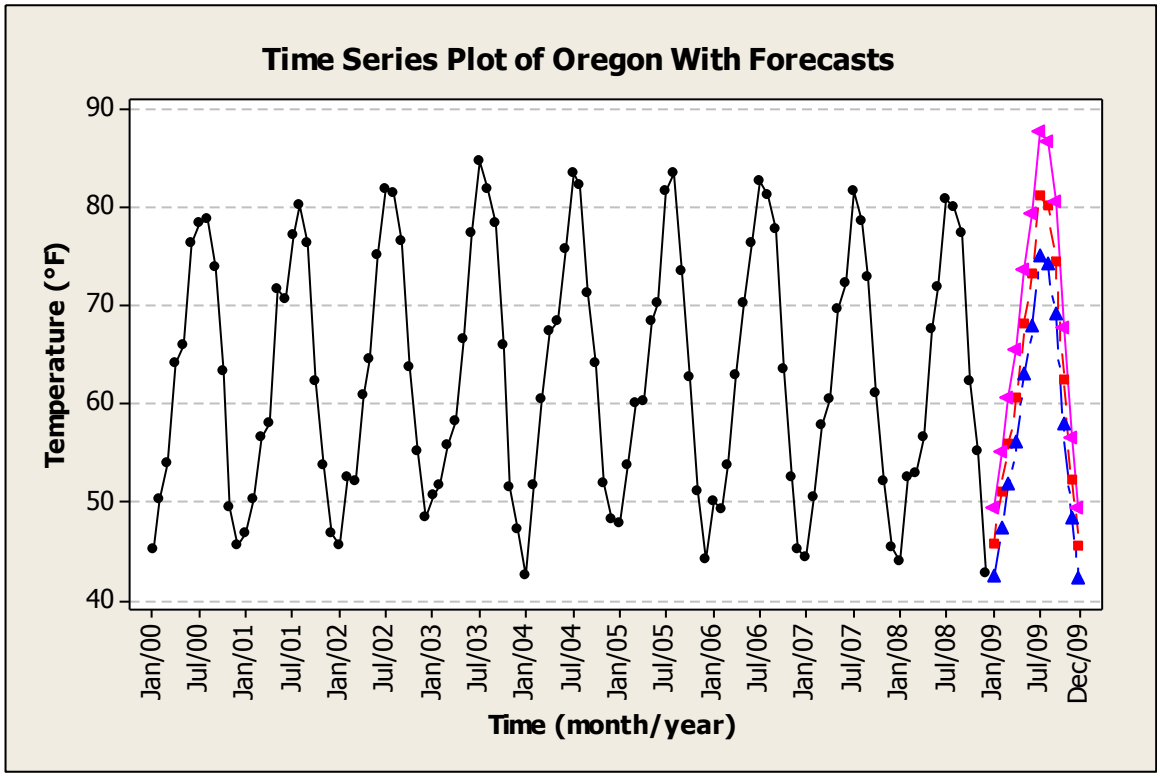


Figure 4.22 *Time series plot for Oregon's monthly temperature. Forecasted values (with confidence intervals) are plotted for the year 2009.*

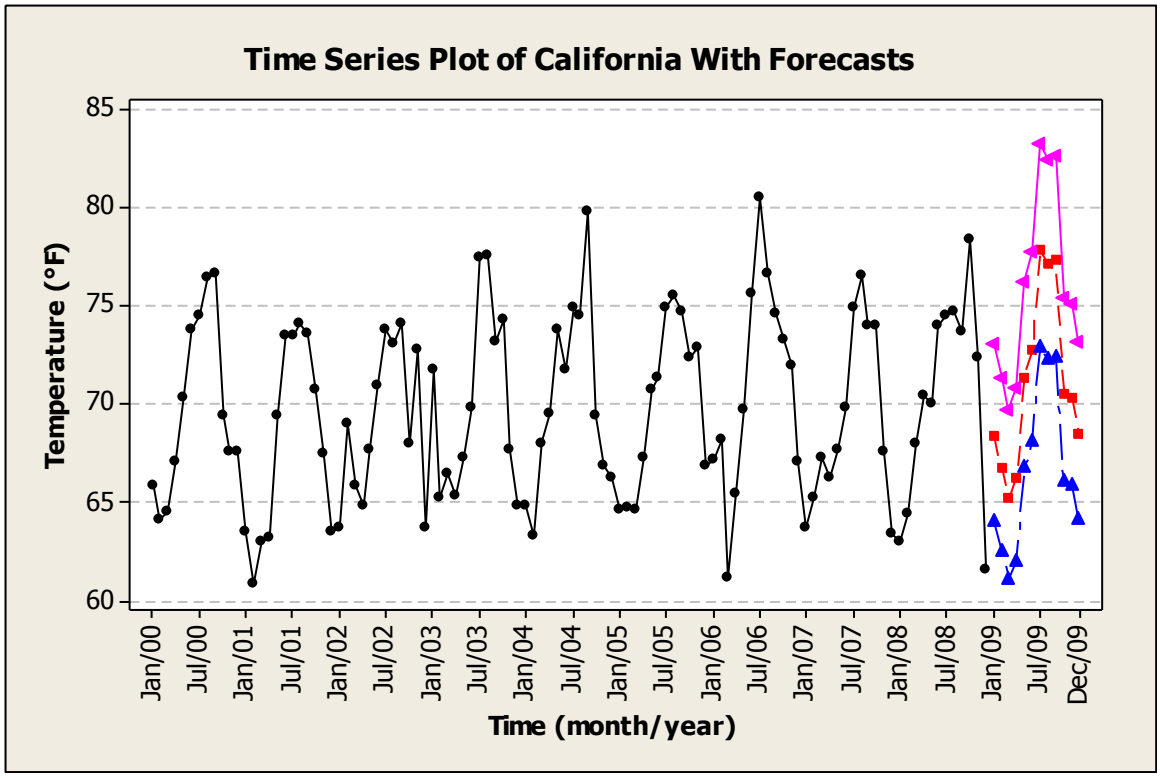


Figure 4.23 *Time series plot for California's monthly temperature. Forecasted values (with confidence intervals) are plotted for the year 2009.*

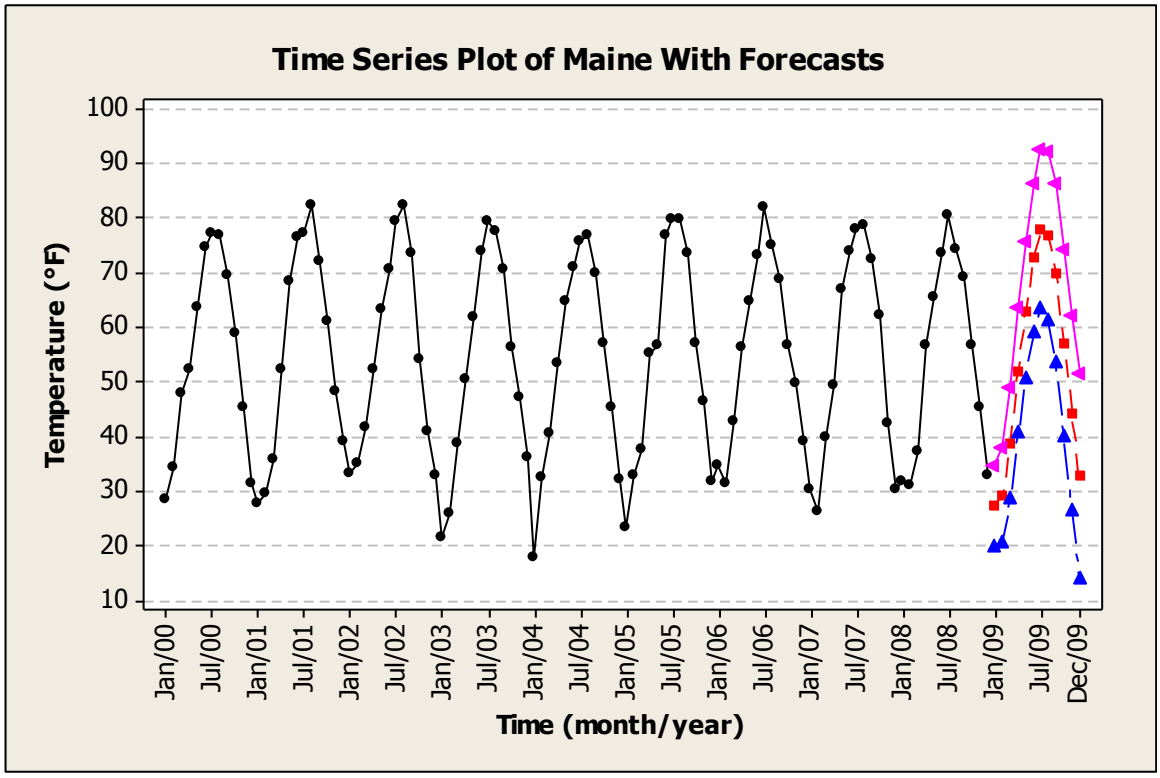


Figure 4.24 *Time series plot for Maine's monthly temperature. Forecasted values (with confidence intervals) are plotted for the year 2009.*

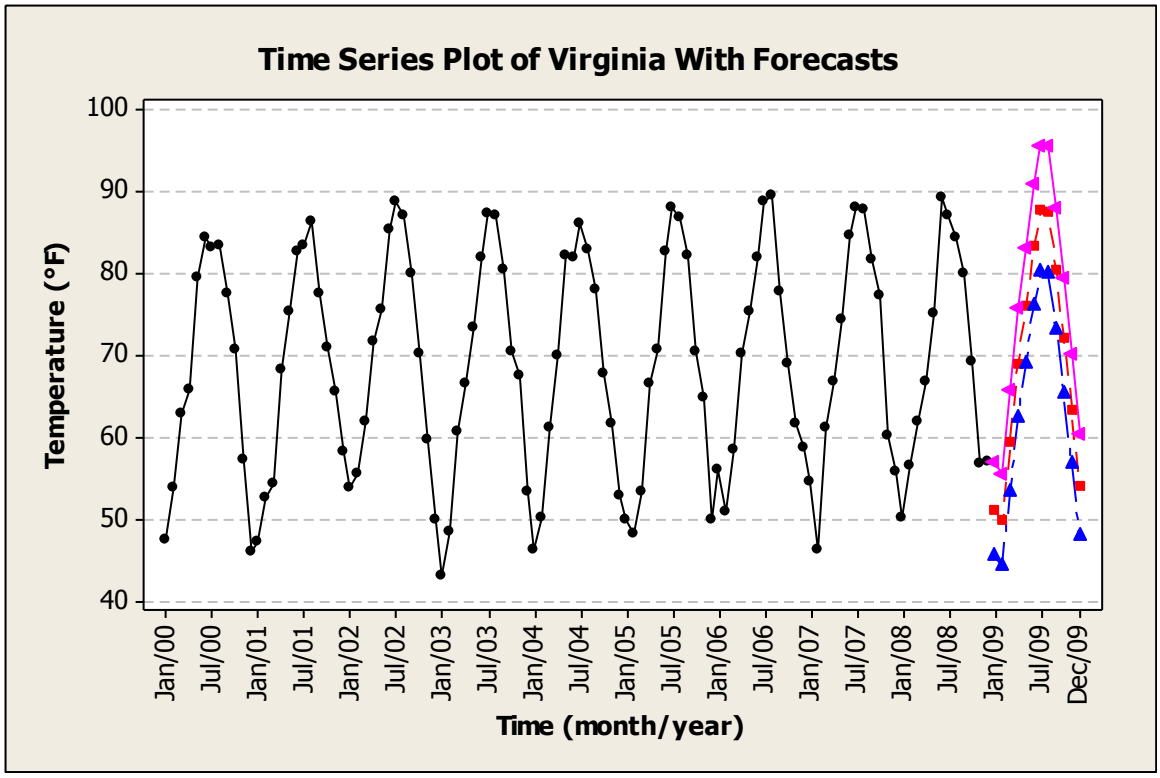


Figure 4.25 *Time series plot for Virginia's monthly temperature. Forecasted values (with confidence intervals) are plotted for the year 2009.*

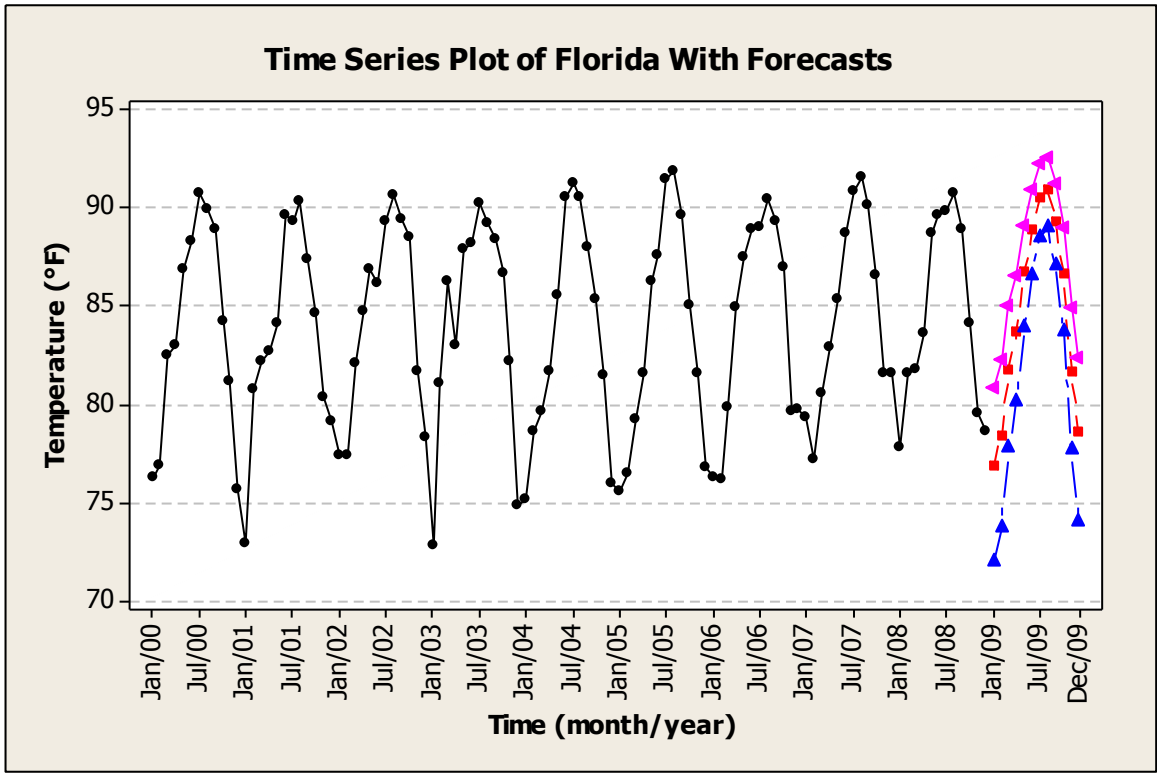


Figure 4.26 *Time series plot for Florida's monthly temperature. Forecasted values (with confidence intervals) are plotted for the year 2009.*

Some of the trends indicated by the ARIMA models are shown in the forecasts. The 2009 monthly forecasts for Washington and Maine are lower than their corresponding 2008 observations. This indicates that the temperatures in those two States have been decreasing (although this may not actually be the case as discussed earlier).

The forecasts for Virginia and Florida are not significantly higher than the 2008 observations (the means of the forecasts were approximately equal to the means of the 2008 observations). So there may not necessarily be a temperature trend for these two States.

To assess the forecasting accuracy of the models the confidence intervals for the forecasts have been examined. There is a 95% chance that the true forecast value will be between the upper and lower confidence limits (together these limits make the confidence interval). A confidence interval is produced to compensate for a model's inaccuracy in forecasting. So a wider confidence interval indicates that the model is less accurate at making forecasts than a model that has a narrow confidence interval. *Table 4.3* shows the average width of the confidence intervals for each State's forecasts.

Table 4.3 *Average confidence interval widths for each State's forecasts*

Washington	Oregon	California	Maine	Virginia	Florida
16.6	9.9	9.4	27.5	13.6	6.1

There is a clear pattern in this table. The more northern states have wider confidence intervals and the Eastern States have more varied confidence interval widths. A similar pattern was seen between the variances of the States in the exploratory analysis. So States with more variable temperatures lead to models which produce less accurate forecasts.

4.5 Summary

The model building process was not without its difficulties. Most data sets required a transformation to normalise the data and several models had to be fitted before the requirement of normally distributed residuals was met. Such problems are not mentioned in other climatological studies such as (*Barrón & Pita, 2004)* and (*Van Hecke, 2009*). It is probable that the difficulties encountered were due to the small sample size of nine years. Whilst ideal models were found eventually it is highly suspect whether they actually do represent the US climate. As a comparison the two studies mentioned earlier used 118 years' and 150 years' worth of data respectively.

Some discoveries were still made via this analysis though. The ARIMA models indicated that the West Coast had a homogeneous climate whereas there is more variation in the climate along the East Coast which is probably due to the varying nature of the Gulf Stream.

The models also revealed that the natures of Washington's and Maine's temperatures are similar despite Maine being a much colder place. It is also possible that there has been an increasing temperature trend within Virginia although the forecasts did cast doubt on this thought. The overall conclusions based on both analyses will feature in the next chapter.

CHAPTER 5
Discussion

5.1 Main Conclusions

The primary aim of this study was to examine the effects that the Atlantic and Pacific Oceans have on the temperatures of the United States' coastal regions through time. It was hoped that the analysis would reveal any differences or similarities between the East Coast and West Coast in regards to their temperatures. Of secondary importance was a comparison between the northern coastal States and southern coastal States.

Much was discovered but the entire investigation was hindered by the small sample size. Few climatological studies are conducted on less than 30 years' worth of data, for instance *Barrón & Pita* (2004) used 118 years. The data was not normally distributed as expected (even after the seasonality had been removed in some cases) and this may have been because there was not enough data available to form the distributions typically seen from monthly temperature data. This made finding adequately fitting models very difficult so the models may not even accurately represent the US climate.

Certain statistical tests required independent observations so the sample size problem only exacerbated when t-tests, F-tests and regression modelling were attempted. Only individual months could be examined which effectively reduced the sample size down to nine so no real conclusions could be made. Even though adequate ARIMA models were found it is difficult to generalise the findings of this study to other periods of time since only nine years were examined.

Some discoveries were still made in this investigation though. The main conclusion drawn from the investigation is that the East Coast has a much more varied climate than the West Coast. The eastern states had much wider temperature ranges and larger variances (the average variance for each state was different as well whilst the western states had similar temperature variances). The t-tests showed that the mean temperature was always different for the eastern states for every month of the year, but this was not always the case for the western states. In general the ARIMA models for the western States were simpler and more alike than the models for the eastern States; this also indicates that the East Coast has a more varied climate. Since the East Coast models had to be more complex their forecasted temperatures were less accurate which implies that the East Coast has a more unpredictable climate.

The more varied climate on the East Coast may be due to the changing nature of the Gulf Stream as it heads northward. Satellite thermal imaging (see figure 1.1) clearly shows that the Gulf Stream does not distribute its heat evenly across the East Coast. Florida is most exposed to the Gulf Stream which is why it was the hottest state and had the lowest temperature variability. The Gulf Stream has very little influence on Maine which is why it was the coldest state and had the highest temperature variability despite it being located on a similar latitude to the West Coast State of Washington. There is no equivalent stream flowing up the West Coast which could be why those states had more similar and less variable temperatures.

In regards to the secondary aim of this study the biggest differences between the north and south were found on the east coast and this was probably due to the nature of the Gulf Stream. There were still some general differences between the North and South of the country. It was shown that when the observations were plotted with time California and Florida had more erratic monthly temperatures than the other states. The southern states also had more complex ARIMA models than the others; they were the only ones to contain second order terms. So it is possible that the Pacific

Ocean also influences the temperatures of the south differently to those of the north. El Niño's effects are more concentrated at the equator, so this may be the reason why the southern State of California had a slightly different nature to its temperature behaviour.

There was no decisive evidence to say whether there was a trend in any of the series. The ARIMA models showed signs of a downward trend for Maine and Washington but it was thought that this was a result of there being curvature in the data rather than an actual trend. Unfortunately ARIMA models are not able to model curvature. There may have been an upward trend (an increasing mean temperature) in Virginia's data as non-seasonal differencing was required to make the data stationary. However the forecasts generated from Virginia's model did not indicate that a trend was present. Perhaps if more data was available the trend would have been more apparent.

5.2 Further Work

There is much that could be done to improve this study. The size of the data set could be increased to match those of other climatological studies (30 -200 years). This would produce more accurate models and make the conclusions more definitive and generalisable. More accurate conclusions could also be made if more locations were used. The west coast only has three states so only three data sets were used. Two of those data sets were recorded at Portland and Seattle which are very close to each other. California is a very large State but only data from Los Angeles was used, it would have been better to have used data from the rest of the state as well in order to gain a better understanding of the west coast. Nearly a thousand miles of the west coast is unaccounted for.

A major problem with the analysis is that the ARIMA models were unable to deal with the curvature in the data. The ARIMA models indicated that a trend was present in the series for Maine and Washington when a simple inspection of the time series graph showed that this was not the case. Models which are able to deal with curvature could be used such as the Gompertz curve or a logistic curve (see *Chatfield chapter 2, 2003*).

The main objective was to compare the two coasts. This study did not do that directly, it only compared the States that are on the two coasts. Perhaps the objective would have been more fulfilled had some kind of multivariate model been used as that would generate a single model that was based on observations from different locations. This would allow for a direct comparison of the east and west coasts; a Vector ARMA (VARMA) model would be ideal for this purpose (see *Chatfield chapter 9, 2003*).

There is another time series model that could be used which is also based on observations from different locations. This model would also consider the relative distances between the observatories. So on the west coast the model would apply less weight to the data from Portland and Seattle since those locations are very close to each other. More weight would be given to the data from Los Angeles so that the entire west coast is better represented. This model is known as the spatial-temporal model (*Lozano et al. 2009*).

There is still much more that could be done to improve our understanding of how the Atlantic and Pacific Oceans influence the USA's coastal temperatures. If the reader wishes to learn more about the US climate, climatology and time series analysis they could examine the sources used to aid this study. A full list of the sources used in this study is given in the references section.

References

A. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking, and N. Abe (2009)
Spatial-temporal Causal Modelling for Climate Change Attribution
T.J. Watson Research Centre
Available from: http://www.niculescu-mizil.org/papers/KDD09Climate-final.pdf

BBC Weather Centre (2009) The Gulf Stream.
Available from: http://www.bbc.co.uk/climate/impact/gulf_stream.shtml

Chris Chatfield (2003) The Analysis of Time Series: An Introduction 6th edition.
Chapman and Hall/CRC

L. García Barrón and M.F. Pita (2004)
Stochastic analysis of time series of temperatures in the south-west of the Iberian Peninsula.
University of Seville
Available from:
http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0187-62362004000400003

LIGO Scientific Collaboration (1997) Laser Interferometer Gravitational-Wave Observatory.
Available from: http://www.ligo.caltech.edu/

NASA Visible Earth (2006) Thermal Image of Gulf Stream.
Available from: http://visibleearth.nasa.gov/
http://commons.wikimedia.org/wiki/File:Golfstrom.jpg

SCRIPPS Institution of Oceanography (1997) So What is an El Niño, Anyway?
Available from: http://meteora.ucsd.edu/~pierce/elnino/whatis.html

Tanja Van Hecke (2009) Time Series Analysis to Forecast Temperature Change.
University College Ghent
Available from:
https://biblio.ugent.be/input/download?func=downloadFile&recordOId=1165208&fileOId=1169141

Worldatlas.com (2013) USA land statistics.
Available from: http://www.worldatlas.com/webimage/countrys/namerica/usstates/uslandst.htm
http://www.worldatlas.com/aatlas/populations/ctyareal.htm

Appendices

Appendix A

The p-values of the t-tests comparing monthly means between all of the states.

T TESTS						
APRIL	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.000108282	6.92715E-06	0.007534984	6.14025E-06	1.10817E-08
Oregon	0.000108282	-	0.001512229	0.000601136	0.000165201	1.04067E-07
California	6.92715E-06	0.001512229	-	2.72572E-07	0.22831896	1.22645E-07
Maine	0.007534984	0.000601136	2.72572E-07	-	4.83256E-07	5.36258E-10
Virginia	6.14025E-06	0.000165201	0.22831896	4.83256E-07	-	7.72777E-09
Florida	1.10817E-08	1.04067E-07	1.22645E-07	5.36258E-10	7.72777E-09	-
AUGUST	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	9.8372E-08	0.986814392	0.033265908	1.76907E-07	2.65039E-09
Oregon	9.8372E-08	-	0.000124117	0.049326709	0.000223208	1.33242E-07
California	0.986814392	0.000124117	-	0.051163673	6.0949E-07	7.28148E-09
Maine	0.033265908	0.049326709	0.051163673	-	9.64838E-05	1.36043E-06
Virginia	1.76907E-07	0.000223208	6.0949E-07	9.64838E-05	-	0.000345836
Florida	2.65039E-09	1.33242E-07	7.28148E-09	1.36043E-06	0.000345836	-
DECEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.61743121	1.86982E-09	1.13255E-05	0.003341382	5.09248E-09
Oregon	0.61743121	-	2.01045E-08	8.72079E-06	0.002121341	1.7214E-09
California	1.86982E-09	2.01045E-08	-	1.14702E-08	0.000295736	5.38556E-06
Maine	1.13255E-05	8.72079E-06	1.14702E-08	-	1.57933E-07	5.40115E-10
Virginia	0.003341382	0.002121341	0.000295736	1.57933E-07	-	2.75248E-08
Florida	5.09248E-09	1.7214E-09	5.38556E-06	5.40115E-10	2.75248E-08	-
FEBRUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.004481238	1.62093E-07	5.09165E-08	0.123597575	1.90827E-09
Oregon	0.004481238	-	4.16802E-07	5.94E-08	0.958003081	6.08909E-10
California	1.62093E-07	4.16802E-07	-	2.65284E-09	8.3377E-06	7.40791E-06
Maine	5.09165E-08	5.94E-08	2.65284E-09	-	3.87753E-08	1.38901E-09
Virginia	0.123597575	0.958003081	8.3377E-06	3.87753E-08	-	1.46738E-08
Florida	1.90827E-09	6.08909E-10	7.40791E-06	1.38901E-09	1.46738E-08	-
JANUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.576632034	1.40224E-09	3.09262E-05	0.114250099	5.82382E-08
Oregon	0.576632034	-	2.54988E-09	1.77635E-05	0.077124389	3.06355E-08
California	1.40224E-09	2.54988E-09	-	2.48895E-07	5.56453E-05	7.36466E-05
Maine	3.09262E-05	1.77635E-05	2.48895E-07	-	6.5821E-08	1.71788E-09
Virginia	0.114250099	0.077124389	5.56453E-05	6.5821E-08	-	5.47297E-09
Florida	5.82382E-08	3.06355E-08	7.36466E-05	1.71788E-09	5.47297E-09	-
JULY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	1.78472E-07	0.563151008	0.009941604	3.20699E-07	9.42058E-08
Oregon	1.78472E-07	-	4.45029E-05	0.03502519	1.15832E-05	3.19168E-06
California	0.563151008	4.45029E-05	-	0.000384827	3.74467E-07	1.44537E-07
Maine	0.009941604	0.03502519	0.000384827	-	9.35E-07	9.26646E-07
Virginia	3.20699E-07	1.15832E-05	3.74467E-07	9.35E-07	-	0.002510613
Florida	9.42058E-08	3.19168E-06	1.44537E-07	9.26646E-07	0.002510613	-
JUNE	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	1.38599E-05	0.055973946	0.015366149	8.61149E-06	5.60571E-08
Oregon	1.38599E-05	-	0.177352421	0.972044819	8.72215E-05	6.19856E-07
California	0.055973946	0.177352421	-	0.090961358	2.26544E-06	8.04126E-09
Maine	0.015366149	0.972044819	0.090961358	-	2.21665E-05	7.59362E-08
Virginia	8.61149E-06	8.72215E-05	2.26544E-06	2.21665E-05	-	0.00103212
Florida	5.60571E-08	6.19856E-07	8.04126E-09	7.59362E-08	0.00103212	-
MARCH	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	3.51128E-05	3.73216E-06	9.37326E-05	0.003365156	1.9025E-08
Oregon	3.51128E-05	-	3.18442E-05	2.87285E-05	0.089435716	1.07961E-07
California	3.73216E-06	3.18442E-05	-	2.42613E-07	0.000356031	1.88614E-07
Maine	9.37326E-05	2.87285E-05	2.42613E-07	-	1.25243E-07	2.21793E-09
Virginia	0.003365156	0.089435716	0.000356031	1.25243E-07	-	6.44213E-08
Florida	1.9025E-08	1.07961E-07	1.88614E-07	2.21793E-09	6.44213E-08	-
MAY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	8.28835E-06	0.000217221	0.962959907	5.4439E-05	1.44869E-08
Oregon	8.28835E-06	-	0.120682351	0.005099398	0.000629479	1.28594E-07
California	0.000217221	0.120682351	-	0.00370027	0.000131475	5.55021E-08
Maine	0.962959907	0.005099398	0.00370027	-	1.22054E-05	1.93502E-07
Virginia	5.4439E-05	0.000629479	0.000131475	1.22054E-05	-	2.9672E-05
Florida	1.44869E-08	1.28594E-07	5.55021E-08	1.93502E-07	2.9672E-05	-
NOVEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.000180311	4.72532E-08	0.012328934	0.000182078	1.02415E-09
Oregon	0.000180311	-	2.66629E-08	0.00055735	0.000256339	4.10902E-10
California	4.72532E-08	2.66629E-08	-	8.41052E-08	0.001443637	4.18259E-06
Maine	0.012328934	0.00055735	8.41052E-08	-	5.54114E-07	8.24122E-10
Virginia	0.000182078	0.000256339	0.001443637	5.54114E-07	-	1.72517E-07
Florida	1.02415E-09	4.10902E-10	4.18259E-06	8.24122E-10	1.72517E-07	-
OCTOBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	8.24849E-07	2.37501E-06	0.541046722	6.14228E-06	1.21156E-10
Oregon	8.24849E-07	-	8.74941E-05	0.002333711	0.000276049	1.49085E-10
California	2.37501E-06	8.74941E-05	-	3.8303E-06	0.269845522	5.47136E-06
Maine	0.541046722	0.002333711	3.8303E-06	-	6.76503E-08	8.0513E-09
Virginia	6.14228E-06	0.000276049	0.269845522	6.76503E-08	-	2.40913E-07
Florida	1.21156E-10	1.49085E-10	5.47136E-06	8.0513E-09	2.40913E-07	-
SEPTEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	2.06182E-07	0.001860156	0.076429645	3.99803E-06	3.56764E-09
Oregon	2.06182E-07	-	0.791810874	0.006331318	0.004214871	3.9604E-07
California	0.001860156	0.791810874	-	0.007904431	0.003116306	1.3101E-07
Maine	0.076429645	0.006331318	0.007904431	-	3.77352E-07	1.89031E-09
Virginia	3.99803E-06	0.004214871	0.003116306	3.77352E-07	-	6.07595E-08
Florida	3.56764E-09	3.9604E-07	1.3101E-07	1.89031E-09	6.07595E-08	-

Appendix B

The p-values of the F-tests comparing monthly variances between all of the states.

F TESTS						
APRIL	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.486152117	0.708845585	0.947247087	0.527878444	0.035080841
Oregon	0.486152117	-	0.288604392	0.446361816	0.190253352	0.007019484
California	0.708845585	0.288604392	-	0.758473288	0.795022871	0.075995743
Maine	0.947247087	0.446361816	0.758473288	-	0.571505216	0.040420521
Virginia	0.527878444	0.190253352	0.795022871	0.571505216	-	0.12461357
Florida	0.035080841	0.007019484	0.075995743	0.040420521	0.1246136	-
AUGUST	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.903881249	0.892472963	0.086897892	0.324201534	0.077867063
Oregon	0.903881249	-	0.798104164	0.109355081	0.385205222	0.061171252
California	0.892472963	0.798104164	-	0.06655717	0.264251516	0.101084004
Maine	0.086897892	0.109355081	0.06655717	-	0.444176121	0.00126623
Virginia	0.324201534	0.385205222	0.264251516	0.444176121	-	0.00872441
Florida	0.077867063	0.061171252	0.101084004	0.00126623	0.0087244	-
DECEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.652888691	0.859868663	0.253817757	0.068834623	0.993017477
Oregon	0.652888691	-	0.784155406	0.117253423	0.026605488	0.659171789
California	0.859868663	0.784155406	-	0.190353158	0.047979143	0.866741977
Maine	0.253817757	0.117253423	0.190353158	-	0.469453716	0.250341275
Virginia	0.068834623	0.026605488	0.047979143	0.469453716	-	0.06763948
Florida	0.993017477	0.659171789	0.866741977	0.250341275	0.0676395	-
FEBRUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.86347518	0.10258766	0.018562175	0.012313433	0.191569273
Oregon	0.86347518	-	0.140774105	0.027300712	0.018329426	0.253494054
California	0.10258766	0.140774105	-	0.417633225	0.325326193	0.727114555
Maine	0.018562175	0.027300712	0.417633225	-	0.85912843	0.250412657
Virginia	0.012313433	0.018329426	0.325326193	0.85912843	-	0.18732104
Florida	0.191569273	0.253494054	0.727114555	0.250412657	0.1873210	-
JANUARY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.919653385	0.878933419	0.069261863	0.280463363	0.425575263
Oregon	0.919653385	-	0.958945815	0.056473776	0.23952467	0.48537057
California	0.878933419	0.958945815	-	0.050787556	0.220439476	0.517662932
Maine	0.069261863	0.056473776	0.050787556	-	0.433344529	0.012013272
Virginia	0.280463363	0.23952467	0.220439476	0.433344529	-	0.06740607
Florida	0.425575263	0.48537057	0.517662932	0.012013272	0.0674061	-
JULY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.895614503	0.753003654	0.501698026	0.672005386	0.010043868
Oregon	0.895614503	-	0.854231977	0.587723661	0.769740031	0.013675544
California	0.753003654	0.854231977	-	0.7191757	0.913058134	0.020844834
Maine	0.501698026	0.587723661	0.7191757	-	0.802050479	0.045699847
Virginia	0.672005386	0.769740031	0.913058134	0.802050479	-	0.02660615
Florida	0.010043868	0.013675544	0.020844834	0.045699847	0.0266061	-
JUNE	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.953349485	0.447876845	0.473935769	0.750317894	0.055958845
Oregon	0.953349485	-	0.414319992	0.439188048	0.706547681	0.049589496
California	0.447876845	0.414319992	-	0.965207104	0.656936289	0.228455944
Maine	0.473935769	0.439188048	0.965207104	-	0.688585617	0.212783729
Virginia	0.750317894	0.706547681	0.656936289	0.688585617	-	0.10476158
Florida	0.055958845	0.049589496	0.228455944	0.212783729	0.1047616	-
MARCH	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.646593965	0.784763289	0.339745415	0.410527973	0.625058126
Oregon	0.646593965	-	0.466176261	0.614882096	0.712218257	0.346940108
California	0.784763289	0.466176261	-	0.223254916	0.276407896	0.828596459
Maine	0.339745415	0.614882096	0.223254916	-	0.892580425	0.154754703
Virginia	0.410527973	0.712218257	0.276407896	0.892580425	-	0.19514834
Florida	0.625058126	0.346940108	0.828596459	0.154754703	0.1951483	-
MAY	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.984394208	0.81455424	0.226772628	0.239480005	0.252392182
Oregon	0.984394208	-	0.79943557	0.234037211	0.247046939	0.244703162
California	0.81455424	0.79943557	-	0.152472143	0.161812466	0.358783234
Maine	0.226772628	0.234037211	0.152472143	-	0.972943053	0.023641539
Virginia	0.239480005	0.247046939	0.161812466	0.972943053	-	0.02549669
Florida	0.252392182	0.244703162	0.358783234	0.023641539	0.0254967	-
NOVEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.515510566	0.744280002	0.72903857	0.220096383	0.016544072
Oregon	0.515510566	-	0.332051615	0.322336713	0.066823209	0.067722081
California	0.744280002	0.332051615	-	0.983814088	0.362077823	0.007662986
Maine	0.72903857	0.322336713	0.983814088	-	0.372583862	0.00729643
Virginia	0.220096383	0.066823209	0.362077823	0.372583862	-	0.00074234
Florida	0.016544072	0.067722081	0.007662986	0.00729643	0.0007423	-
OCTOBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.682377237	0.012461696	0.052822312	0.036768144	0.592170129
Oregon	0.682377237	-	0.031518306	0.117898928	0.085104742	0.898651828
California	0.012461696	0.031518306	-	0.515521201	0.631219221	0.041447178
Maine	0.052822312	0.117898928	0.515521201	-	0.863466526	0.148463983
Virginia	0.036768144	0.085104742	0.631219221	0.863466526	-	0.10851064
Florida	0.592170129	0.898651828	0.041447178	0.148463983	0.1085106	-
SEPTEMBER	Washington	Oregon	California	Maine	Virginia	Florida
Washington	-	0.748228228	0.870520826	0.596993461	0.625947413	0.014464845
Oregon	0.748228228	-	0.628978892	0.398019905	0.421152613	0.006740914
California	0.870520826	0.628978892	-	0.713811271	0.744977741	0.021005443
Maine	0.596993461	0.398019905	0.713811271	-	0.966791434	0.046727902
Virginia	0.625947413	0.421152613	0.744977741	0.966791434	-	0.04280978
Florida	0.014464845	0.006740914	0.021005443	0.046727902	0.0428098	-