Rachel Bowling
Time Series
Summer 2009
Project Documentation

Introduction

For my student project, I wanted to find a topic of interest to me as you suggested on your website. As the middle child of three children, I am the only one that is married. That is all soon to change; however, as in 2010 both of my siblings are getting married: one in March and one in May. Despite numerous reasons for the May wedding date to be changed to June, my sister is insistent that the date be in May. Why? Because she claims that everyone gets married in June and she does not want to be a June bride. So, I decided that for my student project I would examine the frequency of marriages by month. After researching this topic on the internet, the best data I found was from the United Kingdom. Although this data may not be helpful in convincing my sister to change her wedding data, the data was complete and appropriate for my time series analysis. For my student project, I tested the data for white noise, seasonality, and stationarity using the statistical tests and methods outlined in the course. Last, I constructed one possible model using the Yule Walker equations for parameter estimation and ordinary least squares parameter estimation and tested the model using an ex-post forecast.

Data

I found my data on the website www.statistics.gov.uk. The exact link to my data set is: www.statistics.gov.uk/STATBASE/xsdataset.asp?More=Y&vlnk=5286&All=Y&B2.x=137&B2.y=11. The data set provides the number of marriages by month of occurrence from 1947-2003. Both single years and five year bands of data are provided. In constructing my time series model, I used the single year data. The data is from the geographical region of England and Wales. The data set provides indicators for those months that contain five Saturdays in the month as opposed to four. This would be of interest since Saturday is the more popular day for wedding ceremonies. Due to the limited complexity of my model, I did not consider this element in my model. The original data and descriptions of the data as downloaded from the website can be found on the first tab in my spreadsheet labeled “Original Data”.

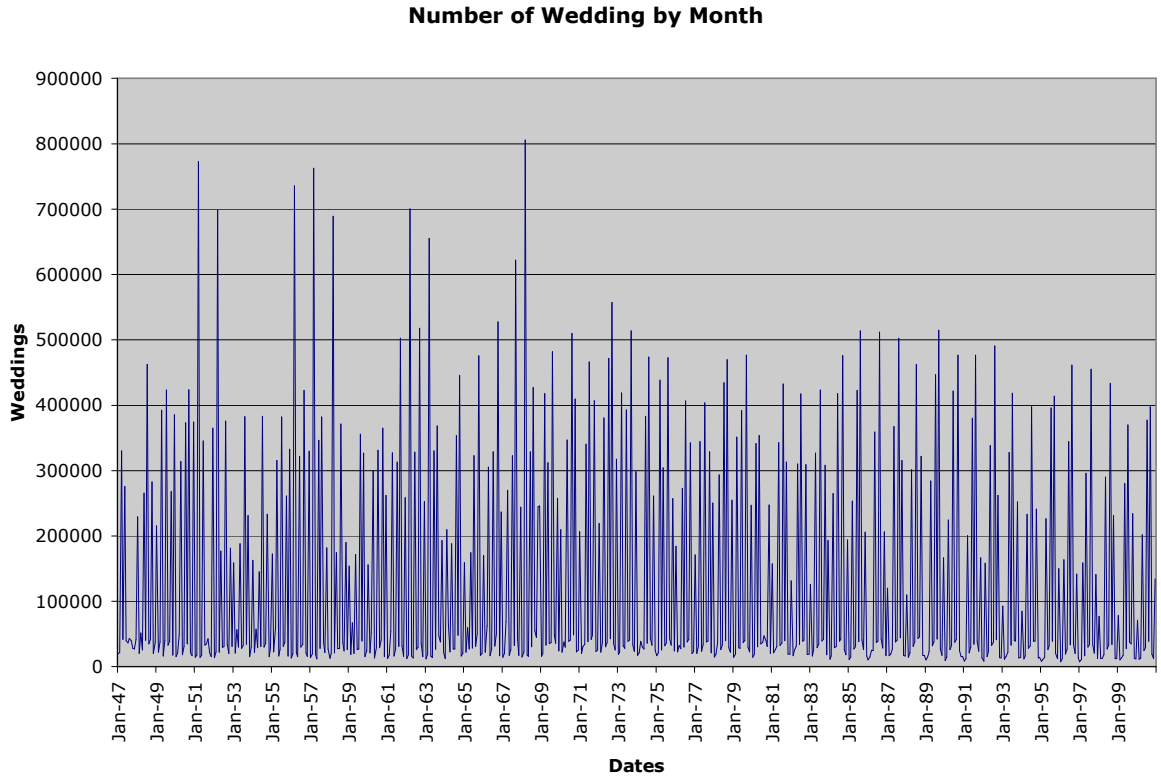
Topic

Originally, I planned to examine and adjust my data for stationarity and seasonality and then construct various possible models on which I could perform diagnostic tests and produce forecasts. However, as I began examining the data for seasonality and stationarity, the model appeared more complex than I anticipated. As a result, I decided to primarily focus on examining the nature of my data and then constructed only one model using the methods of the course. I then performed an ex-post forecast using the model.

Statistical Techniques

Examining the Original Data Set
The original data set had 57 years of data. The graph of this data is shown below:

Rachel Bowling
Time Series
Summer 2009
Project Documentation



Graph 1

After constructing this graph, I realized that it would be difficult to examine the data for patterns such as seasonality and stationarity with so many data points. I also wanted to use some of the data points in ex-post forecasts if I was able to construct a model. For this reason, rather than using years 1947-2003 to construct my model, I decide to reserve 2001-2003 for ex-post forecasts and only use 1991-2000 (10 years of data and 120 data points) to construct my model. Throughout my documentation when I refer to the original data, I am referring to years 1991-2000.

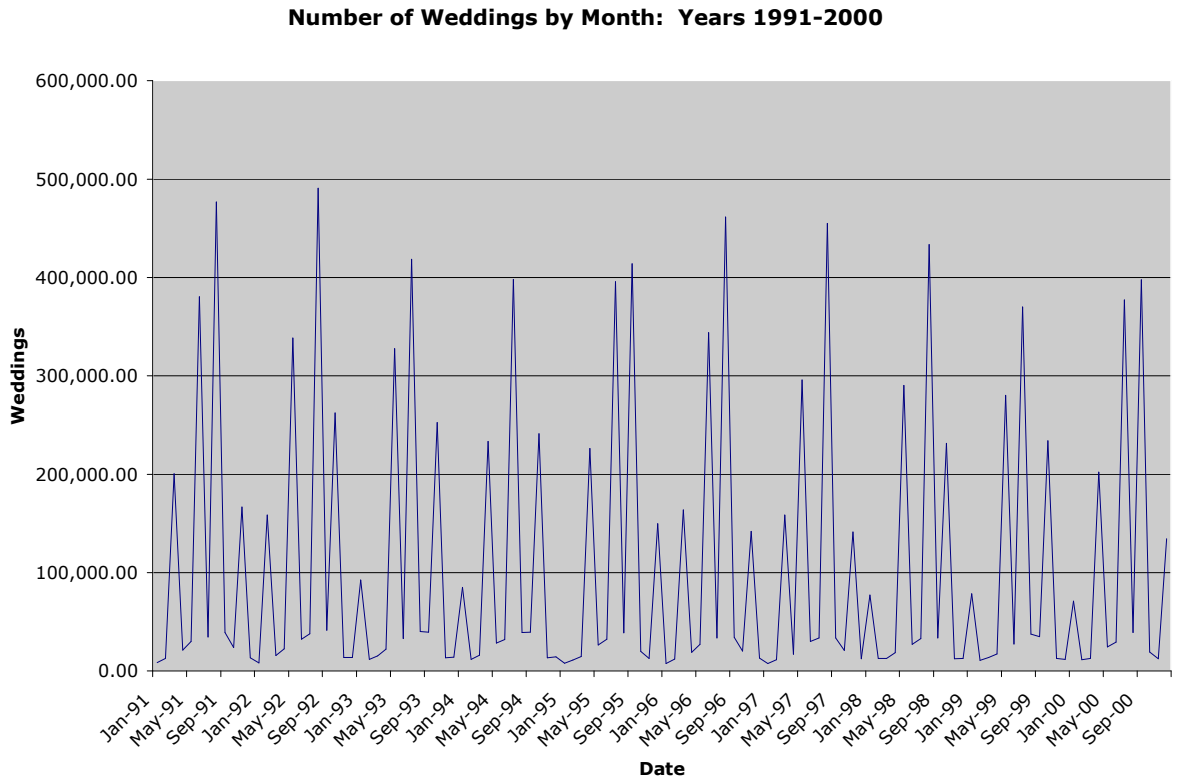
Testing for White Noise: Bartlett Test and Box and Pierce Test

Before beginning construction of my model, I wanted to ensure that my series had not been generated by a white noise process. A white noise process is one in which there is no correlation between data points ($\rho_k = 0$ for all $k<>1$) and therefore little value is gained from using a model to forecast the series. To test if the data had been generated by a white noise process, I used tests by Bartlett and Box and Pierce. The Bartlett test states that if a time series has been generated by a white noise process, then the sample autocorrelations are approximately normally distributed with mean 0 and standard deviation $(1/\sqrt{T})$ where T is the number of observations in the series. For a normal distribution, we know that 95% of all points lie with two standard deviations of the mean. Therefore, using Bartlett's test, if any sample autocorrelation coefficient is greater than two standard deviations away from zero, we can be 95% sure that the true autocorrelation coefficient is not zero and not from a white noise process. My intermediate calculations for the Bartlett test can be found on the "Formatted Data-10 yrs" tab. Using this test, 31 of my sample autocorrelations implied that the data was not generated from a white noise process; however, 88 implied that the series was generated by a white noise process. Clearly I needed

to look into this possibility more thoroughly. To do so, I used the Box and Pierce test. The Box and Pierce Test uses the test statistic: $Q = T \sum \rho_k^2$ where the sum is taken from k=1 to k=K where K is the number of sample autocorrelations used. Box and Pierce showed that this test statistic is an approximate chi square distribution with K degrees of freedom which implies that if the Q test statistic is greater than the 5% critical value (for example), then it is 95% certain that the true autocorrelation functions are not all zero. My intermediate calculations for the Box and Pierce test can be found on the "Formatted Data-10 yrs" tab. Using this test with T=120, K=119, the value of Q is 396.634. I assumed K=120 in order to use the Chi Square table in the text. The 5% critical value for K=120 is 146.57; the 0.5% critical value is 163.64. With Q exceeding both of these values, this implied that I can be 99.5% sure that the true autocorrelation functions are not all zero and therefore the times series has not been generated by a white noise process. Knowing this, I was able to proceed in my model construction for this data.

Stationarity

To learn more about the nature of my data set, I wanted to determine if the data was stationary. A stationary series is assumed to be unchanging with respect to time; that is, the characteristics of the stochastic process such as the mean, variance, and covariance should not change at different points in the series. Before examining the data, I thought that perhaps the data would not be stationary due to population growth over time. I expected to observe an overall upward trend in the data. However; after examining the graphs of the original data, it was clear that an upward trend was not evident; in fact, it appeared that a decreasing trend was evident.

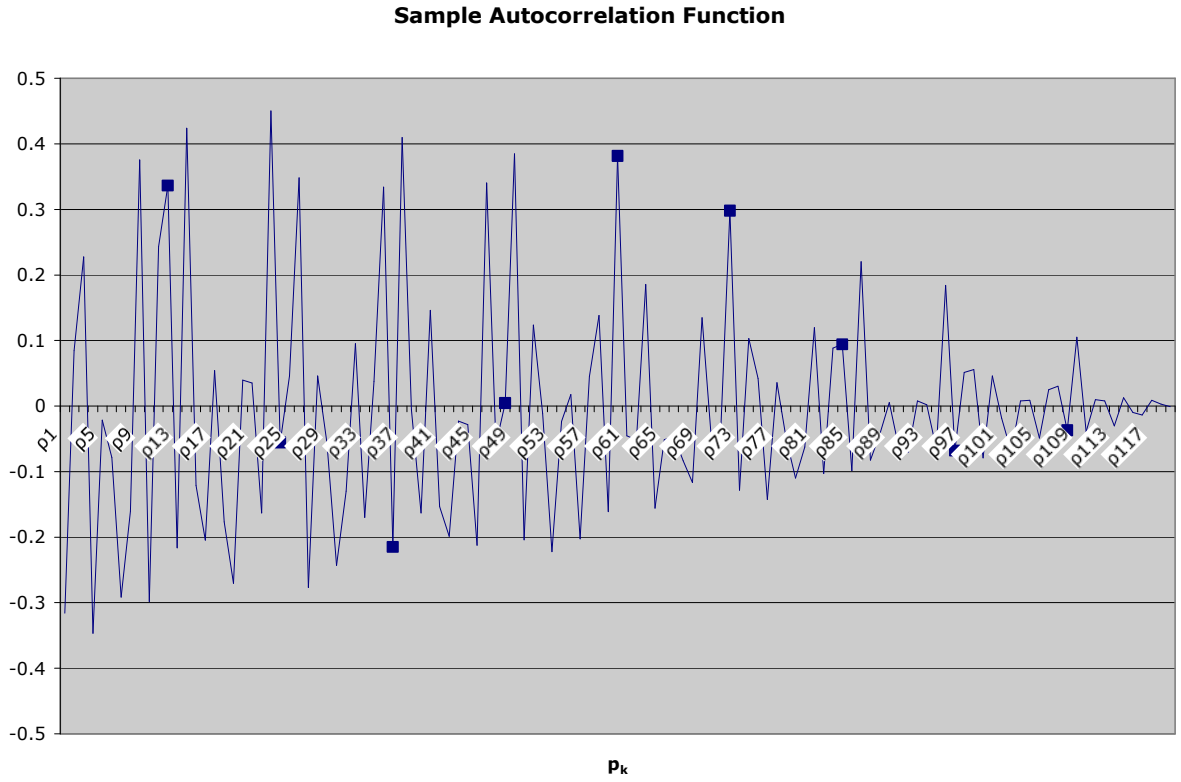


Graph 2

As my first test of stationarity, I estimated the mean and variance of the series at various points in the series: $E(y_t)$, $E(y_{t+12})$, $E(y_{t+24})$, etc. and the corresponding variance. If the series is stationary, these sample means taken at different points should be approximately equal; the sample variances taken at different points should also be approximately equal. The descriptive statistics for each point in the series are shown on the "descriptive stats" tab in my spreadsheet. A summary of the means and variances are shown below:

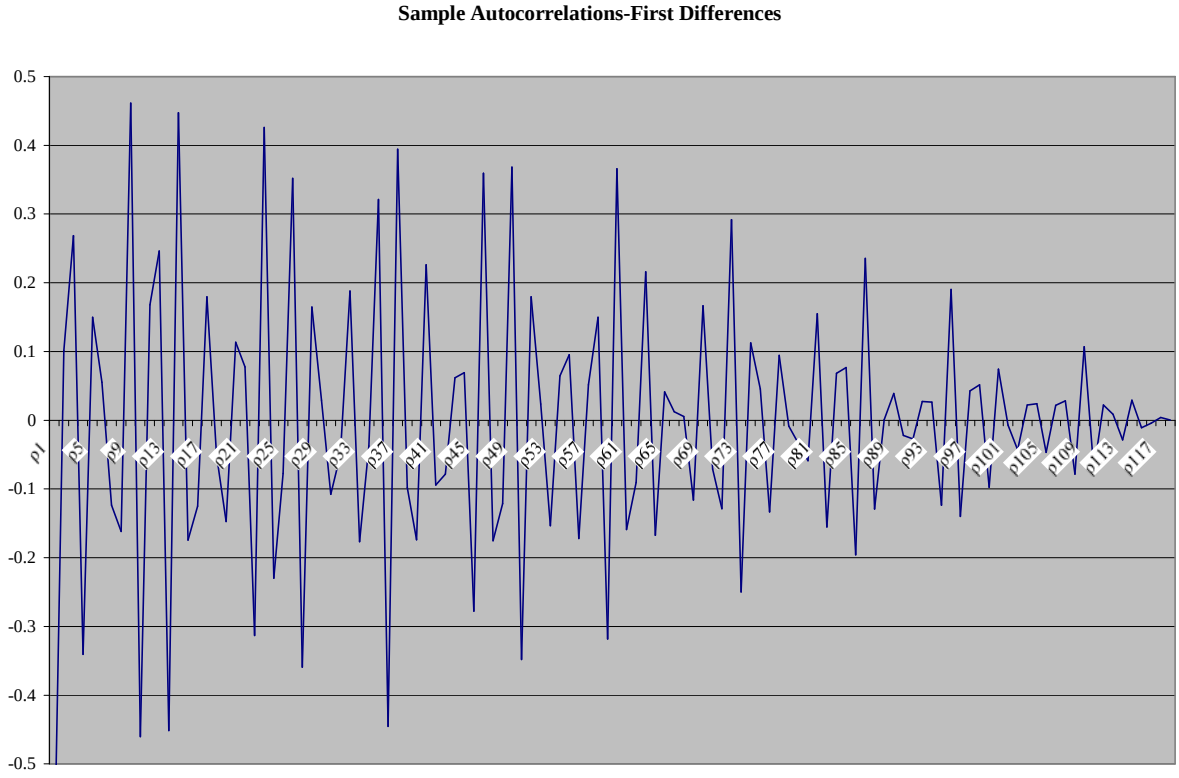
Sample Means		**Corresponding Sample Variances**
$E(y_t)$	106,565.51	19,327,803,281.46
$E(y_{t+12})$	105,359.33	18,865,084,783.20
$E(y_{t+24})$	103,561.91	18,240,601,191.08
$E(y_{t+36})$	103,103.44	18,182,215,140.44
$E(y_{t+48})$	104,274.04	18,805,170,474.07
$E(y_{t+60})$	102,618.92	18,305,417,218.52
$E(y_{t+72})$	101,620.85	17,732,405,332.43
$E(y_{t+84})$	101,647.17	17,514,830,572.09
$E(y_{t+96})$	102,613.46	17,237,765,770.78
$E(y_{t+108})$	111,036.75	20,040,522,614.93

Although the variation is not substantial, it does not appear that the means and variances are invariant with respect to time which implies a nonstationary process. As an additional measure of stationarity, the sample autocorrelation function ρ_k of a stationary series should decrease quickly as k increases. If ρ_k does not decrease quickly as k increases, this implies a nonstationary series. As shown in the graph of the sample autocorrelation below, the absolute value of ρ_k never exceeds 0.5; however, it does not decline to lower values until substantially later in the series.

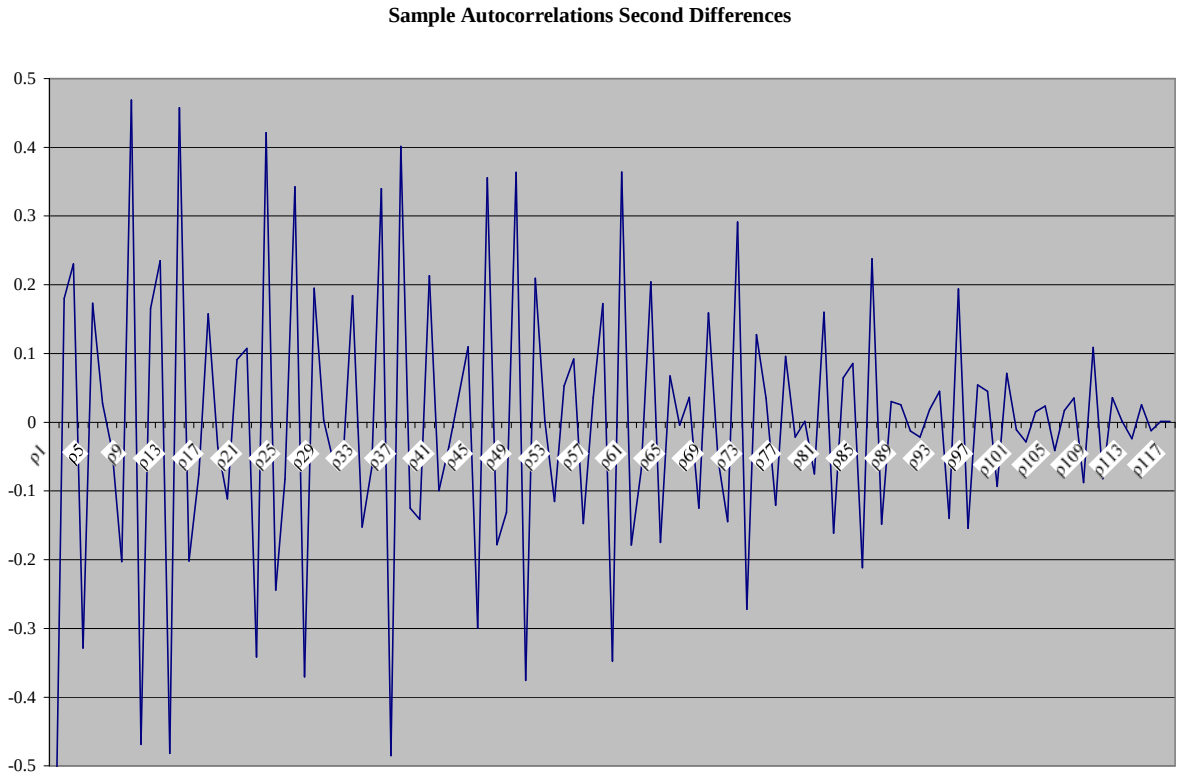


Graph 3

On page 519 of the text, it is stated that if $-1 < \rho_k < 1$ for all $k>0$, this is sufficient condition for our purposes to assume the series is stationary although the actual conditions for stationarity are more complex. Based on this criterion, the series would be considered stationarity; however for completeness I examined first and second differences to determine if this improved the stationarity of the data at all. The graphs of the autocorrelation functions of the first and second differences are shown below. It is clear that these differences do not improve stationarity.


Sample Autocorrelations-First Differences
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
ρ1
ρ5
ρ9
ρ13
ρ17
ρ21
ρ25
ρ29
ρ33
ρ37
ρ41
ρ45
ρ49
ρ53
ρ57
ρ61
ρ65
ρ69
ρ73
ρ77
ρ81
ρ85
ρ89
ρ93
ρ97
ρ101
ρ105
ρ109
ρ113
ρ117

Graph 4



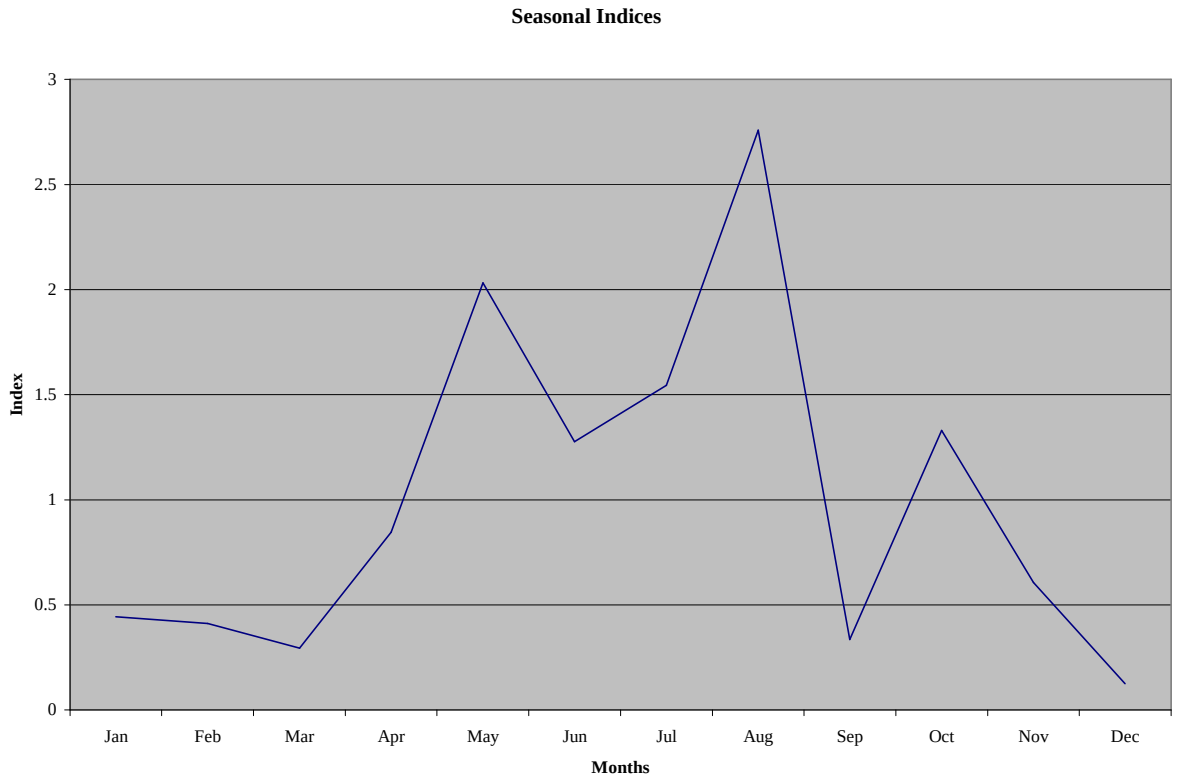
Graph 5

For my project, I assumed that my data was from a stationary process.

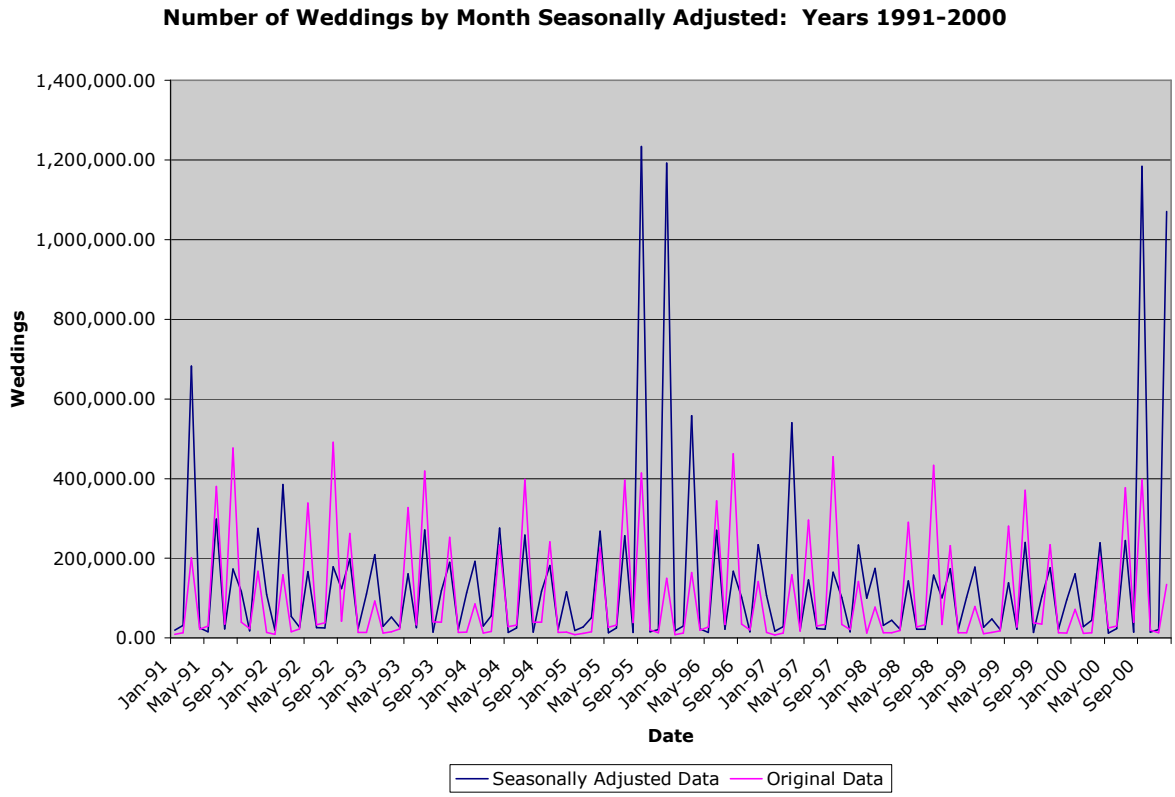
Seasonality

After examining my data for stationarity, I examined it for seasonality. Before examining the graph of my data, I expected the data to exhibit seasonality as it seems more people get married in summer months and perhaps around the Christmas and New Year's holidays. After examining my graph of the data (see Graph 2 above), a pattern was obvious. To further examine and attempt to remove the possible seasonality in my data, I tried a few of the techniques presented in the course. First, I calculated the sample autocorrelations of the data. The autocorrelation function describes how much correlation there is between neighboring data points in the series. I used ρ_k to denote my sample autocorrelations. Using this notation, ρ_1 describes the correlation between neighboring data points, ρ_3 describes the correlation between data points that are separated by three time periods: y_1 and y_4 for example, and ρ_{12} would describe the correlation between data points that are separated by 12 times periods: y_1 and y_{13} for example. For monthly data in which we expect there to be some relationship between months, as I do with my data, ρ_{12}, ρ_{24}, ρ_{36}, and corresponding multiples would be of interest because these values would indicate the relationship between the same months in different years. I expected my sample autocorrelation function to exhibit spikes at these multiples. See Graph 3 above. On my graph, I have indicated the multiples of 12 sample autocorrelations with a square. As you can see, the data does not exhibit the regular peaks I expected. The spikes in the sample autocorrelation function do not exhibit an obvious pattern.

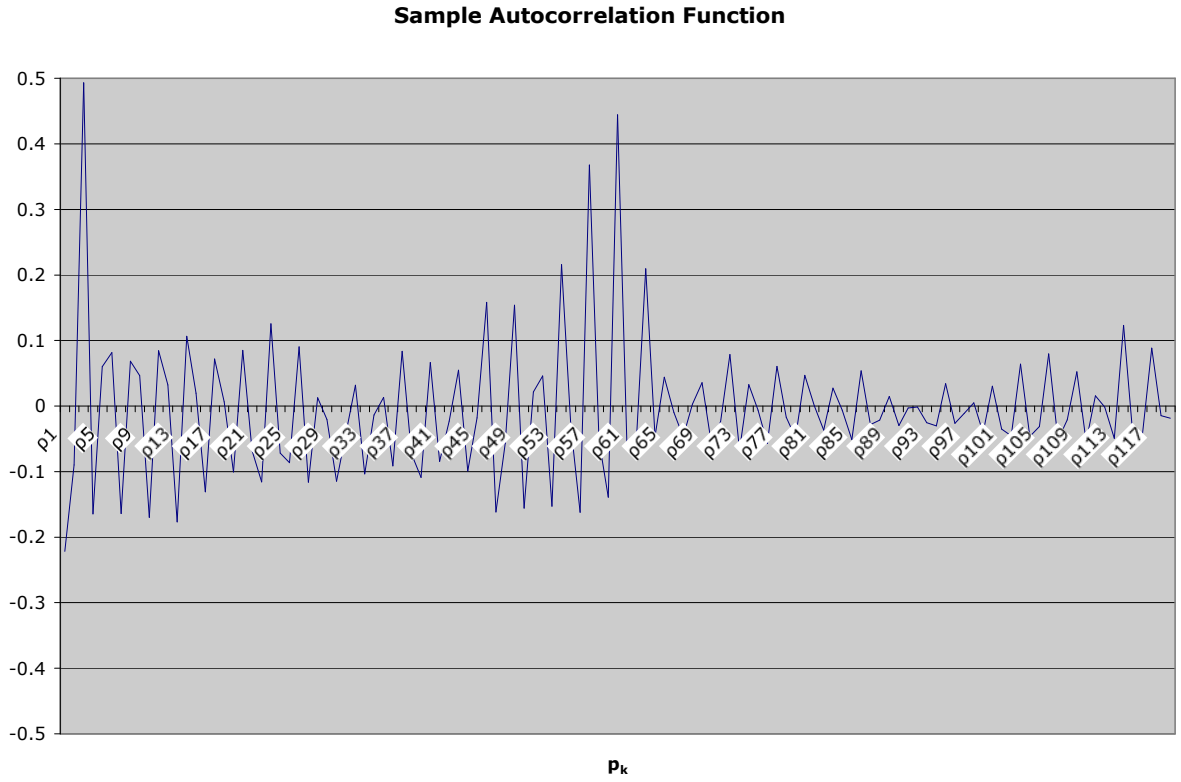
Although the sample autocorrelation function is not as I expected, it is obvious from the graph of the data that some type of seasonality is occurring. In an attempt to remove this seasonality, I decided to use method described in the text as an "ad hoc" method for removing seasonality. Using this method, the time series is thought of as consisting for four components: the long-term trend, a seasonal component, a long term cyclical component, and an irregular component. The goal is to create a new series from the original series that is free from the seasonal component. This method uses averages to create this new series: first to isolate the seasonal and irregular components and then further averaging to remove the irregular component. After the seasonal component is isolated and adjusted, the original series is divided by this seasonal index in an attempt to create a series free of the seasonal component. My intermediate calculations for this method can be found on the tab: "Frmted Data-10 yrs-Seasonal Adj". The graphs of the seasonal indices, my seasonally adjusted series (with the original data), and its autocorrelation function are shown below:



Graph 6


Number of Weddings by Month Seasonally Adjusted: Years 1991-2000
1,400,000.00
1,200,000.00
1,000,000.00
800,000.00
600,000.00
400,000.00
200,000.00
0.00
Weddings
Jan-91
May-91
Sep-91
Jan-92
May-92
Sep-92
Jan-93
May-93
Sep-93
Jan-94
May-94
Sep-94
Jan-95
May-95
Sep-95
Jan-96
May-96
Sep-96
Jan-97
May-97
Sep-97
Jan-98
May-98
Sep-98
Jan-99
May-99
Sep-99
Jan-00
May-00
Sep-00
Date
Seasonally Adjusted Data
Original Data

Graph 7



Graph 8

Although the seasonal indices provide information about the number of weddings in each month, the series appears to be seasonally adjusted, and the sample autocorrelation function is closer to zero for most values, I am not confident enough in my knowledge of this method or in my understanding of the nature of my data to use the new series for my model.

Another method mentioned it the text for removing annual seasonality is considering a new series of $y_t - y_{(t-12)}$. However, after examining the sample autocorrelation function, my data series does not exhibit a strictly annual seasonality. I don't think the 12 month lagged series would be the best choice for removing any seasonality that may be in the data. This method is shown in my spreadsheet for completeness.

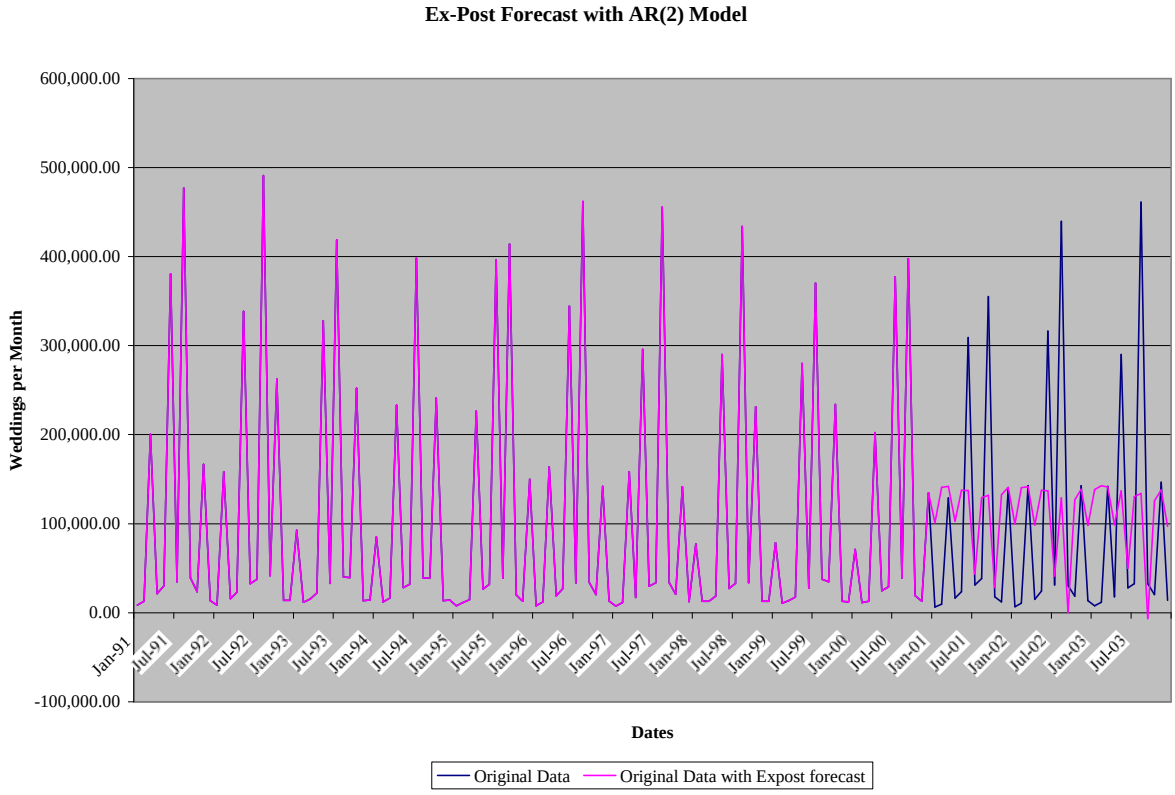
Based primarily on the sample autocorrelation which did not exhibit the annual seasonality I expected, I decided to construct a model without removing any seasonality.

Model Construction

After performing various statistical tests and examining different properties of my data set, it appears it may be more complex than a lower order process. However, I still attempted construction of a time series model for my data. Based on the properties I had already examined, I made the following conclusions:

- Because the series has an oscillating sample autocorrelation function that does not go to zero, this limits the possible models. Strictly moving average models are not a possibility because of their limited memory. If the series was from a moving average model of order q, ρ_k would be zero for all values greater than q. Because the sample autocorrelation function does not go to zero, I eliminated strictly moving average models as a possibility.
- Similarly, I eliminated an autoregressive (AR) model of order 1 as a possible model. If this were the correct model, I would expect the sample autocorrelation function to decline geometrically to zero. Because the sample autocorrelation function is oscillating, I eliminated this as a possibility.
- Although the model is likely more complex, for my project I decided to consider an AR(2) model. I chose the AR(2) model because the autocorrelation function for autoregressive processes of order greater than 1 are typically geometrically dampened, oscillating, sinusoidal functions similar to the sample autocorrelation function of my series.

For the AR(2) model, I first estimated the parameters of my model using the Yule Walker Equations and the conditions for stationarity. Because I assumed the process is autoregressive of order 2 and calculated the sample autocorrelation function, I was able to use the Yule Walker equations to estimate the autoregressive parameters ϕ_1 and ϕ_2. Solving the two algebraic equations gives two possibilities for (ϕ_1, ϕ_2): (0, 1) and (-0.32169231, -0.0171455). The conditions for stationarity require that $\phi_2 + \phi_1 < 1$, $\phi_2 - \phi_1 < 1$, and $|\phi_2| < 1$ which eliminates (0, 1) as a possibility for ϕ_1 and ϕ_2. However, (-0.32169231, -0.0171455) satisfies each of these conditions. Using these parameter estimates, the sample mean of the data, and the formula for the mean of an AR(2) process: $\mu = \delta/(1- \phi_1 - \phi_2)$, I calculated δ to be 142,673.93. The resulting model is given by: $y_t = -0.32169231\ y_{t-1} - 0.0171455 y_{t-2} + 142{,}673.93 + \varepsilon_t$. Although the Yule Walker equations provide reasonable estimates of the parameters, because the model is purely autoregressive, the parameters of the model can be estimated more accurately by ordinary least squares regression. Using ordinary least squares results in the following model: $y_t = -0.327723301\ y_{t-1} - 0.018809833\ y_{t-2} + 145{,}323.4066 + \varepsilon_t$ which is similar to the model produced by the Yule Walker estimates. Notice the parameter estimates satisfy the conditions for stationarity shown above. Also, using the formula for the mean of an AR(2) model, $\mu = \delta/(1- \phi_1 - \phi_2)$, gives $\mu = 107{,}924.1223$ which is close to the sample mean from our data of 106,565.51. A model that is stationary with a mean close the sample mean are indications that the model may be a good fit for our data. However, the R^2 test statistic for this model is only 0.10423 which indicates that the model may be a poor fit, which is more in line with my expectations of the model. Although there are additional diagnostic tests that can be performed, as a final review of my model, I used the data from 2001 to 2003 to conduct an ex-post forecast. The graph below shows both the original data from 1991-2003 and the ex-post forecast from 2001 to 2003.



Graph 9

As shown in the graph, although the model captures some of the oscillation in the model, overall, the model does not provide a good fit.

Interpretations and Conclusions

After working with real data: the number of wedding per month, I have learned that working with time series models and constructing a model that is a good fit is a complex process. The data I chose to work with did not exhibit the characteristics that I expected and the model I constructed was not a good fit for the data. However, I was able to utilize the techniques of the course to learn more about the nature of my data. Ultimately, a different model than the one I constructed is needed to capture all of the complexities of the data. One complexity of the model that I chose to not consider is that some months contain five Saturdays and others four. This skews the number of weddings in those months and is likely a contributing factor as to why my model was not a good fit. Although at the end of this project my knowledge of wedding frequency by month has not increased, my times series knowledge has increased. In order to construct a time series model appropriate for this data, techniques and knowledge beyond what was presented in this introductory course would be needed.