**Predicting cancer counts with state space method the**

**Pros and Cons**

By Kathleen Bonfe

A Masters Project submitted to the Graduate Faculty of Auburn University

in partial fulfillment of the requirements for the Degree of

Master of Science

Auburn, Alabama July 25, 2014

Keywords: State Space Model, Kalman Filter, Incidence, Autoregressive Quadratic trend model

Copyright 2014 by Kathleen Lynette Bonfe

Approved by

Dr. Jerzy Szulga, Professor of Mathematics

Dr. Tin-Yau Tam, Department head, Professor of Mathematics

Dr. Ulrich Albrecht, Professor of Mathematics

Abstract

Cancer is a disease that is caused by abnormal cell growth in which those cells can invade other tissues and spread the disease. It is caused by many different factors and can originate in many different parts of the body. It is the second leading cause of death in the US falling shortly behind heart disease. The America Cancer Society publishes their findings on cancer incidence and mortality predictions yearly for the US. They used the Autoregression quadratic trend model up until 2004 to make these predictions after which they began using the State Space Model. In this paper I will discuss the use of the State Space Model in making incidence predictions.

Table of Contents

# Introduction…………………………………………………..1

# Objectives…………………………………………………….3

# Materials and Methods……………………………………….3

# Results……………………………………………………….11

# Discussion of results………………………………..………..13

# Appendix…………………………………………………….14

Introduction

Time series is sequential data that is collected over a uniform time interval. This is a very common type of data in the real world. Some examples of time series data is the quantity of goods that are shipped monthly, hourly observations made on the yield of a chemical process, or the number of people diagnosed with cancer each year. Time series data is found in many fields such as economics, engineering, biology and the social sciences (2,3,4).

Time Series Analysis are the methods used to analyze the time series data and get statistics and other information from the data. The purpose of Time Series Analysis is to understand the past and predict the future. Time Series Analysis requires the development of stochastic and dynamic models. In a stochastic model there is randomness in the model (a variable that is unseen) and they can only be written as a probability distribution. Dynamic models have time dependence(3,4).

Another aspect of dealing with time series analysis is forecasting. Time Series Forecasting is the use of a model to predict future observations based on the previous observations and in certain cases the future observations as well like. Due to the nature of time series data, they tend to be serially dependent meaning observations close together in time tend to be correlated. Time series can be stochastic or deterministic. If future values can be determined by a mathematical function then the time series is considered to be deterministic (and can be expressed with a closed equation), but if future values of a time series can only be described in terms of a probability distribution then it is called stochastic. An example of deterministic and stochastic models are, Time Trend models, and Autoregressive Moving Average Models, respectively. A stationary process has the property that the mean, variance and autocorrelation structure do not change over time (2,3,4).

State Space Model

The State Space Model (SSM) is a means of modeling a stochastic time series and is a stationary process. It can be used when you have a time-invariant linear time series. The SSM uses the previous states information from the data to model the current state (1,2,4). Because of this and its stochastic attributes, it is sometimes called the Hidden Markov Model (1). In order to predict future observations using the SSM; there are a few options such as smoothing, the Recursive Least Squared Adaptive Filter, the Kalman filter. State Space Method is used for many applications such as target tracking, forecasting income, predicting the number of people with a disease in a given time period (6). A good use of the State Space method and the Kalman is to determine the number of incidences (occurrence of the population being diagnosed) of cancer (7,9,10). I will discuss this model in more detail later on in the paper.

According to the CDC website Cancer is second most common cause of death in the United States falling shortly behind heart disease. Being able to predict cancer incidence is a valuable tool to have to combat this. There are several methods that you could use to predict cancer incidence for upcoming years. The American Cancer Society (ACS) uses the state-space model (SSM) as of 2004 and is continuing to pursue better methods. (7, 9, 10) SSM is a more generalized formula of the previous model an autoregressive quadratic trend model and allows for changes in the short term trends and eliminates subjectivity. (10) This is because SSM is a stochastic process and the autoregressive quadratic trend model is a deterministic process (2,3,4).

Objectives

To use data given from government websites including the Center for Disease Control (CDC) and the National Cancer Institute (NCI) to predict the number of people that will have cancer for a specific year with both the SSM and the autoregressive quadratic trend model. I will discuss the pros and cons of both methods and why the ACS used the SSM up to 2012.

Materials and Methods

Data Gathering

The CDC’s WONDER website gives data that has a time lag of about 4 years. One is able to access the information from 1999 to 2010. The NCI’s Surveillance Epidemiology and End Results (SEER) website gives data for certain cancers from 1973 to 2011 which is a 3 year lag. Using this I have gathered data from both these sites will first see if I can obtain the same results as the ACS for 2004. Also I will be using this data to predict the cancer incidences for 2015.

State Space Method

The State Space Method is a time series model with two equations, a measurement equation and a transition equation, which are as follows:

The equations consist of which is the predicted number of cancer incidence at year t, and which are the state vectors at time t, which is the measurement error term and has a normal Gaussian distribution with mean 0 and variance , and which is the vector error term ~N(0,). The coefficient x and y are the state estimate parameter and measurement estimate parameter respectively (1,2,3,4,7,9,10).

These equations give many quadratic equations over short periods of time versus the autoregressive model which gives a quadratic equation over the entire time period. The measurement equation is a linear model with a state vector of parameters meaning that for this vector we assume it holds all the information needed for the current state. The transition state is a first order Markov process and allows us to model the current state given the immediate past state to provide the number of cancer incidences for the next year via the measurement equation (1,2,3,4).

Assumptions of the SSM include:

1. A linear relationship between the observed values and the vector states.
2. A Normal distribution for the error terms.

(1,2,3,4)

This model can be used to predict future incidences with the Kalman filter, smoothing (which is a backwards recursion algorithm that uses the entire sample), Recursive Least Square Adaptive Filter (recursively finds the filter coefficients that minimize a weighted [linear least squares](http://en.wikipedia.org/wiki/Weighted_least_squares) [cost function](http://en.wikipedia.org/wiki/Loss_function) relating to the input signals used in signal processing), and other nonlinear recursive procedures including the Extended Kalman filter and the Decentralized Kalman filter. The Kalman filter is a linear recursive procedure for calculating the optimal estimator for the state vector given all the information that is available. The Kalman filter uses a recursive algorithm with a set of prediction equations and update equations(6).

Update equations:

You can compute the estimate using the update step where . To predict more than one year ahead just bypass the update equation and change t to the year you would like to be predicted. You can also calculate the mean square estimate and find a confidence interval. For more in depth information on the Kalman filter please refer to the appendix(6).

Modeling

I used SAS PROC STATESPACE to model the number of cancer incidences given by the CDC website and to predict the cancer incidence for 2015. SAS PROC STATESPACE uses the Kalman filter to predict future outcomes. The data provided only consists of the years 1999 to 2010 therefore there is only 11 years to model the data as well as predicting 5 years in the future. Also because I only had 11 years in the CDC data I could not use SAS PROC FORECAST because you must have at least 15 observations to run this procedure.

Because of the lack of years available on the CDC WONDER website I gained permission to use the Surveillance Epidemiology and End Results (SEER) website and gather information going as far back as 1973 and get access to the data from 2011. Using this data I gathered better results for the SSM because I had more observations and the SSM is very sensitive to change. Since there are 39 observations this data is able to be used to run SAS PROC FORECAST.

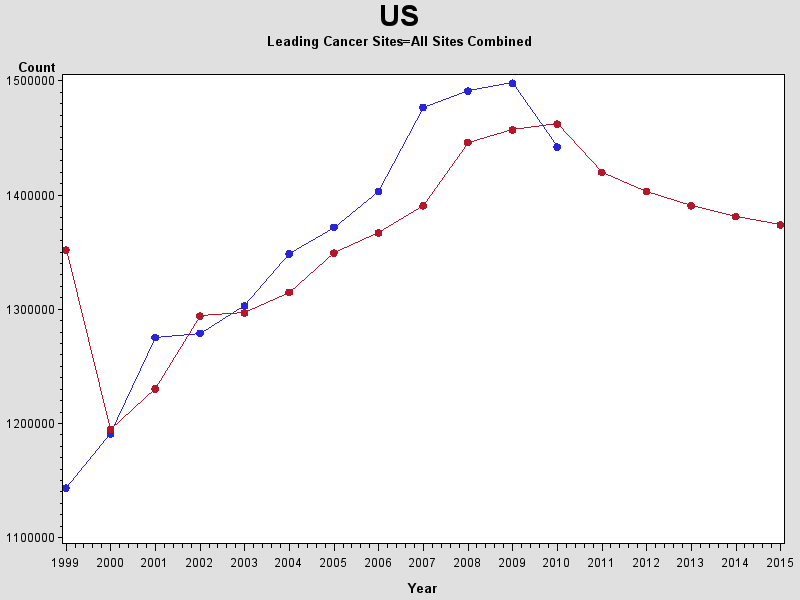
Results

The CDC WONDER site allows a user to obtain data from 1999 to 2010. I used this data set to obtain the SSM results using SAS PROC STATESPACE. The following variables were gathered from the CDC WONDER website: Leading Cancer Sites, State, Year, and Count. The cancer sites I obtained were of leading cancer sites in the US, including data for all the cancer sites combined, colon and rectum, ovary, and breast as well as many others. The variable defined as state is the US State from where the information was drawn. The variable Year is the year in which the information was drawn from. The Count is the number of incidences or mortality (depending on the data set) for those variables.

|  |
| --- |
| Incidence Counts for All Cancer Sites in the United States |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Obs** | **Count** | **FOR1** | **RES1** | **STD1** | **Year** |
| **1** | 1143590 | 1352031.42 | -208441.42 | 76788.54 | 1999 |
| **2** | 1190819 | 1194779.33 | -3960.33 | 76788.54 | 2000 |
| **3** | 1275368 | 1230409.77 | 44958.23 | 76788.54 | 2001 |
| **4** | 1279159 | 1294195.11 | -15036.11 | 76788.54 | 2002 |
| **5** | 1303012 | 1297055.11 | 5956.89 | 76788.54 | 2003 |
| **6** | 1348553 | 1315050.26 | 33502.74 | 76788.54 | 2004 |
| **7** | 1371923 | 1349407.23 | 22515.77 | 76788.54 | 2005 |
| **8** | 1403404 | 1367038.00 | 36366.00 | 76788.54 | 2006 |
| **9** | 1476897 | 1390787.85 | 86109.15 | 76788.54 | 2007 |
| **10** | 1491268 | 1446232.34 | 45035.66 | 76788.54 | 2008 |
| **11** | 1498164 | 1457074.09 | 41089.91 | 76788.54 | 2009 |
| **12** | 1442220 | 1462276.56 | -20056.56 | 76788.54 | 2010 |
| **13** | . | 1420071.36 | . | 76788.54 | 2011 |
| **14** | . | 1403362.02 | . | 96189.64 | 2012 |
| **15** | . | 1390756.18 | . | 105652.67 | 2013 |
| **16** | . | 1381246.10 | . | 110677.82 | 2014 |
| **17** | . | 1374071.52 | . | 113438.52 | 2015 |

This table gives us the observation of the number of cancer incidence for the year for all of the United States which is labeled count. It also gives the forecast for the number of cancer incidences for the years past the years that have data which is labeled for1. Also given in this table are the residuals and standard deviation for each observation and forecast.



The graph above shows the actual count of incidences of cancer in blue and the line given by the state space model in red over the period from 1999 to 2015.

As you can see from the graph the SSM is very perceptive to change through the years. In order to get a better model I used the Surveillance Epidemiology and End Results (SEER) data given from the SEER website which is part of the National Cancer Institute (NCI) because it allowed me to have more observatons. With this website I was able to model and predict Breast Cancer data. SAS PROC STATESPACE uses maximum likelihood to get the parameter estimates. In table below is the parameter estimate for the breast cancer data retrieved from the SEER website.

|  |  |  |  |
| --- | --- | --- | --- |
| **Parameter Estimates** | | | |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** |
| **F(1,1)** | 0.918695 | 0.063209 | 14.53 |

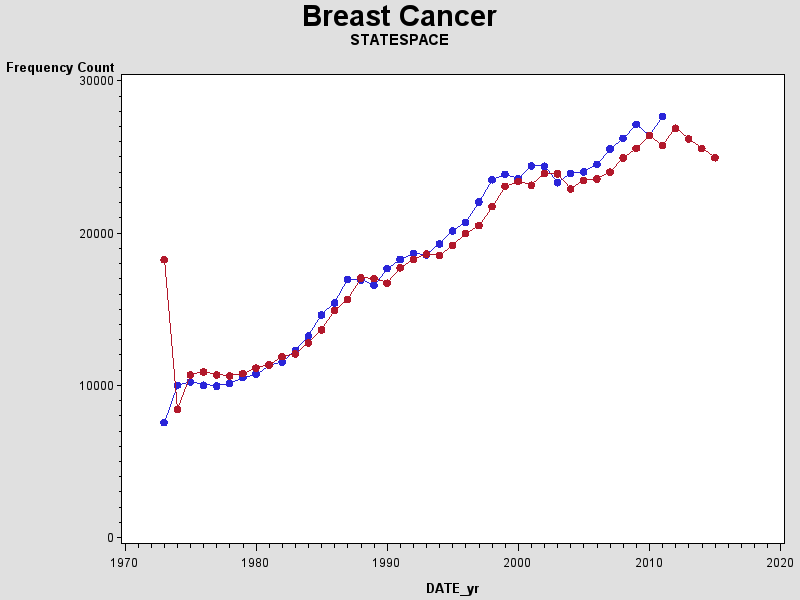
PROC STATESPACE came up with this fitted State Space Model:

|  |
| --- |
| Breast Cancer |
| STATESPACE |

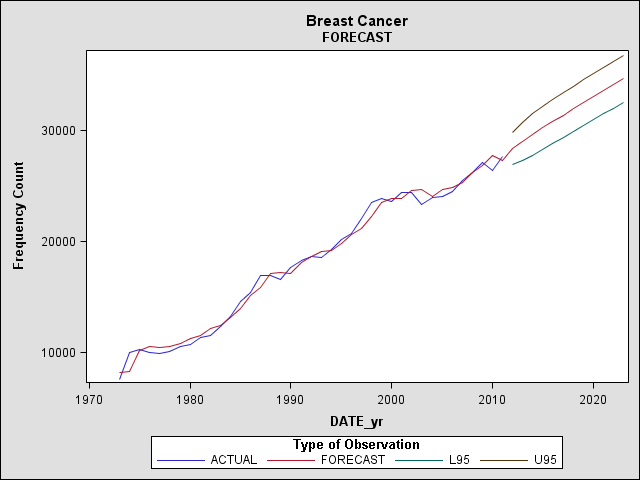
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Obs** | **Count** | **Forecast** | **Residuals** | **Std** | **Year** |
| **22** | 19303 | 18555.27 | 747.73 | 2412.95 | 1994 |
| **23** | 20160 | 19218.57 | 941.43 | 2412.95 | 1995 |
| **24** | 20710 | 20005.89 | 704.11 | 2412.95 | 1996 |
| **25** | 22064 | 20511.18 | 1552.82 | 2412.95 | 1997 |
| **26** | 23519 | 21755.09 | 1763.91 | 2412.95 | 1998 |
| **27** | 23872 | 23091.79 | 780.21 | 2412.95 | 1999 |
| **28** | 23596 | 23416.09 | 179.91 | 2412.95 | 2000 |
| **29** | 24425 | 23162.53 | 1262.47 | 2412.95 | 2001 |
| **30** | 24410 | 23924.13 | 485.87 | 2412.95 | 2002 |
| **31** | 23324 | 23910.35 | -586.35 | 2412.95 | 2003 |
| **32** | 23946 | 22912.65 | 1033.35 | 2412.95 | 2004 |
| **33** | 24036 | 23484.08 | 551.92 | 2412.95 | 2005 |
| **34** | 24529 | 23566.76 | 962.24 | 2412.95 | 2006 |
| **35** | 25540 | 24019.67 | 1520.33 | 2412.95 | 2007 |
| **36** | 26237 | 24948.48 | 1288.52 | 2412.95 | 2008 |
| **37** | 27162 | 25588.81 | 1573.19 | 2412.95 | 2009 |
| **38** | 26441 | 26438.60 | 2.40 | 2412.95 | 2010 |
| **39** | 27672 | 25776.22 | 1895.78 | 2412.95 | 2011 |
| **40** | *.* | 26907.13 | . | 2412.95 | 2012 |
| **41** | . | 26204.46 | . | 3276.65 | 2013 |
| **42** | . | 25558.91 | . | 3857.96 | 2014 |
| **43** | . | 24965.85 | . | 4287.70 | 2015 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Obs** | **Count** | **Forecast** | **Residuals** | **Std** | **Year** |
| **1** | 7562 | 18264.59 | -10702.59 | 2412.95 | 1973 |
| **2** | 10016 | 8432.17 | 1583.83 | 2412.95 | 1974 |
| **3** | 10237 | 10686.65 | -449.65 | 2412.95 | 1975 |
| **4** | 10024 | 10889.68 | -865.68 | 2412.95 | 1976 |
| **5** | 9963 | 10694.00 | -731.00 | 2412.95 | 1977 |
| **6** | 10128 | 10637.96 | -509.96 | 2412.95 | 1978 |
| **7** | 10520 | 10789.54 | -269.54 | 2412.95 | 1979 |
| **8** | 10745 | 11149.67 | -404.67 | 2412.95 | 1980 |
| **9** | 11333 | 11356.38 | -23.38 | 2412.95 | 1981 |
| **10** | 11537 | 11896.57 | -359.57 | 2412.95 | 1982 |
| **11** | 12320 | 12083.98 | 236.02 | 2412.95 | 1983 |
| **12** | 13246 | 12803.32 | 442.68 | 2412.95 | 1984 |
| **13** | 14639 | 13654.03 | 984.97 | 2412.95 | 1985 |
| **14** | 15417 | 14933.78 | 483.22 | 2412.95 | 1986 |
| **15** | 16964 | 15648.52 | 1315.48 | 2412.95 | 1987 |
| **16** | 16922 | 17069.74 | -147.74 | 2412.95 | 1988 |
| **17** | 16590 | 17031.16 | -441.16 | 2412.95 | 1989 |
| **18** | 17683 | 16726.15 | 956.85 | 2412.95 | 1990 |
| **19** | 18282 | 17730.29 | 551.71 | 2412.95 | 1991 |
| **20** | 18664 | 18280.58 | 383.42 | 2412.95 | 1992 |
| **21** | 18581 | 18631.53 | -50.53 | 2412.95 | 1993 |

This is the table for the count, forecast, and residuals for the breast cancer data given by the State Space Model.



This is the graph given from the plot of count vs time for both the observed count and the forecasted count by the SAS PROC STATESPACE. I also ran SAS PROC FORECAST on the SEER breast cancer data and came up with these results. Remember PROC FORECAST uses the autoregressive quadratic time trend to analyze the data.



This is a graph of the observed data vs the forecasted data for SAS PROC FORECAST.

Discussion of Results

The use of the CDC WONDER website data to predict future incidence was convenient because there was more information on all cancer sites put together as well as the individual sites. The data however could only went back to 1999 which gave us only 11 observations. After modeling the data and testing for normality and linearity I found that it did not meet either assumption. Also I could not rely on the model of the data because this was not a large enough sample to get accurate results.

The SEER website gave data for only a few cancer sites of which I choose breast cancer and ran my models and forecasts on it. When looking at the assumptions this data came much closer to the normality and linearity being valid. With 39 observations instead of 11 the model can be considered more accurate than that of the other data. Using PROC STATE SPACE I modeled the SEER data. I was able to plot the modeled data and the observed data which showed how close the forecasted data is to the real observations for any given year. I also modeled the data using PROC FORECAST and plotted those data points to compare the two models. In comparing the two models I found that the SSM changed closely with the data even when the data changed quickly. The Autoregressive quadratic trend model did not change as quickly and when there was a quick change the model was off by a large amount. The forecasted data for the predicted years are going in opposite directions for the two different models. PROC STATE SPACE predicts that breast cancer incidence will continue to decline over the next few years while PROC FORECAST predicts that breast cancer incidence will rise over time.

I mentioned earlier on that ACS provides cancer statistics for the current year. In the CA and in Cancer Facts & Figures, they estimated 1,665,540 new cancer incidences for all cancer in 2014. When I predicted the incidence for all cancer with the data from the CDC WONDER website I got 1,381,246. Again I will point out that this was not a very good estimate due to the lack of information that was available and the invalidity of the assumptions. The data given by the SEER website covers approximately 9.4% of the US Population. It would be difficult to compare the results I gathered to those estimated by the ACS.

References

1. Akaike, Hirotugu. "Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes." Annals of the Institute of Statistical Mathematics 26.1 (1974): 363-387.
2. Anderson, Theodore Wilbur. The statistical analysis of time series. Vol. 19. John Wiley & Sons, 2011.
3. Box, George EP, Gwilym M. Jenkins, and Gregory C. Reinsel. Time series analysis: forecasting and control. John Wiley & Sons, 2013.
4. Harvey, A. C. Time Series Models. Cambridge, Mass: MIT Press, 1993. eBook Collection (EBSCOhost). Web. 1 June 2014.
5. Kalman, Rudolph Emil. "A new approach to linear filtering and prediction problems." Journal of Fluids Engineering 82.1 (1960): 35-45.
6. Kovvali, Narayan; Banavar, Mahesh; Spanias, Andreas. An Introduction to Kalman Filtering with MATLAB Examples. San Rafael: Morgan & Claypool Publishers, 2013. Ebook Library. Web. 30 Jun. 2014.
7. Pickle, Linda W., et al. "A New Method of Estimating United States and State level Cancer Incidence Counts for the Current Calendar Year." CA: a cancer journal for clinicians 57.1 (2007): 30-42.
8. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER\*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2013 Sub (1973-2011) <Katrina/Rita Population Adjustment> - Linked To County Attributes - Total U.S., 1969-2012 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission.
9. Tiwari, Ram C., et al. "A New Method of Predicting US and State Level Cancer Mortality Counts for the Current Calendar Year." CA: a cancer journal for clinicians 54.1 (2004): 30-40.
10. Zhu, Li, et al. "Predicting US and state level cancer counts for the current calendar year." Cancer 118.4 (2012): 1100-1109.

Appendix

Kalman filter

The Kalman filter is a solution to the problem of Sequential Bayesian Estimation of a state in a dynamic system such as in the State Space Model. To use the Kalman filter you must have a linear normal (Gaussian) distribution for both the measurement and the transition equations in SSM. Refer to the equations of the State Space Method.

What we need to accomplish is sequentially estimating the state using a set of measurements . The Sequential Bayesian Estimation procedure states that the posterior pdf of a distribution is the probability of the parameter given the data instead of non-Bayesian methods where the pdf is the probability of the data given the parameters. The Kalman filter utilizes this procedure such that the pdf is and if we look at time step t-1 we get:

Where t|t notates the state at time t computed given the measurements up to time t.

The conditional pdfs for the transition equation and measurement equation from the SSM are and respectively.

Knowing this information we can apply this to the Kalman filter. Now the prediction equations come from the posterior pdf:

The update equations are from the posterior pdf:

Looking back at the final equations we can see the Kalman gain= this is where new information enters the system.

SAS

Data

I obtained the data from the SEER website was large and complicated. The website provided SAS code to load the data into SAS. The WONDER website was easier to gather the data. I downloaded it into Excel and then uploaded that into SAS.

PROC STATESPACE and PROC FORECAST

PROC STATESPACE

Models the data using the State Space Model.

Uses maximum likelihood to calculate the parameter estimates.

Uses Kalman filter to estimate the forecasted data.

PROC FORECAST

Uses Autoregressive Quadratic trend model