A Comparison of Different Methods for Predicting Cancer Mortality Counts at the State Level  
Corinne Wilson  
Department of Mathematics and Statistics  
Old Dominion University  
Norfolk, VA 23529  

ABSTRACT  
Cancer is a major health issue in the United States. Reliable estimates of yearly cancer mortality counts are essential for resourcing and planning. The American Cancer Society has used several methods of forecasting to estimate the future cancer burden and researchers are continually working to develop new methods with improved performance. There have been studies comparing different models for predicting the US cancer mortality counts. This study explores and compares several different models for cancer mortality count predictions at the state level, principally for the state of Virginia. Results of the comparisons appear to show the final improved model to perform better than the others; however, at the state level even the improved model can still produce undesirable results.  

INTRODUCTION  
Cancer is the main cause in one out of every four deaths in the United States; only heart disease causes more deaths each year (ACS 2008). In 2008 the American Cancer Society (ACS) estimates that 565,650 Americans will die from cancer; 13,990 are expected to be Virginians (ACS 2008). In 2007 the National Institutes of Health (NIH) estimates that the total costs associated with cancer reached $219.2 billion: $89 billion for direct medical costs, $18.2 billion for lost productivity due to illness and $112 billion for lost productivity due to premature death (ACS 2008). As a result of these costs it is vital for many agencies to have precise estimates of cancer incidence and mortality counts for resourcing and planning. Agencies need to have reliable predictions in order to budget annually for cancer research, treatment, prevention, and other related expenditures.  
The National Center for Health Statistics (NCHS) of the Centers for Disease Control (CDC) publicly releases the observed mortality data compiled from death certificates certified by attending physicians, funeral directors, medical examiners, and coroners. The latest data available are 3 years old due to the large number of records involved and the complex process of data collection, tabulation, and publication. For instance, in 2007 the NCHS released the actual mortality data for 2004. As a result of this procedural delay it is necessary to predict three years ahead to obtain the current year's numbers to budget and plan accordingly.  
Each year the ACS releases these predicted figures in two publications, Cancer Facts & Figures (CFF) and CA- A Cancer Journal for Clinicians. Included in these publications are the projected number of deaths from site and gender specific cancers and all cancers combined at the national and state level. The ACS has used several methods of forecasting to estimate the future cancer burden and researchers are
continually working to develop new methods with improved performance. Prior to 1995, a model based on linear predictions was used by the ACS to estimate the yearly number of cancer deaths. A quadratic time series model with autoregressive errors called the PF model was used from 1995 to 2003. During this time the ACS would make subjective modifications to the forecasts by choosing from five different forecasts in order to account for recent trend changes in the data that the model was not able to capture. The five possibilities for the published forecasts were the three-year-ahead point predictions, the upper and lower 95% prediction limits, and the midpoints between the prediction limits and the point estimate.

In order to improve forecasts, Tiwari et al. (2004) developed a state space model (SSM) based method and its tuned version (tuned SSM). This method was used to obtain cancer predictions published in Cancer Statistics, 2004 (Jemal et al. 2004). The ACS did extensive research at both the national and state levels, and found the tuned SSM to perform better on average than other methods when comparing mean squared deviations, but at the state level the ACS found the PF model and the tuned SSM to be comparable with a slight advantage for the PF model over the SSM. In part because of its ability to adjust well to rapidly changing trends at the national level, the ACS adopted the tuned SSM for cancer forecast in 2004. Since 2004 the ACS has been using the tuned SSM to predict the yearly cancer mortality counts using the method of moments (MOM) to estimate the error covariance matrices (Tiwari et al. 2004).

In a recent paper, Ghosh et al. (2008) studied the predictions of the 3 methods at the national level and found the tuned SSM to perform better on average, but not uniformly. Apparently, they also studied the models at the state level, and found the results were not as favorable to SSM and tuned SSM as at the national level. However, no specific state level results are reported and data used were only up to 2001. In this article, the interest is to compare the three methods specifically for the state of Virginia’s cancer mortality data. For this, three more years of data are used than Ghosh et al. (2008), that is years 1969 through 2004 are used to compare cancer mortality predictions through 2007 using these methods.

**DATA**

Analysis in this article uses Virginia mortality data from years 1969 through 2004, the latest year available at the time of analysis (SEER 2007). The data is broken down by gender and cancer site where specified. The SEER*Stat Software was used to obtain all Virginia mortality data (NCI 2008). The data is in the form of \( d_t \) where \( t = 1 \) corresponds to the number of cancer deaths in 1969 and \( t = 36 \) corresponds to 2004.

**PF MODEL**

From 1995 until 2003, ACS predictions were based on the PF model using a quadratic time trend with autoregressive errors. This model can be written in the form

\[
d_t = b_0 + b_1 t + b_2 t^2 + u_t
\]

where the \( u_t \)'s are independently distributed with mean zero and constant variance \( \sigma_u^2 \). For all \( t \).

The first step in implementing the PF model is to fit a quadratic time trend model to the series using ordinary least squares. Then the residuals

\[
u_t = d_t - (b_0 + b_1 t + b_2 t^2)
\]

are calculated and an autoregressive process is fit to \( \{u_t\} \) in order to capture the short-term fluctuations of the series. In this autoregressive process the residual at a current time point depends on the residuals at previous time points and a random error term (Harvey 1989, 1993). The combined forecasting model is then used to make future mortality predictions.

The PF model needs at least seven observations consisting of \( d_t \) and \( t \) in order to fit the forecasting model. SAS procedure PROC FORECAST (PF) is used to obtain the three-year-ahead predictions and 95% prediction intervals for each year (SAS 2004). Each year the PF model was applied to gender and site specific groupings (for example male digestive system) and then the overall national-level prediction was a sum of the predictions from all the individual sites. The PF model was also applied at the state level, but to insure that the sum of the state level predictions equaled the national level predictions, the state forecasts were adjusted proportionally when needed.

**STATE SPACE MODEL**

A state space model (SSM) for representing the yearly number of cancer deaths \( d_t \) is

\[
d_t = c_t + 1_t, \quad t = 1, 2, \ldots
\]

where \( c_t \) is the unobserved trend at time \( t \) and \( 1_t \) is the error at time \( t \). Here \( c_t \)'s are assumed to be serially uncorrelated and normally distributed with zero-mean and constant variance \( \sigma_c^2 = \sigma^2 \).

The PF model was slow in capturing sudden year-to-year variations in the series; to improve on this a trend that changes with time can be implemented. There are several time-varying trends available; a local quadratic trend is selected because of its similarity to the quadratic time series model. The local quadratic trend model is

\[
\begin{align*}
c_t &= c_{t-1} + \beta_{c1} 1_{t-1} + \gamma_{c1} 1_{t-2} + \eta_{ct} \\
\beta_t &= \beta_{t-1} + 2\gamma_{c1} 1_{t-1} + \eta_{bct} \\
\gamma_t &= \gamma_{c1} 1_{t-1} + \eta_{ct}
\end{align*}
\]

The errors \( \eta_t \) are assumed to be serially uncorrelated with mean 0 and variance \( \sigma^2 \). They are also assumed to be uncorrelated with each other and with the \( c_t \)'s. Further \( c_t \) is called the measurement error and \( c_t = [1_{ct} \quad 1_{ct} \quad 1_{ct}] \) is the transition error with variance \( W \). The measurement and transition errors are also assumed to be normally distributed so \( V \) and \( W \) can be estimated using maximum likelihood (ML) estimation. If \( \sigma^2 = 0 \) (i = 1, 2, 3) then the local quadratic model reduces to

\[
d_t = \alpha_0 + \beta_0 1_t + \gamma_0 1^2_t + 1_t.
\]

Hence a state space model with a local quadratic trend mimics the PF model.

Following the methods of Ghosh et al. (2008) one can obtain the predicted series...
using this SSM. This type of model prediction can be implemented by various packages including SsfPack2.2. "SsfPack is a suite of C routines for carrying out computations involving the statistical analysis of univariate and multivariate models in state space form" (Koopman et al. 1999).

Figure 1 shows Virginia mortality predictions for female breast cancer using data from years 1969 through 2004. The SSM predictions, PF predictions, and corresponding observed values are all shown. Also shown are the tuned SSM predictions, which will be discussed in the next section. Notice how the SSM adapts faster to the leveling off of the observed series than the PF model which continues to increase for a period of time before it adapts to the new trend. For Virginia female breast cancer the root mean square predicted error (RMSPE) for the SSM is smaller than the RMSPE for the PF model. The SSM is able to adapt faster to trend changes than the PF model. However, small random variations in the observed series are magnified and show up as zigzags in the SSM predictions. This jaggedness is especially noticeable at the state level or in rare cancers. Figure 2 shows female breast cancer mortality predictions for the entire U.S. Notice that even though both SSM predicted series in Figures 1 & 2 are jagged, the predicted series for Virginia's female breast cancer deaths has more severe year-to-year fluctuations than the predicted series for entire U.S.'s female breast cancer deaths.

These exaggerated fluctuations are a weakness of the model, because it creates an uncertainty that can make the predictions useless. Figure 3 shows Virginia testis observed cancer counts and corresponding predictions. Testis cancer has a variable observed series yielding to very erratic predictions from the SSM. For the predicted series shown in the figures, testis cancer has the worst predictions with regards to
RMSPE for both the SSM and the PF model. The SSM model predicts negative mortality counts for 4 different years while the PF model predicts negative counts for 9 years. On the other hand, the SSM predicts -66 people to die from testis cancer in 1982 while the lowest prediction made by the PF method is -5 in 1988. Both of the predicted series for testis cancer are unreasonable.

**TUNED SSM**

To help control the variability of the SSM, tuning parameters can be introduced into the model. The time-invariant error variances V and W are rescaled by the tuning parameters such that the sum of squares of the differences between the predicted mortality counts and the observed mortality counts is minimized.

Let $\tilde{d}$ be the predicted number of deaths at time t. Let $V_t^*$ be the variance and $W_t^*$ be the covariance matrix from the SSM, estimated from observed values $d_1 \cdots d_t$ in order to predict $\tilde{d}_{t+3}$. The added suffix t refers to the portion of the time series that V and W are estimated from, so V and W are still time-invariant. To illustrate, $d_1 \cdots d_7$ are used to estimate $V_7^*$, and $W_7^*$ then to obtain $\tilde{d}_{10}$. Similarly, $d_1 \cdots d_t$ are used to obtain $\tilde{d}_t$ and $V_t^*$ and $W_t^*$ are the corresponding covariance matrices used in the prediction. Likewise, computation of $V_t^*$ and $W_t^*$ continues until the most recent year available that has a corresponding observed value. For example, if 2004 is the latest year for which the observed number of cancer deaths is known then stop with $V_{33}^*$ and $W_{33}^*$. If $V_t^*$ and $W_t^*$ which are used to obtain $\tilde{d}_t$, the estimated number of deaths for 2004. Once $V_7^*$, ..., $V_t^*$, $W_t^*$ ..., $W_{36}^*$ have been estimated, replace each $V_t^*$ and $W_t^*$ with $\kappa V_t^*$ and $\kappa W_t^*$ where $\kappa$ and $\kappa_w$ are unknown constants in the interval (0,1) called tuning parameters. Note that if $\kappa$ and $\kappa_w$ were known, $d_1 \cdots d_t$, variance $\kappa V_t^*$, and covariance matrix $\kappa W_t^*$ could be used to fit a SSM to obtain $\tilde{d}_t$. Let SSPE be the sum of the squares of the prediction errors. Then SSPE is a function of $\kappa$ and $\kappa_w$. These are estimated by minimizing

$$SSPE = \sum_{t=7}^{33} \left( \tilde{d}_{t+3} - d_{t+3} \right)^2$$

Once $\hat{\kappa}$ and $\hat{\kappa}_w$ have been obtained recalculate $\tilde{d}_1 \cdots \tilde{d}_7$ using the tuned variance $\kappa V_t^*$ and tuned covariance matrix $\kappa W_t^*$. Variance $V_t^*$, ..., $V_{36}^*$, and covariance matrices $W_t^*$, ..., $W_{36}^*$ were estimated first using SsfPack2.2 as done in the SSM, then the tuning parameters $\hat{\kappa}$ and $\hat{\kappa}_w$ were estimated using the routine optim in "R" (Ihaka and Gentleman 1996).

Figures 1 & 3 show the tuned SSM, SSM, and PF model predictions for the number of cancer deaths in Virginia for female breast cancer and testis cancer years 1978 through 2004. The tuned SSM has corrected some of the pronounced variations of the SSM. For testis cancer, the prediction for 1982 using the tuned SSM is -3, an improvement over the predicted -66 deaths of the SSM. However, the tuned SSM now has 7 negative predictions which is still better than the PF model which has 9 negative predictions. The worst prediction for the tuned SSM is in year 1997 when the model predicts -6 testis cancer deaths.

**DISCUSSION**

Both the SMM and the tuned SSM are able to respond faster to local changes in the series of cancer deaths compared to the PF model as can be seen in predictions for Virginia female breast cancer deaths (Figure 1). But, both the predicted series from the SSM and the tuned SSM are more jagged than the PF model sometimes resulting in more unreasonable results. The tuned SSM is able to smooth some of the SSM's jaggedness, but still produce oscillating predicted series. In some cases, the tuned SSM is able to bring the predictions closer to the observed values.

Table 1 contains the RMSPEs using all 3 models for predictions from several Virginia cancer groups, years 1978 to 2004. The RMSPE is consistently smaller for
the tuned SSM compared to the untuned SSM. The SSM RMSPE is smaller than the PF model RMSPE for only female breast cancer and leukemia. These two cancers have smaller fluctuations in the observed series than the other cancer sites, allowing the SSM to perform better than the PF model. The more oscillatory series of the other cancer sites produce extreme fluctuations in the untuned SSM. The tuned SSM is able to smooth these fluctuations and perform better than the PF model for all but three of the cancer sites. For female breast cancer tuned SSM reduces the RMSPE by 18%. For leukemia tuned SSM reduces the RMSPE by 17%. However, for the brain and other nervous system cancers the PF model RMSPE is 16% smaller than the tuned SSM RMSPE.

Table 2 shows the observed and predicted values for several Virginia cancer sites for the year 2004. Notice for the cancers included in table 2, the predictions for cancers with smaller mortality counts are close if not identical for the 3 methods.

For Virginia's cancer mortality predictions the tuned SSM appears to perform better than the PF method when looking at the predicted series as a whole. This is because the tuned SSM is able to adapt quicker to changes in mortality trends; however this added sensitivity can sometimes cause unwanted results.

There is definite room for improvement in cancer mortality predictions. Both the SSM and tuned SSM assume the errors to be normally distributed. While this may not be a problem at the national level, small mortality counts at the state level and with some rarer cancers might cause this to be a problem. This is especially apparent with Virginia's testis cancer predictions. One could improve on this by assuming a different distribution on the errors, such as a Poisson distribution, and then using Dynamic Generalized Linear Models. Another suggested improvement would be to use different time-varying trend models for different cancers. But, this would require the researcher to choose the best model for each type of cancer. Yet another suggestion is to use a joinpoint model (Tiwari et al. 2004). Finally, Tiwari also suggested the use of preliminary mortality estimates in predictions. Research is ongoing to find the best method of cancer mortality prediction.

LITERATURE CITED