

Surveillance Research Program, NCI, Technical Report #2012-02

## Technical Details of Predicting US and State-Level Cancer Counts for the Current Calendar Year<sup>1</sup>

National Cancer Institute Cancer Counts Prediction Workgroup

### A.1. Joinpoint regression

Suppose that we observe  $(x_i, y_i)$  for  $i=1, 2, \dots, n$ , and consider a piece-wise linear regression model,  $y_i = \beta_0 + \beta_1 x_i + \delta_1(x_i - \tau_1)^+ + \dots + \delta_k(x_i - \tau_k)^+ + \varepsilon_i$ , where  $a^+ = \max(a, 0)$ .

There are unknown joinpoints,  $\tau_1, \dots, \tau_k$ , where two consecutive linear segments are connected and the number of joinpoints,  $\kappa$ , is also assumed to be unknown. Kim et al. (2000) used the least squares method to estimate the  $\beta$ 's and  $\tau$ 's at each given value of  $k$ , and proposed to use the permutation test to estimate the number of joinpoints,  $\kappa$ . We start with testing the null hypothesis of  $k_0$  joinpoints versus the alternative hypothesis of  $k_1$  joinpoints, and then increase  $k_0$  to  $k_0 + 1$  if the null hypothesis is rejected and decrease  $k_1$  to  $k_1 - 1$  otherwise. We conduct such sequential testing until we test the null hypothesis of  $k$  joinpoints versus the alternative hypothesis of  $k+1$  joinpoints for some  $k$  ( $k_0 \leq k \leq k_1$ ). In order to control the overall over-fitting probability under  $\alpha$ , the significance level of the test at each step is adjusted appropriately. A simple Bonferroni type adjustment was used in earlier versions of Joinpoint, and a modification was made to improve the power as proposed in Kim et al. (2009). A traditional F-type test statistic was used to test the null hypothesis of  $k_0$  joinpoints versus the alternative hypothesis of  $k_1$  joinpoints, but the fact that it does not have a well-known distribution, even asymptotically, motivated us to use a permutation procedure to estimate its P-value.

Regarding the fitting of a piecewise linear regression model at a given  $k$ , we first estimate the regression parameters for given locations of joinpoints and then search for the joinpoint locations that minimize the residual sum of squares. To estimate the unknown joinpoints, we used the grid search proposed by Lerman (1980) and implemented the Hudson's algorithm (Hudson (1966))

---

<sup>1</sup> This report provides technical details for methods used in:

Zhu L, Pickle LW, Ghosh K, Naishadham D, Portier K, Chen HS, Kim HJ, Zou Z, Cucinelli J, Kohler B, Edwards BK, King J, Feuer EJ, Jemal A. Predicting US- and state-level cancer counts for the current calendar year: Part II: evaluation of spatiotemporal projection methods for incidence. *Cancer* 2012 Feb 15;118(4):1100-9.

in later versions to accommodate a continuous fitting where estimated joinpoints can be anywhere in the data range. Strengths of the Hudson's algorithm include that it provides more accurate estimates of the model parameters and is computationally more efficient than fine grid searches, but it is slower than the annual grid search and there were some technical issues to be taken care of. For further details on the Hudson's algorithm, refer to Kim et al. (2008) and Yu et al. (2007). The weighted least squares fitting can be made to handle heteroscedastic errors as well as autocorrelated errors, and Joinpoint provide several options for the weight specification.

Inferences following the least squares fitting were conducted by using asymptotic normal distributions for the slope parameters and the likelihood method for the joinpoints. In order to provide accurate standard error estimates of the slope parameters, Joinpoint incorporated suggestions made in literature: (i) to estimate the standard errors based on non-constrained model and (ii) to delete offending data points observations. See Kim et al. (2008) for further details. In later versions of Joinpoint, we also implemented the point and interval estimates of the annual percent change (APC) and the average annual percent change (AAPC).

The problem of selecting the number of joinpoints is similar to the classical problem of regression model selection, and we pursued both the hypothesis testing and information criterion approaches. The permutation test procedure described above is conservative in nature but has been used as a default with a goal to find a most parsimonious model. It is time consuming since a resampling distribution is generated to estimate the P-value of the test, and we implemented sequential stopping methods to improve its computational efficiency. See Fay et al. (2007) for further details. Another method of regression model selection is an information criterion based method such as Bayesian Information Criterion (BIC) or Akaike Information Criterion. We implemented the BIC as a faster alternative to the permutation procedure, where the model with  $k$  to minimize  $BIC(k) = \ln\left(\frac{RSS_k}{n}\right) + 2k \frac{\ln n}{n}$  is selected as a final model. The simulation results summarized in Kim et al. (2009) indicate that the BIC tends to over-fit the model and the performance of the BIC is close to that of the permutation procedure with the over-fitting probability controlled to be under 0.15. Recently, Zhang and Siegmund (2007) proposed a modified BIC (MBIC) to select the number of mean changes in a sequence of random variables, and provided theoretical and empirical evidences to support its superiority over other selection

methods. Their idea was applied to the joinpoint regression setting, and the modified BIC for a model with k-joinpoints was derived as an asymptotic approximation of the Bayes factor:

$$\text{MBIC}(k) = \text{BIC}(k) + \frac{\ln |X'_k(\hat{\tau})X_k(\hat{\tau})|}{n} - \frac{2}{n} \ln \Gamma\left(\frac{n-k-3}{2}\right) - \frac{k+3}{n} \ln(\text{RSS}(k)),$$

Where  $\text{RSS}(k)$  denote the residual sum of squares for the model with k-joinpoints,  $\Gamma(z)$  is the gamma function,  $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ , and

$$X_k(\hat{\tau}) = \begin{pmatrix} 1 & x_1 & (x_1 - \hat{\tau}_1)^+ & \dots & (x_1 - \hat{\tau}_k)^+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & (x_n - \hat{\tau}_1)^+ & \dots & (x_n - \hat{\tau}_k)^+ \end{pmatrix}.$$

Compared to the BIC, MBIC assigns a harsher penalty for a larger value of k, and it is expected to be more conservative than the BIC is. Our preliminary simulations indicate that the MBIC performs very well for situations with moderate or large effect sizes, but its performance is even worse than that of the permutation test when effect sizes are very small.

## A.2 NordPred

Time trends in incidence and mortality have also been modeled using the age-period-cohort (APC) model [Holford, 1983]. The data used consists of a set of age-specific counts tabulated for several periods of time. Interval widths for age and period are typically assumed to be equal, in this study, 18 five-year age categories [age group  $i$ :  $i=1, \dots, I=18$ ] and seven or three five-year periods [period  $j$ :  $j=1, \dots, J=7$  (or  $J=3$ )] are used depending on the analysis. Cohorts are defined by subtracting subject age from the period that contains the date of the occurrence of the event of interest. So, using five-year age categories and five-year periods, individuals in the age category 60-65 who died from say lung cancer during the period 1999-2003 would belong to the 1934-1943 birth cohort. Birth cohorts will have overlap. Using standard notation, the  $k^{\text{th}}$  cohort is identified by  $k=j + I - i$ .

The typical APC model assumes additive age, period and cohort effects on log rates with Poisson errors. Let  $Y_{ijk}$  be the mortality (or incidence) rate for the  $i^{\text{th}}$  age group,  $j^{\text{th}}$  period and  $k^{\text{th}}$  cohort.

Then  $Y_{ijk}$  is assumed to have a Poisson distribution with mean (and standard deviation),  $\mu_{ijk}$ , that is related to age, period and cohort through a linear link function (2.1).

$$G(\mu_{ijk}) = \alpha_i + \pi_j + \tau_k \quad (2.1)$$

Where  $\mu$  is the expected rate,  $\alpha_i$  is the effect for the  $i^{\text{th}}$  age group,  $\pi_j$  is the effect for the  $j^{\text{th}}$  period and  $\tau_k$  is the effect for the  $k^{\text{th}}$  cohort, and  $G()$  is a monotonic function that links the expected rate to a linear function of the age, group and period effects. The usual constraints,  $\sum_i \alpha_i = \sum_j \pi_j = \sum_k \tau_k$ , are applied. Additional constraints are needed to account for the linear relationship among the subscripts  $k, j$  and  $i$ . This interdependency is accounted for in the generalized inverse used to estimate the parameters and in all formal statistical tests involving these parameters.

A common modification of the traditional APC model is to structure the linear function to incorporate a common drift parameter to facilitate predictions [Clayton and Schiffers, 1987]. With this model, the log link function becomes.

$$G(\mu_{ijk}) = \alpha_i + \delta_j + \tau_k + D \cdot j \quad (2.2)$$

In this formulation, the regression coefficient  $D$  is called the common drift and the  $\delta_j$  measures the deviations from linearity in the period factor.

A log link ( $G()=\log()$ ) leads to exponential growth in the rate over time which typically overestimates true future values. To level off this exponential growth, Engeland *et al.* [1993] and Møller, *et al.* (2003) examined a number of modifications to the common drift form of the APC model and recommended the following power link function.

$$\mu_{ijk} = (\alpha_i + \delta_j + \tau_k + D \cdot j)^5 \quad (2.3)$$

This model is referred to in this paper as the “NordPred” model because it has been used extensively to predict cancer mortality and incidence rates for the Nordic countries.

Model parameters are estimated using standard generalized linear model methodology as implemented in the *glm()* function in R (version 2.12.2, 2011). Input consists of mortality or incidence counts along with age, period and cohort indicator variables, and counts are assumed

to follow a Poisson distribution. Only consecutive age groups with count summed over all periods greater than 45 are included in the estimation routine. The 45 count threshold is needed to ensure that the `glm()` procedure has sufficient data to ensure acceptable parameter estimates. When the total count is less than 45, the average count over the previous two time periods are used as the estimated count. Typically this included counts for the youngest age categories (latest cohorts) where few deaths (or incidence) are observed and hence how these are estimated has little impact on the overall predictability of the model. Population counts are also used by the estimation procedure to compute expected rates from counts and for prediction.

Prediction is performed in two steps. First the fitted generalized linear model is used to predict rates for two periods beyond the last period of actual data. This step requires that population counts for these two projection periods are available. Next, predicted rates are used with linear interpolation to provide estimated rates for predicting three and four years beyond the last year of data. These rates are then converted to counts using estimated individual year population values also computed using linear interpolation from the five-year period estimates.

Future predictions assume cohort and age effects equal to the last estimated values. The fitted model assumes a linear trend for period drift. The experience of Møller, *et al.* (2003) indicates that future predictions are best if the effect of drift is assumed to fade over time. For this analysis, predictions for the first five-year time period beyond the last period of data used only 0.25 of  $\widehat{D}$  and the second five-year time period prediction used no drift. Period deviations ( $\delta_j$ ) for future predictions are assumed zero.

### **A.3 State space method**

#### **A.3.1 Model specification**

As before, we assume that the observed mortality (or incidence) count at time  $t$  is given by  $d_t$ , which is subject to uncertainty due to measurement error. This is quantified using

$$d_t = \alpha_t + \varepsilon_t, \quad t=1, \dots,$$

where  $\alpha_t$  is the trend and  $\varepsilon_t$  is the measurement error with mean 0 and variance  $V_t$ . We call this the measurement equation.

Next, we model the year-to-year variation in the trend in the form of a local-quadratic model. This is achieved by the following set of equations:

$$\left. \begin{aligned} \alpha_t &= \alpha_{t-1} + \beta_{t-1} + \gamma_{t-1} + \eta_{1t} \\ \beta_t &= \beta_{t-1} + 2\gamma_{t-1} + \eta_{2t} \\ \gamma_t &= \gamma_{t-1} + \eta_{3t} \end{aligned} \right\}, \quad t=1, \dots$$

Here  $\beta_t$  and  $\gamma_t$  can be interpreted as the local slope and acceleration respectively of the trend of the mortality series and  $\eta_{it}$  are the random transition errors, assumed to be serially uncorrelated with mean 0. We call this the transition equation.

We can rewrite the measurement and transition equations in the following compact notation

$$d_t = F_t \Theta_t + \varepsilon_t$$

$$\Theta_t = G_t \Theta_{t-1} + \eta_t$$

where  $\Theta_t = (\alpha_t, \beta_t, \gamma_t)'$  is called the state vector,  $F_t = (1, 0, 0)'$  is called the measurement matrix,

$$G_t = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{pmatrix}$$

is called the transition matrix and  $\eta_t = (\eta_{1t}, \eta_{2t}, \eta_{3t})'$  is the vector of transition errors assumed to have covariance  $W_t$ .

We assume that the measurement errors  $\varepsilon_t$  and transition errors  $\eta_t$  are uncorrelated with each other and with themselves at different points in time. To complete the model specification, we assume that the initial state  $\Theta_0$  has mean  $a_0$  and covariance  $C_0$ . This formulation is called a state-space model. For more details on state space models, see Harvey (1989), Harvey (1993) and West and Harrison (1997).

### A.3.2 Estimation and prediction

When  $F_t$ ,  $G_t$ ,  $V_t$ ,  $W_t$ ,  $a_0$  and  $C_0$  are completely known, the Kalman Filter algorithm (Kalman, 1960, Kalman and Bucy, 1961, Meinhold and Singpurwalla, 1983, Harvey, 1989, 1993) can be

applied recursively to calculate optimal estimator of the state vector at time  $t$ . Once the end of the series is reached, further application of the Kalman Filter allows one to obtain optimal predictions of the future observations. The algorithm is briefly described below.

### A.3.2.1 Kalman Filter

Let  $\hat{\Theta}_t$  be the optimal estimator of the state vector  $\Theta_t$  based on  $d_1, \dots, d_t$  and  $C_t$  be the corresponding  $p \times p$  MSE matrix (for our case,  $p=3$ ). We then have

$$C_t = E[(\Theta_t - \hat{\Theta}_t)(\Theta_t - \hat{\Theta}_t)'].$$

Suppose we are at time  $t$  and  $\hat{\Theta}_t$  and  $C_t$  are available. Then, based on data up to and including time  $t$ , the optimal estimator of  $\Theta_{t+1}$  is

$$\hat{\Theta}_{t+1|t} = G_{t+1} \hat{\Theta}_t \quad (3.1)$$

and the updated MSE matrix for  $\hat{\Theta}_{t+1|t}$  is

$$C_{t+1|t} = G_{t+1} C_t G_{t+1}' + W_{t+1}. \quad (3.2)$$

Equations (3.1) and (3.2) are called the prediction equations. The corresponding estimator of  $d_{t+1}$ , called the predicted value,  $y_{t+1|t}$  is then

$$y_{t+1|t} = F_{t+1} \hat{\Theta}_{t+1|t}$$

Let the prediction error of  $d_{t+1}$  based on data upto  $t$  (also called the innovation vector) be denoted by  $v_{t+1}$ . Then,

$$v_{t+1} = d_{t+1} - y_{t+1|t} = F_{t+1} (\Theta_{t+1} - \hat{\Theta}_{t+1|t}) + \varepsilon_{t+1},$$

and its MSE is given by

$$K_{t+1} = F_{t+1} C_{t+1|t} F_{t+1}' + V_{t+1}.$$

Once a new observation  $d_{t+1}$  becomes available, the estimator  $\hat{\Theta}_{t+1|t}$  of the state vector  $\Theta_{t+1}$  and its corresponding MSE can be updated. The updating equations, known as the KF updating equations, are given by

$$\hat{\Theta}_{t+1|t} = \hat{\Theta}_{t+1|t} + C_{t+1|t} F_{t+1}' K_{t+1}^{-1} (d_{t+1} - F_{t+1} \hat{\Theta}_{t+1|t})$$

and

$$C_{t+1|t} = C_{t+1|t} - C_{t+1|t} F_{t+1}' K_{t+1}^{-1} F_{t+1} C_{t+1|t}$$

Starting with initial conditions  $\hat{\Theta}_0$  and  $C_0$ , the above equations are used recursively for  $t=0,1,\dots,T-1$  to finally get  $\hat{\Theta}_T$ , which contains all the information for predicting future values of  $d_t$ ,  $t>T$ . The  $l$ -step-ahead estimator of  $\Theta_{T+l}$ , given information upto  $T$  is then

$$\hat{\Theta}_{T+l|T} = G_{T+l} \hat{\Theta}_{T+l-1|T}, \quad l=1,2,\dots$$

with  $\hat{\Theta}_{T|T} = \hat{\Theta}_T$ . The associated MSE matrix is given by

$$C_{T+l|T} = G_{T+l} C_{T+l-1|T} G_{T+l}' + W_{T+l}, \quad l=1,2,\dots$$

with  $C_{T|T} = C_T$ .

The  $l$ -step-ahead predictor of  $d_{t+l}$  given  $d_1, \dots, d_T$  is

$$y_{T+l|T} = F_{T+l} \hat{\Theta}_{T+l|T}$$

with its prediction MSE being

$$\text{MSE}(y_{T+l|T}) = F_{T+l} C_{T+l|T} F_{T+l}' + V_{T+l}$$

### A.3.2.2 Estimation of $a_0$ , $C_0$ , $V_t$ and $W_t$

Due to lack of any prior information on  $\Theta_0$ , we use a diffuse prior by setting the mean

$a_0 = (\alpha_0, \beta_0, \gamma_0)'$  of the initial state to be the solution of

$$\begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix} = \begin{pmatrix} 111 \\ 124 \\ 139 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \beta_0 \\ \gamma_0 \end{pmatrix}$$

and taking  $C_0 = 10000I_3$  (see Harvey, 1993). We assume that the two covariance matrices are time invariant, given by  $V_t = \sigma_\varepsilon^2$  and  $W_t = \text{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ . Writing  $y_t = \Delta^3 d_t$ , and  $\gamma(k)$  being the autocovariance of order  $k$  for the  $y_t$  series, it can be shown that

$$\left. \begin{aligned} \gamma(0) &= 20\sigma_\varepsilon^2 + 6\sigma_1^2 + 2\sigma_2^2 + 2\sigma_3^2; \\ \gamma(1) &= -15\sigma_\varepsilon^2 - 4\sigma_1^2 - \sigma_2^2 + \sigma_3^2; \\ \gamma(2) &= 6\sigma_\varepsilon^2 + \sigma_1^2; \\ \gamma(3) &= -\sigma_\varepsilon^2; \\ \gamma(k) &= 0; \quad k \geq 4 \end{aligned} \right\}. \quad (3.3)$$

We use Equation (3.3) to estimate the entries of  $V$  and  $W$ , and refer to this method as the method of moments. If any estimates turned out to be negative, they were replaced by 0. We then use the Kalman Filter algorithm described earlier to obtain the predictions  $z_{T+4} = y_{T+4|T}$ .

### A.3.2.3 Tuning

It was observed that for certain cancer sites, this method sometimes resulted in wide year-to-year fluctuations in the predicted counts. To remedy this condition, we used a two-step method of obtaining the predictions. First we estimated the  $V$  and  $W$  matrices using method of moments as before. We then introduced non-negative “tuning parameters”  $\kappa_V$  and  $\kappa_W$ , which are multipliers of the  $V$  and  $W$  matrices respectively. Defining

$$e_{t+4}(\kappa_V, \kappa_W) = z_{t+4} - d_{t+4}$$

as the 4-year-ahead prediction error for  $d_{t+4}$  when using  $\kappa_V$  and  $\kappa_W$  as the tuning parameters, the sum of squares of 4-year-ahead prediction errors is then given by

$$SSPE(\kappa_V, \kappa_W) = \sum_{t=7}^{34} e_{t+4}^2(\kappa_V, \kappa_W).$$

The tuning parameters are chosen so that the above quantity is minimized. The optimal values of  $\kappa_V$  and  $\kappa_W$  were obtained using the Nelder-Mead algorithm (Nelder and Mead, 1965), implemented in R through the routine `optim`. In the second step, the optimal  $(\kappa_V, \kappa_W)$  values in conjunction with the estimated  $V$  and  $W$  matrices were used to obtain the desired 4-year-ahead prediction of  $d_{T+4}$  using the Kalman Filter.

The whole procedure is implemented in R (R Development Core Team, 2008). More details of the State Space method and Kalman Filter algorithm are available in Ghosh et al. (2007).

## A.4 Bayes State Space method

### A.4.1 Model specification

Consider a yearly time series of mortality (or incidence) counts given by  $(d_t)_{t=1}^T$ . At each time point  $t$ , we model the observed counts using a Poisson distribution, namely

$$d_t | \Theta_t \stackrel{indep}{\sim} \text{Poisson}(\Theta_t), \quad t=1, \dots \quad (4.1)$$

Equation (4.1) is called the measurement equation, since it is used to capture the uncertainty in the observations or the measurements. The mean of the Poisson distribution at time  $t$  is assumed to be related to an unknown  $p$ -dimensional vector of regression coefficients  $\mu_t$ , through the link function

$$\Theta_t = \exp(F_t' \mu_t),$$

where  $F_t$  is a completely known  $p$ -dimensional vector, possibly changing with time. The vector of regression coefficients  $\mu_t$  is called the state vector, since it can be used to determine the “average state” of the time series at  $t$ .

Next, we model the year-to-year variation of the time series through the following relation between state vectors at consecutive time points

$$\mu_t = G_t \mu_{t-1} + \varepsilon_t, \quad t=1, \dots, \quad (4.2)$$

where  $G_t$  is a completely known  $p \times p$  transition matrix that is possibly varying with time and  $\varepsilon_t$  is a random error satisfying

$$\varepsilon_t \stackrel{iid}{\sim} N_p(0, \Sigma), \quad t=1, \dots$$

Equation (4.2) is called the transition equation and combined with Equation (1), define a dynamic generalized linear model (DGLM).

#### A.4.2 Model fitting and prediction

We use the Bayesian paradigm to fit the postulated model. The likelihood function is proportional to

$$\frac{1}{|\Sigma|^{T/2}} \exp\left[ \sum_{t=1}^T \left\{ d_t F_t' \mu_t - \frac{(\mu_t - G_t \mu_{t-1})' \Sigma^{-1} (\mu_t - G_t \mu_{t-1})}{2} - e^{F_t' \mu_t} \right\} \right].$$

The Bayesian model specification is completed by specifying the prior distributions of the initial state  $\mu_0$  and the transition covariance matrix  $\Sigma$ . We assume

$$\mu_0 \sim N_p(m_0, C_0),$$

where  $m_0, C_0$  are completely known. Furthermore, we assume that the  $p$ -dimensional covariance matrix  $\Sigma$  is diagonal, with elements

$$\sigma_i^2 \stackrel{iid}{\sim} \text{IG}(a_\sigma, b_\sigma), \quad i=1, \dots, p,$$

where  $\text{IG}(a, b)$  denotes inverse-gamma distribution with parameters  $(a, b)$  whose density is given by

$$f(x) = \frac{1}{\Gamma(a) b^a x^{a+1}} e^{-1/(bx)}, \quad x > 0.$$

We assume  $a_\sigma$  and  $b_\sigma$  are completely known.

We use a combination of various Markov chain Monte Carlo techniques such as the Gibbs sampler and Metropolis-Hastings sampler to estimate the posterior distribution of the model parameters. In particular, using “...” to denote “the rest”, we have

$$\mu_0 | \dots \sim N_p(m_0^*, C_0^*),$$

where

$$C_0^* = (G_1' \Sigma^{-1} G_1 + C_0^{-1})^{-1},$$

and

$$m_0^* = C_0^* (G_1' \Sigma^{-1} \mu_1 + C_0^{-1} m_0).$$

Similarly, we have

$$\sigma_i^2 | \dots \sim \text{IG}(a_\sigma + \frac{T}{2}, \{b_\sigma^{-1} + \frac{1}{2} \sum_{t=1}^T [(\mu_t - G_t \mu_{t-1})]_i^2\}^{-1}), i=1, \dots, p.$$

The remaining state vectors  $(\mu_t)_{t=1}^T$  are updated using Metropolis-Hastings steps with the multivariate normal random walk sampler, whereby the covariance of the proposal distribution is tuned according to the algorithm of Roberts and Rosenthal (2001) to attain optimal acceptance rates. Fitting of similar models is described in Schmidt and Pereira (2011). For more on Bayesian DGLM, see West and Harrison (1997).

At each iteration of the Gibbs sampler, once we obtained an updated value of  $\mu_T$ , we first ran the transition equation (2) four additional steps to obtain an updated value of  $\mu_{T+4}$ . This was then used in the measurement equation (1) to obtain an updated value of  $d_{T+4}$ . Denoting the updated value of  $d_{T+4}$  in the  $m$ th iteration by  $d_{T+4}^{(m)}$ , the estimated value of  $d_{T+4}$  was obtained as

$$z_{T+4} = \frac{1}{M} \sum_{m=1}^M d_{T+4}^{(m)},$$

where  $M$  is the total number of iterations.

For our case, we used  $p=1$  with  $F_t=1$  and  $G_t=1$  (this is a local level model, or a local-polynomial model of first order). We also used  $a_\sigma=3$  and  $b_\sigma=2$ , reflecting lack of information on the transition variance. Furthermore, we chose  $C_0=10$  and  $m_0=0$  to reflect lack of information on the initial state.

The Gibbs sampler was run for 200,000 iterations with the first half discarded as burn-in (for some sites however, it was necessary to run the sampler for 400,000 iterations). Convergence was assessed visually using traceplots of the sampled parameters. On convergence, the remaining iterations were used for posterior calculations, with a thinning of every 100. The code was written in R (R Development Core Team, 2008).

#### A.5. Summary Metrics for Comparing Estimates

Assume  $\hat{\theta}_{sm}$  is the predicted mortality or incidence count for specific scenario  $s$ ,  $s=1, \dots, S$ , via method  $m$ ,  $m=1, \dots, M$ , and that  $\theta_s$  is the true observed count. Let  $\rho_{sm}$  represent the rank of the squared deviation  $\left[ (\hat{\theta}_{sm} - \theta_s)^2 \right]$  for the estimate of scenario  $s$  by method  $m$  among all  $M$  methods. Thus  $1 \leq \rho_{sm} \leq M$ , where two methods that produce the same estimate are assigned an average rank value. The following statistics were computed to support comparison among the different estimation methods.

$$\text{Average Absolute Relative Deviation: } AARD_m = \frac{1}{S} \sum_{s=1}^S \frac{|\hat{\theta}_{sm} - \theta_s|}{\theta_s + .5}$$

$$\text{Maximum Absolute Relative Deviation: } MARD_m = \max \left( \frac{|\hat{\theta}_{sm} - \theta_s|}{\theta_s + .5} \right)$$

$$\text{Mean Relative Sums of Squares: } MRSSD_m = \frac{1}{S} \sum_{s=1}^S \frac{(\hat{\theta}_{sm} - \theta_s)^2}{\theta_s + .5}$$

$$\text{Root Mean Square Error: } RMSE_m = \sqrt{\frac{1}{S} \sum_{s=1}^S (\hat{\theta}_{sm} - \theta_s)^2}$$

$$\text{Normalized Root Mean Square Error: } NRMSE_m = \left( \sqrt{\frac{1}{S} \sum_{s=1}^S (\hat{\theta}_{sm} - \theta_s)^2} \right) / \left( \frac{1}{S} \sum_{s=1}^S \theta_s \right)$$

$$\text{Average Rank of the Relative Sums of Squares: } ARRSS_m = \frac{1}{S} \sum_{s=1}^S \hat{\rho}_{sm}$$

The AARD is interpreted as the average percent deviation from the true value relative to the true value. This measure attempts to take into account the relative differences in observed mortality or incidence counts among different cancers and or geographic areas as we attempt to assess the extent to which the estimates deviate from observed. The MARD is a measure of the worst the deviation from observed might be. The MRSSD is similar to the AARD only deviations are squared resulting in applying higher weights to larger deviations in the average. The RMSE is an estimate of variability of estimates about the true value. The NRMSE is the MSE expressed as a fraction of the mean.

The ARRSS is the average rank of deviations among the methods. A method which has the smallest squared deviations among all methods will have rank of one. If on average the rank for this method is close to one then we would conclude that this method is “best” in the sense that it consistently beats all other methods in getting close to the observed value. A method with smallest average rank is assumed to produce closer estimates to the true value than any of method most often, although there may be situations where it is not always the very best.

#### A.6 List of covariate in incidence prediction

	Variable name	Definition	Data source	Original source
<b>Geographic definition</b>				
	fipscty	fips state/county code (5 digits)	Census	Census
	state	state fips code (2 digits)	Census	Census
	Census Division	9 regions of the country	Census	Census
	inseer	1 if this is a county in the NCI SEER program, 0 if NPCR only	SeerStat	SEER
<b>Medical facilities</b>				
	MDratio	# physicians per 1000 population	ARF	AMA Physician Masterfile
	hosp	# hospitals per 1000 pop	ARF	AMA Physician Masterfile
<b>Ethnicity/origin</b>				
	pcthis	% of Hispanic origin	SeerStat	Census
	pctBlk	% of total pop who are black	SeerStat	Census
	pctAIAN	% of total pop who are American Indian or Alaskan Natives	SeerStat	Census
	pctAPI	% of total pop who are Asian/Pacific Islanders	SeerStat	Census
	pctforeign	% of total pop who are foreign born	SeerStat	Census

	Variable name	Definition	Data source	Original source
	pctlangisol	% of households in which no person ages 14+ speaks only English and who does not speak English very well	Census	Census
<b>Household characteristics</b>				
	pctfemhh	% households headed by female	ARF	Census
	crowded	% of persons living with > 1 person per room on average	SeerStat	Census
<b>Urban/rural indicators</b>				
	pcturban	% urban pop	SeerStat	Census
	popdens	# persons/square mile	SeerStat	Census
<b>Socioeconomic status</b>				
	pctpoor	% living below federal poverty line	SeerStat	Census
	pctlft9ed	% adults over 25 with < 9 years of education	SeerStat	Census
	pctcoled	% adults over 25 with 4+ years of college education	SeerStat	Census
	unemploy	% unemployed	SeerStat	Census
	pctwhtcl	% adults employed in white collar jobs	SeerStat	Census
<b>Cancer screening</b>				
	pctmam	% women ages 50-64 who had a mammogram in past 2 years	BRFSS	BRFSS
	pctpap	% women ages 20+ who had a Pap smear in past 5 years	BRFSS	BRFSS
	pctpsa	% of men ages 40+ who ever had a PSA test	BRFSS	BRFSS
<b>Health insurance</b>				
	pctnoins	% of persons ages 18+ who do not have a health plan or health insurance	BRFSS	BRFSS
<b>Lifestyle</b>				
	pctsmkmales	% of males ages 18+ who ever smoked cigarettes	BRFSS	BRFSS
	pctsmkfem	% of females ages 18+ who ever smoked cigarettes	BRFSS	BRFSS
	pctbmi	% of persons ages 18+ who are >120% of the median body mass index	BRFSS	BRFSS
	overweightOrObese_2007_both	% of persons ages 18+ with a body mass index of 25+	BRFSS	BRFSS
	pctvigorous	% of persons ages 18+ who met guidelines for vigorous exercise in 2001-3	BRFSS	BRFSS
<b>Miscellaneous</b>				
	landarea	land area in square miles	Census	Census geography files
	mortrate	mortality rate	SEER	NCHS

## References

- Clayton D, Schifflers E. (1987), Models for temporal variation in cancer rates. II: Age-period-cohort models. *Statistics in Medicine* 1987; 6:469– 481.
- Engeland A, Haldorsen T, Tretli S, Hakulinen T, Hörte T, Luostarinen T, Magnus K, Schou G, Sigvaldason H, Storm HH, Tulinius H, Vaittinen P. (1993), Prediction of cancer incidence in the Nordic countries up to the years 2000 and 2010. A collaborative study of the five Nordic Cancer Registries. *APMIS Supplementum* 1993; 38:1–124.
- Fay, M. P. , Kim, H.-J., and Hachey, M. (2007), On using Truncated Sequential Probability Ratio Test Boundaries for Monte Carlo Implementation of Hypothesis Tests, *Journal of Computational and Graphical Statistics* 16, 946-967.
- Ghosh, K., Tiwari, R. C., Feuer, E. J., Cronin, K. A., and Jemal, A. (2007), “Predicting US Cancer Mortality Counts Using State Space Models,” in *Computational Methods in Biomedical Research*, eds. Khattree, R. and Naik, D. N., Chapman & Hall/CRC, pp. 131–151.
- Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, UK: Cambridge University Press.
- Harvey, A. C. (1993), *Time Series Models*, Cambridge, MA: The MIT Press, 2nd ed.
- Holford TR. (1983), The estimation of age, period and cohort effects for vital rates. *Biometrics* 1983; 39:311–324.
- Hudson, D.J. (1966), Fitting segmented curves whose join points have to be estimated, *Journal of the American Statistical Association* 61, 1097 – 1129.

- Kalman, R. E. (1960), "A New Approach to Linear Filtering and Prediction Problems," *Journal of Basic Engineering, Transactions ASME, Series D*, 82, 35–45.
- Kalman, R. E. and Bucy, R. S. (1961), "New Results in Linear Filtering and Prediction Theory," *Journal of Basic Engineering, Transactions ASME, Series D*, 83, 95–108.
- Kim, H.-J., Fay, M.P, Feuer, E.J., and Midthune, D.N. (2000), Permutation Tests for Joinpoint Regression with Applications in Cancer Rates, *Statistics in Medicine* 19, 335-351.
- Kim, H.-J. , Yu, B., and Feuer, E. J. (2008), Inference in Segmented Line Regression: A Simulation Study, *Journal of Statistical Computation and Simulation* 78 (11), 1087-1103.
- Kim, H.-J. , Yu, B., and Feuer, E. J. (2009), Selecting the Number of Change-Points in Segmented Line Regression, *Statistica Sinica* 19, 597-609.
- Lerman, P.M. (1980), Fitting Segmented Regression Models by Grid Search, *Applied Statistics* 29, 77-84.
- Meinhold, R. J. and Singpurwalla, N. D. (1983), "Understanding the Kalman Filter," *American Statistician*, 37, 123–127.
- Møller B., Fekjær H., Hakulinen T., Sigvaldason H, Storm H. H., Talbäck M. and Haldorsen T. (2003), "Prediction of cancer incidence in the Nordic countries: Empirical comparison of different approaches" *Statistics in Medicine* 2003; 22:2751-2766
- Nelder, J. A. and Mead, R. (1965), "A Simplex Algorithm for Function Minimization," *Computer Journal*, 7, 308–313.
- R Development Core Team (2011), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Roberts, G. O. and Rosenthal, J. S. (2001), “Optimal Scaling for Various Metropolis-Hastings Algorithms,” *Statistical Science*, 16, 351–367.

Schmidt, A. M. and Pereira, J. B. M. (2011), “Modelling Time Series of Counts in Epidemiology,” *International Statistical Review*, 79, 48–69.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer.

Yu, B., Barrett, M.J., Kim, H.-J., and Feuer, E. J. (2007), Estimating Joinpoints in Continuous Time Scale for Multiple Change-Point Models, *Computational Statistics and Data Analysis* 51, 2420-2427.

Zhang, N.R. and Siegmund, D.O. (2007), A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data, *Biometrics* 63, 22-32.