Technical Report RM 690, Department of Statistics and Probability, Michigan State University

Change Point Analysis of Cancer Mortality Rates for US States using Functional Dirichlet Processes

Sarat C. Dass, Chae Young Lim and Tapabrata Maiti*
Department of Statistics & Probability
Michigan State University, East Lansing, MI 48824
Email: {sdass,lim,maiti}@stt.msu.edu

Abstract

Cancer is a leading cause of mortality in the US. Detecting changes in trend of cancer incidence rates is an important task for analysis and subsequent intervention for the improvement of public health. The National Cancer Institute has developed several joinpoint models to track changes in these incidence rates based on the entire US population. It is known that cancer incidence rates are heterogeneous across geographical regions. The aim of this paper is to supplement the existing tools for analyzing cancer rates from the Surveillance, Epidemiology, and End Results database that are able to find the change points locally. Subsequently, the model can cluster the geographical subregions based on the magnitude and direction of changes of the disease risk. The proposed model to find change-points over time and cluster spatial locations is based on Dirichlet process priors where we consider temporal functions as the random quantities arising from the Dirichlet process prior. Through the analysis of age adjusted lung cancer mortality rates from 1969 to 2006, the proposed model nicely characterized local data features, namely, the local change points, the rate of changes, and clusters of states that exhibited similar trends of cancer incidence rates. This is also an innovative application of Dirichlet process priors on functional spaces.

Keywords: Joinpoint analysis; Disease mapping; Bayesian nonparametrics; Dirichlet process priors.

1 Introduction

Statistical methods for analyzing disease incidence or mortality data over geographical regions and time have gained considerable interest in recent years due to increasing concerns of public health, health disparity and legitimate resource allocations. Cancer is a major threat to public health in the United States and in the world. Cancer accounts for nearly one-quarter of deaths in the United States, exceeded only by heart disease. The American Cancer Society (ACS, www.cancer.org)

*Authors' names are in alphabetical order.

tracks cancer occurrences, including the number of deaths, cases, and survival times after diagnosis. According to the ACS report Cancer Facts and Figures 2010, the expected number of new cancer cases in 2010 is 1,529,560, and about 569,490 individuals are expected to die from cancer in the United States (US) in 2010. In 2008, 7.6 million death from cancer was estimated and by 2030, the global burden is expected to grow to 17.5 million cancer deaths per year. Among many things, the ACS publishes time trends of age-adjusted cancer death rates for different cancer types, and for different sub-populations defined by geographic and socio-demographic characteristics.

Even though several surveillance studies have been undertaken to control cancer, it is a fact that there was an increased number of cancer deaths in 2007 as a result of aging and growth of the US population (ACS 2010). Moreover, the impact of cancer surveillance is not uniformly effective over geographical regions; see, for example, Figure 1 which displays cancer trends for four different US states and the overall trend for the nation. One of the scientific objectives of monitoring cancer rates is to detect changes in the trend over time and identify clusters of sub-populations (generally a set of geographical sub-regions) that are affected by changes (increase or decrease) in risk. A carefully developed procedure that addresses this issue can help administrators find key information for the prevention of cancer.

Several joinpoint models that identify time points associated with a significant change in disease trend have been developed by several authors (See, Carlin et al. (1992), Kim et al. (2000, 2004), Tiwari et al. (2005) and Ghosh et al. (2009)). The models developed by Kim et al. (2000, 2004), for example, are used in cancer statistics review and implemented in the software of the National Cancer Institute (NCI) (Ries et al. 2002). These models focus on detecting joinpoints over time in one time series of disease rates. Ghosh et al. (2009) applied their model to incidence rates of colon and rectum cancer in the US from 1973 to 1999 and incidence rates of prostate cancer among white males in the US from 1975 to 2003. They perform the joinpoint analysis for a *single time series*. Subsequently, by aggregating over all states in the US, they discovered one set of joinpoints (based on a single time series) for the entire nation.

One important question here is whether the rates of cancer incidence before and after the joinpoint is significantly different (statistically speaking) from each other. In other words, we wish to determine if there is a significant *change point* in the cancer incidence rates before and after the joinpoint. Another important concern not addressed by joinpoint modeling is whether there are groups (or, clusters) of states exhibiting similar change-points of cancer incidence rates but with significant variations between and within groups. It is well known that the cancer rates in the US vary widely by geographical area. According to ACS report 2010, lung cancer mortality rates are 3-fold higher in Kentucky, the state with highest rates, than in Utah which has the lowest rates. Geographic variations also reflect differences in environmental exposure and socio-economic factors in population demographics. Figure 1 gives one such illustration based on age-adjusted lung cancer mortality rates for four states: Florida, Arizona, Missouri and Indiana. It is evident that Florida and Arizona share the same change-point and rates of change in each time segment while Missouri and Indiana exhibit different levels of these attributes. Figure 1 also demonstrates that

Florida and Arizona have different levels of variability around its mean value over time. When we are interested in grouping states by the rates of change, variability is nuisance with respect to the clustering criteria. However, the omission of such variability from the model may result in inefficient estimation. The presence of heterogeneity among states can also give a misleading impression for the rates of change corresponding to the overall US. Figure 1 panel (e) for the entire nation does not reveal the two distinct types of change-points exhibited in the first four panels by the four different states. This indicates that a more efficient estimation procedure is possible by taking into account local geographical effects.

To address the above scientific question, we developed a *change-point* model to analyze age-adjusted cancer mortality rates. In essence, our change-point model is supplemental to the existing joinpoint models because while the latter detect changes based on a single time series, the proposed model detects changes in multiple time series in presence of heterogeneity. A distinct difference, however, is that our proposed model does not assume connectedness at the joinpoint but is able to perform the analysis more locally at this expense. The proposed model is also able to cluster geographical regions which have similar rates. Additional model flexibility for grouping is obtained by including model parameters that represent local data characteristics (for example, variability around the mean trend in Figure 1), which are allowed to vary from site to site. We also incorporate the unknown number of change-points into the estimation scheme, to be inferred from the posterior probabilities. Previous work assumed a fixed number of change-points; for example, Ghosh et al. (2009) estimated the number of change-points first and then carried out the subsequent analysis by fixing the number of change-points at its estimated value.

Our approach is to make use of the Dirichlet Process (DP) methodology in an innovative way to cluster spatio-temporal data. The DP was introduced by Ferguson (1973,1974) to provide a random distribution of observables on R^p free from parametric assumptions. Since then, DPs have been extensively studied in the statistics literature, most importantly as the Bayesian equivalent of providing non-parametric inference in a variety of settings. Model fitting in this framework are carried out through MCMC routines and have become standard statistical practice; see, for example, Escobar and West (1998) and MacEachern (1998). Recently, the DP framework has been extended to dependent DPs (DDPs), developed by MacEachern (2000), to describe a stochastic process of random distributions. Subsequently, Gelfand *et al.* (2005) used a special form of DDPs to model dependent data in the spatial context. Model fitting, once again, is carried out via MCMC with the computations becoming slightly more demanding but are straightforward extensions of the existing routines. We note that Ghosh et al. (2009) used DP to perform a non-parametric Bayesian analysis of joinpoints where the baseline distribution G_0 is a distribution on R for the errors; this is not for the purpose of clustering but to robustify the analysis with respect to non-normality.

Our innovative way of using the DP is to consider realizations from G_0 that are in more general object spaces; in this case, it is the space of all functions over time that represent change points in cancer trends. The advantage of extending DPs in this manner is two-fold: First, the change-points are included as unknown model parameters with a prior distribution governed by G_0. This

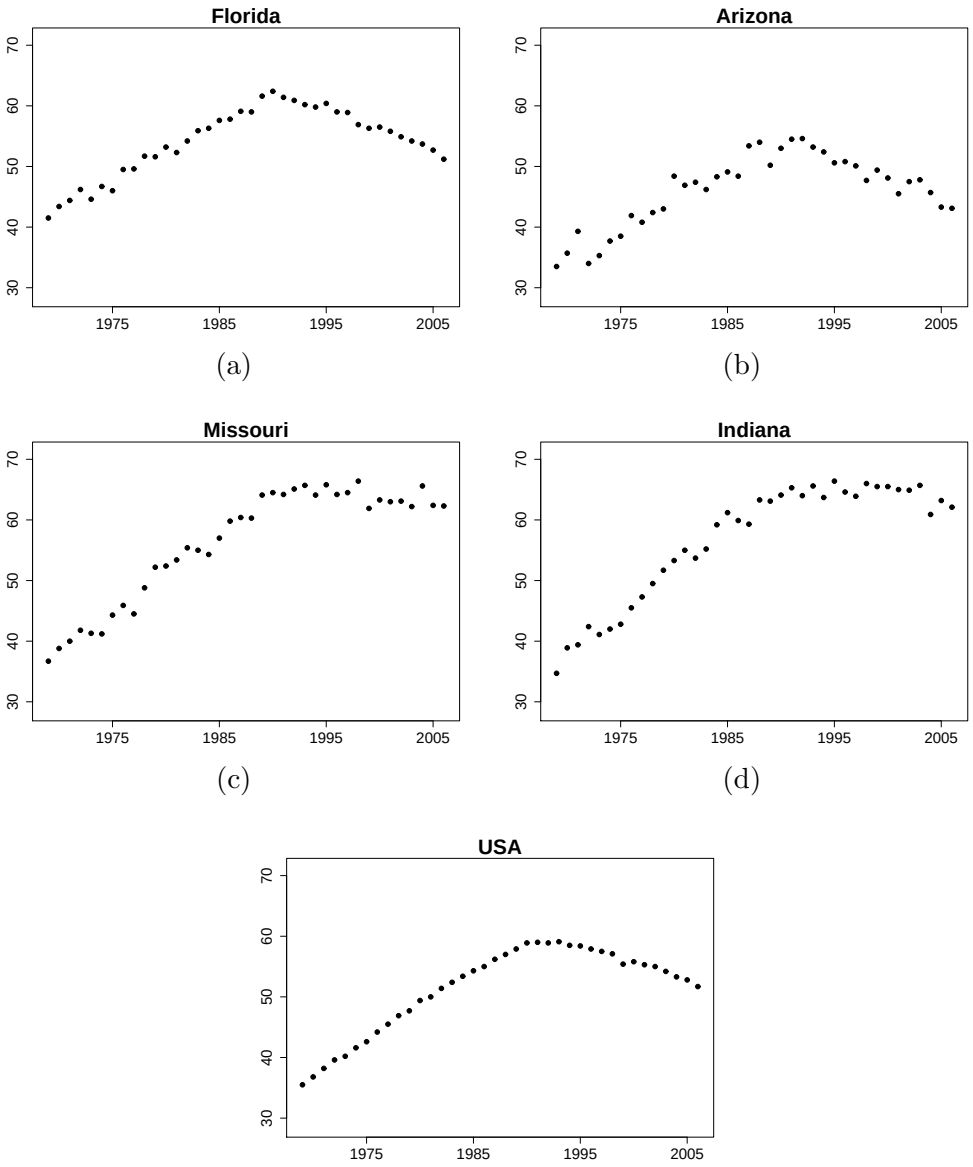

Figure 1: Age-adjusted incidence rates of lung cancer from 1969 to 2006 for four states and the entire US: Florida (a), Arizona (b), Missouri (c), Indiana (d) and entire US (e).

entails that the uncertainty involved in their estimation is taken into account in the inference, and therefore, represents an improvement over the methodology of Ghosh et al. (2009) and previous approaches by others. Second, even for more general object spaces, the intrinsic property of DPs that assign unit mass to all discrete probability distributions can be utilized to enable clustering of sites with respect to similar cancer trends. A number of research articles have utilized this discreteness property of the DP for clustering observables into homogeneous groups according to some pre-specified criteria; see, for example, Gelfand et al. (2005), Escobar and West (1998) and MacEachern (1998). Clustering enables information on different observables be pooled together to obtain smaller estimates of variability and shorter lengths for confidence intervals, in a spirit similar to procedures in small area estimation. Here, of course, the clustering mechanism needs to be flexible enough to capture a variety of clustering characteristics without being forced to concentrate on incorrect specifications, which is achieved by incorporating site specific parameters as mentioned earlier. Extensions of DPs to function spaces, or functional DP methodology, has been carried out in a number of recent research articles; for example, Gelfand et. al. (2005), Duan et al. (2007), Petrone et al. (2009) and Rodriguez et al. (2009). The above research articles utilize the DP prior on the space of functions over a spatial domain with inference based on n independent and identically distributed (iid) realizations of *functions* from this domain. Our application driven methodology of functional DPs is slightly different: For each site on the spatial domain, we have (only one) change-point function. DP-based clustering is obtained for the sites on the spatial domain based on similar change-point functions. Incorporating other aspects of variability via parameters to enhance model flexibility without affecting the DP-based clustering is also an important contribution of this paper.

A primary inferential objective in the analysis of disease data is the summarization and explanation of spatial and spatio-temporal patterns of disease (i.e., disease mapping); see, for example, Elliot et al. (2000), Banerjee et al. (2004) and Lawson (2009) for details and further references. Also of interest is the spatial smoothing, temporal prediction of disease risk and the detection of extremes. Models for inference in this area have been mostly limited by parametric elicitation of dependence structures for pooling spatial information. On the other hand, the proposed DP-based methodology is free of parametric constraints, and its capability of pooling information via data driven clustering can greatly enhance the analysis of spatial and spatio-temporal patterns. As an illustration, we infer the cluster of US states which correspond to the highest drop in cancer trends in Section 4. We are also able to demonstrate statistical significance of the highest drop compared to other clusters of US states. These types of inference can potentially help policy makers identify factors in the top states that contributed to the highest drop, and subsequently, be implemented as policy or programs in the other states.

The rest of the paper is organized as follows. Section 2 gives the details of the data and application while Section 3 presents the proposed change point model and associated Bayesian inference. Section 4 gives two specific model formulations for the cancer data and demonstrates the superiority of incorporating site specific variability. Section 5 gives some validation results. Section

6 gives discussion and future directions for research.

2 A Change Point Model for Cancer Incidence Rates

Cancer incidence rates are obtained from the Surveillance, Epidemiology, and End Results (SEER) program (seer.cancer.gov) of the National Institute of Cancer (NCI). The SEER program is an authoritative source of information on cancer incidence and survival in the US. The SEER program currently collects and publishes cancer mortality and survival data from population-based cancer registries covering approximately 26 percent of the population. An age-adjusted incidence/mortality rate is a primary measure for monitoring cancer trends over time and over geographical locations since cancer is a disease where age is a determining factor. An age-adjusted rate is a weighted average of the age-specific (crude) rates, where the weights are the proportions of persons in the corresponding age groups of a standard population. The potential confounding effect of age is reduced when comparing age-adjusted rates computed using the same standard population. Several sets of standard population data are available in SEER which include the 2000 US standard population as well as the standard US populations for the years 1940, 1950, 1960, 1970, 1980, and 1990. The age-adjusted rate using age groups A through B is calculated using the following formula:

$$aarate_{A-B} = \sum_{i=A}^{B} \left[\left(\frac{count_i}{pop_i} \right) \times 100,000 \times \left(\frac{stdmil_i}{\sum_{i=A}^{B} stdmil_i} \right) \right], \tag{1}$$

where $count_i$, pop_i and $stdmil_i$ are, respectively, the number of incidence/mortality due to a cancer, the population and the choice of a standard population in the age group i. Nineteen age groups and the 2000 US standard population are considered in this study.

We consider lung cancer age-adjusted mortality rates from 1969 to 2006 for the 48 contiguous states in continental United States (excluding Alaska and Hawaii) and Washington D.C. Thus, observations are the age-adjusted lung cancer mortality rates for $t = 1969, 1970, \cdots, 2006$ and $s = 1, 2, \cdots, 49$. Four states and overall USA plots were given in Figure 1 as an example. It is clear from the panels in Figure 1 that there is at least one change-point in the rate of change (i.e., slope) of lung cancer mortality rates for the four states. There are several specific aims of this paper: We would like to determine (i) all possible change-points of slopes of lung cancer mortality rates corresponding to each state, (ii) determine simultaneously if the slopes exhibit some clustering over the states (i.e., different states have identical slope values), and (iii) identify clusters with the highest changes in slope over time, and (iv) whether this highest change is significant compared to the other remaining clusters. To model exponential growth or decay of the age-adjusted rates, we model the logarithm of the age-adjusted rates as a linear function of time as is done in Ghosh et al. (2010), Clegg et al. (2009) and Ghosh et al. (2009). Slopes over the different time segments capture the essential growth rate (positive or negative) pattern of cancer incidence/mortality rates. The change-point model we develop subsequently is in terms of these slopes and the variability of the observations around the log mean trend.

2.1 Change Point Likelihood Based on Observables

The subsequent discussion applies to both cancer incidence and mortality rates, and therefore, we refer to them as just rates. Let W_{st} denote the logarithm of the observed cancer rate at site s and time t for the collection of sites $s = 1, 2, \cdots, N$ and time points $t = U_0, U_0 + 1, U_0 + 2, \cdots, U_1$. Assume that a site s has k change-points in terms of the slope of the log rates; thus, in Figure 1, s may be Florida with $k = 1$ change points. For fixed k, let $[T_{l-1}, T_l)$, $l = 1, 2, \cdots, k+1$ be the time intervals where no changes in the disease trend occur (i.e., no change point). To extract the slope and the variability of the observations around the mean trend in each segment, we consider the following regression model on each $[T_{l-1}, T_l)$:

$$W_{st} = \alpha + \beta t + \epsilon_t, \tag{2}$$

for $t = T_{l-1}, T_{l-1} + 1, \cdots, T_l - 1$ with ϵ_t iid $N(0, \sigma^2)$ for the observed data in $[T_{l-1}, T_l)$; thus, in (2), the log rates are modeled as a linear function of time with intercept and slope α and β, respectively, and σ^2 represents the unknown error variance around the mean linear trend. The dependence on s and l is suppressed for the moment. The following results are well known in regression analysis:

$$(\widehat{\alpha}, \widehat{\beta})^T \sim N\left((\alpha, \beta)^T, \sigma^2 (X^T X)^{-1}\right), \quad \text{and} \tag{3}$$

$$\frac{RSS}{\sigma^2} \sim \chi^2_{T_l - T_{l-1} - 2}, \tag{4}$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the least squares estimators of α and β (which are also the maximum likelihood estimates (MLEs) under the normal error model), RSS is the residual sum of squares given by

$$RSS = \sum_{t=T_{l-1}}^{T_l - 1} (W_{st} - \widehat{\alpha} - \widehat{\beta}\, t)^2, \tag{5}$$

χ^2_ν is the chi-square distribution with ν degrees of freedom, and X is $(T_l - T_{l-1}) \times 2$ matrix whose first and second columns is the vector of ones and $\boldsymbol{t}_l \equiv (T_{l-1}, T_{l-1} + 1, \cdots, T_l - 1)^T$, respectively. Also, in (3) and (4), the statistic $(\widehat{\alpha}, \widehat{\beta})$ is independent of RSS. We emphasize here that the number of joinpoints, k, the time intervals $[T_{l-1}, T_l)$, $l = 1, 2, \cdots, k+1$ and σ^2 are all parameters that are unknown, to be inferred from the subsequent Bayesian analysis. The purpose of introducing these unknown parameters here is to describe the likelihood given the unknown parameters at site s:

$$\widehat{\beta}_l, RSS_l \,|\, \beta_l,\, k, \boldsymbol{t}_l, \sigma_l^2 \overset{ind}{\sim} f_{1l} \times f_{2l} \tag{6}$$

independently for $l = 1, 2, \cdots, k+1$. In (6), $f_{1l}(\widehat{\beta}_l \,|\, \sigma_l^2)$ is the normal pdf with mean β_l and variance $\sigma_l^2 \cdot v$ where v is the $(2,2)$-th entry of $(X^T X)^{-1}$; the explicit form of f_{1l} is

$$f_{1l}(\widehat{\beta}_l \,|\, \beta_l, \sigma_l^2) = \frac{1}{\sqrt{2\,\pi\, v\, \sigma_l^2}} \exp\left\{ -\frac{1}{2\, v\, \sigma_l^2} (\widehat{\beta}_l - \beta_l)^2 \right\}. \tag{7}$$

The density $f_{2l}(RSS_l \,|\, \sigma_l^2)$ in (6) is σ_l^2 times the chi-square density with $m_l = T_l - T_{l-1} - 2$ degrees of freedom whose explicit form is given by

$$f_{2l}(RSS_l \,|\, \sigma_l^2) = \frac{1}{2^{(m_l/2)}\Gamma(m_l/2)} \left(\frac{RSS_l}{\sigma_l^2}\right)^{\frac{m_l}{2}-1} \exp\left\{-\frac{RSS_l}{2\sigma_l^2}\right\} \frac{1}{\sigma_l^2}. \tag{8}$$

Subsequently, we consider the site-wise functions

$$\boldsymbol{\theta}_s(t) = \beta_{sl} \quad \text{if } T_{l-1} \le t \le T_l - 1 \tag{9}$$

where β_{sl} is the true but unknown slope in the interval $[T_{l-1}, T_l)$ at site s. Thus, the functions $\boldsymbol{\theta}_s(t)$ are step-functions of t with k change points at times T_l, $l = 1, 2, \cdots, k$. Denote the set of all observables by $\boldsymbol{Y} = \{\, \boldsymbol{Y}_{sl},\, l = 1, 2, \cdots, k_s,\, s = 1, 2, \cdots, N$ where $\boldsymbol{Y}_{sl} \equiv (\widehat{\beta}_{sl}, RSS_{sl})$ with $\widehat{\beta}_{sl}$ and RSS_{sl} as in (6) for each site s, and k_s is the number of joinpoints for site s. Let $\boldsymbol{\beta}$ denote the collection of all true slope parameters β_{sl}, $l = 1, 2, \cdots, k_s$, $s = 1, 2, \cdots, N\,\}$. Also, denote by $\boldsymbol{K}$, $\boldsymbol{T}$ and $\boldsymbol{\sigma}$ to be the collection of parameters k, $T_l, l = 1, 2, \cdots, k+1$ and σ_l^2 for all the N sites. Assuming independence between the N sites, the likelihood is given by

$$\boldsymbol{f}(\boldsymbol{Y} \,|\, \boldsymbol{\beta},\, \boldsymbol{K},\, \boldsymbol{T},\, \boldsymbol{\sigma}\,) = \prod_{s=1}^{N} \prod_{l=1}^{k_s} f_{1l}^{(s)} \times f_{2l}^{(s)} \tag{10}$$

where $f_{1l}^{(s)}$ and $f_{2l}^{(s)}$ are f_{1l} and f_{2l} corresponding to site s.

As mentioned in the Introduction, the change point analysis here is different from joinpoint modeling. The latter assumes that the cancer incidence rates are continuous at the joinpoints but with different slopes to the left and right of the joinpoint. In our case, we make no assumption on the continuity of the regression at the time points T_l. However, at this expense, the current formulation allows us to infer different slopes for the different sites, and therefore, enable clustering of these slopes based on the DP-methodology. The proposed model also allows the unknown number of change-points and clusters to be inferred concurrently with parameters based on Bayesian posterior probabilities (details in the subsequent sections).

3 Bayesian Inference Using Functional DP Priors

3.1 Functional DP Prior

Let $\boldsymbol{\Theta}$ denote the set of all step functions $\boldsymbol{\theta}_s$ as described in the previous section. We introduce the functional DP as a prior on space of all distributions on $\boldsymbol{\Theta}$. The $DP \equiv DP(\alpha_0 G_0)$ depends on two hyper-parameters, namely, $\alpha_0 > 0$ the precision parameter, and G_0 the baseline (or centering) distribution on $\boldsymbol{\Theta}$. Recall that a randomly generated distribution F from $DP(\alpha_0 G_0)$ is almost surely discrete and admits the representation

$$F = \sum_{i=1}^{\infty} \omega_i\, \delta_{\theta_i}, \tag{11}$$

where δ_z denotes a point mass at z, $\omega_1 = \eta_1$, $\omega_i = \eta_i \prod_{k=1}^{i-1}(1-\eta_k)$, for $i = 2, 3, \cdots$ with $\theta_1, \theta_2 \cdots$ iid from G_0 (Sethuraman, 1994). Traditionally, θ_i was assumed to be scalar or vector-valued taking values in R^p. To model the observational process via change-points, we conceptually extend θ_is in (11) to functions $\boldsymbol{\theta}_i \equiv \{\, \theta_i(t) \,:\, t = U_0, U_0+1, \cdots, U_1\}$. The notations $\boldsymbol{\theta}$, $\boldsymbol{\theta}(t)$ and θ, therefore, will be taken to denote, respectively, a function, the value of $\boldsymbol{\theta}$ evaluated at time t and a possible realization taken by $\boldsymbol{\theta}(t)$. These notations will be used throughout the paper subsequently. For an integer $k \geq 0$, $\boldsymbol{\theta}$ with k change-points has the form

$$\boldsymbol{\theta}(t) = \theta_l \qquad \text{if} \qquad T_{l-1} \leq t < T_l, \tag{12}$$

for $l = 1, 2, \cdots, k+1$ with $U_0 \equiv T_0 < T_1 < \cdots T_k < T_{k+1} \equiv U_1$ as seen earlier. The notation $F \sim DP(\alpha_0 G_0)$ in this context will be taken to mean

$$F = \sum_{i=1}^{\infty} \omega_i \, \delta_{\boldsymbol{\theta}_i}, \tag{13}$$

where $\delta_{\boldsymbol{z}}$ is now a point mass on the step function $\boldsymbol{z}$, ω_is are as before, and the $\boldsymbol{\theta}_i$s are iid from a distribution G_0 on $\boldsymbol{\Theta}$. To specify G_0, the baseline distribution on $\boldsymbol{\Theta}$, it is convenient to utilize a hierarchical structure: (1) Let $K \sim Poisson(\lambda)$. (2) Fix an integer $w > 0$. Given $K = k$, let

$$(n_1, \cdots, n_{k+1}) \sim \text{Multinomial}\left(n_0, \frac{1}{k+1}, \cdots, \frac{1}{k+1}\right),$$

where $n_0 = U_1 - U_0 - (k+1)w = n - 1 - (k+1)w$. (3) Define T_l recursively as $T_0 = U_0$, $T_l = n_l + T_{l-1} + w$ for $l = 1, 2, \cdots, k+1$. Given $T_1, \cdots, T_k$, generate $\theta_1, \cdots, \theta_{k+1}$ iid from the (univariate or multivariate) density π_0 on R^d, and set

$$\boldsymbol{\theta}(t) = \theta_l \;\text{ if } T_{l-1} \leq t < T_l, \tag{14}$$

for $l = 1, \cdots, k+1$. Note that $T_k \leq t \leq T_{k+1}$ for $l = k+1$, K is the number of change-points, T_ls for $l = 1, \cdots, K$ are the time points when a change is made and n_l is the number of time points in the interval $[T_{l-1}, T_l)$ for $l = 1, \cdots, K+1$. Note that again, for $l = K+1$, the interval becomes $[T_K, T_{K+1}]$. By introducing $w > 0$, we avoid zero-length interval since each time interval $[T_l, T_{l+1})$ is at least w units. From the hierarchical specification above, it follows that the infinitesimal measure is given as

$$G_0(d\boldsymbol{\theta}_s) = \left(\frac{e^{-\lambda}\lambda^k}{k!}\right)\left(\frac{\Gamma(n_0+1)}{\prod_{i=1}^{k}\Gamma(n_i+1)}\left(\frac{1}{k+1}\right)^{n_0}\right)\prod_{l=1}^{k+1}\pi_0(\theta_l)\,d\theta_l. \tag{15}$$

3.2 Incorporating Site-specific Variability

The prior development thus far has been on the change point functions $\boldsymbol{\theta}_s$. The variance parameters σ^2_{sl} represent the extent of variability of the log rates around the mean trend. Note from Figures 1 (a) and (b) that although Florida and Arizona have the same cancer trends, the variability around this common mean trend is different for the two states. This necessitates the incorporation of σ^2_{sl}

as site specific parameters independent of the clustering. In fact, we demonstrate in Section 4, the exclusion of such consideration (that is, allowing σ to be common for all the sites in a cluster but different for the different time segments) results in poor clustering of cancer trends. Thus, for additional flexibility, the likelihood component of $\boldsymbol{Y}_{sl}$ incorporates a site-specific variability parameter $\xi_s = \sigma_{sl}^2$ for all $l = 1, 2, \cdots, k$ (that is, one common site-wise variance parameter), for each $s = 1, 2, \cdots, N$. For the subsequent Bayesian analysis, the parameters $\xi_s \in \Xi$, where Ξ is its parameter space, are assumed to be iid from the pdf π_1. Note that ξ_s can be different for each s, and therefore, are not subject to site-based clustering as the change-point functions $\boldsymbol{\theta}_s$. The infinitesimal measure in (15) is now extended to include the site-wise parameters ξ_s and is given by

$$\tilde{G}_0(d\boldsymbol{\theta}_s, d\xi_s) = \left(\frac{e^{-\lambda}\lambda^k}{k!}\right)\left(\frac{\Gamma(n_0+1)}{\prod_{i=1}^{k}\Gamma(n_i+1)}\left(\frac{1}{k+1}\right)^{n_0}\right)\left(\prod_{l=1}^{k+1}\pi_0(\theta_l)\,d\theta_l\right)\pi_1(\xi_s)\,d\xi_s. \tag{16}$$

In what follows, it will be useful to make the following definition: For fixed $\boldsymbol{\theta}_s$, the infinitesimal measure

$$\delta(\boldsymbol{\theta}_s, d\boldsymbol{\xi}_s) = \delta_{\boldsymbol{\theta}_s} \times \pi_1(\xi_s)\,d\xi_s \tag{17}$$

is the product of the point mass measure on $\boldsymbol{\theta}_s$ and the infinitesimal measure $\pi_1(\xi_s)\,d\xi_s$.

Based on the likelihood in (10), the complete hierarchical model specification can now be stated as follows:

$$\begin{aligned}
\boldsymbol{Y} \,|\, \boldsymbol{\beta},\, \boldsymbol{K},\, \boldsymbol{T},\, \boldsymbol{\sigma} &\sim \boldsymbol{f} && (18)\\
\boldsymbol{\theta}_s &\overset{iid}{\sim} F, \quad \text{and} && (19)\\
F &\sim DP(\alpha_0\, G_0). && (20)
\end{aligned}$$

Note that the set $(\boldsymbol{\beta},\, \boldsymbol{K},\, \boldsymbol{T},\, \boldsymbol{\sigma})$ is in one-to-one correspondence with $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_N, \boldsymbol{\xi})$ where $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_N)$.

3.3 Bayesian Inference Methodology

To infer $\boldsymbol{\theta}$s, the standard practice in DP posterior analysis is to integrate out F from the hierarchical specification of (18)-(20) (see, for example, Dey et al. (1998)). The likelihood corresponding to the observables $\boldsymbol{Y}$ in (18) is given by $\ell(\boldsymbol{Y} \,|\, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_N, \boldsymbol{\xi}) = \prod_{s=1}^{N}\prod_{l=1}^{k+1} f(\,\boldsymbol{Y}_{s,l} \,|\, \boldsymbol{\theta}_s,\, \xi_s)$ where the subscript s on k is suppressed. The conditional posterior distribution of the pair $(\boldsymbol{\theta}_s,\, \xi_s)$ given the other pairs $(\boldsymbol{\theta}_{-s},\, \boldsymbol{\xi}_{-s}\,)$ can be derived as

$$\begin{aligned}
(\boldsymbol{\theta}_s,\, \xi_s \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{\xi}_{-s}) &\propto \prod_{l=1}^{k+1} f(\boldsymbol{Y}_{s,l} \,|\, \theta_l,\, \xi_s)\left[\alpha_0\frac{\tilde{G}_0(d\boldsymbol{\theta}_s,\, d\xi_s)}{\alpha_0+N-1} + \frac{1}{\alpha_0+N-1}\sum_{s'\neq s}\delta(\boldsymbol{\theta}_{s'},\, d\boldsymbol{\xi}_{s'})\right],\\
&= \frac{q_{s,0}\tilde{G}_0^*(d\boldsymbol{\theta}_s,\, d\boldsymbol{\xi}_s) + \sum_{s'\neq s} q_{s,s'}\delta(\boldsymbol{\theta}_{s'},\, d\xi_{s'})}{q_{s,0} + \sum_{s'\neq s} q_{s,s'}}, \qquad (21)
\end{aligned}$$

where the second line is obtained from the first after normalization. The quantities $q_{s,0}$ and $q_{s,s'}$ in (21) have the expressions

$$q_{s,0} = \alpha_0 \int_{\mathcal{S}} \int_{\Xi} \prod_{l=1}^{k+1} f(\boldsymbol{Y}_{s,l} \,|\, \theta_l,\, \xi_s)\, \tilde{G}_0(d\boldsymbol{\theta}_s,\, d\boldsymbol{\xi}_s), \quad \text{and} \tag{22}$$

$$q_{s,s'} = \int_{\Xi} \prod_{l=1}^{k^*+1} f(\boldsymbol{Y}_{s,l} \,|\, \theta_l,\, \xi_s)\, \delta(\boldsymbol{\theta}_{s'}, d\boldsymbol{\xi}_{s'}), \tag{23}$$

where k^* is the number of change-points in $\boldsymbol{\theta}_{s'}$. The distribution

$$\tilde{G}_0^*(d\boldsymbol{\theta}_s, d\xi_s) = \frac{\alpha_0 \prod_{l=1}^{k+1} f(\boldsymbol{Y}_{s,l} \,|\, \theta_l,\, \xi_s)\, \tilde{G}_0(d\boldsymbol{\theta}_s, d\boldsymbol{\xi}_s)}{q_{s,0}}$$

is that of $(\boldsymbol{\theta}_s, \xi_s)$ when a new realization of $(\boldsymbol{\theta}_s, \xi_s)$ (i.e., not belonging to any of the previous clusters) has to be generated. An alternative way of writing (21) in terms of the distinct clusters is

$$(\,\boldsymbol{\theta}_s,\, \xi_s \,|\, \boldsymbol{\theta}_{-s},\, \boldsymbol{\xi}_{-s}\,) = \frac{q_{s,0}\tilde{G}_0^*(d\boldsymbol{\theta}_s, d\xi_s) + \sum_{j=1}^{N^*} N_j\, q_{s,j}\, \delta_{\boldsymbol{\theta}_j}}{q_{s,0} + \sum_{j=1}^{N^*} N_j\, q_{s,j}}, \tag{24}$$

where $\boldsymbol{\theta}_j$, $j = 1, 2, \cdots, N^*$ are the distinct change-point functions for the N^* different clusters, N_j is the number of sites s' for which $\boldsymbol{\theta}_{s'}$ is equal to $\boldsymbol{\theta}_j$, and $q_{s,j}$ is $q_{s,s'}$ in (23) with $\boldsymbol{\theta}_{s'}(t)$ replaced by $\boldsymbol{\theta}_j(t)$. Note that $\sum_{j=1}^{N^*} N_j = N - 1$ since the site s is left out.

Expression (24) explicitly demonstrates the clustering capability of DP. The current value of $\boldsymbol{\theta}_s$ can be selected to be one of the other $\boldsymbol{\theta}_{s'}$ with probability $\sum_{j=1}^{N*} N_j\, q_{s,j}/(q_{s,0} + \sum_{j=1}^{N*} N_j\, q_{s,j})$, this positive probability being the reason for possible clustering of sites in terms of $\boldsymbol{\theta}_s$. Expression (24) also allows for a new $\boldsymbol{\theta}_s^*$ to be generated from the posterior distribution G_0^*; this is the likely scenario if the temporal observations at site s, $\boldsymbol{W}_s = \{\, W_{st},\, t = U_0, U_0 + 1, \cdots, U_1 \,\}$, strongly support a different change-points function compared to the existing $\boldsymbol{\theta}_{s'}$ functions for $s' \neq s$. We note that the above treatment is similar to Gelfand et al. (2005) who extended θ_l to a realization of a random field by replacing it with a surface function on a spatial domain. However, Gelfand et al. (2005) do not consider joinpoint extensions as is done here; see also the related discussion in the Introduction.

The DP prior introduces two other hyper-parameters, namely α_0 and λ, into the inferential framework. In our analysis α_0 is fixed at a known value. We take the prior on λ to be π_2. The priors π_0, π_1 and π_2 are taken to be

$$\pi_0(\theta_l) \propto 1, \quad \pi_1(\sigma^2) = \texttt{igamma}(a_1, b_1) \quad \text{and} \quad \pi_2(\lambda) = \texttt{gamma}(a_2, b_2), \tag{25}$$

where `gamma` and `igamma` are the Gamma and inverse Gamma distributions with shape and scale parameters (a_1, b_1) and (a_2, b_2), respectively. The above choices are conjugate to their respective likelihoods enabling the posteriors to be obtained in closed forms. The reader is referred to the Discussion section of this paper for the motivation of using a flat prior for θ_l from the conjugacy

perspective. It turns out that using a common normal prior for θ_l does not allow the integrals in $q_{s,0}$ to be computed in closed form.

For a complete update of all the unknown parameters, it is convenient to append the current parameter space with additional variables arising in the definition of G_0. The N^* clusters are denoted by $\mathcal{C}_j$, $j = 1, 2, \cdots, N^*$. The additional variables (nested within each cluster $\mathcal{C}_j$) are L_s, K_j, $\boldsymbol{T}_j$ and $\boldsymbol{R}_j$. L_s is the cluster labels of each site with $L_s \in \{1, 2, \cdots, N^*\}$, K_j is the number of change-points for cluster j, $\boldsymbol{T}_j$ is the set of change-points in time such that $\boldsymbol{T}_j = (T_1, T_2, \cdots, T_k)$ when $K_j = k$, and $\boldsymbol{R}_j$ is the change-points levels such that $\boldsymbol{R}_j = (\theta_1, \theta_2, \cdots, \theta_{k+1})$; the subscript j is omitted from the ks, Ts, and θs to avoid notational complexity when there is no confusion. Ignoring the computational details for the moment, the updating steps below, at least in principle, are incorporated into the Gibbs sampler for full posterior inference. Each updating step implicitly assumes that all other parameters are given, and the update is performed based on the conditional distribution of the current parameter(s) given the rest. Section 4 and Appendix give explicit expressions for these conditional distributions as well as procedures for generating samples from them for specific choices of the likelihood and priors. The updating steps of the Gibbs sampler are:

(1) Update $(\boldsymbol{\theta}_s, \xi_s)$ (and simultaneously, L_s): The update of $(\boldsymbol{\theta}_s, \xi_s)$ is carried out via (24). First, a Bernoulli experiment, generating '0' and '1' with probabilities $p = q_{s,0}/(q_{s,0} + \sum_{j=1}^{N^*} N_j\, q_{s,j}$ and $1 - p$, respectively, is carried out. If '0' results, a new pair $(\boldsymbol{\theta}_s^*, \xi_s^*)$ is generated from $\tilde{G}_0^*$, N^* is increased to $N^* + 1$ and a new label is given to L_s. If '1' results, the existing cluster label j is sampled with probability $p_j = N_j\, q_{s,j} / \sum_{j=1}^{N^*} N_j\, q_{s,j}$, for $j = 1, 2, \cdots, N^*$. Subsequently, if j^* is sampled, $\boldsymbol{\theta}_s$ is set to $\boldsymbol{\theta}_{j^*}$, ξ_s is generated from the density

$$\frac{(\prod_{l=1}^{k^*+1} f(\boldsymbol{Y}_{s,l} \,|\, \theta_l^*, \xi_s))\, \delta(\boldsymbol{\theta}_{j^*}, d\xi_s)}{q_{s,j^*}},$$

and L_s is set to j^*; k^* and θ_l^*s are, respectively, the number of change points and mean levels corresponding to $\boldsymbol{\theta}_{j^*}$. N^* remains the same unless s was in a cluster with a singleton element in which case N^* changes to $N^* - 1$. This updating step is cycled once through all the N sites.

(2) Update K_j. Since $\boldsymbol{T}_j$ and $\boldsymbol{R}_j$ are nested within each K_j specification, this update really means updating all of $(K_j, \boldsymbol{T}_j, \boldsymbol{R}_j)$. This is an update conditional on all the site-specific variability parameters $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_N)$. We first update K_j from the posterior marginal of K_j, and then update $\boldsymbol{T}_j \,|\, K_j$, and finally $\boldsymbol{R}_j \,|\, \boldsymbol{T}_j$, K_j from their respective conditional distributions. The posterior marginal probability of $K_j = k$ is proportional to

$$e^{-\lambda}\, \frac{\lambda^k}{k!} \sum_{(n_1, n_2, \cdots, n_{k+1})} v(n_1, n_2, \cdots, n_k, n_{k+1}) \tag{26}$$

with

$$v(n_1, n_2, \cdots, n_{k+1}) = \exp\left\{ \sum_{l=1}^{k+1} \tilde{H}_l(n_l) \right\} \frac{\Gamma(n_0 + 1)}{\prod_{l=1}^{k+1} \Gamma(n_l + 1)}, \tag{27}$$

where $\tilde{H}_l(n_l)$ is defined as

$$\tilde{H}_l(n_l) \equiv \log\left[\int_{R^d} \prod_{s\in\mathcal{C}_j} f(\boldsymbol{Y}_{s,l}\,|\,\theta_l,\,\xi_s)\pi_0(\theta_l)\,d\theta_l\right]; \tag{28}$$

the summation in (26) is over all non-negative integers $n_1, n_2, \cdots, n_{k+1}$ such that $\sum_{l=1}^{k+1} n_l = n_0 \equiv U_1 - U_0 - (k+1)w$. Obtaining the posterior probability of $K_j = k$ requires evaluation of (27) for each value of $k \geq 0$. This could require significant amount of computational time and drastically reduce the efficiency of the Gibbs chain, but this did not occur for our application. The Appendix gives more details of these evaluation and generation steps. However, we note that depending on the choice of the likelihood $f(\cdot)$, the integration in (28) may not have a closed form. Then, alternative numerical integration methods, such as Laplace approximation, can be considered.

To update $\boldsymbol{T}_j$ given $K_j = k$, note that this is equivalent to updating $(n_1, \cdots, n_{k+1})$ with probabilities $p(n_1, \cdots, n_{k+1}) \propto v(n_1, n_2, \cdots, n_{k+1})$. This is carried out by exhaustively listing of all such combinations and numerically computing the corresponding probabilities. The update $\boldsymbol{R}_j$ given $\boldsymbol{T}_j$ and K_j is done based on the conditional distribution

$$(\boldsymbol{R}_j\,|\,\cdots) \propto \prod_{l=1}^{k+1}\left[\prod_{s\in\mathcal{C}_j} f(\boldsymbol{Y}_{s,l}\,|\,\theta_l,\,\xi_s)\pi_0(\theta_l)\right] \tag{29}$$

with the $k+1$ components of R_j generated independently of each other from their respective component densities $(\theta_l\,|\,\cdots) \propto \prod_{s\in\mathcal{C}_j} f(\boldsymbol{Y}_{s,l}\,|\,\theta_l,\xi_s)\pi_0(\theta_l)$.

(3) Update $\boldsymbol{\xi}$. This is carried out using the conditional distribution

$$(\xi_s\,|\,\cdots) \propto \prod_{l=1}^{k+1} f(\boldsymbol{Y}_{s,l}\,|\,\theta_l,\xi_s)\pi_1(\xi_s) \tag{30}$$

independently for each $s = 1, 2, \cdots, N$; in (30), k and θ_ls are the number of change-points and mean levels corresponding to cluster $\mathcal{C}_j$ to which site s belongs. Finally,

(4) Update λ using

$$\pi(\lambda\,|\,\cdots) \propto \left(\prod_{j=1}^{N^*} e^{-N_j^*\lambda}\frac{\lambda^{N_j^* k_j}}{(k_j!)^{N_j^*}}\right)\pi_2(\lambda), \tag{31}$$

where k_j is the number of change-points corresponding to $\boldsymbol{\theta}_j$ in cluster $\mathcal{C}_j$, N_j^* is the number of sites in $\mathcal{C}_j$ for $j = 1, 2, \cdots, N^*$.

3.4 Inference based on Posterior Samples

After convergence is established, we take B samples from the posterior distribution to make inference on all unknown quantities. Let $\mathcal{X}_b^*$, $b = 1, 2, \cdots, B$ be B samples of the posterior obtained from the Gibbs sampler. Components of $\mathcal{X}_b^*$ include N realizations of step functions θ_s and $\boldsymbol{\xi}$ (or equivalently, $\boldsymbol{\beta}, \boldsymbol{K}, \boldsymbol{T}, \boldsymbol{\sigma}$). Thus, marginal posterior inference can be carried out for each of these

components. For example, to infer $\boldsymbol{\theta}_s(t)$ for a particular site s and time point t, we extract all $\boldsymbol{\theta}_s(t)$ components from each $\mathcal{X}_b^*$, $b = 1, 2, \cdots, B$. The B realizations of $\boldsymbol{\theta}_s(t)$ are then used to compute the posterior mean, variance and confidence interval. A similar procedure also works for N^* where we can obtain marginal probabilities of $N^* = n^*$ for all non-negative integers n^*. Results for simulation experiments and real data are given in the subsequent sections.

A more challenging inference problem is to obtain results for the clustering tendencies, for example, the "average" clusters. Note that the output of the Gibbs sampler at each iteration is a clustering of the N states, and therefore, it is difficult to obtain a summary posterior measure, such as mean and variance, for the clustering of sites. To get some idea about average clustering tendencies reflected by the posterior distribution, the following methodology is developed: For every pair of sites (s_1, s_2) in $\{1, 2, \cdots, N\}$, define $D_b(s_1, s_2) = 1$ if s_1 and s_2 belong to the same cluster in $\mathcal{X}_b^*$, and 0, otherwise, for $b = 1, 2, \cdots, B$. Subsequently, we construct the average distance measure between the sites s_1 and s_2 using

$$dist(s_1, s_2) = 1 - \bar{D}(s_1, s_2)$$

where $\bar{D}(s_1, s_2) = \sum_{b=1}^{B} D_b(s_1, s_2)/B$. Based on $dist$, an agglomerative clustering algorithm is performed with the maximum number of clusters threshold in the algorithm fixed at the value of N^* for which the posterior probability has the maximum value. The clustering outputs from this procedure match with our expected scenario. Subsequent sections give results based on real and validation data.

4 Analysis of Cancer Incidence Rates Revisited

We consider two specific choices of models. The site-specific variability model is given by Model 1 below. In Model 2, we assume that σ^2 is cluster-dependent (not site-specific), that is, σ^2 is same over all states in the same cluster (but different for the different clusters). Based on previous discussion, we can write these two models as follows:
Model 1: $\boldsymbol{Y}_{sl} = (\widehat{\beta}_l, RSS_l)^T$, $\theta_l = \beta_l$, $\xi_s = \sigma^2$
Model 2: $\boldsymbol{Y}_{sl} = (\widehat{\beta}_l, RSS_l)^T$, $\theta_l = (\beta_l, \sigma_l^2)$,
suppressing the subscript s on $\widehat{\beta}_l$ and RSS_l.

Note that Model 2 is not a subset of Model 1 or vice versa. In Model 2, σ_l^2 is common to all sites within a cluster but can vary for the different time intervals $[T_{l-1}, T_l)$. In Model 1, one common σ_s^2 is assumed for each site which does not change within each time segment.

The Appendix gives the model specific expressions used for the Bayesian inference. We run three Gibbs chains for 10,000 iterations. The convergence is established after 5,000 iterations and we take 2,000 samples from each chain after convergence so that total 6,000 samples are used for further posterior analysis. Specific values of hyper-parameters are set to $a_1 = b_1 = 1$ for π_1, $a_2 = b_2 = 1$ for π_2. α_0 is set to 1/100. The number of clusters of states based on the highest posterior probability

Number of clusters	4	5	6	7	8
Posterior Prob	0.0688	0.5153	0.3820	0.0335	0.0003

Table 1:
Posterior probabilities of number of clusters for Model 1

Change-Points	Arizona	Florida	Indiana	Missouri
No Change-Points	0.0293	0.0042	0	0
$T_1 = 1994$	0	0	0	0.0010
$T_1 = 1993$	0.0010	0.0002	0.0002	0.0002
$T_1 = 1992$	0.0147	0.0355	0.0355	0.0005
$T_1 = 1991$	0.0657	0.1723	0.1723	0.0028
$T_1 = 1990$	0.0768	0.1533	0.1533	0.0172
$T_1 = 1989$	0.5807	0.6033	0.3892	0.0200
$T_1 = 1988$	0.0662	0.0258	0.3382	0.0975
$T_1 = 1987$	0.0432	0.0007	0.0048	0.7212
$T_1 = 1986$	0.0042	0	0	0.1285
$T_1 = 1985$	0.0010	0	0	0.0032
Two Change-Points	0.1033	0	0.0113	0.0025
Three Change-Points	0.0118	0.0045	0.0172	0.0055

Table 2:
Posterior probabilities of a change-point for Model 1

is found to be $N^* = 5$; see Table 1 for posterior probabilities. Using the posterior estimate of N^*, we use the clustering methodology explained in section 3.4 to cluster states into 5 groups.

As mentioned in Introduction, we expect Florida and Arizona to belong to the same cluster while Indiana and Missouri to belong in another. This is what is revealed from the analysis. Marginal posterior analysis on the number of change-points for each state revealed that one change-point corresponds to the highest probability. Further, the posterior probabilities of the time intervals corresponding to no change-point and a single change-point are given in Table 2 for each of the four states. The entries in Table 2 is the marginal posterior probabilities corresponding to the most significant partitions of the interval $[1969, 2006]$ based on output of the Gibbs sampler. Note that both Arizona and Florida showed one change-point, $T_1 = 1989$ while for Missouri and Indiana, the change-point was $T_1 = 1987$. Corresponding to these change-points, the mean posterior estimates of σ_s (site-wise) and β_l (cluster-wise) is given in Table 3.

Next, we demonstrate the superiority of Model 1 over Model 2 based on predictive analysis. A new realization of W_{st}, W^*_{st}, is obtained by sampling from the normal distribution with mean $\alpha_l + \beta_l\, t$ and variance σ^2_s where β_l and σ^2_s are posterior realizations from the Gibbs chain and α_l is given from the data for the corresponding time interval and site. The B values of W^*_{st} are then used

State	σ_s	Change-Point(s)	(β_1, β_2)
Arizona	0.0416	$T_1 = 1989$	$(0.0209, -0.0107)$
Florida	0.0182	$T_1 = 1989$	$(0.0196, -0.0118)$
Indiana	0.0297	$T_1 = 1987$	$(0.0297, 0.00008)$
Missouri	0.0292	$T_1 = 1987$	$(0.0295, 0.00006)$

Table 3:
Posterior outputs for Model 1

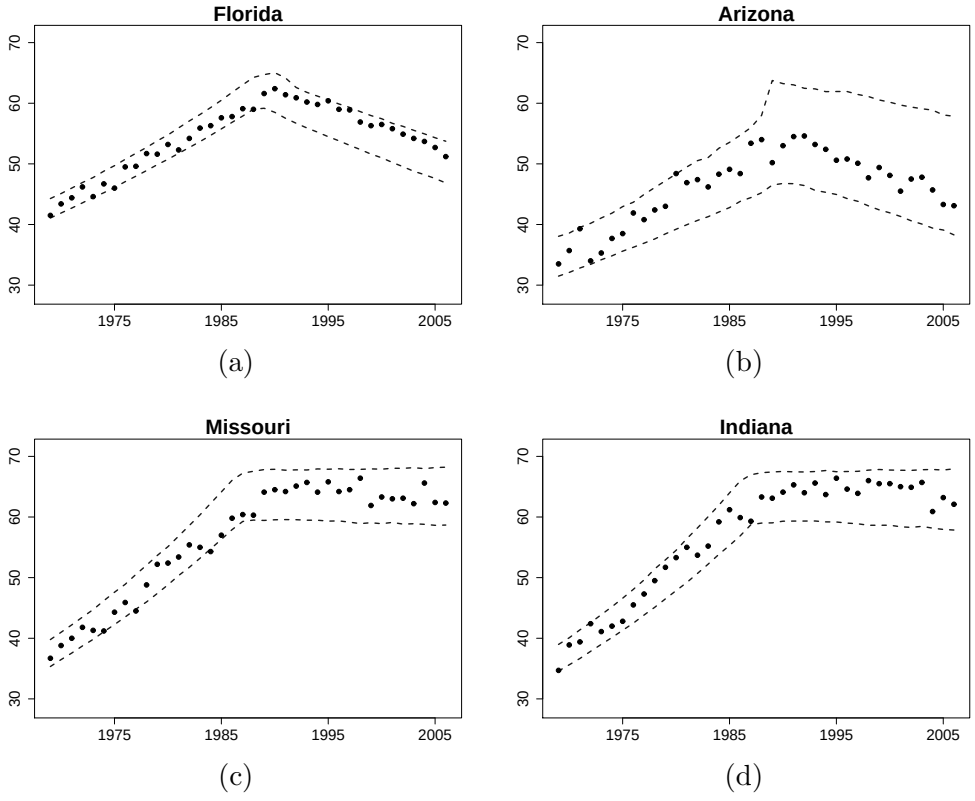

Figure 2: Examples of states belonging to different clusters from the implementation of the change-point methodology. The bands around the observed values (age-adjusted cancer rates) are the 95% predictive credible intervals based on Model 1.

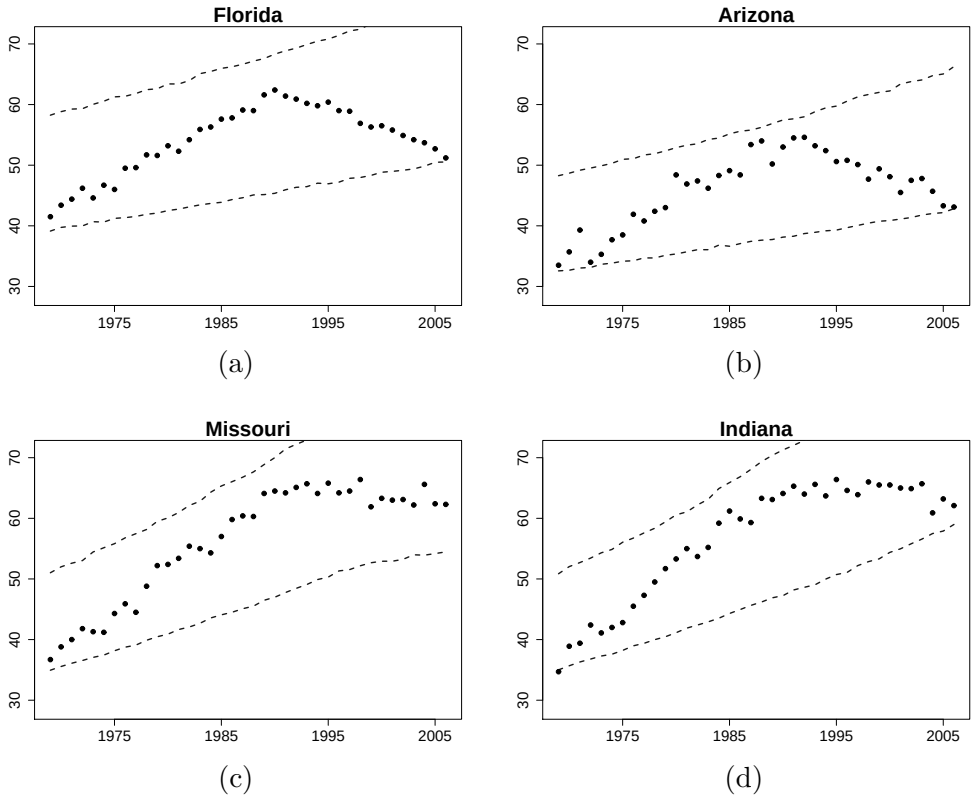

Figure 3: The bands around the observed values (age-adjusted cancer rates) are the 95% predictive credible intervals based on Model 2.

to construct the 95% credible predictive interval. The confidence bands generated are shown in Figure 2 in the original scale. A similar procedure is repeated for Model 2 to obtain the confidence bands shown in Figure 3. The better model will be the one that detects at least one change-point and that gives narrower confidence bands. Note that change-points are not detected and the width of the predictive confidence bands are too large for Model 2. These results indicate that there is significant evidence from the data to suggest heterogeneous (i.e., site-specific) variability around the mean within clusters.

Next, the cluster with the highest drop in cancer incidence rate is identified. The difference $\beta_2 - \beta_1$ in Table 4 is computed using posterior samples for each of the 5 clusters based on Model 1. Table 4 also gives the corresponding 95% credible intervals of the 5 clusters for $\beta_2 - \beta_1$. Note that the top cluster has a drop in rates that is significantly different from clusters 2, 4 and 5. States in this cluster consists of Colorado, Georgia, Oregon and Virginia. One subsequent investigation may, therefore, be to identify the underlying reasons for the highest drop in cancer rates, and to identity and implement effective policies or programs in these states to the other states in the nation.

Cluster	Posterior Mean of $\beta_2 - \beta_1$	95% Credible Interval
1	-0.0370	$(-0.0397, -0.0336)$
2	-0.0318	$(-0.0341, -0.0298)$
3	-0.0316	$(-0.0385, -0.0294)$
4	-0.0314	$(-0.0337, -0.0298)$
5	-0.0307	$(-0.0329, -0.0291)$

Table 4:
Clusters (from the agglomerative procedure) with the highest drop in cancer incidence rates measured in terms of $\beta_2 - \beta_1$.

5 Validation Results

To validate our data analysis, a simulation experiment is carried out with a total of $N = 100$ sites on a 10×10 lattice, $\mathcal{L} = \{\,(r,c) \,:\, 1 \le r, c \le 10,\, r, c \text{ integers}\}$. The lattice is partitioned into 4 sub-regions, that is, $\mathcal{L} = \cup_{j=1}^{4} \mathcal{L}_j$, where $\mathcal{L}_j$, for $j = 1, 2, 3$ and 4 represent the true clusters. The sub-regions $\{\,\mathcal{L}_j\,\}_{j=1}^{4}$ are given as follows: $\mathcal{L}_1 = \{\,(r,c) \,:\, 1 \le r \le 7 \text{ and } 1 \le c \le 7\,\}$, $\mathcal{L}_2 = \{\,(r,c) \,:\, 1 \le r \le 7 \text{ and } 8 \le c \le 10\,\}$, $\mathcal{L}_3 = \{\,(r,c) \,:\, 8 \le r \le 10 \text{ and } 1 \le c \le 7\,\}$ and $\mathcal{L}_4 = \{\,(r,c) \,:\, 8 \le r \le 10 \text{ and } 8 \le c \le 10\,\}$. It follows that the sub-regions are rectangular with $49, 21, 21$ and 9 sites, respectively, corresponding to $j = 1, 2, 3$ and 4. The number of time points taken is 20 with $U_0 = 1$ and $U_1 = 20$. The functional form of $\boldsymbol{\theta}_s$ for $s \in \mathcal{L}_j$ is taken to be a step function

$$\boldsymbol{\theta}_s(t) = \theta_{j\,l}$$

for the l-th time subinterval, $l = 1, 2, \cdots, K_j + 1$; recall that K_j denotes the total number of change-points (corresponding to $K_j + 1$ change-point time intervals) in $\mathcal{L}_j$. The following choices are made corresponding to each $\mathcal{L}_j$. For $\mathcal{L}_1$, $K_1 = 2$, $\theta_{1,1} = 20$, $\theta_{2,1} = 15$ and $\theta_{3,1} = 10$. The corresponding change-points are $T_1 = 6, T_2 = 14$. For $\mathcal{L}_2$, $K_2 = 2$, $\theta_{1,2} = 10.03$, $\theta_{2,2} = 20.02$ and $\theta_{3,2} = 30.05$. The corresponding change-points are $T_1 = 6, T_2 = 14$. For $\mathcal{L}_3$, $K_3 = 1$, $\theta_{1,3} = 10.02$, $\theta_{2,3} = 25.04$ and the corresponding change-point is $T_1 = 8$. Finally, for $\mathcal{L}_4$, $K_4 = 1$, $\theta_{1,4} = 10$, $\theta_{2,4} = 25$ and the corresponding change-point is $T_1 = 8$. The site-wise variance parameter is common for all the sites in $\mathcal{L}$ and taken to be $\sigma^2 = 2^2$. In each site s and time t, we generate a data from $\mathcal{N}(\theta_s(t), \sigma^2)$ independently. With this rather artificial specification of $\theta_s(t)$ and σ^2, we expect that $\mathcal{L}_3$ and $\mathcal{L}_4$ are merged into one cluster.

Specific values of hyper-parameters for π_2 are set to $a_2 = b_2 = 1$ and α_0 is set to $1/100$. Three Gibbs chains are started from initial estimates of clusters that represent over-dispersion. Our choice of the monitoring statistic is $\theta_s(t)$ of some fixed sites and time. $\theta_s(t)$ retains the same interpretation across different realizations of k_j and $\theta_{j\,l}$. The assessment of convergence is carried out based on the methodology of Gelman and Rubin and convergence is achieved after $2,000$ iterations. On a computer with Intel Core i7 CPU with 2.93GHz with 4GB RAM, each $1,000$ iterations of the Gibbs

site	time	mean	standard error	95% Credible interval
(1,1)	4	20.0079	0.1068	(19.8091, 20.2170)
	10	15.0418	0.0632	(14.9206, 15.1638)
	18	10.0893	0.0703	(9.9523, 10.2241)
(9,2)	4	9.9038	0.1845	(9.5570, 10.2166)
	10	24.9587	0.0792	(24.8198, 25.1135)
	18	24.9587	0.0792	(24.8198, 25.1135)
(8,8)	4	9.9038	0.1845	(9.5570, 10.2166)
	10	24.9587	0.0792	(24.8198, 25.1135)
	18	24.9587	0.0792	(24.8198, 25.1135)

Table 5:
Posterior estimates of $\theta_s(t)$ for the simulated data

chain took about 73 minutes.

Table 5 gives posterior means, posterior standard errors (square root of posterior variance) and 95% credible intervals for three example sites $(1,1), (9,2)$ and $(8,8)$ and the time points $t = 4, 10$ and 18 to demonstrate the methodology. Initially, $(1,1)$ is in $\mathcal{L}_1$, (9,2) is in $\mathcal{L}_3$ and (8,8) is in $\mathcal{L}_4$. Note that all true values lie inside their respective credible intervals and $\theta_s(t)$ values for site (9,2) and (8,8) are same since they became in the same cluster.

6 Discussion

In this paper, we propose change-point models for spatio-temporal data that can detect change-points over time and group spatial sites into several clusters with respect to their change-point functions. Clustering is achieved by using a Dirichlet process prior on the space of step functions over time. The model was developed to analyze state-wise age adjusted rates to find local change-points and clusters that have similar changes.

Our analysis based on predictive distribution demonstrate that Model 1 is superior to Model 2. Thus, model flexibility is achieved far more by incorporating site specific parameters which are nuisance to the clustering compared to adding extra parameters for clustering. The latter action may in fact distort true underlying trends as evidenced by Figure 3. Model 2 was our initial extension since all the expressions for $q_{s,0}$, $q_{s,j}$, $\tilde{H}_l(n_l)$ and $H_l(n_l)$ (see Appendix) could be obtained in closed form based on conjugate prior specifications. This was not the case for Model 1 and thus, two separate conditional steps (namely, (29) and (30)) had to be introduced for Model 1. It was fortunate that in the case of a singleton cluster, namely, site s (again due to conjugate prior specification), the integration with respect to θ_ls and ξ_s was obtainable in closed form. Generating a new realization from $q_{s,0}$ necessitated the marginal distribution of ξ_s be obtained even if numerically. The use of the flat prior for θ_l (or, β_ls) in (25) was motivated from this perspective. In the case

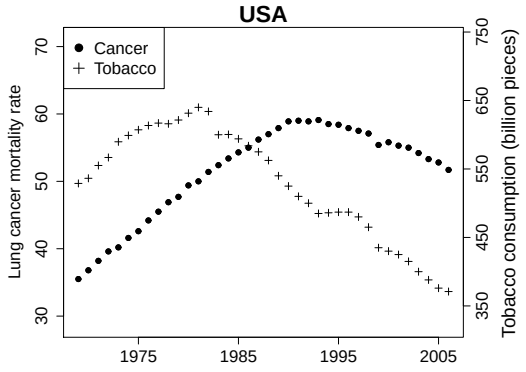

Figure 4: Age-adjusted lung cancer mortality rates and Tobacco consumption (billion pieces) in the US from 1969 to 2006. Tobacco consumption data is from Tobacco Yearbook, The US Department of Agriculture.

of non-conjugacy, we expect the Bayesian computations to be more involved and time consuming. The Laplace approximation of the integration is a possible alternative to avoiding time consuming sampling approaches such as the (multidimensional) griddy grid. However, we expect the Laplace approximation to work well when the j-th cluster $\mathcal{C}_j$ consists of a moderate to large number of sites, but not for singleton clusters. To investigate this issue for more general non-conjugate likelihood and priors is one avenue for future research.

For the real application, we find that state-level and national level age-adjusted lung cancer mortality rates show a clear change-point around late 1980s to early 1990s. Some states like Florida and Arizona follow similar patterns as national level rates while some states like Missouri and Indiana show different patterns from the national level rates (see Figure 1). In particular, Missouri and Indiana have smaller rate of changes after the change-point compared to Florida and Arizona as well as national level (see Table 3). Indeed, we can argue that lung cancer mortality rates have not changed much after 1990s for these states, while the national level seems to significantly decrease. This further indicates that we need different attention on each individual state. Another avenue for future research is to incorporate covariate information into the clustering mechanism. For example, tobacco consumption is related to lung cancer mortality rates with a certain (possibly heterogeneous) time lag. Tobacco consumption has decreased starting the early 1980s while age-adjusted lung cancer mortality rates have decreased starting from the early 1990s (see Figure 4). We intend to develop models to determine if this national level comparison still holds at the state or cluster levels. The analysis may reveal different relationships between tobacco consumption and lung cancer mortality rates in different sub-groups, or it may provide overall proof of the tobacco-cancer relationship at the national level.

References

[1] American Cancer Society (2010) "Cancer Facts & Figures 2010", American Cancer Society, Atlanta, GA.

[2] Antoniak,C. (1974) "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems", *Ann. Statist.*, vol. **2**, pp 1152-1174.

[3] Banerjee S., Gelfand, A. E., and Carlin, B. P. (2004), Hierarchical Modeling and Analysis for Spatial Data, Chapman & Hall/CRC.

[4] Carlin, B. P., Gelfand, A. E. and Smith, A. F. M. (1992), "Hierarchical Bayesian Analysis of Changepoint Problems," Applied Statistics, 41, 389-405.

[5] Clegg, L. X., Hankey, B. F. , Tiwari, R., Feuer, E. J. and Edwards, B. K., "Estimating average annual per cent change in trend analyis", Statist. Med. 2009, 28, pp. 3670–3682.

[6] Elliott,P., Wakefield, J., Best, N., and Briggs, D. (Eds.) (2000), "Spatial epidemiology: methods and applications," Publisher: Oxford University Press, Oxford.

[7] Escobar, M. D. and West, M. (1998), "Computing Nonparametric Hierarchical Models," in Practical Nonparametric and Semiparametric Bayesian Statistics, eds. D. Dey, P. Muller, D. Sinha.

[8] Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," Ann. Statist., vol. 1, pp 209-230.

[9] Ferguson, T. (1974), "Prior distributions on spaces of probability measures," Ann. Statist., vol 2, pp 615-629.

[10] Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," J. Amer. Statist. Assoc, 100, 1021-1035.

[11] Ghosh, P., Basu, S. and Tiwari, R. C. (2009), "Bayesian Analsysis of Cancer Rates from SEER Program using Parametric and Semiparametric Jointpoint Regression Models," J. Amer. Statist. Assoc., 104, 439-452.

[12] Ghosh, P. and Kaushik, G. and Tiwari, R. (2011), "Bayesian approach to cancer-trend analysis using age-stratified Poisson regression models", Statist. Med., 30, pp. 127–139.

[13] Kim, H. J., Fay, M. P., Feuer, E. J., and Midthune, D. N. (2000), "Permutation Tests for Joinpoint Regression with Applications to Cancer Rates," Statistics in Medicine, 19, 335-351.

[14] Kim, H. J., Fay, M. P., Yu, B., Barrett, M. J., and Feuer, E. J. (2004), "Comparability of Segmented Line Regression Models," Biometrics, 60, 1005-1014.

[15] Lawson, A. B. (2009) Bayesian disease mapping: hierarchical modeling in spatial epidemiology, Boca Raton:CRC Press

[16] MacEachern, S. N. (1998), "Computational Methods for Mixture of Dirichlet Process Models," in Practical Nonparametric and Semiparametric Bayesian Statistics, eds. D. Dey, P. Muller, D. Sinha.

[17] MacEachern, S. N. and Muller, P. (2000), "Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models," in Robust Bayesian Analysis, eds. D. Rios Insua and F. Ruggeri, 295-315, Springer-Verlag.

[18] Ries, L. A. G., Eisner, M. P., Kosary, C. L., Hanley, B. F., Miller, B. A., Clegg, L., and Edwards, B. K. (2002), SEER Cancer Statistics Review, National Cancer Institute, Bethesda, MD, available at `http://seer.cancer.gov/csr/1973-1999`.

[19] Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," Statistica Sinica, 4, 639-650.

[20] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database, "Mortality - All COD, Aggregated With State, Total U.S. (1969-2007) <Katrina/Rita Population Adjustment>", National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released June 2010. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

[21] Tiwari, R. C., Cronin, K. A., Davis, W., Feuer, E. J., Yu, B., and Chib, S. (2005), "Bayesian Model Selection for Joinpoint Regression with Application to Age-adjusted Cancer Rates," J. R. Statist. Soc. Ser. C, 54, 919-939.

Appendix A: Calculations for Model 1

We first consider the update of $(\boldsymbol{\theta}_s, \xi_s \,|\, \cdots)$ according to (24). The expression for $q_{s,0}$ in (24) and (22) can be obtained as

$$q_{s,0} \;=\; \alpha_0 \sum_{k=0}^{\infty} \sum_{(n_1,\cdots,n_{k+1})} \left(\prod_{l=1}^{k+1} \exp\{\, H_l(n_l)\} \right) \frac{n_0!}{n_1! \cdots n_{k+1}!} \left(\frac{1}{k+1} \right)^{n_0} P(K=k), \qquad (32)$$

where $H_l(n_l)$ is given by

$$H_l(n_l) = \log \left\{ \int_0^{\infty} \int_{-\infty}^{\infty} f(\boldsymbol{Y}_{sl} \,|\, \beta_l, \sigma_s^2) \pi_0(\beta_l) \pi_1(\sigma_s^2)\, d\beta_l\, d\sigma_s^2 \right\}. \qquad (33)$$

Under Model 1 and prior choices given in Section 4, $H_l(n_l)$ has a closed form expression, namely,

$$\begin{aligned} H_l(n_l) \;=\; & \left(\frac{m_l}{2} - 1\right) \log RSS_{sl} + \log \Gamma\left(\frac{m_l}{2} + a_1\right) - \left(\frac{m_l}{2} + a_1\right) \log \left(\frac{RSS_{sl}}{2} + b_1^{-1}\right) \\ & - \frac{m_l}{2} \log 2 - \log \Gamma\left(\frac{m_l}{2}\right) - \log \Gamma\left(a_1\right) - a_1 \log b_1, \end{aligned}$$

where RSS_{sl} is the site-specific residual sum of squares on the time interval $[T_{l-1}, T_l)$ and $m_l = T_l - T_{l-1} - 2$. The outer two sums in (32) are numerically evaluated based on an exhaustive listing of $(n_1, n_2, \cdots, n_{k+1})$ given k, and then summed over $k \geq 0$. For a new realization from the first component in (24), the generation procedure is according to equations (26-30) but with two distinct differences: First, cluster $\mathcal{C}_j$ is taken to be the singleton site $\{s\}$ and second, ξ_s is also integrated out in the expression of $\tilde{H}_l(n_l)$ in (28); thus, $\tilde{H}_l(n_l)$ is replaced by $H_l(n_l)$ above. The number of change points k is generated from (26) based on exhaustive numerical tabulation and summation over different combinations of $(n_1, n_2, \cdots, n_k)$. Given k, $(n_1, n_2, \cdots, n_k)$ is generated from (27) based on the stored values of $v(n_1, n_2, \cdots, n_k)$. To generate $(\beta_1, \beta_2, \cdots, \beta_{k+1}, \sigma_s^2)$, based on (29) and (30), we note that two separate conditional updates are unnecessary. The marginal of σ_s^2 can be explicitly determined by integrating out $(\beta_1, \cdots, \beta_{k+1})$ giving $\pi(\sigma_s^2) \sim \texttt{igamma}(a_0, b_0)$, where $a_0 = \sum_{l=1}^{k+1} \frac{m_l}{2} + a_1$ and $b_0 = \left(\sum_l \frac{RSS_{sl}}{2} + b^{-1}\right)^{-1}$. Thus, we generate σ_s^2 from $\texttt{igamma}(a_0, b_0)$ distribution first, and then given σ_s^2, generate $\beta_l \sim N(\widehat{\beta}_l, v_l \sigma_s^2)$ independently for $l = 1, 2, \cdots, k+1$. This simplification is not available for a non-singleton cluster of sites $\mathcal{C}_j$ and in this case, the two separate conditional updates are needed.

In case one of the other $q_{s,j}$ components in (24) is selected, the site s is included into the jth cluster and we generate the site-specific variability parameter σ_s^2 from the density $\pi(\sigma_s^2) \propto \left(\prod_{l=1}^{k+1} f(\boldsymbol{Y}_{sl} \,|\, \beta_l,\, \sigma_s^2)\right) \pi_1(\sigma_s^2)$ (for fixed β_ls) which can be seen to be the $\texttt{igamma}(a_0, b_0)$ distribution with $a_0 = \sum_l \frac{m_l+1}{2} + a_1$ and $b_0 = \left(\sum_{l=1}^{k+1} \frac{1}{2v_l}(\widehat{\beta}_l - \beta_l)^2 + \sum_l \frac{RSS_{sl}}{2} + b_1^{-1}\right)^{-1}$. The analytic expression of $q_{s,j}$ for the jth cluster is

$$\begin{aligned}
\log q_{s,j} &= \sum_{l=1}^{k^*+1} \left(\left(\frac{m_l}{2} - 1\right) \log RSS_{s,l} - \frac{1}{2}\log(2\pi) - \frac{1}{2}\log v_l - \frac{m_l}{2}\log 2 - \log\Gamma\left(\frac{m_l}{2}\right)\right) \\
&+ \log\Gamma\left(\sum_l \frac{m_l+1}{2} + a_1\right) - \left(\sum_l \frac{m_l+1}{2} + a_1\right) \log\left(\sum_l \frac{RSS_{s,l}}{2} + \frac{1}{2}\sum_l \frac{(\widehat{\beta}_l - \beta_l)^2}{v_l} + b_1^{-1}\right) \\
&- \log\Gamma(a_1) - a_1 \log b_1,
\end{aligned}$$

where k^* is the number of change points in the jth cluster; v_l, $RSS_{s,l}$ and $\widehat{\beta}_l$ are obtained from the observations based on the site s and the time interval $[T_{l-1}, T_l)$ while β_l is from the jth cluster information.

We now give the details of updating $(K_j, \boldsymbol{T}_j, \boldsymbol{R}_j)$ in the jth cluster given $\boldsymbol{\xi}$. We have the following analytic expression for $\tilde{H}_l(n_l)$ in (28):

$$\begin{aligned}
\tilde{H}_l(n_l) &= -\frac{N_j^*}{2}\log(2\pi v_l) - \frac{1}{2}\sum_s \log\sigma_s^2 + \frac{1}{2}\log 2\pi - \frac{1}{2}\log\left(\sum_s (v_l\sigma_s^2)^{-1}\right) + \frac{1}{2}\frac{\left(\sum_s \frac{\widehat{\beta}_{s,l}}{v_l\sigma_s^2}\right)^2}{\sum_s (v_l\,\sigma_s^2)^{-1}} \\
&- \frac{1}{2}\sum_s \left(\frac{\widehat{\beta}_{s,l}^2}{v_l\sigma_s^2}\right) + \sum_s \log f_0(RSS_{s,l}|\sigma_s^2).
\end{aligned}$$

The number of change points, k, and the time widths $(n_1, n_2, \cdots, n_k)$ are generated as before

based on exhaustive enumeration. The generation of β_l is carried out independently for each $l = 1, 2, \cdots, k+1$ using $(\beta_l | \cdots) \sim N(\mu_{\beta_l}, \nu^2_{\beta_l})$, where $\nu^2_{\beta_l} = \left(\sum_s (v_l \sigma^2_s)^{-1}\right)^{-1}$, and $\mu_{\beta_l} = \nu^2_{\beta_1} \sum_s \frac{\widehat{\beta}_{s,l}}{v_l \sigma^2_s}$. Given β_ls, the site specific parameters σ^2_s are updated independently for each $s \in \mathcal{C}_j$ based on the conditional distribution $(\sigma^2_s | \cdots) \propto \left(\prod_{l=1}^{k+1} f(\boldsymbol{Y}_{s,l} \,|\, \beta_l,\, \sigma^2_s)\right) \pi_1(\sigma^2_s)$ which is $\texttt{igamma}(a_0, b_0)$ with $a_0 = \sum_l \frac{m_l+1}{2} + a_1$ and $b_0 = \left(\sum_{l=1}^{k+1} \frac{1}{2v_l}(\widehat{\beta}_{s,l} - \beta_l)^2 + \sum_l \frac{RSS_{s,l}}{2} + b_1^{-1}\right)^{-1}$.

Appendix B: Calculations for Model 2

For Model 2, there are $k+1$ cluster specific variability parameters σ^2_l, $l = 1, 2, \cdots, k+1$. The expression for $q_{s,0}$ in (22) can be re-written as

$$q_{s,0} \;=\; \alpha \sum_{k=0}^{\infty} \sum_{(n_1, \cdots, n_{k+1})} \left(\prod_{l=1}^{k+1} \exp\{\, H_l(n_l)\}\right) \frac{n_0!}{n_1! \cdots n_{k+1}!} \left(\frac{1}{k+1}\right)^{n_0} P(K = k), \qquad (34)$$

where $H_l(n_l)$ is given by

$$H_l(n_l) = \log\left\{\int_0^{\infty} \int_{-\infty}^{\infty} f(\boldsymbol{Y}_{s,l} \,|\, \beta_l, \sigma^2_l) \pi_0(\beta_l) \pi_1(\sigma^2_l)\, d\beta_l\, d\sigma^2_l\right\}. \qquad (35)$$

Under Model 2 and previously mentioned prior choices, the integrals with respect to β_l and σ^2_l can be evaluated in closed form giving

$$\begin{aligned} H_l(n_l) \;=\; & \left(\frac{m_l}{2} - 1\right) \log RSS_s + \log\Gamma\left(\frac{m_l}{2} + a_1\right) - \left(\frac{m_l}{2} + a_1\right) \log\left(\frac{RSS_s}{2} + b_1^{-1}\right) \\ & - \frac{m_l}{2} \log 2 - \log\Gamma\left(\frac{m_l}{2}\right) - \log\Gamma\left(a_1\right) - a_1 \log b_1. \end{aligned}$$

By integrating out β_l, we generate σ^2_ls independently from their marginal distributions: $\pi(\sigma^2_l) \sim \texttt{igamma}(a_0, b_0)$, where $a_0 = \frac{m_l}{2} + a_1$ and $b_0 = \left(\frac{RSS_{sl}}{2} + b^{-1}\right)^{-1}$. After generating σ^2_l, generate $\beta_l \sim N(\widehat{\beta}_l, v_l \sigma^2_l)$ independently for $l = 1, 2, \cdots, k+1$.

The analytic expression of $q_{s,j}$ for the jth cluster is given by

$$\begin{aligned} \log q_{s,j} \;=\; & -\frac{1}{2} \sum_{l=1}^{k^*+1} \log(2\pi v_l) - \frac{3}{2} \sum_{l=1}^{k^*+1} \log \sigma^2_l - \frac{1}{2} \sum_{l=1}^{k^*+1} \frac{(\widehat{\beta}_l - \beta_l)^2}{v_l \sigma^2_l} - \left(\sum_{l=1}^{k^*+1} \frac{m_l}{2}\right) \log 2 \\ & - \sum_{l=1}^{k^*+1} \log\Gamma\left(\frac{m_l}{2}\right) + \sum_{l=1}^{k^*+1} \left(\frac{m_l}{2} - 1\right) \log\left(\frac{RSS_{s,l}}{\sigma^2_l}\right) - \sum_{l=1}^{k^*+1} \frac{RSS_{s,l}}{2\sigma^2_l} \end{aligned}$$

where k^* is the number of change points in the jth cluster; as before, v_l, $RSS_{s,l}$ and $\widehat{\beta}_l$ are obtained from the observations based on the site s and the jth cluster information.

For updating $(K_j, \boldsymbol{T}_j, \boldsymbol{R}_j)$ in the case of Model 2, we can avoid the two separate conditional steps since β_ls can be integrated out to give closed form expressions for the marginal of σ^2_l, for

$l = 1, 2, \cdots, k+1$. The expression for $\tilde{H}_l(n_l)$ is

$$\begin{aligned}\tilde{H}_l(n_l) &= \left(\frac{m_l}{2} - 1\right)\sum_s \log RSS_s - \left(\frac{N_j^* - 1}{2}\right)\log(2\pi v_l) - N_j^*\frac{m_l}{2}\log 2 - N_j^*\log\Gamma\left(\frac{m_l}{2}\right) \\ &\quad - \log\Gamma(a_1) - a_1\log b_1 - \frac{1}{2}\log N_j^* + \log\Gamma\left(N_j^*\frac{m_l+1}{2} + a_1 - \frac{1}{2}\right) \\ &\quad - \left(N_j^*\frac{m_l+1}{2} + a_1 - \frac{1}{2}\right)\log\left(b_1^{-1} + \frac{1}{2}\sum_s RSS_s - \frac{(\sum_s \widehat{\beta}_{s,l})^2}{2v_l N_j^*} + \frac{\sum_s(\widehat{\beta}_{s,l})^2}{2v_l}\right).\end{aligned}$$

Each σ_l^2 is updated independently based on the marginal distribution $(\sigma_l^2 \mid \cdots) \sim \texttt{igamma}(a_0, b_0)$ where $a_0 = N_j^*\frac{m_l+1}{2} + a_1 - \frac{1}{2}$ and $b_0 = \left(b_1^{-1} + \frac{1}{2}\sum_s RSS_s - \frac{(\sum_s \widehat{\beta}_{s,l})^2}{2v_l N_j^*} + \frac{\sum_s(\widehat{\beta}_{s,l})^2}{2v_l}\right)^{-1}$. Then, β_ls are updated independently based on $\beta_l \sim N(\mu_{\beta_l}, \nu_{\beta_l}^2)$ with $\mu_{\beta_l} = \frac{1}{N_j^*}\sum_s \widehat{\beta}_{s,l}$ and $\nu_{\beta_l}^2 = \frac{v_l\sigma_l^2}{N_j^*}$.