Estimation of Median Household Income for Small Areas : A Bayesian Semiparametric Approach

DHIMAN BHADRA[1], MALAY GHOSH[2] AND DALHO KIM[3]

[1] Production and Quantitative Methods Area, Indian Institute of Management Ahmedabad

[2] Department of Statistics, University of Florida, Gainesville, FL 32601

[3] Department of Statistics, Kyungpook National University, Taegu : 702-701, Korea

Email: dhiman@iimahd.ernet.in

Abstract

Estimation of median income of small areas is one of the principal targets of inference of the U.S Bureau of Census. These estimates play an important role in the formulation of various governmental decisions and policies. Since these estimates are collected over time, they often possess an inherent longitudinal pattern. Taking proper account of this time varying pattern may result in better estimates for the current or future median household incomes for a particular state or county. In this study, we put forward a semiparametric modeling procedure for estimating the median household income for all the U.S states. Our models include a nonparametric functional part for accommodating any unspecified time varying income pattern and also a state specific random effect to account for the within-state correlation of the income observations. Model fitting and parameter estimation is carried out in a hierarchical Bayesian framework using Markov chain Monte Carlo (MCMC) methodology. It is seen that the semiparametric model estimates can be superior to both the direct estimates and the Census Bureau estimates. Overall, our study indicates that proper modeling of the underlying longitudinal income profiles can improve the performance of model based estimates of household median income of small areas.

KEY WORDS : Current Population Survey, MCMC, Penalized spline, Random Walk, Semiparametric Modeling.

1 INTRODUCTION

Sample survey methodologies are widely used for collecting relevant information about a population of interest over time. Apart from providing population level estimates, surveys are also

designed to estimate various features of subpopulations or domains. Domains may be geographic areas like state or province, county, school district etc. or can even be identified by a particular socio-demographic characteristic like a specific age-sex group. Sometimes, the domain-specific sample size may be too small to yield direct estimates of adequate precision. This led to the development of small area estimation procedures which specifically deal with the estimation of various features of small domains. Generally, observations on various characteristics of small areas are collected over time, and thus, may possess a complicated underlying time-varying pattern. It is likely that models which exploit the time varying pattern in the observations may perform better than classical small area models which do not utilize this feature. In this study, we present a semiparametric Bayesian framework for the analysis of small area level data which explicitly accomodates for the longitudinal pattern in the response and the covariates.

1.1 SAIPE Program and Related Methodology

The Small Area Income and Poverty Estimates (SAIPE) program of the U.S Census Bureau was established with the aim of providing annual estimates of income and poverty statistics for all states, counties and school districts across the United States. The resulting estimates are generally used for the administration of federal programs and the allocation of federal funds to local jurisdictions. There are also many state and local programs that depend on these estimates. Prior to the creation of the SAIPE program, the decennial census was the only source of income and poverty statistics for households, families and individuals related to small geographic areas like counties, cities and other substate areas. Due to the ten year lag in the release of successive census values, there was a large gap in information concerning fluctuations in the economic situation of the country in general and local areas in particular. The establishment of the SAIPE program has largely mitigated this issue.

The current methodology of the SAIPE program is based on combining state and county estimates of poverty and income obtained from the American Community Survey (ACS) with other indicators of poverty and income using the Fay-Herriot class of models (Fay and Herriot, 1979).

The indicators are generally the mean and median adjusted gross income (AGI) from IRS tax returns, SNAP benefits data (formerly known as Food Stamp Program data), the most recent decennial census, intercensal population estimates, Supplemental Security Income Recipiency and other economic data obtained from the Bureau of Economic Analysis (BEA). Estimates from ACS are being used since January 2005 on the recommendation of the National Academy of Sciences Panel on Estimates of Poverty for Small Geographic Areas (2000). Income and poverty estimates until 2004 were based on data from the Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS).

Apart from various poverty measures, the SAIPE program provides annual state and county level estimates of median household income. For illustrating our methodology, we have considered data from ASEC for the period 1995-1999 in order to estimate the state wide median household income for 1999. We have then compared our estimates with the corresponding census figures for 1999. The SAIPE regression model for estimating the median household income for 1999 use as covariates, the median adjusted gross income (AGI) derived from IRS tax returns and the median household income estimate for 1999 obtained from the 2000 Census. The response variable is the direct estimate of median household income for 1999 obtained from the March 2000 CPS. Bayesian techniques are used to weigh the contributions of the CPS median income estimates and the regression predictions of the median income based on their relative precision. The standard deviations of the error terms are estimated by fitting a model to the estimates of sampling error covariance matrices of the CPS median household income estimates for several years. The mean function in this model is referred to as a “generalized variance function" (Bell, 1999). Noninformative prior distributions are placed on the regression parameter corresponding to the IRS median income since it was found to be statistically significant even in the presence of census data, both in the 1989 and 1999 models.

1.2 Related Research

Estimation of median income for small areas contributes to the policy making process of many Federal and State agencies. Before the establishment of the SAIPE program, the estimation of median income for four-person families was of general interest. The Census Bureau used the ideas suggested by Fay (1987) in this regard. Estimation was carried out in an empirical Bayes (EB) framework suggested by Fay, Nelson and Litow (1993). Later, Datta, Ghosh, Nangia and Natarajan (1996) extended the EB approach of Fay (1987) and also put forward univariate and multivariate hierarchical Bayes (HB) models. The estimates from their EB and HB procedures significantly improved over the CPS median income estimates for 1979. Ghosh, Nangia and Kim (henceforth referred to as GNK) (1996) exploited the repetitive nature of the state-specific CPS median income estimates and proposed a Bayesian time series modeling framework to estimate the statewide median income of four-person families for 1989. In doing so, they used a time specific random component and modeled it as a random walk. They concluded that the bivariate time series model utilizing the median incomes of four and five person families performs the best and produces estimates which are much superior to both the CPS and Census Bureau estimates. In general, the time series model always performed better than its non-time series counterpart.

Semiparametric regression methods have not been used in small area estimation contexts until recently. This was mainly due to methodological difficulties in combining the different smoothing techniques with the estimation tools generally used in small area estimation. The pioneering contribution in this regard is the work by Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008) in which they combined small area random effects with a smooth, non-parametrically specified trend using penalized splines. In doing so, they expressed the non-parametric small area estimation problem as a mixed effects regression model and analyzed it using restricted maximum likelihood. Theoretical results were presented on the prediction mean squared error and likelihood ratio tests for random effects. Inference was based on a simple non-parametric bootstrap approach. The methodology was used to analyze a non-longitudinal, spatial dataset concerning the estimation of mean acid neutralizing capacity (ANC) of lakes in the north eastern states of U.S.

1.3 Motivation and Overview

The motivation of our work also originates from the repetitive nature of the CPS median income estimates. But, in contrast to the approach of GNK (1996), we have viewed the state specific annual household median income values as longitudinal profiles or "income trajectories". This gained more ground because we used the state wide CPS median household income values for only five years (1995 - 1999) in our estimation procedure. Figure 1 shows sample longitudinal CPS median household income profiles for six states spanning 1995 to 2004 while Figures 2a. and 2b. shows the plots of the CPS median income against the IRS mean and median incomes for all the states for the years 1995 through 1999. It is apparent that CPS median income may have an underlying non-linear pattern with respect to IRS mean income, specially for large values of the latter. The above two features motivated us to use a semiparametric regression approach. In doing so, we have modeled the income trajectory using penalized spline (or P-spline) (Eilers and Marx, 1996) which is a commonly used but powerful function estimation tool in non-parametric inference. The P-spline is expressed using truncated polynomial basis functions with varying degrees and number of knots, although other types of basis functions like B-splines or thin plate splines can also be used. We have worked with two types of models viz. a regular semiparametric model and a semiparamteric random walk model. For each of these models, analysis has been carried out using a hierarchical Bayesian approach. Since we chose non-informative improper priors for the regression parameters, propriety of the posterior has been proved before proceeding with the computations. Markov chain Monte Carlo methodologies, specifically, Gibbs sampling (Gelfand and Smith, 1990) has been used to obtain the parameter estimates.

We have compared the state-specific estimates of median household income for 1999 with the corresponding decennial census values in order to test for their accuracy. In doing so, we observed that the semiparametric model estimates improve upon both the CPS and the SAIPE estimates. Interestingly, the positioning of the knots had significant influence on the results as will be discussed later on. We want to mention here that the SAIPE model had a considerable advantage over ours in that they used the census estimates of the median income for 1999 as a predictor. In

small area estimation problems, the census estimates are regarded as the “gold standard" since these are the most accurate estimates available with virtually negligible standard errors. So, using those as explanatory variables was an added advantage of the SAIPE state level models. The fact that our estimates still improve on the SAIPE model based estimates is a testament to the flexibility and strength of the semiparametric methodology specially when observations are collected over time. It also indicates that it may be worthwhile to take into account the longitudinal income patterns in estimating the current income conditions of the U.S states.

The remaining sections are arranged as follows. In Section 2 we introduce the two types of semiparametric models we have used. Section 3 goes over the hierarchical Bayesian analysis we performed. In Section 4, we describe the results of the data analysis with regard to the median household income dataset. In Section 5, we discuss the Bayesian model assessment procedure we used to test the goodness-of-fit of our models. We end with a discussion and some references towards future work in Section 6. The appendix contains the proofs of the posterior propriety and the expressions of the full conditional distributions.

2 MODEL SPECIFICATION

2.1 General Notation

Let $\mathbf{Y}_{ij} = (Y_{ij1}, ..., Y_{ijs})'$ be the sample survey estimators of some characteristics $\boldsymbol{\theta}_{ij} = (\theta_{ij1}, ..., \theta_{ijs})'$ for the i^{th} small area at the j^{th} time ($i = 1, 2, ..., m; j = 1, 2, ..., t$). The target of inference is generally $\boldsymbol{\theta}_{ij}$ or some function of it. Specifically, in our context, $\boldsymbol{\theta}_{ij} = \theta_{ij}$ which denotes the median household income of the i^{th} state at the j^{th} year. We are interested in estimating $(\theta_{1u}, ..., \theta_{mu})'$ i.e the median household income for all the states at time u. We may also want to estimate the difference in median household incomes at times v and u i.e $(\theta_{1v} - \theta_{1u}, ..., \theta_{mv} - \theta_{mu})'$. We denote by X_{ij} the covariate corresponding to the i^{th} state and j^{th} year.

2.2 Semiparametric Income Trajectory Models

We assume the following two semiparametric models :

2.2.1 Model I : Basic Semiparametric Model (SPM)

Let Y_{ij} and X_{ij} denote the CPS median household income and the IRS mean (or median) income recorded for the i^{th} state at the j^{th} year. The basic semiparametric model can be expressed as

$$Y_{ij} = f(x_{ij}) + b_i + u_{ij} + e_{ij} \tag{1}$$

where $f(x_{ij})$ is an unspecified function of x_{ij} reflecting the unknown response-covariate relationship. We approximate $f(x_{ij})$ using a P-spline and rewrite (1) as

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 x_{ij} + ... + \beta_p x_{ij}^p + \sum_{k=1}^{K} \gamma_k (x_{ij} - \tau_k)_+^p + b_i + u_{ij} + e_{ij} \\ &= \mathbf{X}_{ij}'\boldsymbol{\beta} + \mathbf{Z}_{ij}'\boldsymbol{\gamma} + b_i + u_{ij} + e_{ij} \\ &= \theta_{ij} + e_{ij} \end{aligned} \tag{2}$$

where $\theta_{ij} = \mathbf{X}_{ij}'\boldsymbol{\beta} + \mathbf{Z}_{ij}'\boldsymbol{\gamma} + b_i + u_{ij}$ is our target of inference.

Here $\mathbf{X}_{ij} = (1, x_{ij}, ..., x_{ij}^p)', \mathbf{Z}_{ij} = \{(x_{ij} - \tau_1)_+^p, ..., (x_{ij} - \tau_K)_+^p\}', \boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$ is the vector of regression coefficients while $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_K)'$ is the vector of spline coefficients. The above spline model with degree p can adequately approximate any unspecified smooth function. Typically, linear ($p = 1$) or quadratic ($p = 2$) splines serves most practical purposes since they ensure adequate smoothness in the fitted curve. m and t respectively denote the number of small areas and the number of time points at which the response and covariates are measured. Thus, in our case, $m = 51$, for the 50 U.S states and the District of Columbia and $t = 5$ for the years 1995-1999. b_i is a state-specific random effect while u_{ij} represents an interaction effect between the i^{th} state and the j^{th} year. We assume $b_i \sim^{i.i.d} N(0, \sigma_b^2)$ and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 I_K)$. σ_γ^2 controls the amount of smoothing of the underlying income trajectory. Moreover, it is assumed that u_{ij} and e_{ij}

are mutually independent with $u_{ij} \sim N(0, \psi_j^2)$ and $e_{ij} \sim N(0, \sigma_{ij}^2)$. The σ_{ij}^2's are the sampling standard deviations corresponding to the CPS direct median income estimates obtained using the "generalized variance function" technique mentioned in Section 1.1. In the datasets provided by the Census Bureau, these estimates are given for all the states at each of the time points. The knots $(\tau_1, ..., \tau_K)$ are usually placed on a grid of equally spaced sample quantiles of x_{ij}'s.

From (1) and (2), we have

$$\theta_{ij} = f(x_{ij}) + b_i + u_{ij}$$

which reflects our basic assumption that the true unknown household median income may have an unspecified variational pattern with the IRS mean (or median) income. Thus, the covariate effect is expressed by the unspecified nonparametric function $f(x_{ij})$ which reflects the possible nonlinear effect of x_{ij} on θ_{ij}.

2.2.2 Model II : Semiparametric Random Walk Model (SPRWM)

Since, for each state, the response and the covariates are collected over time for each state, there may be a definite trend in their behavior. Thus, we added a time specific random component to (1) and modeled it as a random walk as follows

$$\begin{aligned} Y_{ij} &= \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\boldsymbol{\gamma} + b_i + v_j + u_{ij} + e_{ij} \\ &= \theta_{ij} + e_{ij} \end{aligned} \tag{3}$$

where $\theta_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\boldsymbol{\gamma} + b_i + v_j + u_{ij}$

Here, v_j denotes the time specific random component. We assume that, $(v_j|v_{j-1}, \sigma_v^2) \sim N(v_{j-1}, \sigma_v^2)$ with $v_0 = 0$. Alternatively, we may write, $v_j = v_{j-1} + w_j$ where $w_j \sim^{i.i.d} N(0, \sigma_v^2)$. This is the so-called random walk model and is similar to the systems equations used in dynamic linear models.

Before proceeding to the next section, we may note that unlike the models of GNK (1996), the models given in (2) and (3) incorporate state specific random effects (b_i). This rectifies a limitation

of the former as pointed out in Rao (2003).

3 HIERARCHICAL BAYESIAN INFERENCE

3.1 Likelihood Function

Let $\mathbf{Y}_i = (Y_{i1}, ..., Y_{it})'$ be the response and $\mathbf{X}_i = (\mathbf{X}_{i1}, ..., \mathbf{X}_{it})'$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}, ..., \mathbf{Z}_{it})'$ be the covariates for the i^{th} state. Let $\mathbf{\Omega}_i = (\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \boldsymbol{\psi}^2, \sigma_b^2, \sigma_\gamma^2)$ be the parameter space corresponding to the i^{th} state where $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{it})'$ and $\boldsymbol{\psi}^2 = (\psi_1^2, ..., \psi_t^2)'$. Thus, the full parameter space will be given by $\mathbf{\Omega} = \mathbf{\Omega}_1 \times \mathbf{\Omega}_2 \times ... \times \mathbf{\Omega}_m$. For the i^{th} state, the likelihood corresponding to Model I (SPM) can be written as

$$\begin{aligned} L(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i | \mathbf{\Omega}_i) &\propto L(\mathbf{Y}_i | \boldsymbol{\theta}_i) L(\boldsymbol{\theta}_i | \boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \boldsymbol{\psi}^2, \mathbf{X}_i, \mathbf{Z}_i) L(b_i | \sigma_b^2) L(\boldsymbol{\gamma} | \sigma_\gamma^2) \\ &= \prod_{j=1}^{t} \left\{ L(Y_{ij} | \theta_{ij}, \sigma_{ij}^2) L(\theta_{ij} | \mathbf{X}_{ij}' \boldsymbol{\beta} + \mathbf{Z}_{ij}' \boldsymbol{\gamma} + b_i, \psi_j^2) \right\} L(b_i | \sigma_b^2) L(\boldsymbol{\gamma} | \sigma_\gamma^2) \end{aligned} \tag{4}$$

Here, $L(U|a, b)$ denotes a normal density with mean a and variance b while $L(b_i|\sigma_b^2)$ and $L(\boldsymbol{\gamma}|\sigma_\gamma^2)$ denotes a normal distribution with mean 0 and variances σ_b^2 and σ_γ^2 respectively.

For the random walk model, the parameter space for the i^{th} state would be $\mathbf{\Omega}_i = (\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \mathbf{v}, \boldsymbol{\psi}^2, \sigma_b^2, \sigma_\gamma^2, \sigma_v^2)$ where $\mathbf{v} = (v_1, ..., v_t)$ is the vector of time specific random effects. Thus, the likelihood function for the i^{th} state will have an extra component corresponding to $\mathbf{v}$ as follows

$$\begin{aligned} L(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i | \mathbf{\Omega}_i) = \prod_{j=1}^{t} \left\{ L(Y_{ij} | \theta_{ij}, \sigma_{ij}^2) L(\theta_{ij} | \mathbf{X}_{ij}' \boldsymbol{\beta} + \mathbf{Z}_{ij}' \boldsymbol{\gamma} + b_i, \psi_j^2) L(v_j | v_{j-1}, \sigma_v^2) \right\} \times \\ \times L(b_i | \sigma_b^2) L(\boldsymbol{\gamma} | \sigma_\gamma^2) \end{aligned} \tag{5}$$

where $L(v_j|v_{j-1}, \sigma_v^2)$ denotes a normal distribution with mean v_{j-1} and variance σ_v^2 where $v_0 = 0$.

3.2 Prior Specification

To complete the Bayesian specification of our model, we need to assign prior distributions to the unknown parameters. We assume noninformative improper uniform prior for the polynomial coefficients (or fixed effects) $\boldsymbol{\beta}$ and proper conjugate gamma priors on the inverse of the variance components $(\psi_1^2, ..., \psi_t^2, \sigma_b^2, \sigma_\gamma^2, \sigma_v^2)$. The prior distributions are assumed to be mutually independent. We choose small values (0.001) for the gamma shape and rate parameters to make the priors diffuse in nature so that inference is mainly controlled by the data distribution.

Thus, we have the following priors : $\boldsymbol{\beta} \sim$ uniform(R^{p+1}), $(\psi_j^2)^{-1} \sim G(c_j, d_j)(j = 1, ..., t)$, $(\sigma_b^2)^{-1} \sim G(c, d)$, $(\sigma_\gamma^2)^{-1} \sim G(c_\gamma, d_\gamma)$ and $(\sigma_v^2)^{-1} \sim G(c_v, d_v)$. Here $X \sim G(a, b)$ denotes a gamma distribution with shape parameter a and rate parameter b having the expression $f(x) \propto x^{a-1}\text{exp}(-bx), x \geq 0$.

3.3 Posterior Distribution and Inference

The full posterior of the parameters given the data is obtained in the usual way by combining the likelihood and the prior distribution as follows

$$p(\boldsymbol{\Omega}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) \propto \prod_{i=1}^{m} L(\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i|\boldsymbol{\Omega}_i)\pi(\boldsymbol{\beta})\pi(\sigma_b^2)\pi(\sigma_\gamma^2) \prod_{j=1}^{t} \pi(\psi_j^2) \tag{6}$$

For the random walk model, there will be an additional term $\pi(\sigma_v^2)$. By the conditional independence properties, we can factorize the full posterior as

$$[\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \sigma_b^2, \sigma_\gamma^2, \{\psi_1^2, ..., \psi_t^2\}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}] \propto [\mathbf{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \{\psi_1^2, ..., \psi_t^2\}, \mathbf{X}, \mathbf{Z}][\mathbf{b}|\sigma_b^2] \times$$
$$\times[\boldsymbol{\gamma}|\sigma_\gamma^2][\boldsymbol{\beta}][\sigma_\gamma^2][\sigma_b^2] \prod_{j=1}^{t} [\psi_j^2]$$

Our target of inference is $\{\theta_{ij}, i = 1, ..., m; j = 1, ...t\}$, the true median household income of all the states. Since the marginal posterior distribution of θ_{ij} is analytically intractable, high dimensional integration needs to be carried out in a theoretical framework. However, this task can

be easily accomplished in an MCMC framework by using Gibbs sampler to sample from the full conditionals of θ_{ij} and other relevant parameters. In implementing the Gibbs sampler, we follow the recommendation of Gelman and Rubin (1992) and run $n\ (\geq 2)$ parallel chains. For each chain, we run $2d$ iterations with starting points drawn from an overdispersed distribution. To diminish the effects of the starting distributions, the first d iterations of each chain are discarded and posterior summaries are calculated based on the rest of the d iterates. The full conditionals for both the models are given in the appendix.

4 DATA ANALYSIS

We applied the semiparametric models in Section 2.2. to analyze the median household income dataset referred to in Section 1.3. The response variable Y_{ij} and the covariates X_{ij} denote respectively the CPS median household income estimate and the corresponding IRS mean (or median) income estimate for the i^{th} state at the j^{th} year ($i = 1, ..., 51; j = 1, ..., 5$). The state-specific mean or median income figures are obtained from IRS tax return data. The Census Bureau gets files of individual tax return data from the IRS for use in specifically approved projects such as SAIPE. For each state, the IRS mean (median) income is the mean (median) adjusted gross income (AGI) across all the tax returns in that state. Like other SAIPE model covariates obtained from administrative records data, these variables do not exactly measure the median income across all households in the state. One of the reasons for this is that the AGI would not necessarily be the same as the exact income figure and the tax return universe does not cover the entire population i.e some households do not need to file tax returns, and those that do not are likely to differ in regard to income than those that do. However, the use of the mean or median AGI as a covariate only requires it to be correlated with median household income, not necessarily be the same thing. Specifically for this study, we have used IRS mean income as our covariate. This is because, it seems to possess an underlying non-linear relationship with the CPS median income (Figure 2a), and so it is more suited to a semiparametric analysis.

4.1 Comparison Measures and Knot Specification

Our dataset originally contained the median household income of all the U.S states and the District of Columbia for the years 1995-2004. However, we only used the information for the five year period 1995-1999 since our target of inference are the state specific median household incomes for 1999. We evaluated the performance of our estimates by comparing them to the corresponding census figures for 1999. This is because, in small area estimation problems, the census estimates are often treated as "gold standard" against which all other estimates are compared. However, such a comparison is only possible for those years which immediately precede the census year e.g. 1969, 1979, 1989 and 1999.

In order to check the performance of our estimates, we plan to use four comparison measures. These were originally recommended by the panel on small area estimates of population and income set up by the Committee on National Statistics in July 1978 and are available in their July 1980 report (p. 75). These are

- Average Relative Bias (ARB) = $(51)^{-1} \sum_{i=1}^{51} \frac{|c_i - e_i|}{c_i}$
- Average Squared Relative Bias (ASRB) = $(51)^{-1} \sum_{i=1}^{51} \frac{|c_i - e_i|^2}{c_i^2}$
- Average Absolute Bias (AAB) = $(51)^{-1} \sum_{i=1}^{51} |c_i - e_i|$
- Average Squared Deviation (ASD) = $(51)^{-1} \sum_{i=1}^{51} (c_i - e_i)^2$

Here c_i and e_i respectively denote the census and model based estimate of median household income for the i^{th} state ($i = 1, ..., 51$). Clearly, lower values of these measures would imply a better model based estimate.

The basic structure of our models would remain the same as in Section 2.2. We have used truncated polynomial basis for the P-spline component in both the models. Since Fig 2a doesnot indicate a high degree of non-linearity, we have restricted ourselves to a linear spline ($p = 1$). The selection of knots is always a subjective but tricky issue in these kind of problems. Sometimes experience on the subject matter may be a guiding force in placing the knots at the "optimum"

locations where a sharp change in the curve pattern can be expected. Too few or too many knots generally create problems in terms of worsening the fit. This is because, if too few knots are used, the complete underlying pattern may not be captured properly, thus resulting in a biased fit. On the other hand, once there are enough knots to fit important features of the data, further increase in the number of knots have little effect on the fit and may even degrade the quality of the fit (Ruppert, 2002). Generally, at most 35 to 40 knots are recommended for effectively all sample sizes and for nearly all smooth regression functions. Following the general convention, we have placed the knots on a grid of equally spaced sample quantiles of the independent variable (IRS mean income).

4.2 Computational Details

We implemented and monitored the convergence of the Gibbs sampler following the general guidelines given in Gelman and Rubin (1992). We ran three independent chains each with a sample size of 10,000 and with a burn-in sample of another 5,000. We initially sampled the θ_{ij}'s from t-distributions with 2 df having the same location and scale parameters as the corresponding normal conditionals given in the Appendix. This is based on the Gelman-Rubin idea of initializing certain samples of the chain from overdispersed distributions. However, once initialized, the successive samples of θ_{ij}'s are generated from regular univariate normal distributions. Convergence of the Gibbs sampler was monitored by visually checking the dynamic trace plots, acf plots and by computing the Gelman-Rubin diagnostic. The comparison measures deviated slightly for different initial values. We chose the least of those as the final measures presented in the tables that follows.

4.3 Analytical Results

Data on CPS median income and IRS mean incomes were available for 50 states and the District of Columbia for the time span 1995-2004. CPS median income ranged from $ 24,879.68 to $ 52,778.94 with a mean of $ 36,868.48 and standard deviation of $5954.94 while IRS mean annual income ranged from $ 27,910 to $ 72,769.38 with a mean of $ 41,133.45 and standard deviation of $7196.56.

We fitted Model I (SPM) with all possible knot choices from 0 to 40 but the best results were achieved with 5 knots. The estimates (with 5 knots) improved significantly over the CPS estimates based on all the four comparison measures. Addition of more knots seemed to degrade the fit of the model. This may happen as pointed out in Ruppert (2002). On the other hand, the SAIPE model based estimates were slightly superior to the SPM estimates.

Next, we fitted the semiparametric random walk model (SPRWM) to our data. Overall, the random walk structure lead to some improvement in the performance of the estimates. However, for the model with 5 knots, the performance of the estimates remained nearly the same. This may be because 5 knots is sufficient to capture the underlying pattern in the income trajectory and the random walk component doesnot lead to any further improvement. Last but not the least, the random walk model estimates, although generally better than those of the basic semiparametric model, still cannot claim to be superior to the SAIPE estimates for all the comparison measures. Table 1 reports the posterior mean, median and 95% CI for the parameters of the SPRWM with 5 knots.

It is of interest that the 95% CI for γ_1, γ_4 and γ_5 doesnot contain 0 indicating the significance of the first, fourth and fifth knots. This is indicative of the relevance of knots in the penalized spline fit on the CPS median income observations. The same is true for the coefficients of SPM.

4.4 Knot Realignment

As mentioned in Section 1.1, the SAIPE state models use the census estimates of median income (for 1999) as one of the predictor which essentially gives them a big edge over us. This may be one of the reasons why the estimates obtained from the semiparametric models are atmost comparable, but not superior to the SAIPE estimates. But that doesn't rule out the fact that the semiparametric models have room for improvement. In this section, we will look for any possible deficiencies in the our models and will try to come up with some improvements, if there is any.

As mentioned in Section 4.1, selection and proper positioning of knots plays a pivotal role in capturing the true underlying pattern in a set of observations. Poorly placed knots does little in this

regard and can even lead to an erroneous or biased estimate of the underlying trajectory. Ideally, a sufficient number of knots should be selected and placed uniformly throughout the range of the independent variable to accurately capture the underlying observational pattern.

Figures 3a. and 3b. shows the exact positions of 5 and 7 knots in the plot of CPS median income against IRS mean income. In both the cases, the knots are placed on a grid of equally spaced sample quantiles of IRS mean income. In both the figures, the knots lie on the left of IRS mean = 50000, the region where the density of observations is high. The knots tend to lie in this region because they are selected based on quantiles which is a density-dependent measure. Thus, in both the figures, the coverage area of knots (i.e the part of the observational pattern which is captured by the knots) is the region to the left of the dotted vertical lines. On the other hand, the non-linear pattern is tangible only in the low density area of the plot i.e the region lying to the right of IRS mean = 50000. Evidently, none of the knots lie in this part of the graph. Thus, we can presume that in both the cases (5 and 7 knots), the underlying non-linear observational pattern is not being adequately captured.

As a natural solution to this issue, we decided to place half of the knots in the low density region of the graph while the other half in the high density region. The exact boundary line between the high density and low density regions is hard to determine. We tested different alternatives and came up with IRS mean = 47000 as a tentative boundary because it gave the best results. In both the regions, we placed the knots at equally spaced sample quantiles of the independent variable. Figure 4 shows the new knot positions for 5 knots.

It is clear from Figure 4 that the new knots are more dispersed throughout the range of IRS mean than the old ones. The region between the bold and dashed vertical lines denotes the additional coverage that has been achieved with the knot rearrangement. Based on the number of data points inside this region, it is clear that a much larger proportion of observations has been captured with the knot realignment. No knots are in the region beyond the bold vertical lines (i.e beyond IRS mean 56000) possibly due to the very low density of the observations in that area. Overall, it seems that, the new knots can capture the underlying non-linear pattern in the dataset which the old knots

failed to achieve. We also experimented by placing all the knots in the low density region (beyond IRS mean = 47000) but the results were not satisfactory. This indicates that the knots should be uniformly placed throughout the range of the independent variable to get an optimal fit.

We have worked with 5 knots because it performed consistently well for both the SPM and SPRW models. On fitting the semiparametric models with the new knot alignment, we did achieve some improvement in the results. Table 2 reports the comparison measures for the raw CPS estimates, SAIPE estimates and the semiparametric estimates with the knot realignment while Table 3 depicts the percentage improvement of the semiparametric estimates over the CPS and SAIPE estimates. Here, SPM(5)* and SPRWM(5)* respectively denote the semiparametric models with the realigned 5 knots.

It is clear that, with the knot realignment, the comparison measures corresponding to the semiparametric estimates have decreased substantially, specially so for the SPM. The new comparison measures for the semiparametric models are quite lower than those corresponding to the SAIPE estimates. Thus, we may say that the semiparametric model estimates performs better than the SAIPE estimates with the realigned knots. This improvement is apparently due to the additional coverage of the observational pattern that is being achieved with the relocation of the knots. As a result of this increased coverage, the new knots are possibly capturing the underlying nonlinear pattern in the observations which the old knots failed to achieve. Although we have done this exercise with only 5 knots, it would be interesting to experiment with other types of knot alignment and with different number of knots. Tables 4 and 5 report the posterior mean, median and 95% CI for the parameters in SPM(5)* and SPRWM(5)* respectively.

It is of interest to note that, with the knot realignment, all the knot coefficients (i.e the γ's) are significant for both SPM and SPRWM. For the old configuration, some of the knot coefficients were not significant for the models. This corroborates the fact that, with the knot realignment, all the five knots are significantly contributing to the curve fitting process in terms of capturing the true underlying non-linear pattern in the observations.

4.5 Comparison with an Alternate Model

We also compared the semiparametric models (with 5 knots) with the model proposed by GNK (1996), henceforth referred to as the GNK model. Their univariate model is as follows

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_j + u_{ij} + e_{ij} \tag{7}$$

where $(b_j|b_{j-1}) \sim N(0, \sigma_b^2)$, $u_{ij} \sim N(0, \psi_j^2)$ and $e_{ij} \sim N(0, \sigma_{ij}^2)$.

One of the major qualitative difference between the above model and our semiparametric models is that the former doesnot have a state specific random effect. In fact, it would also be interesting to compare the above model with the basic semiparametric model (SPM) with 0 knots i.e

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + u_{ij} + e_{ij} \tag{8}$$

where $b_i \sim^{i.i.d} N(0, \sigma_b^2)$ while u_{ij} and e_{ij} have the same distribution as above. Clearly, the only difference between (7) and (8) is that the former contains a time specific random component while the latter contains a area specific random component. GNK (1996) showed that the estimates from the bivariate version of the GNK model (7) performs much better than the census bureau estimates in estimating the median household income of 4-person families in the United States. Table 6 depicts the comparison measures corresponding to the above models.

It is clear that, although the estimates from the GNK model perform slightly better than the CPS, those are quite inferior to the semiparametric and SAIPE estimates. This may be because the state specific random effects in the semiparametric models can account for the within-state correlations in the income values, something which the GNK model fails to do. Since the comparison measures for SPM(0) are much lower than those for the GNK model, we can also conclude that the area specific random effect is much more critical than a time specific random component in this situation.

5 MODEL ASSESSMENT

To examine the goodness-of-fit of the semiparametric models, we used a Bayesian Chi-square goodness-of-fit statistic (Johnson, 2004). This is essentially an extension of the classical Chi-square goodness-of-fit test where the statistic is calculated at every iteration of the Gibbs sampler as a function of the parameter values drawn from the respective posterior distribution. Thus, a posterior distribution of the statistic is obtained which can be used for constructing global goodness-of-fit diagnostics.

To construct this statistic, we form 10 equally spaced bins $((k-1)/10, k/10)$, $k = 1, ..., 10$, with fixed bin probabilities, $p_k = 1/10$. The main idea is to consider the bin counts $m_k(\tilde{\Theta})$ to be random where $\tilde{\Theta}$ denotes a posterior sample of the parameters. At each iteration of the Gibbs sampler, bin allocation is made based on the conditional distribution of each observation given the generated parameter values i.e Y_{ij} would be allocated to the k^{th} bin if $F(Y_{ij}|\tilde{\Theta}) \in ((k-1)/10, k/10)$, $k = 1, ..., 10$. Johnson's Bayesian chi-square statistic is then calculated as

$$R^B(\tilde{\Theta}) = \sum_{k=1}^{10} \left[\frac{m_k(\tilde{\Theta}) - np_k}{\sqrt{np_k}} \right]^2$$

For the purpose of model assessment, two summary measures can be used, both derived from the posterior distribution of $R^B(\tilde{\Theta})$. First one is the proportion of times the generated values of R^B exceeds the 0.95 quantile of a χ^2_9 distribution. Values quite close to 0.05 would suggest a good fit. The second diagnostic is the probability that $R^B(\tilde{\Theta})$ exceeds a χ^2_9 deviate i.e

$$A = P_{\tilde{\Theta}|\mathbf{y}}(R^B(\tilde{\Theta}) > X), \quad X \sim \chi^2_9$$

Since the nominal value of this probability is 0.5, values close to 0.5 would suggest a good fit.

The only assumptions for this statistic to work are that the observations should be conditionally independent and the parameter vector should be finite dimensional. The second assumption naturally holds in our case. Regarding the first one, since we have multiple observations over time for every state, there may be within-state dependence between those. Thus, instead of taking all the

observations (i.e the CPS median income values), we decided to use the last observation for each state. For the basic semiparametric model (SPM), the above summary measures were respectively 0.049 and 0.5 while for the random walk model (SPRWM), these were 0.047 and 0.51. These measures suggest that both SPM and SPRWM fits the data quite well. Figure 5a. and 5b. shows the quantile-quantile plots of R^B values obtained from 10000 samples of SPM and SPRWM with 5 knots. Both the plots demonstrate excellent agreement between the distribution of R^B and that of a $\chi^2(9)$ random variable.

Johnson points out that the Bayesian chi-square test statistic is also an useful tool for code verification. If the posterior distribution of R^B deviates significantly from its null distribution, it may imply that the model is incorrectly specified or there are coding errors. Since the summary measures are quite close to the corresponding null values, we think that our models provide a satisfactory fit to the data set and also that there are no coding errors.

6 DISCUSSION

The proper estimation of median household income for different small areas is one of the principal goals of the U.S Census Bureau. These estimates are frequently used by the Federal Government for the administration and maintenance of different federal programs and also for the allotment of federal grants to local jurisdictions. Although these estimates are available annually for every state, the U.S Census Bureau generally uses a non-longitudinal approach in their estimation procedure based on the Fay-Herriot model (Fay and Herriot, 1979). In this study, we have proposed a semiparametric class of models which exploit the longitudinal trend in the state-specific income observations. In doing so, we have modeled the CPS median income observations as an "income trajectory" using penalized splines (Eilers and Marx, 1996). We have also extended the basic semiparametric model by adding a time series random walk component which can explain any specific trend in the income levels over time. We have used as our covariate, the mean adjusted gross income (AGI) obtained from IRS tax returns for all the states. Analysis has been carried out in a hierarchical Bayesian framework. Our target of inference has been the state wide median

household incomes for the year 1999. We have evaluated our estimates by comparing those with the corresponding census estimates of 1999 using some commonly used comparison measures.

Our analysis has shown that information of past median income levels of different states do provide strength towards the estimation of state specific median incomes for the current period. In fact, if there is an underlying non-linear pattern in the median income levels, it may be worthwhile to capture that pattern as accurately as possible and use that in the inferential procedure. In terms of modeling the underlying observational pattern, the positioning of knots proved to be both important and interesting. The quality (in terms of their "closeness" to the census estimates) of the estimates tended to improve as the knots were positioned more uniformly throughout the range of the independent variable. It became apparent that the contribution of the knots towards deciphering the underlying observational pattern improved substantially when those were properly placed with an optimal coverage area. This in turn improved the approximation of the curve *vis-a-vis* the true unknown observational pattern. This proved interesting because, still now, there is no absolute rule which controls the positioning of knots. Our final estimates proved to be superior, not only to the raw CPS estimates, but also to the current U.S Census Bureau (SAIPE) estimates. Although the basic semiparametric model performed much better that the semiparametric random walk model with 5 knots, more experiments need to be done with different knot positions and number before anything conclusive can be said about their relative performance as a whole. But, it seems that, if adequate knots are used and if those are placed uniformly throughout the range of the independent variable, then a random walk component may not improve the fit any further provided there is no strong trend in the income levels. The main advantage of our modeling procedure is that it can be used for any possible patterns in the response (income, poverty etc) observations of small areas. In a subsequent work related to the estimation of median incomes of 4-person families, we have shown that the multivariate version of the basic semiparametric model perform quite well too and provide estimates which are consistently superior to the U.S Census Bureau estimates.

The above models can be extended in various ways based on the nature of the observational pattern and the quality (or richness) of the dataset. Some obvious extensions are given as fol-

lows : (1) In the models considered above, the spline structure $f(x_{ij})$ represents the population mean income trajectory for all the states combined. The deviation of the i^{th} state from the mean is modeled through the random intercept b_i. This implies that the state-specific trajectories are parallel. A more flexible extension would be to model the state-specific deviations as unspecified non-parametric functions as follows

$$\begin{aligned} Y_{ij} &= f(x_{ij}) + g_i(x_{ij}) + u_{ij} + e_{ij} \\ \text{where} \quad g_i(x_{ij}) &= b_{i1} + b_{i2}x_{ij} + \sum_{k=1}^{K^*} w_{ik}(x_{ij} - \kappa_k)_+ \end{aligned} \tag{9}$$

Here $g_i(x)$ is an unspecified nonparametric function representing the deviation of the i^{th} state from the population mean trajectory $f(x)$. $g_i(x)$ is also modeled using P-spline with a linear part, $b_{i1} + b_{i2}x$ and a non-linear one, $\sum_{k=1}^{K^*} w_{ik}(x - \kappa_k)_+$ thus allowing for more flexibility. Both these components are random with $(b_{i1}, b_{i2})' \sim N(\mathbf{0}, \Sigma)$ (Σ being unstructured or diagonal) and $w_{ik} \sim N(0, \sigma_w^2)$. This extension is particularly relevant in situations where the state-specific income trajectories are quite distinct from the population mean curve and thus need to be modeled explicitly. We plan to pursue this extension if we can procure a richer dataset with longer state specific income trajectories. (2) Sometimes the function to be estimated (here the median income pattern) may have varying degrees of smoothness in different regions. In that case, a single smoothing parameter may not be proper and a spatially adaptive smoothing procedure can be used (Ruppert and Carroll, 1999). (3) We used the truncated polynomial basis function to model the income trajectory but other types of bases like B-splines, radial basis functions etc can also be used. (4) Although we used a parametric normal distributional assumption for the random state and time specific effects, a broader class of distributions like the Dirichlet process or Polya trees may be tested.

Last but not the least, we think that semiparametric modeling approach holds a lot of promise for small domain problems specially where observations for each domain are collected over time. The associated class of semiparametric models can well be an attractive alternative to the models

generally employed by the U.S Census Bureau.

APPENDIX A : PROOFS

A.1 : Proof of Posterior Propriety

The proof of posterior propriety for the basic semiparametric model (Model I) is outlined below. The necessary changes to the proof for the random walk model are mentioned at the end.

Theorem 1. Let $\psi^2_{max} = \max(\psi^2_1, ..., \psi^2_t) = \psi^2_k$, say, for some $k \in [1, ..., t]$. Then, posterior propriety holds if the following conditions are satisfied

1. $(m - p - 5)/2 + c_k > 0$ and $d_k > 0$

2. $m/2 + c_j - 2 > 0$ and $d_j > 0$, $j = 1, ..., t; j \neq k$

Proof. The basic parameter space is $\Omega = (\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{b}, \sigma^2_b, \sigma^2_\gamma, \{\psi^2_1, ..., \psi^2_t\})$ where $\boldsymbol{\theta} = (\boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_m)'$ and $\mathbf{b} = (b_1, ..., b_m)'$. Let

$$\begin{aligned} I &= \int ... \int p(\boldsymbol{\Omega}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) d\boldsymbol{\Omega} \\ &= \int ... \int \prod_{i=1}^{m} \Big\{ L(\mathbf{Y}_i|\boldsymbol{\theta}_i) L(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \boldsymbol{\psi}^2, \mathbf{X}_i, \mathbf{Z}_i) L(b_i|\sigma^2_b) \Big\} L(\boldsymbol{\gamma}|\sigma^2_\gamma) \pi(\boldsymbol{\beta}) \pi(\sigma^2_b) \pi(\sigma^2_\gamma) \prod_{j=1}^{t} \pi(\psi^2_j) d\boldsymbol{\Omega} \end{aligned} \tag{10}$$

We have to show that $I \leq M$ where M is any finite positive constant.
Integrating first w.r.t $\boldsymbol{\beta}$, we have

$$\begin{aligned} I_{\boldsymbol{\beta}} &= \int \pi(\boldsymbol{\beta}) \prod_i L(\boldsymbol{\theta}_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \boldsymbol{\psi}^2, \mathbf{X}_i, \mathbf{Z}_i) d\boldsymbol{\beta} \\ &= \int \exp\big[-\frac{1}{2} \sum_i (\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\gamma} - b_i\mathbf{1})' \Psi^{-1} (\boldsymbol{\theta}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\boldsymbol{\gamma} - b_i\mathbf{1}) \big] d\boldsymbol{\beta} \\ &= |\sum_i \mathbf{X}'_i \Psi^{-1} \mathbf{X}_i|^{-1/2} \exp\Big[-\frac{1}{2} \sum_i \mathbf{W}'_i \Psi^{-1} \mathbf{W}_i + \frac{1}{2} \Big(\sum_i \mathbf{W}'_i \Psi^{-1} \mathbf{X}_i \Big) \Big(\sum_i \mathbf{X}'_i \Psi^{-1} \mathbf{X}_i \Big)^{-1} \Big(\sum_i \mathbf{X}'_i \Psi^{-1} \mathbf{W}_i \Big) \Big] \end{aligned} \tag{11}$$

where $\mathbf{W}_i = \boldsymbol{\theta}_i - \mathbf{Z}_i\boldsymbol{\gamma} - b_i\mathbf{1}$ and Ψ^{-1} = diag$(\psi_1^{-2}, \psi_2^{-2}, ..., \psi_t^{-2})$.

Now, $\mathbf{W}_i'\Psi^{-1}\mathbf{W}_i = \mathbf{W}_i'\Psi^{-1/2}\Psi^{-1/2}\mathbf{W}_i = \mathbf{S}_i'\mathbf{S}_i$ where $\mathbf{S}_i = \Psi^{-1/2}\mathbf{W}_i$. Similarly, $\mathbf{W}_i'\Psi^{-1}\mathbf{X}_i = \mathbf{S}_i'\mathbf{T}_i$, $\mathbf{X}_i'\Psi^{-1}\mathbf{W}_i = \mathbf{T}_i'\mathbf{S}_i$ and $\mathbf{X}_i'\Psi^{-1}\mathbf{X}_i = \mathbf{T}_i'\mathbf{T}_i$ where $\mathbf{T}_i = \Psi^{-1/2}\mathbf{X}_i$. On replacing these, the expression in the exponent of (11) becomes

$$\begin{aligned} & -\frac{1}{2}\Big[\sum_i \mathbf{S}_i'\mathbf{S}_i - \big(\sum_i \mathbf{S}_i'\mathbf{T}_i\big)\big(\sum_i \mathbf{T}_i'\mathbf{T}_i\big)^{-1}\big(\sum_i \mathbf{T}_i'\mathbf{S}_i\big)\Big] \\ = \; & -\frac{1}{2}\Big[\mathbf{S}'\mathbf{S} - \mathbf{S}'\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{S}\Big] \\ = \; & -\frac{1}{2}\mathbf{S}'\Big[\mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\Big]\mathbf{S} = Q, \text{ say} \end{aligned}$$

where $\mathbf{S} = (\mathbf{S}_1', ..., \mathbf{S}_m')'$ and $\mathbf{T} = (\mathbf{T}_1', ..., \mathbf{T}_m')'$. Since $(\mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}')$ is idempotent, $\mathbf{S}'\Big[\mathbf{I} - \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\Big]\mathbf{S}$ is non-negative, implying $Q \leq 0$ and thus exp(Q) ≤ 1.

Next, we consider integration w.r.t $\boldsymbol{\psi}^2$ i.e

$$\begin{aligned} I_{\boldsymbol{\psi}} &= \int ... \int |\sum_i \mathbf{X}_i'\Psi^{-1}\mathbf{X}_i|^{-1/2} \prod_{j=1}^{t} (\psi_j^2)^{-m/2-c_j+1}\exp(-d_j/\psi_j^2)d\psi_1^2...d\psi_t^2 \\ &= \int ... \int |\sum_{i,j} \mathbf{X}_{ij}\psi_j^{-2}\mathbf{X}_{ij}'|^{-1/2} \prod_{j=1}^{t} (\psi_j^2)^{-m/2-c_j+1}\exp(-d_j/\psi_j^2)d\psi_1^2...d\psi_t^2 \end{aligned} \tag{12}$$

Assuming $\psi_{max} = \max(\psi_1, ..., \psi_t)$, we have, $\forall j = 1, ..., t$, $\psi_j^{-2} \geq \psi_{max}^{-2} \Rightarrow \mathbf{X}_{ij}\psi_j^{-2}\mathbf{X}_{ij}' \geq \mathbf{X}_{ij}\psi_{max}^{-2}\mathbf{X}_{ij}' \Rightarrow \sum_{i,j}\mathbf{X}_{ij}\psi_j^{-2}\mathbf{X}_{ij}' \geq \psi_{max}^{-2}\sum_{i,j}\mathbf{X}_{ij}\mathbf{X}_{ij}'$ and thus

$$|\sum_{i,j} \mathbf{X}_{ij}\psi_j^{-2}\mathbf{X}_{ij}'|^{-1/2} \leq (\psi_{max}^2)^{(p+1)/2}|\sum_{i,j}\mathbf{X}_{ij}\mathbf{X}_{ij}'|^{-1/2} \tag{13}$$

Combining (12) and (13), we have

$$I_{\boldsymbol{\psi}} \leq |\sum_{i,j}\mathbf{X}_{ij}\mathbf{X}_{ij}'|^{-1/2}\int ... \int (\psi_{max}^2)^{(p+1)/2}\prod_{j=1}^{t}(\psi_j^2)^{-m/2-c_j+1}\exp(-d_j/\psi_j^2)d\psi_1^2...d\psi_t^2$$

Assuming $\psi^2_{max} = \psi^2_k$ for some $k \in [1, ..., t]$, we have,

$$
\begin{aligned}
I_{\boldsymbol{\psi}} &\leq |\sum_{i,j} \mathbf{X}_{ij}\mathbf{X}'_{ij}|^{-1/2} \int ... \int \Big[\int (\psi^2_k)^{\frac{p-m}{2}-c_k+\frac{3}{2}} \exp\Big(-\frac{d_k}{\psi^2_k}\Big) d\psi^2_k\Big] \prod_{j=1,j\neq k}^{t} (\psi^2_j)^{-\frac{m}{2}-c_j+1} \exp\Big(-\frac{d_j}{\psi^2_j}\Big) d\psi^2_1 ... d\psi^2_t \\
&= |\sum_{i,j} \mathbf{X}_{ij}\mathbf{X}'_{ij}|^{-1/2} \frac{\Gamma((m-p-5)/2+c_k)}{d_k^{(m-p-5)/2+c_k}} \prod_{j=1,j\neq k}^{t} \frac{\Gamma(m/2+c_j-2)}{d_j^{m/2+c_j-2}} = \mathbb{W}, \text{say} \qquad (14)
\end{aligned}
$$

where $\mathbb{W}$ is finite if $(m-p-5)/2 + c_k > 0$, $d_k > 0$, $m/2 + c_j - 2 > 0$ and $d_j > 0$ for $j = 1, ..., t; j \neq k$.

Combining (10) and (14), we have

$$
I \leq \mathbb{W} \int ... \int \prod_{i=1}^{m} \Big\{L(\mathbf{Y}_i|\boldsymbol{\theta}_i) L(b_i|\sigma^2_b)\Big\} L(\boldsymbol{\gamma}|\sigma^2_\gamma) \pi(\sigma^2_b) \pi(\sigma^2_\gamma) d\boldsymbol{\Omega}^* \qquad (15)
$$

where $\boldsymbol{\Omega}^* = (\boldsymbol{\Omega} - \boldsymbol{\beta} - \boldsymbol{\psi})$. Since all the components of the integrand in (15) have proper distributions, the above integral would be finite thus proving posterior propriety.

For the random walk model, the integrand in (10) will have an additional likelihood term $\prod_{j=1}^{t} L(v_j|v_{j-1}, \sigma^2_v)$ and a prior term $\pi(\sigma^2_v)$. The derivation would then proceed exactly as above and the integrand in (15) will also contain these additional terms. But since both of these are proper distributions (normal and inverse gamma respectively), I will still be finite under the conditions stated in the theorem.

APPENDIX B : FULL CONDITIONAL DISTRIBUTIONS

The full conditional distributions of the parameters for the basic semiparametric model are as follows :

$$
\begin{aligned}
\Big[\theta_{ij}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}^2, \mathbf{b}, \mathbf{X}, \mathbf{Z}\Big] &\sim N\Big[\Big(\frac{1}{\sigma^2_{ij}} + \frac{1}{\psi^2_j}\Big)^{-1} \Big(\frac{y_{ij}}{\sigma^2_{ij}} + \frac{(\mathbf{X}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\boldsymbol{\gamma} + b_i)}{\psi^2_j}\Big), \Big(\frac{1}{\sigma^2_{ij}} + \frac{1}{\psi^2_j}\Big)^{-1}\Big] \\
\Big[b_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\psi}^2, \sigma^2_b, \mathbf{X}, \mathbf{Z}\Big] &\sim N\Big[\Big(\frac{1}{\sigma^2_b} + \sum_{j=1}^{t} \frac{1}{\psi^2_j}\Big)^{-1} \Big(\sum_{j=1}^{t} \frac{1}{\psi^2_j}(\theta_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta} - \mathbf{Z}'_{ij}\boldsymbol{\gamma})\Big), \Big(\frac{1}{\sigma^2_b} + \sum_{j=1}^{t} \frac{1}{\psi^2_j}\Big)^{-1}\Big]
\end{aligned}
$$

$$
\begin{aligned}
\Big[\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\theta},\mathbf{b},\boldsymbol{\psi}^2,\mathbf{X},\mathbf{Z}\Big] &\sim N\Big[\Big(\sum_{i=1}^{m}\sum_{j=1}^{t}\frac{\mathbf{X}_{ij}\mathbf{X}'_{ij}}{\psi_j^2}\Big)^{-1}\Big(\sum_{i=1}^{m}\sum_{j=1}^{t}\frac{\mathbf{X}_{ij}}{\psi_j^2}(\theta_{ij}-\mathbf{Z}'_{ij}\boldsymbol{\gamma}-b_i)\Big),\Big(\sum_{i=1}^{m}\sum_{j=1}^{t}\frac{\mathbf{X}_{ij}\mathbf{X}'_{ij}}{\psi_j^2}\Big)^{-1}\Big]\\
\Big[\boldsymbol{\gamma}|\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{b},\boldsymbol{\psi}^2,\sigma_\gamma^2,\mathbf{X},\mathbf{Z}\Big] &\sim N\Big[\Big(\sum_{i,j}\frac{\mathbf{Z}_{ij}\mathbf{Z}'_{ij}}{\psi_j^2}+\frac{1}{\sigma_\gamma^2}I\Big)^{-1}\Big(\sum_{i,j}\frac{\mathbf{Z}_{ij}}{\psi_j^2}(\theta_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}-b_i)\Big),\Big(\sum_{i,j}\frac{\mathbf{Z}_{ij}\mathbf{Z}'_{ij}}{\psi_j^2}+\frac{1}{\sigma_\gamma^2}I\Big)^{-1}\Big]\\
\Big[(\sigma_\gamma^2)^{-1}|\boldsymbol{\gamma}\Big] &\sim G\Big[\frac{K}{2}+c_\gamma,\frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{\gamma}+d_\gamma\Big]\\
\Big[(\psi_j^2)^{-1}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta},\mathbf{b},\mathbf{X},\mathbf{Z}\Big] &\sim G\Big[c_j+\frac{m}{2},\frac{1}{2}\sum_{i=1}^{m}(\theta_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}-\mathbf{Z}'_{ij}\boldsymbol{\gamma}-b_i)^2+d_j\Big]\\
\Big[(\sigma_b^2)^{-1}|\mathbf{b}\Big] &\sim G\Big[\frac{m}{2}+c,\frac{1}{2}\sum_{i=1}^{m}b_i^2+d\Big]
\end{aligned}
$$

Here $G(a,b)$ denotes a gamma distribution with shape = a and rate = b.

The full conditional distribution of the parameters for the semiparametric random walk model will follow similarly as above. In this case, $\mathbf{v}$ and σ_v^2 will have normal and inverse gamma full conditionals respectively while the full conditionals of the other parameters will depend on $\mathbf{v}$.

Acknowledgements

The research was partially supported by National Science Foundation grant SES-0631426. The authors gratefully acknowledge William Bell of the Bureau of the Census for providing them with the SAIPE dataset and for many helpful suggestions and advice throughout the course of the work.

REFERENCES

1. Bell, W. R. (1999), "Accounting for Uncertainty About Variances in Small Area Estimation, " Conference Paper, U.S Bureau of Census.

2. Datta, G. S., Ghosh, M., Nangia, N., and Natarajan, K. (1996), "Estimation of Median Income of Four-Person Families : A Bayesian Approach," in *Bayesian Analysis in Statistics and Econometrics : Essays in Honor of Arnold Zellener*, eds. D. A. Berry, K. M. Chaloner, and J. K. Geweke, New York : Wiley, pp. 129-140.

3. Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-splines and penalties (with Discussion),"*Statistical Science*, 11, 89-121.

4. Fay, R. E. (1987), "Application of Multivariate Regression to Small Domain Estimation," in *Small Area Statistics*, eds. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, New York : Wiley, pp. 91-102.

5. Fay, R. E., Nelson, C. T., and Litow, L. (1993), "Estimation of Median Income for 4-Person Families by State," in *Indirect Estimators in Federal Programs*, Statistical Policy Working Paper 21, Washington, DC : Statistical Policy Office, Office of Management and Budget, pp. 901-917.

6. Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398-409.

7. Gelman, A. E., and Rubin, D. (1992), "Inference from Iterative Simulation (with discussion)," *Statistical Science*, 7, 457-511.

8. Ghosh, M., Nangia, N., and Kim, D. (1996), "Estimation of Median Income of Four-Person Families : A Bayesian Time Series Approach," *Journal of the American Statistical Association*, 91, 1423-1431.

9. Opsomer, J. D., Claeskens, G., Ranalli, M. G., and Breidt, F. J. (2008), "Non-parametric Small Area Estimation using Penalized Spline Regression," *Journal of the Royal Statistical Society, Series B*, 70, 265-286.

10. Rao, J. N. K. (2003), *Small Area Estimation*, Wiley Inter Science.

11. Ruppert, D.(2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational and Graphical Statistics*, 11, 735-757.

12. Ruppert, D., and Carroll, R. (1999), "Spatially-Adaptive Penalties for Spline Fitting," *Australian and New Zealand Journal of Statistics*, 42, 205-223.

13. Ruppert, D., Wand, M. P., and Carroll, R. (2003), *Semiparametric Regression*. Cambridge, U.K. : Cambridge University Press.

Table 1: Parameter Estimates of SPRWM with 5 Knots

Parameter	Mean	Median	95% CI
β_0	4677.71	4660.08	(4633.31, 4758.7)
β_1	0.8156	0.816	(0.814, 0.817)
γ_1	-0.154	-0.154	(-0.158, -0.149)
γ_2	0.02	0.024	(-0.016, 0.040)
γ_3	-0.008	-0.016	(-0.056, 0.066)
γ_4	-0.093	-0.119	(-0.127, -0.037)
γ_5	-0.165	-0.173	(-0.187, -0.139)

Table 2: Comparison Measures for SPM(5)* and SPRWM(5)* Estimates with Knot Realignment

Estimate	ARB	ASRB	AAB	ASD
CPS	0.0415	0.0027	1,753.33	5,300,023
SAIPE	0.0326	0.0015	1,423.75	3,134,906
SPM(5)*	0.028	0.0012	1173.71	2,334,379
SPRWM(5)*	0.0295	0.0013	1256.08	2,747,010

Table 3: Percentage Improvements of SPM(5)* and SPRWM(5)* Estimates over SAIPE and CPS Estimates

Estimate	Model	ARB	ASRB	AAB	ASD
SAIPE	SPM(5)*	14.11%	20.00%	17.56%	25.54%
	SPRWM(5)*	9.51%	13.33%	11.78%	12.37%
CPS	SPM(5)*	32.53%	55.55%	33.06%	55.96%
	SPRWM(5)*	28.92%	51.85%	28.36%	48.17%

Table 4: Parameter Estimates of SPM(5)*

Parameter	Mean	Median	95% CI
β_0	4767.48	4769.04	(4743.33, 4791.67)
β_1	0.811	0.810	(0.809, 0.812)
γ_1	-0.189	-0.191	(-0.198, -0.180)
γ_2	0.0389	0.0395	(0.0189, 0.059)
γ_3	0.104	0.102	(0.099, 0.126)
γ_4	-0.240	-0.253	(-0.305, -0.179)
γ_5	-0.127	-0.155	(-0.181, -0.081)

Table 5: Parameter Estimates of SPRWM(5)*

Parameter	Mean	Median	95% CI
β_0	4826.28	4824.39	(4806.77, 4860.56)
β_1	0.806	0.809	(0.801, 0.810)
γ_1	-0.159	-0.156	(-0.183, -0.151)
γ_2	0.014	0.012	(0.004, 0.039)
γ_3	0.08	0.08	(0.027, 0.123)
γ_4	-0.237	-0.244	(-0.369, -0.125)
γ_5	-0.225	-0.183	(-0.538, -0.085)

Table 6: Comparison Measures for Time Series and other Model Estimates

Estimate	ARB	ASRB	AAB	ASD
CPS	0.0415	0.0027	1,753.33	5,300,023
SAIPE	0.0326	0.0015	1,423.75	3,134,906
GNK	0.0397	0.0025	1709.58	5,229,869
SPM(0)	0.0337	0.0017	1408.7	3,137,978
SPM(5)*	0.028	0.0012	1173.71	2,334,379
SPRWM(5)*	0.0295	0.0013	1256.08	2,747,010

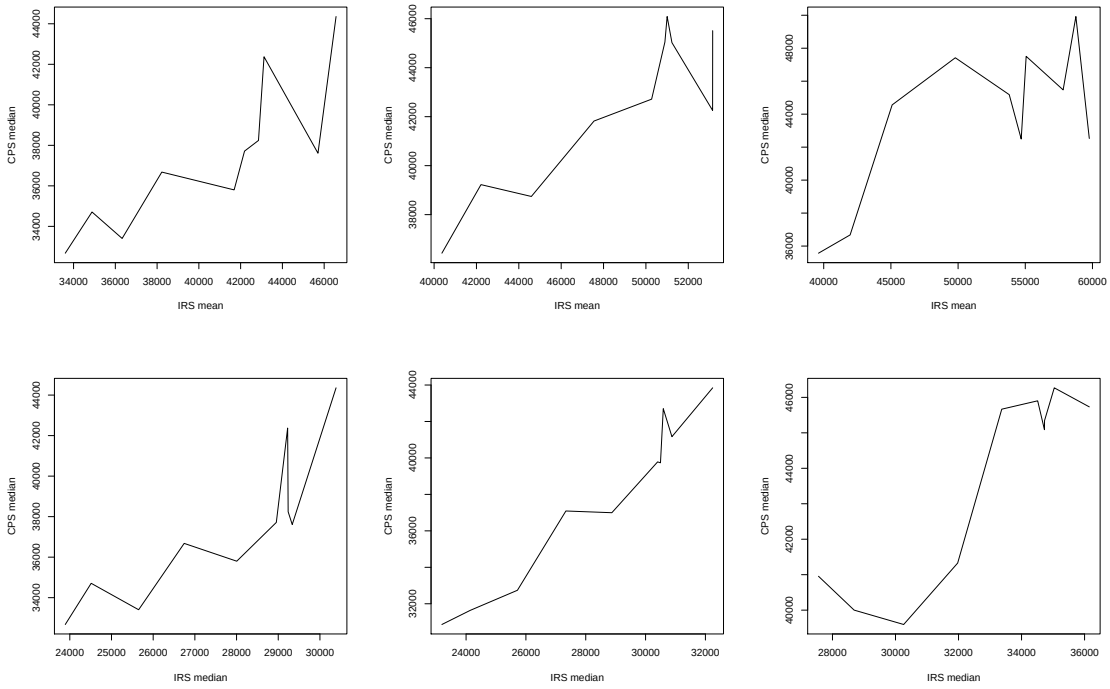
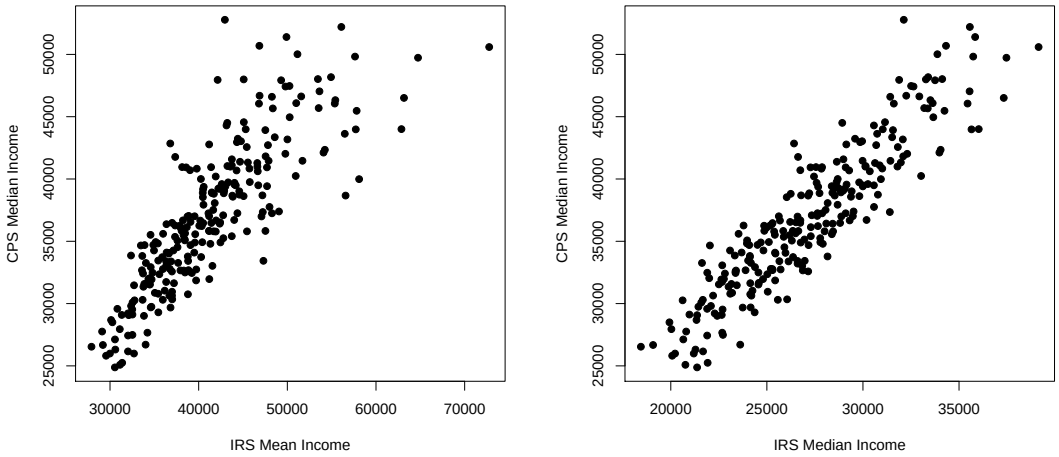
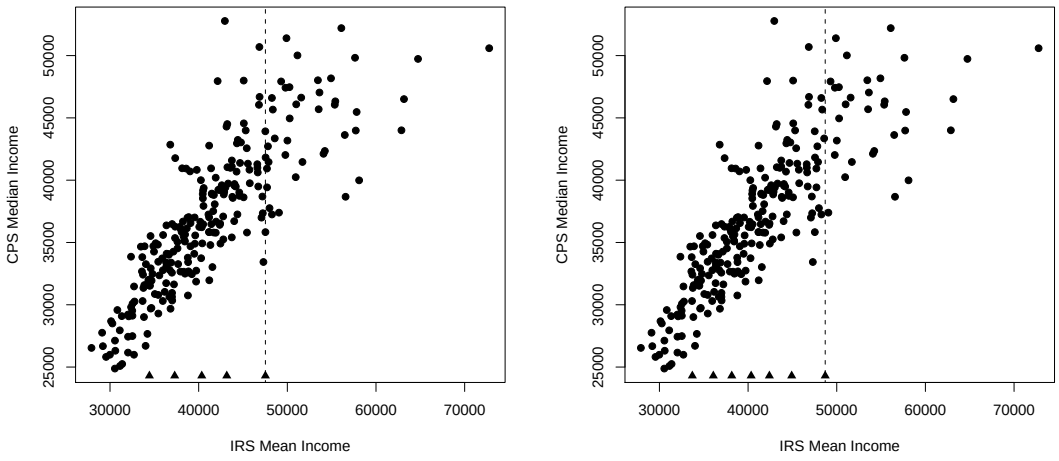


Figure 1: Longitudinal CPS median income profiles for 6 states plotted against IRS mean and median incomes. (1st row : IRS Mean Income; 2nd row : IRS Median Income).



(a) IRS mean income plot (b) IRS median income plot

Figure 2: Plots of CPS median income against IRS mean and median incomes for all the U.S states from 1995 to 1999.



(a) Positioning of 5 Knots (b) Positioning of 7 Knots

Figure 3: Exact positions of 5 and 7 knots in the plot of CPS median income against IRS mean income. The knots are depicted as the bold faced triangles at the bottom.

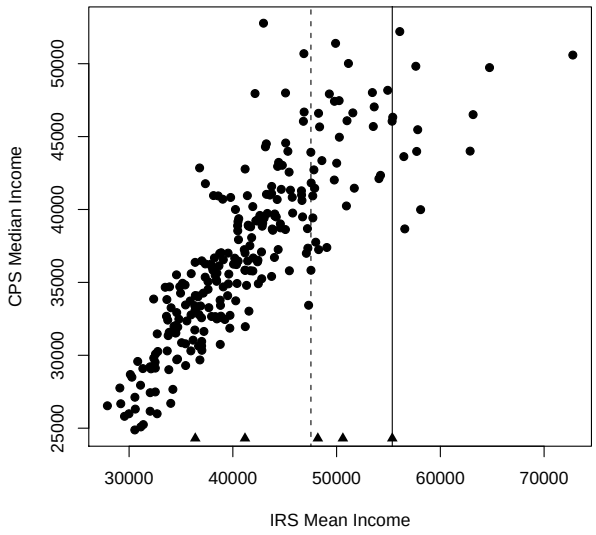
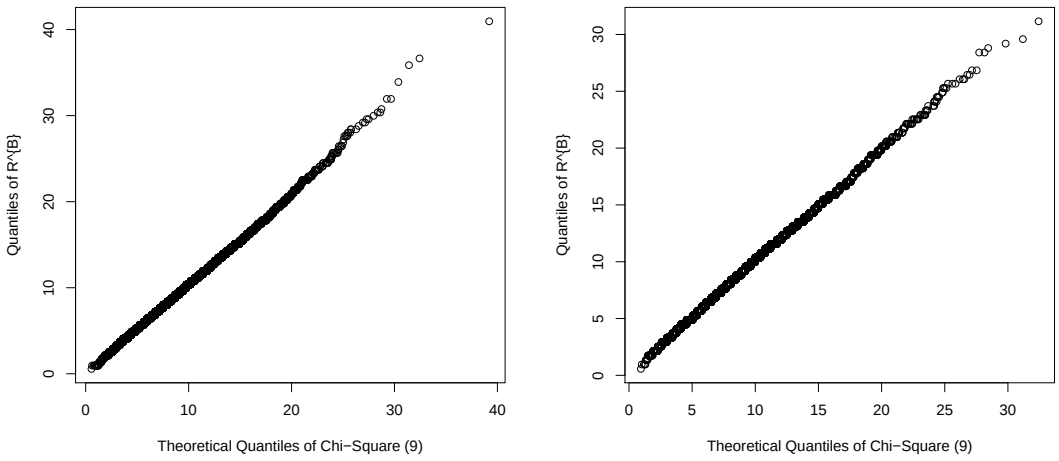


Figure 4: Positions of 5 knots after realignment. The knots are the bold faced triangles at the bottom. The region between the dashed and bold lines is the additional coverage area gained from the realignment.



(a) Basic Semiparametric Model

(b) Semiparametric RW Model

Figure 5: Quantile-quantile plot of R^B values for 10000 draws from the posterior distribution of the basic semiparametric and semiparametric random walk models. The X-axis depicts the expected order statistics from a χ^2 distribution with 9 degrees of freedom.