# Estimation of Median Incomes of Small Areas: A Bayesian Semiparametric Approach

Malay Ghosh
University of Florida
Joint work with D. Bhadra and D. Kim

August 13, 2011

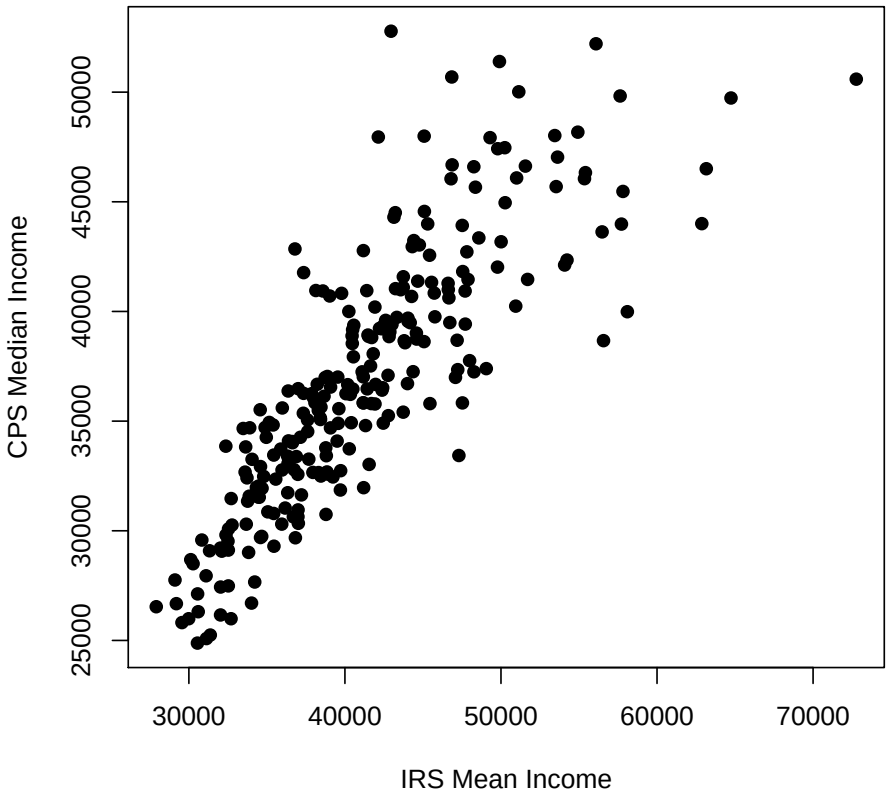
## Outline

- Introduction
- Semiparametric Modeling
- Hierarchical Bayesian Model
- Data Analysis
- Goodness of Fit Test
- Adaptive Knot Selection
- Summary and Conclusion

## Introduction

- Often observations on various characteristics of small areas are collected over time, and thus, may possess an underlying time-varying pattern.
- It is likely that models which exploit this pattern may perform better than those which do not utilize this feature.
- In this study, we present a semiparametric Bayesian framework for the analysis of small area data, while explicitly accommodating for the longitudinal pattern in the response and the covariates.

- Estimation of median household income of small areas is one of the principal targets of inference of the U.S Bureau of Census under its Small Area Income and Poverty Estimation (SAIPE) program.
- The above estimates play an important role in the administration of federal programs and allocation of federal funds to local jurisdictions.
- Since these estimates are collected over time, they often possess an underlying longitudinal pattern.
- In this talk, I will use the household income data for all the U.S states for the period 1995 through 1999 to estimate the true state specific median household income for 1999.
- Fig I. plots the CPS (Current Population Survey) median incomes against the IRS mean income for all the states spanning 1995-1999.


CPS Median Income
25000
30000
35000
40000
45000
50000
30000
40000
50000
60000
70000
IRS Mean Income

- The Small Area Income and Poverty Estimates (SAIPE) program of the U.S Census Bureau provides annual estimates of income and poverty statistics for all states, counties and school districts across the United States.
- They use the Fay-Herriot class of models (Fay and Herriot, 1979) in combining state and county estimates of poverty and income obtained from different sources.
- 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.

- Data: IRS median income and CPS state median household income estimates for 1995-1999. In addition, we have the 1999 state median household income estimates from 2000 census data.
- We have used data from CPS for the period 1995-1999 in order to estimate the state wide median household income for 1999.
- This is because, the most recent census estimates correspond to the year 1999 and these census values can be used for comparison purposes.

- Ghosh, Nangia and Kim (1996) proposed a Bayesian time series modeling framework to estimate the statewide median income of four-person families for 1989.
- Opsomer et.al (2008) pioneered the use of nonparametric regression methodology in small area estimation context.
- They combined small area random effects with a smooth, non-parametrically specified trend using penalized splines.
- They applied their model to analyze a non-longitudinal, spatial dataset concerning the estimation of mean acid neutralizing capacity (ANC) of lakes.

## Semiparametric Modeling

- The annual state specific median income estimates can be looked upon as a longitudinal profile or "income trajectory".
- Moreover, the median income estimates may possess an underlying non-linear pattern with respect to the covariates.
- These characteristics motivated us to use a semi-parametric modeling approach for our problem.
- Our main objective is to estimate the 1999 state median household income using a semi-parametric approach and to compare these estimates with the CPS as well as the SAIPE model based estimates.

- Sometimes the relationship between two variables is too complicated to be expressed using a known functional form.
- Non-parametric statistical methods uses the data, but not any prespecified function to determine the true underlying functional relationship between the variables.
- For example, suppose $Y$ and $X$ are related as $y_i = f(x_i) + e_i$, $i = 1, 2, ..., m$. where $e_i \sim N(0, \sigma_e^2)$ and $f(x)$ is unspecified.
- In a non-paramteric setting, $f(x)$ is often estimated using Penalized splines (P-splines).

- In the P-spline framework, $f(x)$ is represented as $f(x;\boldsymbol{\beta}) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{p+k}(x-\tau_k)_+^p$.
- Here, $p$ is the degree of the spline, $(x)_+^p = x^p I(x>0)$ and $(\tau_1 < \tau_2 < ... < \tau_K)$ is a fixed set of knots.
- The spline coefficients $(\beta_{p+1}, ..., \beta_{p+K})$ measure the jumps of the spline at the knots $(\tau_1, ..., \tau_K)$.
- Smoothness of the resulting fit is achieved by "penalizing" or restricting these jumps.
- Provided the knots are evenly spread out over the range of $x$, the functions $f(x;\boldsymbol{\beta})$ can accurately estimate a very large class of smooth functions $f(\cdot)$.

- Let $Y_{ij}$ be the sample survey estimators of some characteristics $\theta_{ij}$ for the $i^{th}$ small area at the $j^{th}$ time $(i = 1, 2, ..., m; j = 1, 2, ..., t)$.
- The inferential target is usually $\theta_{ij}$ or some function of it.
- In our context, $\theta_{ij}$ denotes the true median household income of the $i^{th}$ state at the $j^{th}$ year.
- We denote by $X_{ij}$, the covariate corresponding to the $i^{th}$ state and $j^{th}$ year.
- In our problem, $X_{ij}$ is the IRS mean income recorded for the $i^{th}$ state and $j^{th}$ year.

- Our basic semiparametric model (SPM) is

$$Y_{ij} = f(x_{ij}) + b_i + u_{ij} + e_{ij}.$$

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

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{K} \gamma_k (x_{ij} - \tau_k)_+ + b_i + u_{ij} + e_{ij} \\ &= \theta_{ij} + e_{ij}, \ i = 1, ..., m; j = 1, ..., t. \end{aligned}$$

- Here, $b_i$ is a state-specific random effect while $u_{ij}$ represents an interaction effect between the $i^{th}$ state and the $j^{th}$ year.
- $u_{ij}$ and $e_{ij}$ are assumed to be mutually independent with $u_{ij} \sim N(0, \psi_j^2)$ and $e_{ij} \sim N(0, \sigma_{ij}^2)$. $\sigma_{ij}^2$'s are assumed to be known.
- We assume $b_i \sim^{i.i.d} N(0, \sigma_b^2)$ and $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 I_K)$ where $\sigma_\gamma^2$ controls the amount of smoothing of the underlying income trajectory.
- Generally, the knots $(\tau_1, ..., \tau_K)$ are placed on a grid of equally spaced sample quantiles of $X_{ij}$'s.

- A second model, a semiparametric random walk model (SPRWM) introduces in addition a trend component (over time) in the model.
-

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

- Here, $v_j$ denotes the time specific random component.
- We assume $v_j | v_{j-1} \sim N(v_{j-1}, \sigma_v^2)$.
- An alternate representation is $v_j = v_{j-1} + w_j$, where $w_j \overset{iid}{\sim} N(0, \sigma_v^2)$.

## Hierarchical Bayesian Model

- Our Hierarchical Bayesian model is

1. $(Y_{ij}|\theta_{ij}) \overset{ind}{\sim} N(\theta_{ij}, \sigma_{ij}^2), \qquad i = 1, ..., m; j = 1, ...t$
2. $(\theta_{ij}|\boldsymbol{\beta}, \boldsymbol{\gamma}, b_i, \psi_j^2) \overset{ind}{\sim} N(\mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{Z}'_{ij}\boldsymbol{\gamma} + b_i, \psi_j^2)$
3. $b_i \overset{iid}{\sim} N(0, \sigma^2), \qquad i = 1, ..., m$
4. $\boldsymbol{\gamma} \sim N(\mathbf{0}, \sigma_\gamma^2 I_K)$

- We use a noninformative uniform improper prior for $\boldsymbol{\beta}$ while proper but diffuse inverse gamma priors for the variance parameters.
- We use Gibbs sampler in an MCMC framework to sample from the full conditionals of $\theta_{ij}$, our target of inference.

- Since we have used improper prior for $\boldsymbol{\beta}$, posterior propriety was proved before doing any computation.
- We use Gibbs sampler in MCMC framework to sample from the full conditionals of $\theta_{ij}$, our target of inference.
- 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.
- The first $d$ iterations of each chain are discarded and posterior summaries are calculated based on the rest of the $d$ iterates.

## Data Analysis

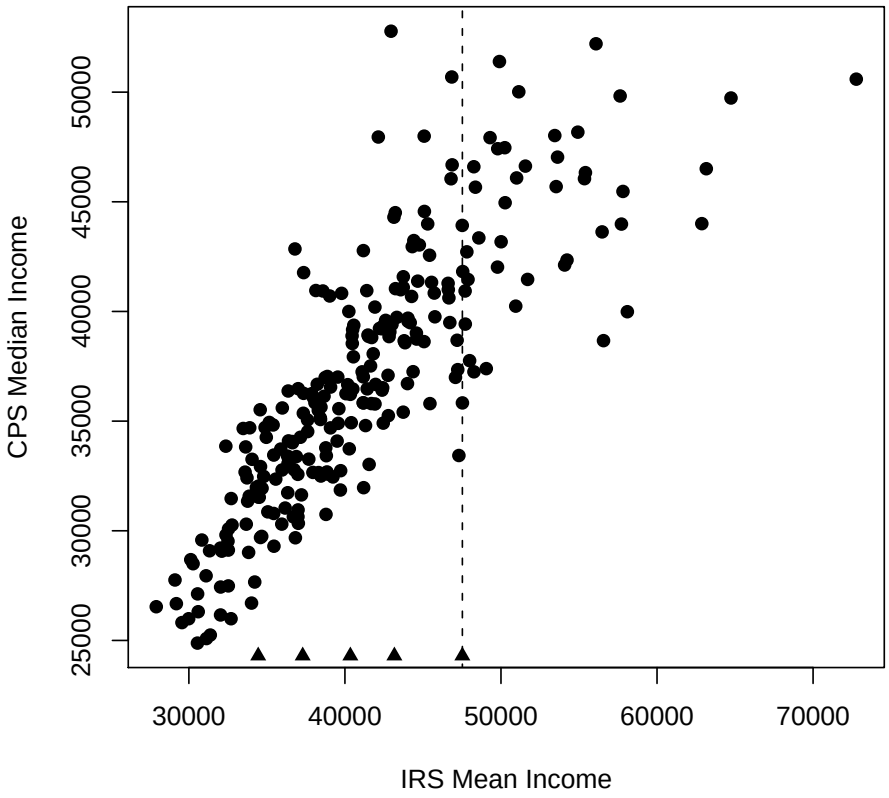
- We fitted the semi-parametric small area model (SPM) with IRS mean as predictor and varying number of knots.
- Decennial census values are used as the "gold standard".
- Comparison Measures:
- Average Relative Bias (ARB) $= \frac{1}{51}\sum_{i=1}^{51}\frac{|c_i - e_i|}{c_i}$;
- Average Squared Relative Bias (ASRB) $= \frac{1}{51}\sum_{i=1}^{51}\frac{|c_i - e_i|^2}{c_i^2}$;
- Average Absolute Bias (AAB) $= \frac{1}{51}\sum_{i=1}^{51}|c_i - e_i|$;
- Average Squared Deviation (ASD) $= \frac{1}{51}\sum_{i=1}^{51}(c_i - e_i)^2$.

The model with 5 knots in the income trajectory produced the best estimates.

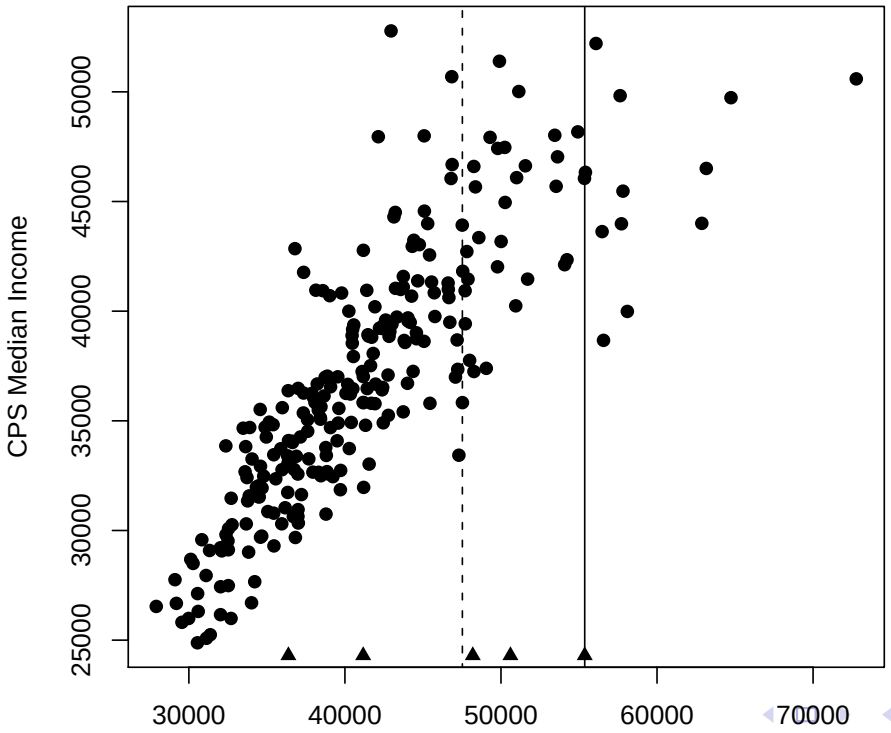
| Estimate | ARB | ASRB | AAB | ASD |
|---|---|---|---|---|
| CPS | 0.0415 | 0.0027 | 1,753.33 | 5,300,023 |
| SAIPE | 0.0326 | 0.0015 | 1,423.75 | 3,134,906 |
| SPM(5) | 0.0326 | 0.0016 | 1,398.46 | 3,287,368 |

The 95% CI for $\gamma_1$, $\gamma_4$ and $\gamma_5$ does not contain 0 indicating the significance of the first, fourth and fifth knots.

- It is clear that the SPM estimates are superior to the CPS median income estimate but they are equivalent to the SAIPE model estimates.
- In a semiparametric regression framework, 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.
- The exact positions of the 5 knots in our setup are shown in the following figure.


CPS Median Income
25000
30000
35000
40000
45000
50000
30000
40000
50000
60000
70000
IRS Mean Income

- It is clear that the knots mostly lie in the high density region of the graph while the non-linearity is mainly visible in the low density region.
- Thus, 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 following figure shows the new pattern.



- It is clear that a much larger proportion of observations has been captured with the knot realignment.
- The region between the bold and dashed vertical lines denotes the additional coverage that has been achieved with the knot rearrangement.
- Since the new coverage area overlaps the region of non-linearity, it seems that the new knots are able to capture the underlying non-linear pattern in the dataset which the old knots failed to achieve.
- On fitting the semiparametric model with the new knot alignment, we did achieve some improvement in the results as shown below.

| 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 | 0.0400 | 0.0025 | 1709.58 | 5,229,869 |
| SPM(5) | 0.0326 | 0.0016 | 1,398.46 | 3,287,368 |
| SPM(5)* | 0.0280 | 0.0012 | 1173.71 | 2,334,379 |
| SPRWM(5)* | 0.0300 | 0.0013 | 1256.08 | 2, 747,010 |

- The new comparison measures for SPM are quite lower than those of the SAIPE model.
- The percentage improvements of the SPM estimates over the SAIPE estimates are respectively 14.11%, 20%, 17.56% and 25.54%.
- This improvement is apparently due to the additional coverage of the observational pattern that is being achieved with the relocation of the knots.

## Goodness of Fit Test

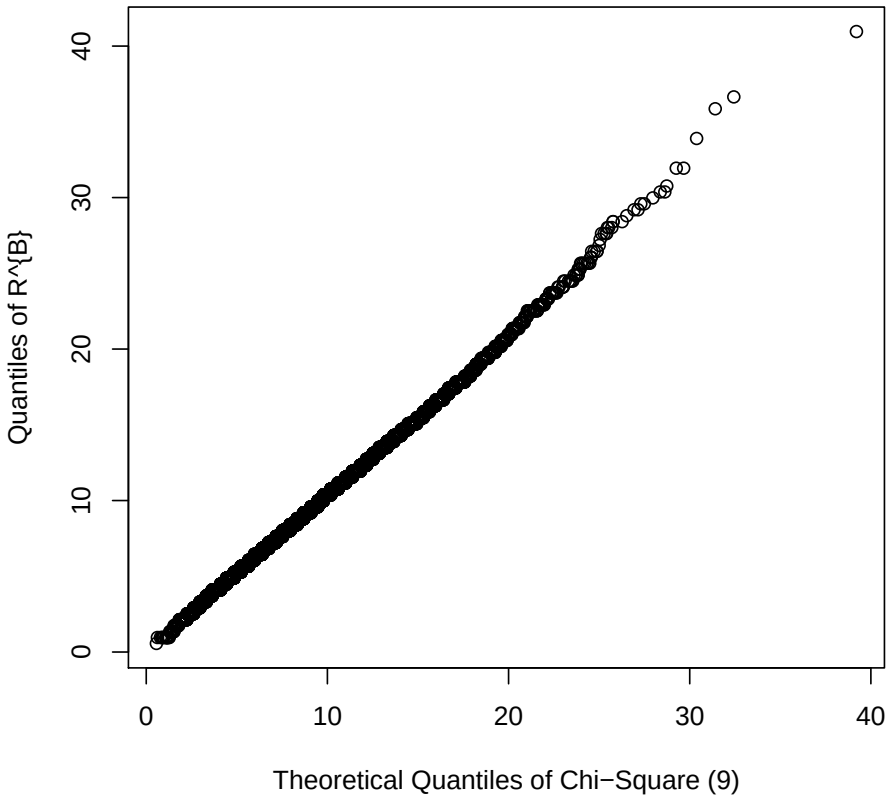
- To examine the goodness-of-fit of the semiparametric models, we used a Bayesian Chi-square goodness-of-fit statistic.
- This is an extension of the classical Chi-square goodness-of-fit test where the statistic is calculated at every iteration of the Gibbs sampler as a function of the parameter values drawn from the posterior distribution.
- We form 10 equally spaced bins $((k-1)/10, k/10)$, $k = 1, ..., 10$, with fixed bin probabilities, $p_k = 1/10$.
- 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. $F(y_{ij}|\theta_{ij})$.

- The Bayesian chi-square statistic is then calculated as

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

- Here $m_k(\tilde{\Theta})$ is the random bin count given the posterior sample $\tilde{\Theta}$.
- For model assessment, we use two summary measures. First one is the proportion of times the generated values of $R^B$ exceeds the 0.95 quantile of a $\chi^2_9$ 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 $\chi^2_9$ deviate. Values close to 0.5 would suggest a good fit.

- For the SPM, the above summary measures were respectively 0.049 and 0.5 indicating a good fit.
- The QQ plots of $R^B$ values shown below also demonstrate good agreement between the distribution of $R^B$ and that of a $\chi^2(9)$ random variable.



## Adaptive Knot Selection

- Recall the model
  $Y_{ij} = f(x_{ij}) + b_i + u_{ij} + e_{ij}$, where
  $f(x_{ij}) = \beta_0 + \beta_1 x_{ij} + \sum_{k=1}^{K}(x_{ij} - \tau_k)_+$.
- However, now we do not require either a fixed number or fixed locations of the knots.
- So now the posterior will involve two additional sets of parameters- (i) the number of knots and (ii) the locations of the knots.
- We consider the prior under which $K$, the number of knots, follows a Poisson distribution with some mean, say, $\mu$.
- Conditional on $K = k$, we consider the locations $\tau_1 < \tau_2 < \ldots < \tau_k$ as order statistics from a uniform $(a, b)$ distribution.
- In addition to sampling the earlier parameters, we now need to sample the the knot number and the knot locations $(k, \tau)$ at each iteration of the Gibbs sampler.

- As a result, the dimension of the parameter space changes at every iteration.
- Thus we need a Reversible Jump Markov chain Monte Carlo (RJMCMC) which accounts for varying dimension of the parameter space.
- The RJMCMC algorithm consists of three types of transition.
- Knot selection (birth step), Knot deletion (death step) and Knot relocation (relocation step).
- The probabilities for these moves are denoted by $b_k$, $d_k$ and $\zeta_k$ respectively.
- $b_k = c \min\{1, \frac{\pi_{k+1}}{\pi_k}\}$, $d_k = c \min\{1, \frac{\pi_{k-1}}{\pi_k}\}$, $\zeta_k = 1 - b_k - d_k$.
- $\pi_k = \exp(-\mu)\mu^k/k!$.
- $c$ is a preassigned constant.

- $M_\tau(k) = \{k, \tau_1, \ldots, \tau_k\}$: current model as identified by $k$ and $\tau$.
- Birth step: $M_\tau(k) \to M_\tau(k+1)$ with prob. $b_k$.
- Death step: $M_\tau(k) \to M_\tau(k-1)$ with prob. $d_k$.
- Relocation $M_\tau(k) \to M_\tau(k^*)$ with prob. $\zeta_k$.

## Summary and Conclusion

- Information on 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.
- The contribution of the knots towards deciphering the underlying observational pattern improved substantially when they were placed with an optimal coverage area.
- Our final estimates proved to be superior, to both the CPS estimates, and also to the current U.S Census Bureau (SAIPE) estimates.

- Some possible extensions are as follows :
- The state specific deviations can be modeled as unspecified nonparametric functions instead of just a random intercept.
- If the median income pattern have varying degree of smoothness, spatially adaptive smoothing procedures can be used.
- Other kinds of basis functions like B-splines or radial bases etc can also be used to model the income trajectory.
- Instead of a parametric normal distributional assumption for the random effects, a broader class of distributions like the Dirichlet process or Polya trees may be tested.

- The theorem for posterior propriety is as follows :
- Theorem. 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 (i) $(m - p - 5)/2 + c_k > 0$ and $d_k > 0$ and (ii) $m/2 + c_j - 2 > 0$ and $d_j > 0$, $j = 1, ..., t; j \neq k$.