NISS

Bayesian Multiscale Multiple Imputation with Implications to Data Confidentiality

Scott H. Holan, Daniell Toth,
Marco A.R. Ferreira and Alan F. Karr

Technical Report Number 171
November 2008

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Bayesian Multiscale Multiple Imputation with Implications to Data Confidentiality

Scott H. Holan[1], Daniell Toth[2], Marco A. R. Ferreira[3] and Alan F. Karr [4]

Abstract

Many scientific, sociological and economic applications present data that are collected on multiple scales of resolution. One particular form of multiscale data arises when data are aggregated across different scales both longitudinally and by economic sector. Frequently, such data sets experience missing observations in a manner that they can be accurately imputed using the method we propose known as *Bayesian multiscale multiple imputation*. This method borrows information both longitudinally and across different levels of aggregation to produce accurate imputations of missing observations as well as estimates that respect the constraints imposed by the multiscale nature of the data. Our approach couples dynamic linear models with a novel imputation step based on singular normal distribution theory. Although our method is of independent interest, one important implication of such methodology is its potential effect on confidential databases protected by means of cell suppression. In order to demonstrate the proposed methodology and to assess the effectiveness of disclosure practices in longitudinal databases, we conduct a large scale empirical study using the U.S. Bureau of Labor Statistics Quarterly Census of Employment and Wages (QCEW). During the course of our empirical investigation it is determined that several of the predicted cells are within 1% accuracy, thus causing potential concerns for data confidentiality.

KEY WORDS: Cell suppression; Disclosure; Dynamic linear models; Missing data; Multiscale modeling; QCEW.

[1](to whom correspondence should be addressed) Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100, holans@missouri.edu

[2]Office of Survey Methods Research, Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 1950, Washington, D.C. 20212, Toth.Daniell@bls.gov

[3]Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO, 65211-6100, ferreiram@missouri.edu

[4]National Institute of Statistical Sciences, 19 T.W. Alexander Drive, Research Triangle Park, NC 27709-4006, karr@niss.org

1 Introduction

Given the public's concerns about data confidentiality there is an ever-increasing need for identifying and controlling disclosure risks. Typically, disclosure risks arise in the course of disseminating microdata on individual units, such as people or establishments, to researchers or other statistical agencies. In fact, statistical agencies often face conflicting missions. On the one hand they seek to release data suitable for a wide range of statistical analyses, while on the other hand they wish to protect the confidentiality of their respondents. Agencies that fail to protect confidentiality may face serious consequences, including legal action. Moreover, the statistical agency may lose public trust and thus create an atmosphere in which respondents are less willing to participate in studies or to provide accurate information (Gomatam et al. 2005).

To reduce disclosure risk, statistical agencies often alter the data prior to its release. For example, it is common for agencies to perturb, coarsen or swap data values prior to release (Willenborg and De Waal 2001). However, decreasing risk necessarily also decreases data utility, and increasingly statistical disclosure limitation (SDL) techniques are employed that explicitly account for risk-utility tradeoffs.

One particular path to disclosure is through linkages across multiple databases. In particular, when agencies release microdata to the public it may be possible for "intruders" to link records across databases in such a way as to compromise the confidentiality of the data (Fienberg 2006). Failure to release data in ways that prevent such identifications may be a breach of law and may discredit the statistical agency involved (Reiter 2005). As databases become more extensive and record linkage techniques improve, it is possible that releasing microdata may no longer be feasible. Under these circumstances, a viable alternative is to release only data summaries. Unfortunately, this type of release is often less useful for complex analyses and may still suffer from disclosure risks (Dobra, Karr, Sanil and Fienberg 2002; Dobra, Karr and Sanil 2003).

Another approach for protecting against disclosure is to release synthetic data (i.e., sim-

ulated microdata). Although synthetic data may have low risk of disclosure, they have correspondingly reduced utility. In this context both risk and utility depend on the model used for synthesis (see Gomatam et al. 2005 and the references therein).

An alternative framework for protecting against disclosure is to release only the results of statistical analyses of the data, with no release of microdata. Remote analysis servers would permit users to submit requests for analyses and be provided some form of output (for instance, estimated parameters and standard errors) (Keller-McNulty and Unger 1998; Duncan and Mukerjee 2000; Schouten and Cigrang 2003). Such servers are not free from risk of disclosure. In fact, it may be possible for intruders to discover identities or other attributes of interest through "targeted" queries (Gomatam et al. 2005).

Despite the multiplicity of SDL methods available to statistical agencies, it is still common practice within many surveys to protect against disclosure through the use of "cell suppression": cell entries in tables that are deemed risky (usually because they represent only a few data subjects) are simply suppressed. One example is in the Bureau of Labor Statistics' (BLS) Quarterly Census of Employment and Wages (QCEW). In order to protect against disclosure risks that arise from additive relationships within a table, additional, "secondary" cell suppressions are required. For a comprehensive discussion regarding data confidentiality as it pertains to QCEW, see Section 2 and Cohen and Li (2006).

Optimal cell suppression is an NP-hard problem, and most implemented algorithms rely on heuristics. Assuming that all of the risks of disclosure are accounted for through primary and secondary cell suppressions is problematic, as unforeseen disclosure risks may remain. This is especially true for complex data releases where there are both multiscale aggregations (for example, to both county and state levels, or to both fine- and coarse-grained industry classifications) and longitudinal data (whether from panels or repeated cross-sectional data collections). Together these data attributes potentially enable a data intruder to estimate the values of suppressed cells more accurately than might be anticipated.

In this paper, we propose a method of Bayesian multiscale multiple imputation that uti-

lizes the additive relationships (multiscale attributes) along with inherent serial correlation to impute missing values. While the method is of independent interest as a means of imputing missing data, this paper focuses on how it can be used to improve understanding of disclosure risk associated with cell suppression on longitudinal, multiscale data. Possibly disconcertingly, the framework can be extremely effective. In many instances, we are able to impute suppressed cells to within 1% accuracy. Moreover, the imputed values simultaneously respect the constraints imposed through the multiscale properties of the data. In addition, the Bayesian framework provides measures of uncertainty for the imputed values, which might not be true of other methods,

Our approach couples dynamic linear models (DLMs) (West and Harrison 1997) with multiple imputation techniques through the use of properties for normally distributed random variables with singular covariance matrices (Marsaglia 1964). As noted above, the method proposed here is quite general and can be modified to handle a wide array of multiscale (constrained) data structures. Our framework produces estimates of missing values that are close to the true unobserved values, but is also capable of producing estimates of trend, seasonality and regression effects along with associated measures of uncertainty. Further, the method is computationally feasible, and can be implemented in practical situations. Finally, the method requires no specific unknown "problem-specific" parameters. Instead, we employ a flexible set of default priors that require little or no subjective specification of problem specific parameters up to choice of the particular DLM. In fact, the only assumption made in the prior specification is that the signal-to-noise ratio is moderate.

A related approach proposed by Ansley and Kohn (1983) uses a method for computing the exact likelihood of a vector autoregressive-moving average process with missing or aggregated data. The two approaches differ in several respects. Most notably, our approach couples the flexibility of DLMs with properties of normally distributed random variables with singular valued covariance matrices. This produces a versatile framework that allows us to take advantage of, rather than be hampered by, the constraints present in the data. In the

Ansley and Kohn framework, by contrast, imputation in our context is impossible, at least without modification of their methodology or substantial bookkeeping on the part of the practitioner to eliminate redundant information. The multiscale aspect of our approach is crucial: the singular value covariance matrix allows us to systematically accommodate any redundant information present in the data in a mathematically rigorous and fully automatic manner.

The remainder of this paper is organized as follows. Section 2 provides a brief description of the Quarterly Census of Employment and Wages (QCEW). In Section 3 our method is formally developed and an illustration to the QCEW is provided. Section 4 quantifies the performance of our method through a large empirical study. Specifically, we apply our method to 11 real QCEW data sets. This empirical study demonstrates the effectiveness of our methodology and in doing so exposes the vulnerability of "cell suppression" as a method for eliminating disclosure risk in longitudinal databases. Finally, Section 5 concludes.

2 QCEW - Data Structure

The BLS conducts a census that collects data under a cooperative program between BLS and the State Employment Security Agencies known as the Quarterly Census of Employment and Wages (QCEW). The data contained in QCEW consist of broad employment and wage information for all U.S. workers covered by the Unemployment Compensation for Federal Employee program. Tabulations of QCEW outcomes are available by 6-digit North American Industrial Classification Systems (NAICS) industry, county, ownership, and size groups under several formats such as BLS Internet ftp servers. The detailed coverage and easy accessibility make it especially vulnerable to confidentiality disclosure risks (Cohen and Li 2006). To protect this tabular data against disclosure risks, Cell Suppression (CS) is imposed. Although the BLS consistently applies both primary and secondary cell suppressions, additional risks arise from additive relationships in the table along with serial correlation. As noted in Section 1, the problem is NP-hard (Kelly 1990), and several solutions have been proposed

(see Cohen and Li (2006) and the references therein).

As a matter of practice the CS problem and its solutions are addressed contemporaneously. This shortcoming increases the data's susceptibility to attack. In general, the QCEW data contains many different levels of aggregation and patterns of suppression. For example, suppose we have six years of quarterly data for three series and the aggregate of the three series. Let y_{jt} denote the j^{th} sub-series $j = 1, \ldots, k$ and t^{th} quarter $t = 1, \ldots, T$. Here k denotes the number of aggregate sub-series and T denotes the number of quarters; in our example we have $k = 3$ and $T = 24$. In some years two or more quarterly values are missing (i.e., primary and secondary cell suppressions) for two of the three series, but the aggregate value is often present for all quarters, so for each quarter $t = 1 \ldots 24$ we have either the full set of values $\mathbf{y}_t = (y_{1t}, y_{2t}, y_{3t})'$ or a set where some of the values have been suppressed, as for example $(\mathcal{S}, y_{2t}, \mathcal{S})$ in a quarter where the first and third series values are suppressed, as indicated by the letter $\mathcal{S}$. In addition, for many series we have annual totals for all three series. Let $q_t = y_{1t} + y_{2t} + y_{3t}, t = 1, \ldots, T$, be the total for quarter t. Further, let $\mathbf{a}_{t'} = (a_{1t'}, a_{2t'}, a_{3t'})'$ denote the annual totals for each of the six years, $t' = 1 \ldots 6$, where $a_{jt'} = y_{j\,(4t'-3)} + y_{j\,(4t'-2)} + y_{j\,(4t'-1)} + y_{j\,(4t')}, j = 1, \ldots, 3$. Then the complete time series is given by $\{\mathbf{y}_1, q_1, \mathbf{y}_2, q_2, \mathbf{y}_3, q_3, \mathbf{y}_4, q_4, \mathbf{a}_1, \mathbf{y}_5, q_5, \ldots, \mathbf{y}_{24}, q_{24}, \mathbf{a}_6\}$. However, in our case, we do not have the complete time series because some of the observations have been suppressed; an example of this is shown in Tables 1 and 2.

It is important to note that the data displayed in Tables 1 and 2 only constitute two example data sets from QCEW. In many instances the suppressed cells can be an annual total (i.e. Table 2) or even an aggregate total (not displayed). Additionally, the multiscale nature can have an aggregate along with k sub-series where k does not necessarily equal 3; in fact, we only require $k \geq 2$. Further, each of the k-sub-series can be an aggregate of l_k ($l_k \geq 2$) additional sub-series. Nevertheless, the framework we propose effectively accommodates these different data structures.

3 Multiscale Multiple Imputation

In recent years, multiple imputation, the practice of "filling in" missing data with plausible values, has emerged as powerful tool for analyzing data with missing values. More formally, multiple imputation (MI) refers to the procedure of replacing each missing value by a vector of imputed values. Upon completion of the imputation, standard-complete data methods can be used to analyze each data set. In addition, when $D \geq 2$ sets of imputation are formed and constitute repeated draws from the predictive distribution of the missing values under a specified model then the D complete-data sets can be combined to form one inference that properly accounts for the uncertainty due to nonresponse under that model. For a comprehensive discussion see Little and Rubin (2002) and the references therein.

Bayesian approaches to MI have experienced increased popularity due to their usefulness in complicated realistic problems. Rubin (1987) describes methods for generating MIs using parametric Bayesian models in the context of simple problems. In general, suppose that $Y = (Y_{obs}, Y_{mis})$ follows a parametric model $P(Y|\theta)$ where θ has a prior distribution and the missing data mechanism for Y_{mis} is ignored. Then we can write

$$P(Y_{mis}|Y_{obs}) = \int P(Y_{mis}|Y_{obs}, \theta)P(\theta|Y_{obs})d\theta.$$

Imputation for Y_{mis} can be obtained through a two step procedure. The first step is to sample the unknown parameters from their observed-data posterior $\theta^* \sim P(\theta|Y_{obs})$. Then given θ^*, the next step is to sample Y_{mis} from their conditional predictive distribution

$$Y^*_{mis} \sim P(Y_{mis}|Y_{obs}, \theta^*).$$

Typically this approach is facilitated by taking advantage of MCMC algorithms. For further discussion on MI see Schafer (1999).

Treating the suppressed data as missing and the additive structure as a multiscale problem provides a powerful environment for conducting multiscale multiple imputation. However, longitudinal - multiscale data inherently produce redundant information. Thus, to

systematically accommodate and take advantage of these redundancies, without substantial bookkeeping on the part of the practitioner, requires innovative methods involving singular value covariance matrices. This section formally develops the Bayesian multiscale multiple imputation scheme and provides an illustration by applying our method on two representative QCEW series. The main point of this illustration is to demonstrate our approach through two detailed examples. Subsequently we provide a comprehensive assessment of our methods performance in Section 4.

3.1 The multiscale multiple imputation scheme

The Bayesian multiscale multiple imputation scheme can be viewed as a two stage iterative procedure. In the first stage all of the sub-series (i.e. all of the series other than the aggregate series) are modeled individually, conditional on the missing values, using DLMs (West and Harrison 1997). Thus considering the example series in Section 2 we have 3 DLMs each modeling a series of 6 years of quarterly data (excluding the annual totals). It is important to note that although we model each series individually our procedure can be modified in a straight forward manner to include correlation between series. However, this is typically unnecessary as much of the between series correlation is accounted for through the multiscale (aggregation) constraints. The second step of our procedure performs imputation of the missing values for each years worth of data after accounting for all of the additive constraints.

Formally our procedure proceeds as follows. We assume that the complete data $\mathbf{y}_1, \ldots, \mathbf{y}_T$ follow a general linear state-space model which can be written as (West and Harrison 1997)

$$\begin{aligned} \mathbf{y}_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \boldsymbol{\varepsilon}_t, & \boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_t), \\ \boldsymbol{\theta}_t &= \mathbf{G_t}\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t & \boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_t). \end{aligned}$$

The first equation is known as the observation equation and the second equation is known as the system equation. In this context $\boldsymbol{\theta}_t$ is the latent process, $\mathbf{F}_t$ relates the observations to

the latent process, $\mathbf{G}_t$ describes the evolution of the latent process through time and $\mathbf{V}_t$ and $\mathbf{W}_t$ are the observational covariance matrix and covariance matrix of the system equation innovation respectively. The general state-space model has become commonplace in the time series literature owing to its versatility in accommodating a wide array of data structures such as seasonality and regression effects among others. For a comprehensive discussion on state-space models see Durbin and Koopman (2001), Harvey (1989), West and Harrison (1997) and the references therein.

Typically $\mathbf{F}_t$, $\mathbf{G}_t$, $\mathbf{V}_t$ and $\mathbf{W}_t$ are known up to a few hyperparameters as is the case in the models we employ for illustration. As such estimation can be performed using Markov chain Monte Carlo (MCMC) (Robert and Casella 2004; Gamerman and Lopes 2006). Each iteration of the MCMC algorithm is then divided into three blocks: simulation of the unknown hyperparmeters, simulation of the latent process, and simulation (imputation) of the missing values. The details of the simulation of the hyperparameters is model-specific while the latent process can be efficiently simulated using the forward filter backward sampler (FFBS) (Früwirth-Schnatter 1994; Carter and Kohn 1994).

In our particular case, $\mathbf{y}_t$ contains k sub-series related to different economic sectors. In order to model the joint evolution of those sub-series through time, $\mathbf{F}_t'\boldsymbol{\theta}_t$ may contain regression terms, seasonality, first- and second-order trends, common latent factors, etc. However, in our experience, we have noticed that many of those terms are already captured by the aggregated series and are automatically accounted for when we sample the missing data conditional on the aggregated series. For this reason, for the remainder of this paper we assume y_{jt}, $j = 1, \ldots, k$, follow a first-order DLM. Specifically, we have

$$\begin{aligned} y_{jt} &= \theta_{jt} + \varepsilon_{jt} \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma_j^2), \\ \theta_{jt} &= \theta_{j,t-1} + \omega_{jt} \qquad \omega_t \sim \mathcal{N}(0, W_j). \end{aligned} \tag{1}$$

This model can be thought of as a first order Taylor approximation of a smooth function representing the time trend of the series. Typically the variances σ_j^2 and W_j are strongly correlated *a posteriori*. Therefore it is often computationally beneficial to reparametrize W_j

in terms of a signal-to-noise ratio. In this direction, we define $W_j = \xi_j \sigma_j^2$ and as a result, the hyperparameters σ_j^2 and ξ_j will be much less correlated *a posteriori*. This reduction in correlation helps both in terms of speed of convergence of the MCMC algorithm and in terms of choosing a prior distribution for the hyperparameters. Finally, the model (1) is completed with a prior $\theta_0 \sim \mathcal{N}(a, R)$, where a and R are user defined and usually taken to be noninformative.

The next step in estimation is the imputation step. Let $\mathbf{z}_{t'}$ denote the observations and their aggregates for year t'. Assuming $k = 3$ then $\mathbf{z}_{t'} = (y_{1,4t'-3}, \ldots, y_{1,4t'}, y_{2,4t'-3}, \ldots, y_{2,4t'}, y_{3,4t'-3}, \ldots, y_{3,4t'}, q_{4t'-3}, \cdots, q_{4t'}, a_{1t'}, a_{2t'}, a_{3t'})'$. Further, let $\boldsymbol{\theta}^*_{t'} = (\theta_{1,4t'-3}, \ldots, \theta_{1,4t'}, \theta_{2,4t'-3}, \ldots, \theta_{2,4t'}, \theta_{3,4t'-3}, \ldots, \theta_{3,4t'})'$and $\mathbf{H}$ denote the matrix that operates on the individual observations and returns the individual observations and along with the several aggregate totals. Then it follows, from (1), that

$$\mathbf{z}_{t'} | \boldsymbol{\theta}^*_{t'} \sim \mathcal{N}(\boldsymbol{\mu}_{t'}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu}_{t'} = \mathbf{H}\boldsymbol{\theta}^*_{t'}$ and $\boldsymbol{\Sigma} = \mathbf{H}\mathbf{V}\mathbf{H}'$, with $\mathbf{V} = \text{diag}(\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_3^2, \sigma_3^2, \sigma_3^2, \sigma_3^2)$. For example, in the case considered in Section 2 (Table 1)

$$\mathbf{H} = \begin{pmatrix} & \mathbf{I}_{12} & \\ \mathbf{I}_4 & \mathbf{I}_4 & \mathbf{I}_4 \\ & \mathbf{I}_3 \otimes \mathbf{1}_4' & \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product, $\mathbf{I}_m$ denotes the $m \times m$ identity matrix and $\mathbf{1}_m$ is the vector of ones having length m; thus $\mathbf{H}$ has dimension 19×12.

Typically, several elements of $\mathbf{z}_{t'}$, either individual or aggregated values, may have been suppressed; let $\mathbf{z}_{t',o}$ and $\mathbf{z}_{t',m}$ be the observed and missing values of $\mathbf{z}_{t'}$, respectively. Then, the covariance matrix $\boldsymbol{\Sigma}$ can further be partitioned in terms of the missing and observed values. Specifically, define $\boldsymbol{\Sigma}_{mm}$, $\boldsymbol{\Sigma}_{mo} = \boldsymbol{\Sigma}'_{om}$ and $\boldsymbol{\Sigma}_{oo}$ to be the covariance matrix of the missing data, the missing data with the observed and of the observed data respectively. Then

the covariance matrix $\mathbf{\Sigma}$ can be written

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{oo} & \mathbf{\Sigma}_{om} \\ \mathbf{\Sigma}_{mo} & \mathbf{\Sigma}_{mm} \end{pmatrix}.$$

Further, consider the spectral decomposition of $\mathbf{\Sigma}_{oo}$; that is, $\mathbf{\Sigma}_{oo} = \mathbf{PDP}'$ where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p]$ has orthonormal columns given by the normalized eigenvectors of $\mathbf{\Sigma}_{oo}$ and $\mathbf{D} = \text{diag}(d_1, \ldots, d_p)$, with $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ corresponding to the eigenvalues of $\mathbf{\Sigma}_{oo}$. In addition, let q denote the number of zero eigenvalues of $\mathbf{\Sigma}_{oo}$. Note that q is equal to the number of redundancies found in the observed data due to having knowledge about the aggregated values. In order to eliminate these redundancies define $\mathbf{D}^* = \text{diag}(d_1, \ldots, d_{p-q})$ to be the diagonal matrix with diagonal equal to the positive eigenvalues of $\mathbf{\Sigma}_{oo}$ and $\mathbf{P}^* = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{p-q}]$ the matrix of corresponding normalized eigenvectors. Then the *pseudoinverse*, also known as the Moore-Penrose inverse (Searle 1992), can be computed as $\mathbf{\Sigma}_{oo}^{+} = \mathbf{P}^*(\mathbf{D}^*)^{-1}\mathbf{P}^{*\prime}$. Ultimately to impute the missing (suppressed) data we need to find the conditional distribution of missing values given the observed values. Following Marsaglia (1964) we have that

$$\mathbf{z}_{t',m} | \mathbf{z}_{t',o} \sim \mathcal{N}(\boldsymbol{\gamma}_{t,m}, \mathbf{\Omega}_m), \tag{2}$$

where

$$\boldsymbol{\gamma}_{t,m} \;=\; \boldsymbol{\mu}_{t',m} - \mathbf{\Sigma}_{mo}\mathbf{\Sigma}_{oo}^{+}(\mathbf{z}_{t',o} - \boldsymbol{\mu}_{t',o}) \tag{3}$$

and

$$\mathbf{\Omega}_m = \mathbf{\Sigma}_{mm} - \mathbf{\Sigma}_{mo}\mathbf{\Sigma}_{oo}^{+}\mathbf{\Sigma}_{om}. \tag{4}$$

Remark 1 In the case were $\mathbf{\Sigma}_{oo}$ is full rank, (2), (3) and (4) reduce to the familiar formulas from the standard theory of multivariate normal distributions (Mardia, Kent and Bibby 1979).

Remark 2 Alternatively, one could eliminate the redundant information by eliminating some redundant elements from $\mathbf{z}_{t',o}$, but this would require substantial bookkeeping. Conversely, our procedure is fully automatic.

Remark 3 Usually, the covariance matrix $\mathbf{\Omega}_m$ is singular. In order to simulate from (2), we first compute the spectral decomposition $\mathbf{\Omega}_m = \mathbf{P_\Omega D_\Omega P_\Omega}'$. Let r be the rank of $\mathbf{\Omega}_m$. Additionally, let $\mathbf{D}^*_\Omega$ be the diagonal matrix with diagonal equal to the positive eigenvalues of $\mathbf{\Sigma}_{oo}$ and $\mathbf{P_\Omega}^*$ the matrix of corresponding eigenvectors. The next step is to simulate $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_r)$. Finally, a vector simulated from (2) is computed as $\boldsymbol{\gamma}_{t,m} + \mathbf{P_\Omega}^*(\mathbf{D}^*_\Omega)^{1/2}\mathbf{u}$.

Estimation of the model and the multiscale imputation is performed using MCMC in a fully Bayesian analysis. In this direction we need to assign prior distributions for the signal-to-noise ratio ξ_j and variance parameters σ_j^2 $(j = 1, \ldots, k)$ described above. First, we note that the signal-to-noise ratio parameters ξ_j are most likely small. Otherwise, the components of the latent process would vary too much over time and ultimately this would make it difficult to predict the suppressed cells. As a result, we expect ξ_j to be significantly smaller than 1. Therefore we assume that the prior distribution for each ξ_j is $IG(\alpha_j, \beta_j)$ with density

$$f(\xi_j) \propto \xi_j^{-(\alpha_j+1)} \exp\left(-0.5\frac{\beta_j}{\xi_j}\right),$$

where α_j and β_j are fixed *a priori* such that there is high probability that ξ_j is less than 0.3. Finally, we assume that $\sigma_j^2 \sim IG(\tau_j, \kappa_j)$, with $\tau_j = \kappa_j = .01$, $j = 1, \ldots, k$, which is a noninformative conjugate prior for σ_j^2 in this context.

In order to explore the posterior distribution, we use the Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990). This requires the full conditional distributions for ξ_j and σ_j^2, $j = 1, \ldots, k$, which are both of standard form. Specifically, the parameter $\xi_j|\boldsymbol{\theta}_{jt}, \sigma_j^2 \sim IG(\alpha_j^*, \beta_j^*)$ where $\alpha_j^* = \alpha_j + (T-1)/2$ and

$$\beta_j^* = \beta_j + 0.5\sum_{t=2}^{T}(\theta_{jt} - \theta_{j,t-1})^2/\sigma_j^2,$$

and T denotes the length of the series (in our example $T = 24$). Sampling the parameter σ_j^2 is equally straightforward as the full conditional for $\sigma_j^2|\xi_j, \mathbf{y}_{jt}, \boldsymbol{\theta}_{jt}$ is $IG(\tau_j^*, \kappa_j^*)$ where

$\tau_j^* = \tau_j + (2T-1)/2$ and

$$\kappa_j^* = \kappa_j + \sum_{t=1}^{T}(\mathbf{y}_{jt} - \boldsymbol{\theta}_{jt})^2 + .5\sum_{t=2}^{T}(\boldsymbol{\theta}_{jt} - \boldsymbol{\theta}_{j,t-1})^2/\xi_j.$$

Simulation of $\boldsymbol{\theta}_{jt}$ is performed with the usual FFBS as introduced and described in Carter and Kohn (1994) and Früwirth-Schnatter (1994). This step is fairly standard, therefore we omit the exact equations for the sake of brevity. For a comprehensive discussion, see Gamerman and Lopes (2006).

As we have seen, the overall algorithm for Bayesian multiscale multiple imputation consists of three components. First, conditional on the missing values, we sample the hyperparameters associated with each dynamic linear model. Second, conditional on the missing values, we estimate the latent process using the FFBS algorithm. Finally, we perform multiscale multiple imputation. In order to start the Gibbs sampler, we transform data from yearly format $\mathbf{z}_{jt}^*$ to $\mathbf{y}_{jt}$ $(j = 1, \ldots, k)$ and replace any missing cells by their series mean. After choosing starting values and defining all MCMC parameters, the algorithm can be summarized as follows:

Step 1: For $j = 1, \ldots, k$, sample the latent process $\boldsymbol{\theta}_{jt}$ using the FFBS algorithm.

Step 2: For $j = 1, \ldots, k$, sample the hyperparameters ξ_j and σ_j^2 from their full conditional distributions.

Step 3: Transform data from $\mathbf{y}_{jt}$ format to $\mathbf{z}_t^*$ format and sample $\mathbf{z}_{t',m}$ from (2).

Step 4: Transform data back to $\mathbf{y}_{jt}$ $(j = 1, \ldots, k)$ format and replace any missing cells by their the values obtained in Step 3.

Step 5: Repeat Steps 1–4.

3.2 Illustration - QCEW

To illustrate the imputation scheme proposed in Section 3.1 we provide a limited case study. Since the analyses were performed on a confidential version of the QCEW data (in

order to compare imputed to true values), we report only measures of performance of our imputed values. In other words, we can not simultaneously report the estimated values while providing specific measures of performance though we do impart a qualitative assessment here. Subsequently, in Section 4, we provide a detailed evaluation on the efficacy of our approach.

The data used here are the 6 years of quarterly data described in Section 2 and shown in Tables 1 and 2. As discussed in Section 3.1, for both data sets, at the beginning of the Gibbs sampler the missing values are imputed using the series mean. Further, for both data sets, the prior mean and variance for $\boldsymbol{\theta}_{j0}$ are set at $a = 0$ and $R = 10^{10}$ respectively. In terms of ξ_j and σ_j^2, $\alpha_j = 3$, $\beta_j = .1$ and $\tau_j = \kappa_j = .01$ for $j = 1, 2, 3$. Next, we run a single MCMC chain for 10,000 iterations discarding the first 5000 iterations for burn in. Convergence of the MCMC is verified through trace plots of the posterior.

Tables 3 and 4 provide the imputed values along with their associated 95% pointwise credible intervals. Additionally, Figures 1 and 2 show the aggregate series along with the 3 sub-series being estimated. It is important to note that in the majority of cases these series contain annual totals; however these totals are not portrayed in Figures 1 and 2.

Although we do not provide a measure of accuracy in this illustration we can see from Figures 1 and 2 that the imputation seems to have estimated plausible values for the suppressed data. Moreover, from Tables 3 and 4 it is also apparent that the multiscale (aggregation) constraints are preserved using our approach. In fact, even though we consider the imputed values after rounding to the nearest whole dollar the multiscale constraints are still exactly preserved. Finally, even for the case of the QCEW Data Set 2 (cf. Table 2 and Figure 2) where we have a substantially higher percentage of missing and less supporting aggregate information our method appears to provide reasonable performance in spite of what appears to be a challenging pattern of missingness.

4 Empirical Study - QCEW

To evaluate the effectiveness of our approach we conducted an empirical study using real data from the U.S. Bureau of Labor Statistics Quarterly Census of Employment and Wages (QCEW). Specifically we consider 11 data sets and impute the suppressed cells. Owing to disclosure practices, the authors outside of BLS have no knowledge regarding the values of the suppressed data (i.e. suppressed cells) neither before nor after the imputation. In fact, we apply the Bayesian multiscale imputation method to data that can be readily obtained by the public via BLS Internet ftp servers. Post imputation, the estimated missing values are compared at BLS (on site) to determine their accuracy. In other words, the analyses are conducted using only publicly available data (i.e. under the suppressed conditions) and then subsequently compared with the complete data by BLS employees in order to quantify the departures between the imputed and the actual values.

For all of the analyzes considered here the prior mean and variance for $\boldsymbol{\theta}_{j0}$ are set at $a = 0$ and $R = 10^{10}$ respectively. In terms of ξ_j and σ_j^2, $\alpha_j = 3$, $\beta_j = .1$ and $\tau_j = \kappa_j = .01$ for $j = 1, 2, 3$. Next, we run a single MCMC chain for 10,000 iterations discarding the first 5000 iterations for burn in. Convergence of the MCMC is verified through trace plots of the posterior.

In keeping with the disclosure practices of the BLS we do not present imputed values and measures of accuracy simultaneously. Instead we display the cumulative percentage of values that fall within 1%, 2%, 5% and 10% of their true values (Table 5). Further, the pattern of missingness is not the same for each data set. Therefore, we present the percentage missing for each sub-series for each data set (Table 6). In addition, Table 6 also indicates which data sets have sub-series missing the annual total.

As depicted in Table 5 we are able to impute at least 20% of the suppressed values to within 1% of their true values in over half of the data sets considered. Additionally, in 5 of the 11 data sets we are able to impute suppressed values to within 2% of their true value at least 50% of the time. Similarly, in 7 of the 11 data sets we are able to impute the suppressed

values to within 5% of their true values over 50% of the time. In fact, in 3 of these 11 data sets we are able to impute all of the missing values to within 5% of their true values. Finally, in 8 of the 11 data sets we can impute the data to within 10% of the their true values over 50% of the time. In all cases the 95% credible interval contained the true value.

Of course, we do not expect our method to perform well in all circumstances. For example, the pattern and/or percentage of missingness may be such that the multiscale nature and serial correlation of the data may afford little added benefit. One such example is given by QCEW6 where values are suppressed for years 5 and 6 (including the annual totals) for both sub-series 2 and 3. In this case, the imputation method is essentially trying to forecast 2 years ahead (up to 8 steps ahead) based on 4 years of data (16 data points). Moreover, judging by the spectral decomposition of the observation covariance matrix we are not acquiring much additional information as a result of the additive relationships.

5 Discussion

The imputation approach that we present provides a natural framework for serially correlated multiscale data. The method is flexible and can be applied across a broad array of multiscale data structures. Further, our method provides estimates of attributes of the data that may be of interest to the practitioner utilizing the data for applied research. For example, in addition to accurately imputing "missing" values, our framework can provide estimates of trend, seasonality and regression effects along with associated measures of uncertainty.

In addition, the estimates produced by our approach are computationally feasible and produce estimates sufficiently rapidly to allow imputation in practical situations. In fact, in our illustrations (Section 3.2) and empirical study (Section 4), we implemented our procedure using the same signal-to-noise prior specification throughout and each analysis ran in a matter of a few minutes on a laptop computer (MacBook Pro 2.5 GHz Intel Core Duo Processor - 4 GB 667 MHz DDR2 SDRAM). Another computational benefit is that our method

does not require any unknown "problem-specific" parameters. That is, our method requires little or no subjective specification of problem specific parameters up to the particular choice of DLM.

Further, we propose an approach to multiple imputation that couples DLMs with normally distributed random variables having singular valued covariance matrices. This produces a flexible framework capable of taking advantage of both inherent constraints present in multiscale (aggregated) data as well as serial correlation. In this context the multiscale aspect of our approach, in conjunction with the singular value covariance matrix, is critical as it allows us to effectively capitalize on redundant information in a mathematically rigorous and fully automatic manner.

In general, no imputation method can be expected to perform well in situations where the percentage of missingness is excessive. Although in many instances our approach can overcome a high percentage of missing data by borrowing strength through aggregate relationships. However, there are equally as many cases where the pattern of missingness precludes such benefit. In those cases, without any additional information *a priori*, the performance of our method suffers.

Nevertheless, the effectiveness of our approach is demonstrated through an illustration (Section 3.2) and an extensive empirical study (Section 4). In particular we apply our method to 11 QCEW data sets and show that in many instances we are able to impute suppressed (missing) cells to within 1% accuracy. In doing so we expose the vulnerability of "cell suppression" as a method for eliminating disclosure risks in longitudinal databases.

Importantly, the approach can be used to assess the vulnerability of longitudinal confidential databases when the method of protection is cell suppression. We envision that it will be of great importance to federal statistical agencies employing cell suppression. An agency can implement our approach prior to releasing data to determine if there are any unsuspected disclosure risks. Releases deemed to have high disclosure risks can be addressed prior to dissemination. The method is applicable to any multiscale temporal data protected

under cell suppression and, because of the computational efficiency of our approach, can be implemented on large scale databases in real time.

Acknowledgements This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical, methodological, technical, and operational issues are those of the authors and not necessarily those of the U.S. Bureau of Labor Statistics. Holan's research was supported by ASA/NSF/BLS and NISS research fellowships. Additionally, this research was supported by NSF grant EIA–0131884 to the National Institute of Statistical Sciences (NISS). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. Finally, we thank Michael Buso (Bureau of Labor Statistics) for his assistance with the QCEW data.

References

Ansley, C.F., and Kohn, R. (1983), "Exact Likelihood of Vector Autoregressive-Moving Average Process with Missing or Aggregated Data," *Biometrika*, 70, 275–278.

Carter, C.K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541–553.

Cohen, S., and Li, B.T. (2006), "A Comparison on Data Utility between Publishing Fixed Intervals versus Traditional Cell Suppression on Tabular Employment Data," U.S. Bureau of Labor Statistics. http://www.bls.gov/ore/abstract/st/st060100.htm

Dobra, A., Karr, A., and Sanil, A. (2003), "Preserving Confidentiality of High-dimensional Tabulated Data: Statistical and Computational Issues," *Statist. Comput.*, 13, 363–370.

Dobra, A., Karr, A., Sanil, A., and Fienberg, S. (2002), "Software systems for tabular data releases," *International J. Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 529–544.

Duncan, G., and Mukherjee, S. (2000), "Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks Through Additive Noise," *Journal of the American Statistical Association*, 95, 720–729.

Durbin, J., and Koopman, S. (2001), *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press.

Fienberg, S. (2006), "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation," *Statistical Science*, 21, 143–154.

Früwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202.

Gamerman, D., and Lopes, H.F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, 2nd ed.*, Boca Raton: Chapman & Hall/CRC.

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

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

Gomatam, S., Karr, A., Reiter, J., and Sanil, A. (2005), "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk-Utility Framework for Remote Access Analysis Servers," *Statistical Science*, 20, 163–177.

Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press.

Keller-McNulty, S., and Unger, E. (1998), "A Database System Prototype for Remote Access to Information Based on Confidential Data," *Journal of Official Statistics*, 14, 347–360.

Kelly, J. (1990), *Confidentiality Protection in Two and Three-Dimensional Tables*, Unpublished - Ph.D. Thesis: University of Maryland - College Park.

Little, R.J., and Rubin, D.B. (2002), *Statistical Analysis With Missing Data*, New York: Wiley.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, New York: Academic Press.

Marsaglia, G. (1964), "Conditional Means and Covariances of Normal Variables with Singular Covariance Matrix," *Journal of the American Statistical Association*, 59, 1203–1204.

Reiter, J. (2005), "Estimating Risks of Identification Disclosure for Microdata," *Journal of the American Statistical Association*, 100, 1103–1113.

Robert, C.P., and Casella, G. (2004), *Monte Carlo Statistical Methods, 2nd ed.*, New York: Springer-Verlag.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Schafer, J. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15.

Schouten, B., and Cigrang, M. (2003), "Remote Access Systems for Statistical Analysis of Microdata," *Statistics and Computation.*, 13, 381–389.

Searle, S.R. (1982), *Matrix Algebra Useful for Statisticians*, New York: Wiley.

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models: 2nd ed.*, New York: Springer.

Willenborg, L., and De Waal, T. (2001), *Elements of Statistical Disclosure Control*, New York: Springer.

	Total	Series 1	Series 2	Series 3
wage01-1	399688	49201	197316	153171
wage01-2	714639	$\mathcal{S}$	$\mathcal{S}$	479513
wage01-3	688482	54039	233588	400855
wage01-4	447404	$\mathcal{S}$	$\mathcal{S}$	198231
wage01-a	2250213	204177	814266	1231770
wage02-1	462232	49039	226622	186571
wage02-2	706801	$\mathcal{S}$	226219	$\mathcal{S}$
wage02-3	679498	$\mathcal{S}$	265220	$\mathcal{S}$
wage02-4	553380	$\mathcal{S}$	216504	$\mathcal{S}$
wage02-a	2401911	150107	934565	1317239
wage03-1	453892	$\mathcal{S}$	235871	$\mathcal{S}$
wage03-2	627605	$\mathcal{S}$	222709	$\mathcal{S}$
wage03-3	492338	28911	260932	202495
wage03-4	488352	29535	224213	234604
wage03-a	2062187	116585	943725	1001877
wage04-1	628245	122516	265484	240245
wage04-2	796096	130296	240055	425745
wage04-3	643023	134871	262762	245390
wage04-4	759910	138567	272218	349125
wage04-a	2827274	526250	1040519	1260505
wage05-1	650100	164995	232009	253096
wage05-2	715893	185907	228384	301602
wage05-3	733692	187186	274578	271928
wage05-4	731393	191415	275615	264363
wage05-a	2831078	729503	1010586	1090989
wage06-1	811330	313003	209979	288348
wage06-2	883901	315194	250611	318096
wage06-3	841881	323209	224255	294417
wage06-4	865273	325835	249976	289462
wage06-a	3402385	1277241	934821	1190323

Table 1: Disclosed QCEW Data Set 1 with suppressed cells denoted by $\mathcal{S}$.

	Total	Series 1	Series 2	Series 3
wage01-1	35247480	6456128	27555264	1236088
wage01-2	29085928	5638595	22425971	1021362
wage01-3	29331857	6362500	21759797	1209560
wage01-4	32320399	6729254	24490149	1100996
wage01-a	125985664	25186477	96231181	4568006
wage02-1	25233545	6191050	17743550	1298945
wage02-2	22103990	$\mathcal{S}$	15493524	$\mathcal{S}$
wage02-3	23647695	$\mathcal{S}$	16199098	$\mathcal{S}$
wage02-4	27900353	$\mathcal{S}$	19592672	$\mathcal{S}$
wage02-a	98885583	25314368	69028844	4542371
wage03-1	26571054	$\mathcal{S}$	17599297	$\mathcal{S}$
wage03-2	25017823	$\mathcal{S}$	17289908	$\mathcal{S}$
wage03-3	26713862	$\mathcal{S}$	17302366	$\mathcal{S}$
wage03-4	32011096	8794890	$\mathcal{S}$	$\mathcal{S}$
wage03-a	110313835	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
wage04-1	23082164	8096669	$\mathcal{S}$	$\mathcal{S}$
wage04-2	22773180	7932895	$\mathcal{S}$	$\mathcal{S}$
wage04-3	23269552	8620975	$\mathcal{S}$	$\mathcal{S}$
wage04-4	28673482	9383772	$\mathcal{S}$	$\mathcal{S}$
wage04-a	97798378	34034311	$\mathcal{S}$	$\mathcal{S}$
wage05-1	21721426	7358822	$\mathcal{S}$	$\mathcal{S}$
wage05-2	21716384	7582785	$\mathcal{S}$	$\mathcal{S}$
wage05-3	25895877	9134881	15689149	1071847
wage05-4	30344595	9667405	19318854	1358336
wage05-a	99678282	33743893	61309507	4624882
wage06-1	23653708	8605217	13883597	1164894
wage06-2	23924694	9082470	13514676	1327548
wage06-3	21323373	8405353	11707047	1210973
wage06-4	28035179	9988831	16537826	1508522
wage06-a	96936954	36081871	55643146	5211937

Table 2: Disclosed QCEW Data Set 2 with suppressed cells denoted by $\mathcal{S}$.

	Total	Series 1	Series 2	Series 3
wage01-1	399688	49201	197316	153171
wage01-2	714639	47043 (19086, 75041)	188083 (160085, 216040)	479513
wage01-3	688482	54039	233588	400855
wage01-4	447404	53894 (25896, 81851)	195279 (167322, 223277)	198231
wage01-a	2250213	204177	814266	1231770
wage02-1	462232	49039	226622	186571
wage02-2	706801	48763 (0, 107340)	226219	431819 (373242, 487940)
wage02-3	679498	34427 (0, 89542)	265220	379851 (324736, 433178)
wage02-4	553380	17878 (0, 72974)	216504	318998 (263902, 378480)
wage02-a	2401911	150107	934565	1317239
wage03-1	453892	11872 (0, 58126)	235871	206149 (159895, 261640)
wage03-2	627605	46267 (13, 101578)	222709	358629 (303318, 404883)
wage03-3	492338	28911	260932	202495
wage03-4	488352	29535	224213	234604
wage03-a	2062187	116585	943725	1001877
wage04-1	628245	122516	265484	240245
wage04-2	796096	130296	240055	425745
wage04-3	643023	134871	262762	245390
wage04-4	759910	138567	272218	349125
wage04-a	2827274	526250	1040519	1260505
wage05-1	650100	164995	232009	253096
wage05-2	715893	185907	228384	301602
wage05-3	733692	187186	274578	271928
wage05-4	731393	191415	275615	264363
wage05-a	2831078	729503	1010586	1090989
wage06-1	811330	313003	209979	288348
wage06-2	883901	315194	250611	318096
wage06-3	841881	323209	224255	294417
wage06-4	865273	325835	249976	289462
wage06-a	3402385	1277241	934821	1190323

Table 3: Imputed suppressed cells corresponding to data in Table 1 along with 95% credible intervals with values rounded to the nearest whole dollar.

	Total	Series 1	Series 2	Series 3
wage01-1	35247480	6456128	27555264	1236088
wage01-2	29085928	5638595	22425971	1021362
wage01-3	29331857	6362500	21759797	1209560
wage01-4	32320399	6729254	24490149	1100996
wage01-a	125985664	25186477	96231181	4568006
wage02-1	25233545	6191050	17743550	1298945
wage02-2	22103990	5550102 (5323311, 5786374)	15493524	1060364 (824092, 1287155)
wage02-3	23647695	6368924 (6139738, 6597414)	16199098	1079673 (851183, 1308859)
wage02-4	27900353	7204292 (6965088, 7434342)	19592672	1103389 (873339, 1342593)
wage02-a	98885583	25314368	69028844	4542371
wage03-1	26571054	7796647 (7476104, 8097539)	17599297	1175110 (874218, 1495653)
wage03-2	25017823	6602844 (6302386, 6922295)	17289908	1125071 (805620, 1425529)
wage03-3	26713862	8233839 (7912920, 8548211)	17302366	1177657 (863285, 1498576)
wage03-4	32011096	8794890	22044744 (21727171, 22363838)	1171462 (852368, 1489035)
wage03-a	110313835	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
wage04-1	23082164	8096669	13824717 (13510783, 14151890)	1160778 (833605, 1474712)
wage04-2	22773180	7932895	13679769 (13362913, 13992710)	1160516 (847575, 1477372)
wage04-3	23269552	8620975	13482879 (13170373, 13796721)	1165698 (851856, 1478204)
wage04-4	28673482	9383772	18108725 (17771926, 18429733)	1180985 (859977, 1517784)
wage04-a	97798378	34034311	$\mathcal{S}$	$\mathcal{S}$
wage05-1	21721426	7358822	13268901 (13061204, 13478068)	1093703 (884536, 1301400)
wage05-2	21716384	7582785	13032603 (12823436, 13240300)	1100996 (893299, 1310163)
wage05-3	25895877	9134881	15689149	1071847
wage05-4	30344595	9667405	19318854	1358336
wage05-a	99678282	33743893	61309507	4624882
wage06-1	23653708	8605217	13883597	1164894
wage06-2	23924694	9082470	13514676	1327548
wage06-3	21323373	8405353	11707047	1210973
wage06-4	28035179	9988831	16537826	1508522
wage06-a	96936954	36081871	55643146	5211937

Table 4: Imputed suppressed cells corresponding to data in Table 2 along with 95% credible intervals with values rounded to the nearest whole dollar.

Data	<1%	<2%	<5%	<10%
QCEW1	7.14	42.86	57.14	71.43
QCEW2	0.00	0.00	5.00	50.00
QCEW3	25.00	50.00	100.00	100.00
QCEW4	10.00	50.00	50.00	60.00
QCEW5	48.39	70.97	83.87	93.55
QCEW6	0.00	0.00	0.00	10.00
QCEW7	21.43	28.57	50.00	57.14
QCEW8	30.00	30.00	30.00	50.00
QCEW9	0.00	9.09	13.64	31.82
QCEW10	50.00	87.50	100.00	100.00
QCEW11	62.50	75.00	100.00	100.00

Table 5: Percentage of imputed values within 1,2,5 and 10% of the true values. Note that QCEW1 - QCEW11 denotes the 11 different QCEW data sets used in this empirical investigation.

	Percent Missing			
Data	Aggregate	Series 1	Series 2	Series 3
QCEW1	0.00	23.3	6.70	16.7
QCEW2	0.00	0.00	33.3	33.3
QCEW3	0.00	6.70	6.70	0.00
QCEW4*	0.00	0.00	16.7	16.7
QCEW5*	0.00	23.3	30.0	50.0
QCEW6*	0.00	0.00	33.3	33.3
QCEW7*	0.00	0.00	23.3	23.3
QCEW8*	0.00	0.00	33.3	33.3
QCEW9*	0.00	36.7	0.00	36.7
QCEW10	0.00	13.3	0.00	13.3
QCEW11	0.00	0.00	26.7	26.7

Table 6: Percentage of missing values by series within a particular data set. Note that * indicates that the data set is missing some of the annual totals in addition to the suppressed disaggregated values. Note that QCEW1 - QCEW11 denotes the 11 different QCEW data sets used in this empirical investigation.

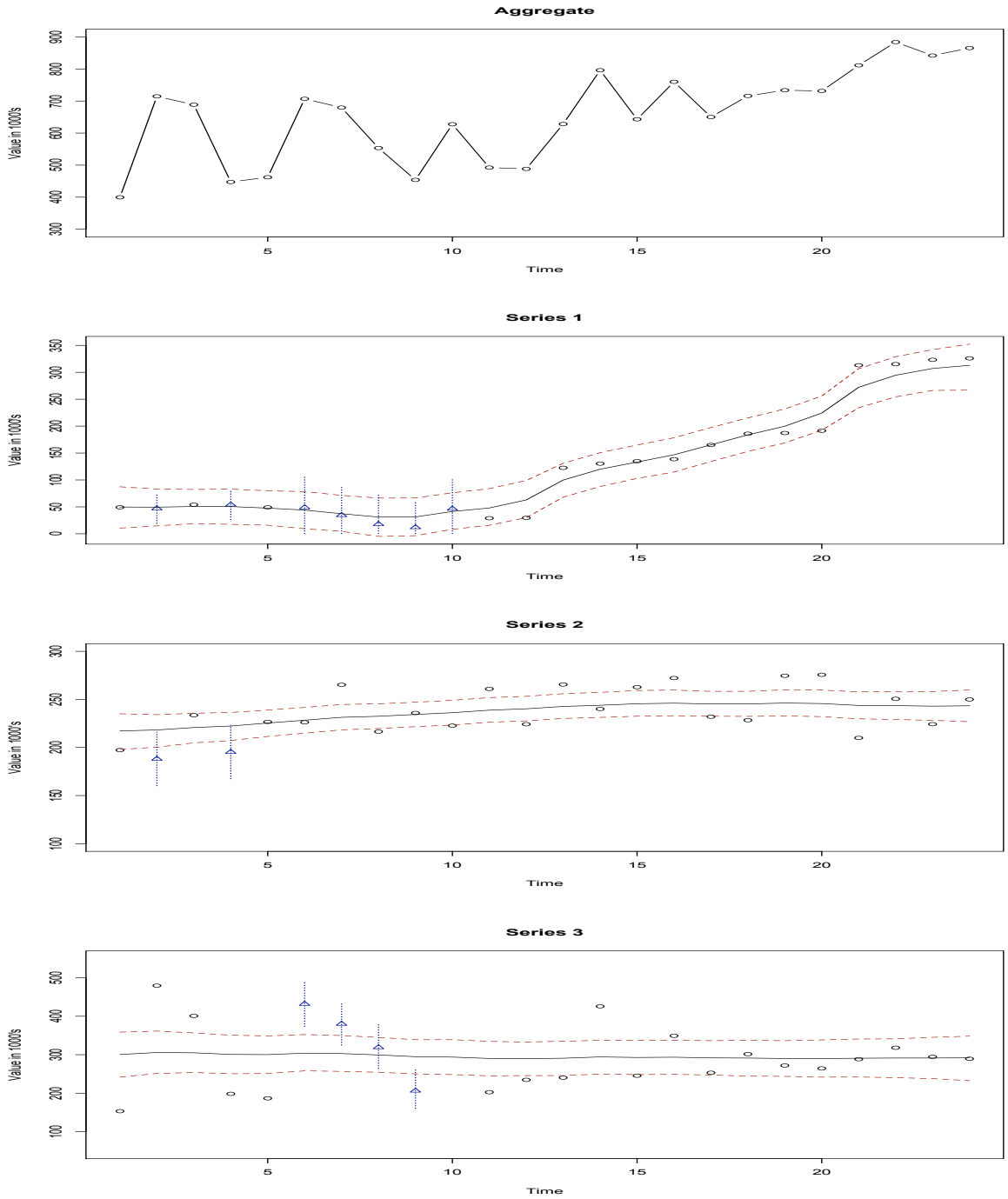

Figure 1: Aggregate series along with 3 sub-series corresponding to QCEW Data Set 1. The circles are the observed data and the triangles are the imputed suppressed cells. The solid line represents the estimated latent process whereas the horizontal dashed lines and vertical dashed lines correspond to the 95% point-wise credible interval for the latent process and imputed suppressed cells respectively.

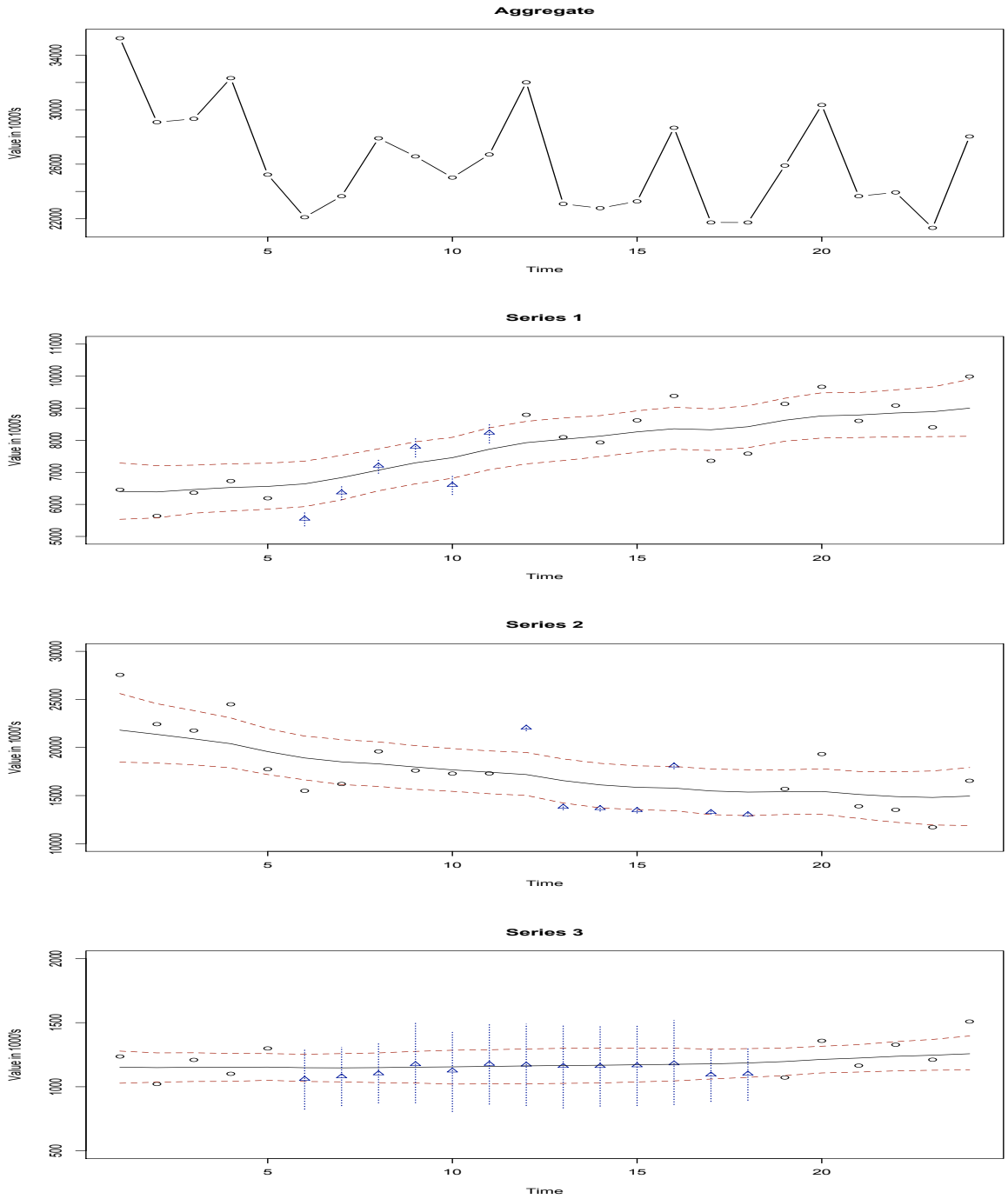

Figure 2: Aggregate series along with 3 sub-series corresponding to QCEW Data Set 2. The circles are the observed data and the triangles are the imputed suppressed cells. The solid line represents the estimated latent process whereas the horizontal dashed lines and vertical dashed lines correspond to the 95% point-wise credible interval for the latent process and imputed suppressed cells respectively.