Bayesian Multiscale Multiple Imputation With Implications for Data Confidentiality

Scott H. HOLAN, Daniell TOTH, Marco A. R. FERREIRA, and Alan F. KARR

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 datasets experience missing observations in a manner that they can be accurately imputed, while respecting the constraints imposed by the multiscale nature of the data, using the method we propose known as *Bayesian multiscale multiple imputation*. 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. 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 when microdata on individual units, such as people or establishments, are disseminated 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. 2006).

To reduce disclosure risk, statistical agencies often alter the data prior to release. For example, it is common for agencies to perturb, coarsen, or swap data values (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 (Karr et al. 2006).

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 et al. 2002; Dobra, Karr, and Sanil 2003).

Another approach for protecting against disclosure is to release synthetic data (i.e., simulated 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 Reiter and Raghunathan 2007; Machanavajjhalla et al. 2008; 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 (i.e., 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; Karr et al. 2006).

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":

Scott H. Holan is Assistant Professor, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100 (E-mail: *holans@missouri.edu*). Daniell Toth is Research Mathematical Statistician, Office of Survey Methods Research, Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Room 1950, Washington, DC 20212 (E-mail: *Toth.Daniell@bls.gov*). Marco A. R. Ferreira is Assistant Professor, Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100 (E-mail: *ferreiram@missouri.edu*). Alan F. Karr is Director, National Institute of Statistical Sciences, 19 T. W. Alexander Drive, Research Triangle Park, NC 27709-4006 (E-mail: *karr@niss.org*). This article 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 National Science Foundation (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. The authors thank the editor, associate editor, and two anonymous referees for their insightful comments that helped substantially improve this article. Finally, the authors thank Michael Buso (Bureau of Labor Statistics) for his assistance with the QCEW data.

Journal of the American Statistical Association
June 2010, Vol. 105, No. 490, Applications and Case Studies
DOI: 10.1198/jasa.2009.ap08629

cell entries in tables that are deemed risky (usually because they represent only a few data subjects) are simply suppressed. In general, cell suppression is imposed on both contingency tables and on microdata. However, in this context, we consider microdata rather than contingency tables. 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 a nondeterministic polynomial-time (NP)-hard problem and implemented algorithms rely on heuristics (Cox 1980, 1995; Fischetti and Salazar 2001). 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-grained 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 article, we propose a method of Bayesian multiscale multiple imputation (BMMI) that utilizes the additive relationships (multiscale attributes) along with inherent serial correlation to impute suppressed values. While the method is of independent interest as a means of imputing missing data, this article 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 such as "carry-forward" or "equal proportions" (cf. Section 4).

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 (Muirhead 1982; Siotani, Hayakawa and Fujikoshi 1985). Specifically, we make use of two properties of singular normal distributions. First, the number of zero eigenvalues of the singular covariance matrix is equal to the number of hard constraints in the data due to having knowledge about the aggregated values. Second, the conditional distribution of subvectors is also normally distributed with covariance matrix, which depends on a generalized inverse of the singular covariance matrix of the entire random vector.

As noted previously, the method proposed here is applicable to 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 parameter specification by the user. Given the choice of the particular DLM, we handle the unknown "problem-specific" parameters by employing a set of default priors that require no subjective specification. This set of default priors has performed very well for all the datasets analyzed in Sections 3 and 4.

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 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 covariance matrix allows us to systematically accommodate any redundant information present in the data in a mathematically rigorous and fully automatic manner.

Our multiscale multiple imputation methodology is related to the multiscale time series modeling approach of Ferreira et al. (2006). In particular, for a two-level model Ferreira et al. (2006) used an initial univariate time series process at the fine resolution level and a stochastic linear equation based on longitudinal aggregation relating fine and coarse levels to obtain the conditional distribution of the fine level given the coarse level. Subsequently, they use Jeffrey's rule of conditioning (Jeffrey 1992) to revise the process at the coarse level to produce a coherent joint model for the two resolution levels. See chapter 11 of Ferreira and Lee (2007) for a comprehensive discussion regarding these multiscale time series models. Conversely, here we use deterministic equations linking the fine resolution multivariate time series with their longitudinal and subseries aggregated coarse time series. Ultimately our objective is imputation at the fine level, thus we require the conditional distributions of the suppressed (missing) cells given the observed cells and aggregated series. Departing from Ferreira et al. (2006), our link equations are deterministic and necessitate the use of conditional distributions, which are singular normal distributions.

The remainder of this article is organized as follows. Section 2 provides a brief description of the QCEW. In Section 3 our method is formally developed and an illustration employing 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 datasets. 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 the 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 State unemployment insurance laws and

for civilian workers covered by the program of Unemployment Compensation for Federal Employees. Tabulations of QCEW outcomes are available by six-digit North American Industrial Classification Systems (NAICS) industry, county, ownership, and size groups under several formats, for example, via 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 heuristic solutions have been proposed (see Cox 1980, 1995; Fischetti and Salazar 2001; 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. The QCEW data contain 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 jth subseries $j = 1, \ldots, k$ and tth quarter $t = 1, \ldots, T$. Here k denotes the number of aggregate subseries 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 were 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 were 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 datasets from the QCEW. In many instances the suppressed cells can be an annual total (e.g., Table 2) or even a subseries aggregate total (not displayed). Additionally, the multiscale nature can have an aggregate along with k subseries where k does not necessarily equal 3; in fact, we only require $k \geq 2$. Further, each of the k subseries can be an aggregate of l_k ($l_k \geq 2$) additional subseries. Nevertheless, the framework we propose effectively accommodates these multiple data structures.

Table 1. Disclosed QCEW dataset 1, with suppressed cells denoted by $\mathcal{S}$

	Total	Series 1	Series 2	Series 3
wage01-1	399,688	49,201	197,316	153,171
wage01-2	714,639	$\mathcal{S}$	$\mathcal{S}$	479,513
wage01-3	688,482	54,039	233,588	400,855
wage01-4	447,404	$\mathcal{S}$	$\mathcal{S}$	198,231
wage01-a	2,250,213	204,177	814,266	1,231,770
wage02-1	462,232	49,039	226,622	186,571
wage02-2	706,801	$\mathcal{S}$	226,219	$\mathcal{S}$
wage02-3	679,498	$\mathcal{S}$	265,220	$\mathcal{S}$
wage02-4	553,380	$\mathcal{S}$	216,504	$\mathcal{S}$
wage02-a	2,401,911	150,107	934,565	1,317,239
wage03-1	453,892	$\mathcal{S}$	235,871	$\mathcal{S}$
wage03-2	627,605	$\mathcal{S}$	222,709	$\mathcal{S}$
wage03-3	492,338	28,911	260,932	202,495
wage03-4	488,352	29,535	224,213	234,604
wage03-a	2,062,187	116,585	943,725	1,001,877
wage04-1	628,245	122,516	265,484	240,245
wage04-2	796,096	130,296	240,055	425,745
wage04-3	643,023	134,871	262,762	245,390
wage04-4	759,910	138,567	272,218	349,125
wage04-a	2,827,274	526,250	1,040,519	1,260,505
wage05-1	650,100	164,995	232,009	253,096
wage05-2	715,893	185,907	228,384	301,602
wage05-3	733,692	187,186	274,578	271,928
wage05-4	731,393	191,415	275,615	264,363
wage05-a	2,831,078	729,503	1,010,586	1,090,989
wage06-1	811,330	313,003	209,979	288,348
wage06-2	883,901	315,194	250,611	318,096
wage06-3	841,881	323,209	224,255	294,417
wage06-4	865,273	325,835	249,976	289,462
wage06-a	3,402,385	1,277,241	934,821	1,190,323

3. MULTISCALE MULTIPLE IMPUTATION

In recent years, multiple imputation, the practice of "filling in" missing data with plausible values, has emerged as a 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 dataset. In addition, when $D \geq 2$ sets of imputations are formed and constitute repeated draws from the posterior predictive distribution of the missing values under a specified model, then the D complete datasets can be combined to form one inference that properly accounts for the uncertainty due to imputation 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) described methods for generating MIs using parametric Bayesian models in the context of simple problems. In general, suppose that $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ follows a parametric model $P(\mathbf{Y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ has a prior distribution and the missing data mechanism for $\mathbf{Y}_{mis}$ is ignored (Little and Rubin 2002, p. 120). Then we can write

$$P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}) = \int P(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathbf{Y}_{obs})\, d\boldsymbol{\theta}.$$

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

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

Table 2. Disclosed QCEW dataset 2, with suppressed cells denoted by $\mathcal{S}$

	Total	Series 1	Series 2	Series 3
wage01-1	35,247,480	6,456,128	27,555,264	1,236,088
wage01-2	29,085,928	5,638,595	22,425,971	1,021,362
wage01-3	29,331,857	6,362,500	21,759,797	1,209,560
wage01-4	32,320,399	6,729,254	24,490,149	1,100,996
wage01-a	125,985,664	25,186,477	96,231,181	4,568,006
wage02-1	25,233,545	6,191,050	17,743,550	1,298,945
wage02-2	22,103,990	$\mathcal{S}$	15,493,524	$\mathcal{S}$
wage02-3	23,647,695	$\mathcal{S}$	16,199,098	$\mathcal{S}$
wage02-4	27,900,353	$\mathcal{S}$	19,592,672	$\mathcal{S}$
wage02-a	98,885,583	25,314,368	69,028,844	4,542,371
wage03-1	26,571,054	$\mathcal{S}$	17,599,297	$\mathcal{S}$
wage03-2	25,017,823	$\mathcal{S}$	17,289,908	$\mathcal{S}$
wage03-3	26,713,862	$\mathcal{S}$	17,302,366	$\mathcal{S}$
wage03-4	32,011,096	8,794,890	$\mathcal{S}$	$\mathcal{S}$
wage03-a	110,313,835	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
wage04-1	23,082,164	8,096,669	$\mathcal{S}$	$\mathcal{S}$
wage04-2	22,773,180	7,932,895	$\mathcal{S}$	$\mathcal{S}$
wage04-3	23,269,552	8,620,975	$\mathcal{S}$	$\mathcal{S}$
wage04-4	28,673,482	9,383,772	$\mathcal{S}$	$\mathcal{S}$
wage04-a	97,798,378	34,034,311	$\mathcal{S}$	$\mathcal{S}$
wage05-1	21,721,426	7,358,822	$\mathcal{S}$	$\mathcal{S}$
wage05-2	21,716,384	7,582,785	$\mathcal{S}$	$\mathcal{S}$
wage05-3	25,895,877	9,134,881	15,689,149	1,071,847
wage05-4	30,344,595	9,667,405	19,318,854	1,358,336
wage05-a	99,678,282	33,743,893	61,309,507	4,624,882
wage06-1	23,653,708	8,605,217	13,883,597	1,164,894
wage06-2	23,924,694	9,082,470	13,514,676	1,327,548
wage06-3	21,323,373	8,405,353	11,707,047	1,210,973
wage06-4	28,035,179	9,988,831	16,537,826	1,508,522
wage06-a	96,936,954	36,081,871	55,643,146	5,211,937

Typically this approach is facilitated by taking advantage of Markov chain Monte Carlo (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 (i.e., annual totals and subseries aggregate totals). Thus, to systematically accommodate and take advantage of these redundancies without substantial bookkeeping on the part of the practitioner requires innovative methods involving singular covariance matrices.

The imputation model we propose presumes suppressed values are "missing at random." However, the data are not missing at random because the values are missing due to a p-percent rule (Willenborg and de Waal 2001) or a more extensive set of criteria. With nonignorable missing data it is often useful, but incorrect, to fit an ignorable model. In fact, enforcing the p-percent rule or more extensive suppression criteria constraints on the imputations might approximate a nonignorable model reasonably well and thus lead to better imputations. However, the suppression rules used by the BLS for the QCEW series are not published and are, in fact, not public information. A general description of the QCEW confidentiality procedures can be found in Statistical Working Paper 22 (*http://www.fcsm.gov/working-papers/spwp22.html*, chapter 3, p. 47); the procedures in chapter 5 of the *Handbook of Methods* on the BLS website are, as of this writing, out of date. Furthermore, for any cell in which either the wage or employment data are marked as sensitive, both items are suppressed. Consequently, approximating a nonignorable model by enforcing confidentiality rules is not possible in this context. Nevertheless, it would be possible for an intruder to take a reasonable guess at the confidentiality rules (the outdated 80% rule may provide a good starting point). One avenue for further research would be to investigate how much reasonable guesses improve or worsen estimates.

This section formally develops the BMMI scheme and provides an illustration by applying our method on two representative QCEW series as well as a simulated example. The main point of this illustration is to demonstrate our approach through several detailed examples. Subsequently, we provide a comprehensive assessment of our method's performance in Section 4.

3.1 The Multiscale Multiple Imputation Scheme

The BMMI scheme can be viewed as a two-stage iterative procedure. In the first stage all of the subseries (i.e., all of the series other than the aggregate of the subseries) are modeled individually, conditional on the missing values, using DLMs (West and Harrison 1997). Thus considering the example series in Section 2 we have three DLMs, each modeling a series of six 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 straightforward manner to include correlation between series. However, this is typically unnecessary as much of the between-series correlation is accounted for through the multiscale (subseries aggregation) constraints. The second step of our procedure performs imputation of the missing values for each year 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)

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

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 a 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. In this case, estimation can be performed using 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 hyperparameters, 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) (Carter and Kohn 1994; Früwirth–Schnatter 1994).

In our particular case, $\mathbf{y}_t$ contains k subseries related to different economic sectors. In order to model the joint evolution of those subseries through time, $\mathbf{F}_t'\boldsymbol{\theta}_t$ may contain regression terms, seasonality, first and second-order trends, common latent factors, and so on. However, in our experience, we 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 article we assume y_{jt}, $j = 1, \ldots, k$, follows 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 reparameterize 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 in Equation (1) is completed with a prior $\theta_{j0} \sim \mathcal{N}(a, R)$, where a and R are user defined and usually taken to imply a vague prior.

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 $\mathbf{z}_{t'} = (y_{1,4t'-3}, \ldots, y_{1,4t'}, y_{2,4t'-3}, \ldots, y_{2,4t'}, y_{3,4t'-3}, \ldots, y_{3,4t'}, q_{4t'-3}, \ldots, q_{4t'}, a_{1t'}, a_{2t'}, a_{3t'})'$. 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 along with the several longitudinal and subseries aggregate totals. Then it follows from Equation (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 $\mathbf{V} = \operatorname{diag}(\sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_1^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_2^2, \sigma_3^2, \sigma_3^2, \sigma_3^2, \sigma_3^2)$. 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 a dimension of 19×12.

Typically, several elements of $\mathbf{z}_{t'}$, either individual or aggregated values, will be 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 $\boldsymbol{\Sigma}$ can be written

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

Further, consider the spectral decomposition of $\boldsymbol{\Sigma}_{oo}$, that is, $\boldsymbol{\Sigma}_{oo} = \mathbf{P}\mathbf{D}\mathbf{P}'$ where $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_p]$ has orthonormal columns given by the normalized eigenvectors of $\boldsymbol{\Sigma}_{oo}$ and $\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_p)$, with $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ corresponding to the eigenvalues of $\boldsymbol{\Sigma}_{oo}$. In addition, let q denote the number of zero eigenvalues of $\boldsymbol{\Sigma}_{oo}$. Note that q is equal to the number of redundancies found in the observed data due to having knowledge about the longitudinal or subseries aggregated values. In order to eliminate these redundancies define $\mathbf{D}^* = \operatorname{diag}(d_1, \ldots, d_{p-q})$ to be the diagonal matrix with diagonal equal to the positive eigenvalues of $\boldsymbol{\Sigma}_{oo}$ 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 1982), can be computed as $\boldsymbol{\Sigma}_{oo}^+ = \mathbf{P}^*(\mathbf{D}^*)^{-1}\mathbf{P}^{*\prime}$. Ultimately to impute the missing (suppressed) data we need to find the conditional distribution of missing values given the observed values. Using standard properties of normal distributions with singular covariance matrices (Muirhead 1982; Siotani, Hayakawa and Fujikoshi 1985), we have

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

where

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

and

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

Remark 1. In the case where $\boldsymbol{\Sigma}_{oo}$ is full rank, Equations (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 can eliminate the redundant information by eliminating some redundant elements from $\mathbf{z}_{t',o}$, but this will require substantial bookkeeping. By contrast, our procedure is fully automatic.

Remark 3. Usually, the covariance matrix $\boldsymbol{\Omega}_m$ is singular. In order to simulate from Equation (2), we first compute the spectral decomposition $\boldsymbol{\Omega}_m = \mathbf{P}_\Omega\mathbf{D}_\Omega\mathbf{P}_\Omega'$. Let r be the rank of $\boldsymbol{\Omega}_m$. Additionally, let $\mathbf{D}_\Omega^*$ be the diagonal matrix with diagonal equal to the positive eigenvalues of $\boldsymbol{\Omega}_m$ 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 Equation (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 are 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 previously. First, we note that the signal-to-noise ratio parameters ξ_j are most likely small. Otherwise, the components of the latent process will vary too much over time and ultimately

this will 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 $\text{IG}(\alpha_j, \beta_j)$ 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 $\sigma_j^2 \sim \text{IG}(\tau_j, \kappa_j)$, with $\tau_j = \kappa_j = 0.01$, $j = 1, \ldots, k$, which is a vague 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, $\xi_j|\boldsymbol{\theta}_{jt}, \sigma_j^2 \sim \text{IG}(\alpha_j^*, \beta_j^*)$, 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 since the full conditional for $\sigma_j^2|\xi_j, \mathbf{y}_{jt}, \boldsymbol{\theta}_{jt}$ is $\text{IG}(\tau_j^*, \kappa_j^*)$ where $\tau_j^* = \tau_j + (2T-1)/2$ and

$$\kappa_j^* = \kappa_j + \sum_{t=1}^{T}(y_{jt} - \theta_{jt})^2 + 0.5\sum_{t=2}^{T}(\theta_{jt} - \theta_{j,t-1})^2/\xi_j.$$

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 BMMI 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 the 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 the $\mathbf{y}_{jt}$ format to the $\mathbf{z}_t^*$ format and sample $\mathbf{z}_{t',m}$ from Equation (2).

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

Step 5. Repeat Steps 1 through 4 until convergence.

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 values to true values), we report only measures of performance of our imputed values. We cannot simultaneously report the estimated values while providing specific measures of performance, although we do impart a qualitative assessment here. Therefore, in addition, we provide the results of a simulated example where the suppressed cells being imputed can be disclosed. Subsequently, in Section 4, we provide a detailed evaluation on the efficacy of our approach.

As a general guideline, if little or no prior information is available then vague or (in some sense) noninformative priors should be used. Here we use vague priors for θ_{j0} and σ_j^2. Specifically, the prior mean and variance for θ_{j0} are set at $a = 0$ and $R = 10^{10}$, respectively, whereas for σ_j^2, we take $\tau_j = \kappa_j = 0.01$ for $j = 1, 2, 3$. For the signal-to-noise ratio parameter ξ_j, we assume $\alpha_j = 3$, $\beta_j = 0.1$ which guarantees the existence of the first two prior moments and at the same time imparts little impact in the analysis. We use this set of prior specifications for both datasets in this section and for all the datasets in Section 4. As this set of priors performs extremely well across the several datasets considered here, we recommend their use as a default choice of prior distributions.

The data used here are the six years of quarterly data described in Section 2 and shown in Tables 1 and 2. As discussed in Section 3.1, for both datasets, at the beginning of the Gibbs sampler the missing values are imputed using the series mean. 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 three subseries 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. In Figure 1 it may appear that some of the imputations are not performing well since the imputed values and their 95% CIs lie outside the 95% CI of the latent process. However, it is important to keep in mind that the credible intervals for the latent process fail to take into account the additive constraints because they model each process separately. Therefore, it is possible to obtain accurate imputations that lie outside the 95% CI for the individual latent processes.

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 dataset 2 (cf. Table 2 and Figure 2) where we have a substantially higher percentage of missing and less supporting longitudinal and subseries aggregate information, our method appears to provide reasonable performance in spite of what appears to be a challenging pattern of missingness.

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	399,688	49,201	197,316	153,171
wage01-2	714,639	47,043	188,083	479,513
		(19,086, 75,041)	(160,085, 216,040)	
wage01-3	688,482	54,039	233,588	400,855
wage01-4	447,404	53,894	195,279	198,231
		(25,896, 81,851)	(167,322, 223,277)	
wage01-a	2,250,213	204,177	814,266	1,231,770
wage02-1	462,232	49,039	226,622	186,571
wage02-2	706,801	48,763	226,219	431,819
		(0, 107,340)		(373,242, 487,940)
wage02-3	679,498	34,427	265,220	379,851
		(0, 89,542)		(324,736, 433,178)
wage02-4	553,380	17,878	216,504	318,998
		(0, 72,974)		(263,902, 378,480)
wage02-a	2,401,911	150,107	934,565	1,317,239
wage03-1	453,892	11,872	235,871	206,149
		(0, 58,126)		(159,895, 261,640)
wage03-2	627,605	46,267	222,709	358,629
		(13, 101,578)		(303,318, 404,883)
wage03-3	492,338	28,911	260,932	202,495
wage03-4	488,352	29,535	224,213	234,604
wage03-a	2,062,187	116,585	943,725	1,001,877
wage04-1	628,245	122,516	265,484	240,245
wage04-2	796,096	130,296	240,055	425,745
wage04-3	643,023	134,871	262,762	245,390
wage04-4	759,910	138,567	272,218	349,125
wage04-a	2,827,274	526,250	1,040,519	1,260,505
wage05-1	650,100	164,995	232,009	253,096
wage05-2	715,893	185,907	228,384	301,602
wage05-3	733,692	187,186	274,578	271,928
wage05-4	731,393	191,415	275,615	264,363
wage05-a	2,831,078	729,503	1,010,586	1,090,989
wage06-1	811,330	313,003	209,979	288,348
wage06-2	883,901	315,194	250,611	318,096
wage06-3	841,881	323,209	224,255	294,417
wage06-4	865,273	325,835	249,976	289,462
wage06-a	3,402,385	1,277,241	934,821	1,190,323

The previous illustrations are necessarily limited. Therefore, we further demonstrate our approach through a simulated example. Specifically, we simulate data that behave like the data found in Table 2. In particular, we use the estimated observation variances and latent processes to simulate data according to Equation (1) and perform the appropriate aggregations. Further, we suppress the same cells as in Table 2. Finally, all MCMC parameters are taken to be identical to those used in the illustrations for QCEW datasets 1 and 2.

Table 5 provides the simulated dataset along with the imputed values and their 95% credible intervals. In general, the imputed values are "close" to the truth (i.e., 50% are within 1%, 65% are within 2%, 69% are within 5%, and 92% are within 10%). Additionally, all of the 95% credible intervals contain the true values with many having a "narrow" width. Equally as important, this example further illustrates that our method preserves the aggregate constraints. This aspect can be crucial to data users interested in making subject matter inference. Lastly, we investigate the effect of the prior specification on ξ_j (the signal-to-noise ratio parameter) by varying the value of β_j between 0.1 and 2. For this example, we find our results to be insensitive to this choice.

4. EMPIRICAL STUDY: QCEW

To evaluate the effectiveness of our approach we conducted an empirical study using real data from the QCEW. Specifically we considered 11 datasets and imputed the suppressed cells. As in Section 3.2, owing to BLS disclosure practices, the authors outside of BLS have no knowledge of the values of the suppressed data. Instead, as would "real intruders," we applied the BMMI method to data obtained from the public BLS internet FTP servers. Post imputation, the estimated missing values were compared at BLS (on site) to determine their accuracy.

For all of the analyses considered here, the prior mean and variance for θ_{j0} are set at $a = 0$ and $R = 10^{10}$, respectively.

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

	Total	Series 1	Series 2	Series 3
wage01-1	35,247,480	6,456,128	27,555,264	1,236,088
wage01-2	29,085,928	5,638,595	22,425,971	1,021,362
wage01-3	29,331,857	6,362,500	21,759,797	1,209,560
wage01-4	32,320,399	6,729,254	24,490,149	1,100,996
wage01-a	125,985,664	25,186,477	96,231,181	4,568,006
wage02-1	25,233,545	6,191,050	17,743,550	1,298,945
wage02-2	22,103,990	5,550,102	15,493,524	1,060,364
		(5,323,311, 5,786,374)		(824,092, 1,287,155)
wage02-3	23,647,695	6,368,924	16,199,098	1,079,673
		(6,139,738, 6,597,414)		(851,183, 1,308,859)
wage02-4	27, 900,353	7,204,292	19,592,672	1,103,389
		(6,965,088, 7,434,342)	(873,339, 1,342,593)	
wage02-a	98,885,583	25,314,368	69,028,844	4,542,371
wage03-1	26,571,054	7,796,647	17,599,297	1,175,110
		(7,476,104, 8,097,539)		(874,218, 1,495,653)
wage03-2	25,017,823	6,602,844	17,289,908	1,125,071
		(6,302,386, 6,922,295)		(805,620, 1,425,529)
wage03-3	26,713,862	8,233,839	17,302,366	1,177,657
		(7,912,920, 8,548,211)		(863,285, 1,498,576)
wage03-4	32,011,096	8,794,890	22,044,744	1,171,462
			(21,727,171, 22,363,838)	(852,368, 1,489,035)
wage03-a	110,313,835	$\mathcal{S}$	$\mathcal{S}$	$\mathcal{S}$
wage04-1	23,082,164	8,096,669	13,824,717	1,160,778
			(13,510,783, 14,151,890)	(833,605, 1,474,712)
wage04-2	22,773,180	7,932,895	13,679,769	1,160,516
			(13,362,913, 13,992,710)	(847,575, 1,477,372)
wage04-3	23,269,552	8,620,975	13,482,879	1,165,698
			(13,170,373, 13,796,721)	(851,856, 1,478,204)
wage04-4	28,673,482	9,383,772	18,108,725	1,180,985
			(17,771,926, 18,429,733)	(859,977, 1,517,784)
wage04-a	97,798,378	34,034,311	$\mathcal{S}$	$\mathcal{S}$
wage05-1	21,721,426	7,358,822	13,268,901	1,093,703
			(13,061,204, 13,478,068)	(884,536, 1,301,400)
wage05-2	21,716,384	7,582,785	13,032,603	1,100,996
			(12,823,436, 13,240,300)	(893,299, 1,310,163)
wage05-3	25,895,877	9,134,881	15,689,149	1,071,847
wage05-4	30,344,595	9,667,405	19,318,854	1,358,336
wage05-a	99,678,282	33,743,893	61,309,507	4,624,882
wage06-1	23,653,708	8,605,217	13,883,597	1,164,894
wage06-2	23,924,694	9,082,470	13,514,676	1,327,548
wage06-3	21,323,373	8,405,353	11,707,047	1,210,973
wage06-4	28,035,179	9,988,831	16,537,826	1,508,522
wage06-a	96,936,954	36,081,871	55,643,146	5,211,937

In terms of ξ_j and σ_j^2, $\alpha_j = 3$, $\beta_j = 0.1$, and $\tau_j = \kappa_j = 0.01$ for $j = 1, 2, 3$. Next, we ran a single MCMC chain for 10,000 iterations discarding the first 5000 iterations for burn-in. Convergence of the MCMC was verified through trace plots of the posterior.

In keeping with the disclosure practices of the BLS we cannot present imputed values and measures of accuracy simultaneously. Instead we display in Table 6 the cumulative percentage of values that fall within 1%, 2%, 5%, and 10% of their true values. Further, the pattern of missingness is not the same for each dataset, for example, see Tables 1 and 2. Therefore, we present the percentage and number of missing values for each subseries for each dataset in Table 7. In addition, Table 7 also indicates which datasets have subseries missing the annual total.

As depicted in Table 6, we are able to impute at least 20% of the suppressed values to within 1% of their true values in over half of the datasets considered. Additionally, in 5 of the 11 datasets 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 datasets 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 datasets we are able to impute all of the missing values to within 5% of their true values. Finally, in 8 of the 11

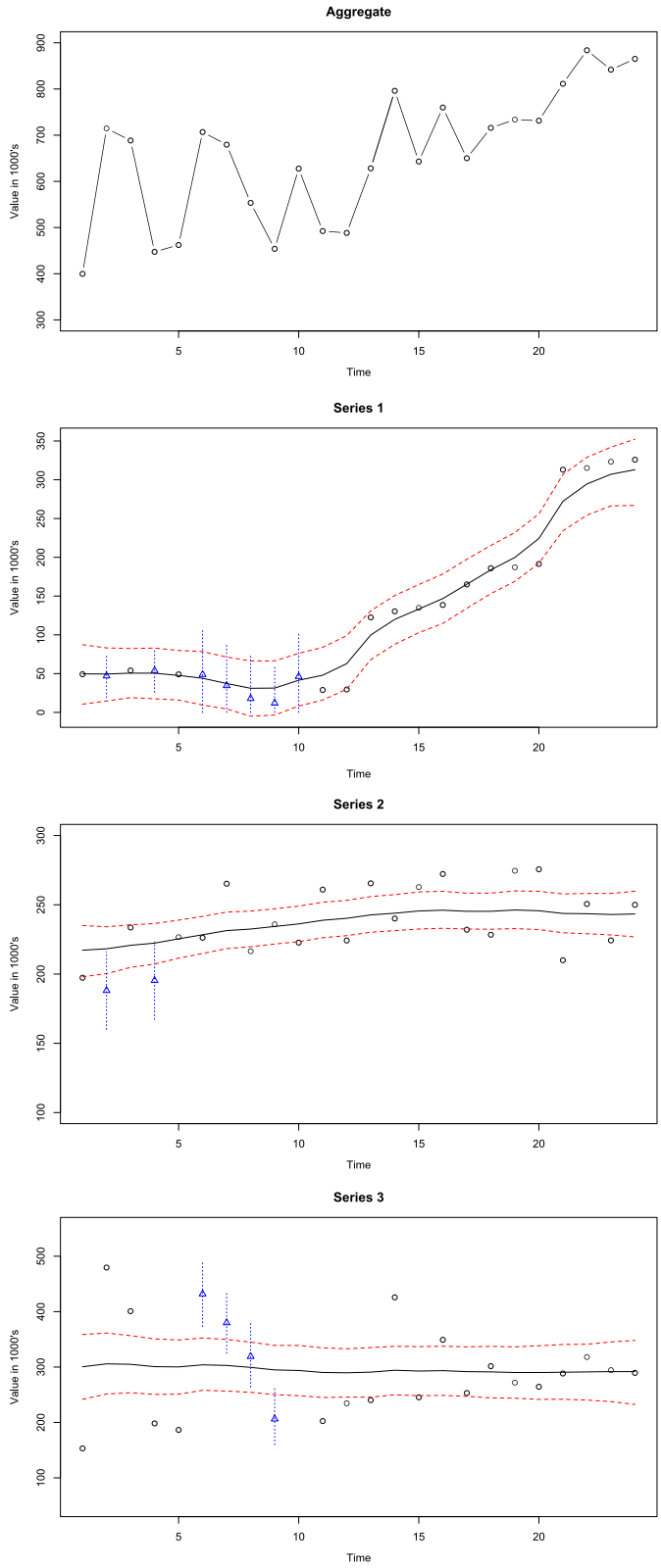

Figure 1. Aggregate series along with three subseries corresponding to QCEW dataset 1. The circles are the observed data and the triangles are the imputed suppressed cells. The solid line represents the estimated latent process whereas the horizontal dashed lines and vertical dashed lines correspond to the 95% pointwise credible interval for the latent process and imputed suppressed cells, respectively. A color version of this figure is available in the electronic version of this article.

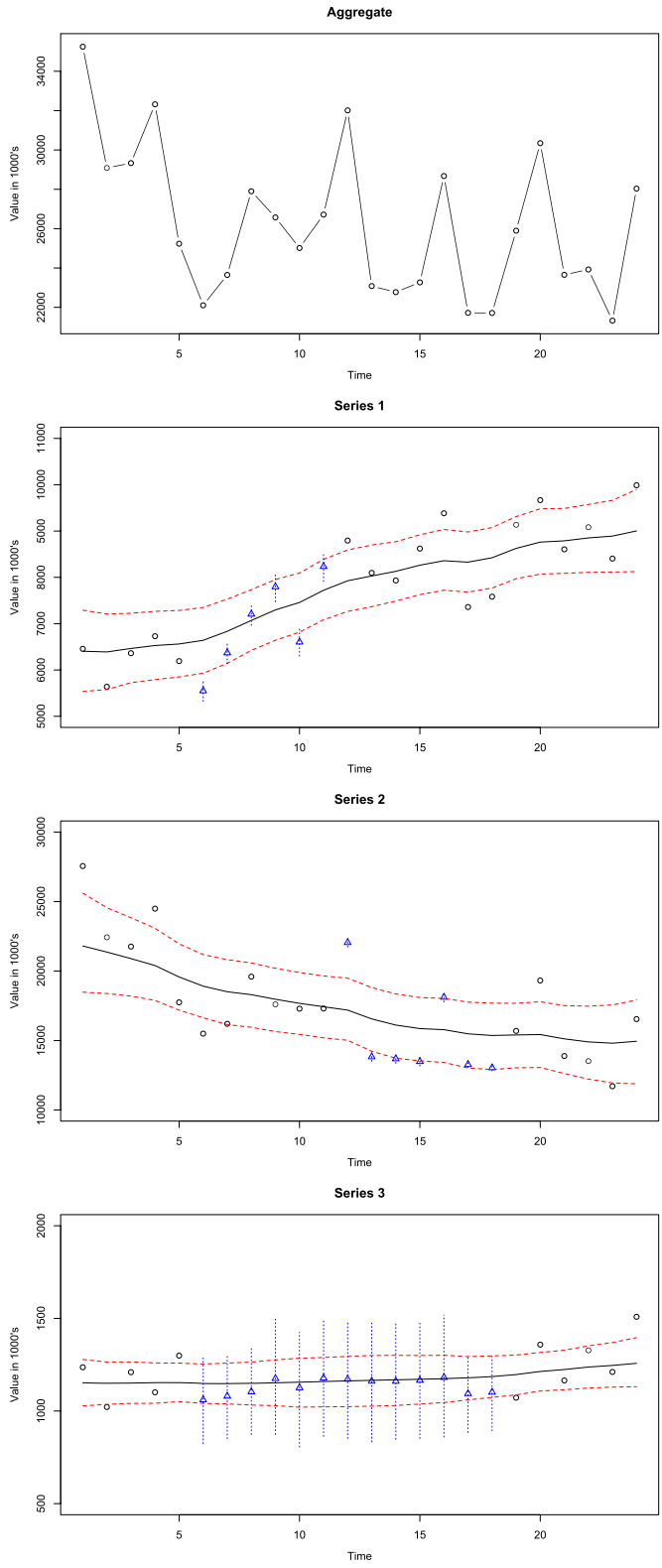

Figure 2. Aggregate series along with three subseries corresponding to QCEW dataset 2. The circles are the observed data and the triangles are the imputed suppressed cells. The solid line represents the estimated latent process whereas the horizontal dashed lines and vertical dashed lines correspond to the 95% pointwise credible interval for the latent process and imputed suppressed cells, respectively. A color version of this figure is available in the electronic version of this article.

Table 5. Simulated QCEW dataset and imputed suppressed cells along with 95% credible intervals, with values rounded to the nearest whole dollar. The data are simulated using the estimated latent processes and observation variances from the model fit to the data from Table 2. $\mathcal{S}$ indicates that the annual total has been suppressed during imputation, imputed cells are in bold face, 95% credible intervals are in parentheses and simulated truth is given below the imputed value

	Total	Series 1	Series 2	Series 3
wage01-1	27,443,858	5,560,324	20,956,820	926,714
wage01-2	34,895,854	5,545,158	28,074,525	1,276,171
wage01-3	23,681,853	7,241,424	15,374,154	1,066,275
wage01-4	30,853,633	5,925,113	23,878,473	1,050,047
wage01-a	116,875,197	24,272,018	88,283,972	4,319,207
wage02-1	26,428,227	5,644,354	19,656,222	1,127,651
wage02-2	34,622,518	**6,992,114**	26,473,969	**1,156,436**
		(6,760,100, 7,219,497)		(929,053, 1,388,450)
		7,008,673		1,139,876
wage02-3	27,218,433	**7,196,340**	18,855,186	**1,166,907**
		(6,969,889, 7,416,209)		(947,038, 1,393,359)
		7,204,472		1,158,775
wage02-4	26,626,885	**7,417,536**	18,029,381	**1,179,969**
		(7,186,112, 7,646,670)		(950,835, 1,411,393)
		7,392,843		1,204,661
wage02-a	114,896,067	27,250,344	83,014,759	4,630,964
wage03-1	27,546,176	**6,260,091**	20,136,569	**1,149,517**
		(5,964,359, 6,579,573)		(830,035, 1,445,249)
		6,260,717		1,148,890
wage03-2	24,529,017	**7,268,469**	16,067,398	**1,193,149**
		(6,972,584, 7,587,713)		(873,905, 1,489,034)
		7,327,113		1,134,506
wage03-3	32,263,843	**7,863,037**	23,183,411	**1,217,395**
		(7,552,511, 8,176,789)		(903,643, 1,527,921)
		7,749,997		1,330,435
wage03-4	25,792,818	7,957,334	**16,623,057**	**1,212,427**
			(16,311,749, 16,943,968)	(891,516, 1,523,735)
			16,547,690	1,287,794
wage03-a	110,131,854	29,295,161, $\mathcal{S}$	75,935,068, $\mathcal{S}$	4,901,625, $\mathcal{S}$
wage04-1	23,366,852	6,374,607	**15,766,228**	**1,226,017**
			(15,447,304, 16,084,415)	(907,831, 1,544,942)
			15,775,969	1,216,267
wage04-2	25,748,661	8,523,937	**15,990,338**	**1,234,386**
			(15,680,075, 16,303,342)	(921,381, 1,544,649)
			16,073,429	1,151,295
wage04-3	28,036,930	8,738,304	**18,052,181**	**1,246,445**
			(17,749,427, 18,381,373)	(917,253, 1,549,200)
			18,265,939	1,032,687
wage04-4	28,627,141	8,648,610	**18,717,428**	**1,261,103**
			(18,405,834, 19,039,923)	(938,607, 1,572,697)
			18,871,679	1,106,582
wage04-a	105,779,584	32,285,458	68,987,016, $\mathcal{S}$	4,507,110, $\mathcal{S}$
wage05-1	20,412,883	8,634,799	**10,500,630**	**1,277,453**
			(10,294,532, 10,709,460)	(1,068,623, 1,483,552)
			10,384,820	1,393,264
wage05-2	27,261,060	9,253,787	**16,711,263**	**1,296,011**
			(16,502,433, 16,917,361)	(1,089,912, 1,504,841)
			16,827,073	1,180,200
wage05-3	25,511,257	8,279,555	15,880,200	1,351,502
wage05-4	20,028,561	7,747,870	11,206,373	1,074,318
wage05-a	93,213,761	33,916,011	54,298,466	4,999,284
wage06-1	21,707,795	8,444,422	11,903,157	1,360,216
wage06-2	22,878,272	8,434,868	13,048,299	1,395,105
wage06-3	33,539,086	9,102,959	22,906,572	1,529,555
wage06-4	29,724,012	9,003,361	19,368,464	1,352,187
wage06-a	107,849,165	34,985,610	67,226,492	5,637,063

Table 6. Percentage of imputed values within 1%, 2%, 5%, and 10% of the true values for both primary and secondary cell suppressions combined. QCEW1–QCEW11 denote the 11 different QCEW datasets used in this empirical investigation. BMMI and MICE denote Bayesian multiscale multiple imputation and multivariate imputation by chained equation (R-contributed package), respectively

	BMMI				Carry forward			
Data	<1%	<2%	<5%	<10%	<1%	<2%	<5%	<10%
QCEW1	7.14	42.86	57.14	71.43	0.00	0.00	7.14	28.57
QCEW2	0.00	0.00	5.00	50.00	0.00	0.00	5.00	15.00
QCEW3	25.00	50.00	100.00	100.00	0.00	25.00	50.00	50.00
QCEW4	10.00	50.00	50.00	60.00	10.00	10.00	20.00	60.00
QCEW5	48.39	70.97	83.87	93.55	0.00	12.90	16.13	29.03
QCEW6	0.00	0.00	0.00	10.00	0.00	0.00	5.00	10.00
QCEW7	21.43	28.57	50.00	57.14	0.00	0.00	7.14	28.57
QCEW8	30.00	30.00	30.00	50.00	0.00	5.00	10.00	30.00
QCEW9	0.00	9.09	13.64	31.82	0.00	0.00	9.09	22.73
QCEW10	50.00	87.50	100.00	100.00	12.50	12.50	62.50	75.00
QCEW11	62.50	75.00	100.00	100.00	12.50	12.50	25.00	50.00
	MICE				Equal proportion			
	<1%	<2%	<5%	<10%	<1%	<2%	<5%	<10%
QCEW1	0.00	0.00	0.00	7.14	0.00	0.00	7.14	7.14
QCEW2	0.00	10.00	25.00	35.00	0.00	0.00	7.14	7.14
QCEW3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	50.00
QCEW4	0.00	0.00	10.00	20.00	0.00	0.00	10.00	10.00
QCEW5	0.00	0.00	10.00	20.00	3.23	16.13	35.48	67.74
QCEW6	0.00	0.00	5.00	10.00	0.00	0.00	0.00	0.00
QCEW7	0.00	7.14	14.29	35.71	0.00	0.00	21.43	42.86
QCEW8	0.00	0.00	5.00	10.00	0.00	0.00	0.00	10.00
QCEW9	0.00	0.00	4.55	9.09	22.73	22.73	40.91	54.55
QCEW10	12.50	12.50	25.00	62.50	0.00	0.00	0.00	0.00
QCEW11	12.50	12.50	18.75	56.25	6.25	18.75	37.50	81.25

datasets 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.

Several observations are worth noting. First, visual inspection of the patterns of suppressed cells in conjunction with the performance of the BMMI method did not reveal any strong patterns. However, it does seem that the BMMI method performs better on series where there are many more secondary suppressions than primary. Further, it seems that the BMMI method less accurately imputes primary suppressed cells when they are suppressed due to a small number of establishments or employers than for other suppression reasons.

Table 7. Percentage of missing values by series within a particular dataset. The number in parenthesis denotes the number of missing values out of the 24 quarters and number missing out of 6 annual totals respectively. QCEW1–QCEW11 denote the QCEW datasets used in this empirical investigation

	Percent and number missing			
Data	Aggregate	Series 1	Series 2	Series 3
QCEW1	0.00	23.33 (7, 0)	6.67 (2, 0)	16.67 (5, 0)
QCEW2	0.00	0.00 (0, 0)	33.33 (8, 2)	33.33 (8, 2)
QCEW3	0.00	6.67 (2, 0)	6.67 (2, 0)	0.00 (0, 0)
QCEW4	0.00	0.00 (0, 0)	16.67 (4, 1)	16.67 (4, 1)
QCEW5	0.00	23.33 (6, 1)	30.00 (7, 2)	50.00 (13, 2)
QCEW6	0.00	0.00 (0, 0)	33.33 (8, 2)	33.33 (8, 2)
QCEW7	0.00	0.00 (0, 0)	23.33 (6, 1)	23.33 (6, 1)
QCEW8	0.00	0.00 (0, 0)	33.33 (8, 2)	33.33 (8, 2)
QCEW9	0.00	36.67 (8, 3)	0.00 (0, 0)	36.67 (8, 3)
QCEW10	0.00	13.33 (4, 0)	0.00 (0, 0)	13.33 (4, 0)
QCEW11	0.00	0.00 (0, 0)	26.67 (8, 0)	26.67 (8, 0)

By construction the BMMI method respects the aggregate constraints. As a consequence, when only two suppressions are imposed, which is typically the case in situations considered here, the error for one imputed cell is the negative of the error of its imputed complementary cell. Therefore, the percentage of error to the true value is smaller for large cells than it is for small cells. The primary suppressed cells are often, but far from always, the smaller cell; thus, in general, the imputations are often closer percentage-wise to the secondary suppressions than they are to the primary suppressions.

Agencies considering the use of cell suppression may exhibit little concern if secondary suppressions are accurately imputed (assuming they are not sensitive) so long as primary suppressions are poorly imputed. Therefore, it is of particular interest to evaluate the performance of the BMMI methodology in the context of only primary suppressions. For primary suppressions, Table 8 displays the cumulative percentage of values that fall within 1%, 2%, 5%, and 10% of their true values. In this case, we are able to impute over 30% of the values to within 1% of their true values for 4 of the 11 datasets. Additionally, we can impute over 50% of the values to within 2% of their true values in 4 of the 11 datasets and over 80% of the values to

Table 8. Percentage of imputed values within 1%, 2%, 5%, and 10% of the true values for primary cell suppressions only. QCEW1–QCEW11 denote the 11 different QCEW datasets used in this empirical investigation. BMMI and MICE denote Bayesian multiscale multiple imputation and multivariate imputation by chained equation (R-contributed package), respectively

	BMMI				Carry forward			
Data	<1%	<2%	<5%	<10%	<1%	<2%	<5%	<10%
QCEW1	0.00	16.67	33.33	33.33	0.00	0.00	0.00	16.67
QCEW2	0.00	0.00	0.00	20.00	0.00	0.00	0.00	0.00
QCEW3	0.00	0.00	100.00	100.00	0.00	0.00	100.00	100.00
QCEW4	0.00	0.00	0.00	20.00	0.00	0.00	20.00	40.00
QCEW5	0.00	50.00	50.00	100.00	0.00	50.00	50.00	50.00
QCEW6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QCEW7	33.33	50.00	83.33	83.33	0.00	0.00	0.00	33.33
QCEW8	30.00	30.00	30.00	50.00	0.00	0.00	0.00	40.00
QCEW9	0.00	9.09	9.09	18.18	0.00	0.00	9.09	18.18
QCEW10	50.00	50.00	100.00	100.00	50.00	50.00	50.00	50.00
QCEW11	33.33	66.67	100.00	100.00	0.00	0.00	0.00	16.67
	MICE				Equal proportion			
	<1%	<2%	<5%	<10%	<1%	<2%	<5%	<10%
QCEW1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QCEW2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QCEW3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00
QCEW4	0.00	0.00	0.00	0.00	0.00	0.00	20.00	20.00
QCEW5	0.00	0.00	0.00	0.00	0.00	50.00	50.00	100.00
QCEW6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
QCEW7	0.00	0.00	0.00	33.33	0.00	0.00	0.00	0.00
QCEW8	0.00	0.00	10.00	20.00	0.00	0.00	0.00	10.00
QCEW9	0.00	0.00	0.00	0.00	45.45	45.45	81.82	100.00
QCEW10	0.00	0.00	50.00	50.00	0.00	0.00	0.00	0.00
QCEW11	0.00	0.00	0.00	33.33	16.67	33.33	33.33	50.00

within 5% of their true values for 4 of the 11 datasets. Finally, in 3 of the 11 datasets we are able to impute 100% of the values to within 5% of their true values whereas in 4 of the 11 datasets we are able to impute 100% of the values to within 10%.

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 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 subseries 2 and 3. In this case, the imputation method is essentially trying to forecast two years ahead (eight steps ahead) based on four 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.

To evaluate the effectiveness of our method, we compared our approach with two naïve approaches and one more or less "standard" approach that can be easily implemented using the R programming language (R Development Core Team 2009). The first approach we considered was "carry forward." For this method we took the last nonmissing value to be our imputed value. In the case that there were no previous nonmissing values we used the first subsequent nonmissing value as our imputed value. The second method we considered was imputation by "equal proportion." For this imputation method we set our imputed value equal to the average proportion of the subseries aggregate total, taken over each quarter with nothing missing. For the series we considered in this comparison, this was possible since no subseries aggregate totals were missing. Finally, we considered imputation using "multivariate imputation by chained equations" (MICE) (Van Buren and Oudshoorn 1999) found in the R-contributed package "mice" (R Development Core Team 2009). For a complete description of this method see Little and Rubin (2002, p. 217).

In general, our method substantially outperformed all three of the other methods, as shown in Tables 6 and 8. Additionally, it is important to stress that our method preserves the inherent additive structure, which the naïve approaches and MICE fail to do. If a data intruder simply wants to estimate missing cells then this is not a problem. However, for the typical "data user," interested in conducting analyses, preserving these additive relationships is crucial.

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, our approach is computationally feasible and produces estimates sufficiently rapidly to allow imputation in practical situations. In fact, in our illustrations in Section 3.2 and empirical study in 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 requires no parameter specification by the user. Given the choice of the particular DLM, we handle the unknown "problem-specific" parameters by employing a set of default priors that require no subjective specification.

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

In general, no imputation method can be expected to perform well in situations where the percentage of missingness is excessive. In many instances our approach can overcome a high percentage of missing data by borrowing strength through longitudinal and subseries 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 datasets 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.

Several SDL techniques are available to statistical agencies and can be used for disclosure protection within QCEW. In particular, remote analysis servers can be established that will release only the results of statistical analyses of the data. In this case there will be no release of the QCEW data. Alternatively, it will be possible to release only carefully aggregated data that will limit the risk of disclosure. Finally, our method can be used to replace sensitive (plus complementary) cells with partial synthesis (Reiter 2003). However, this will need to be accompanied by additional synthesis; otherwise, this will present the same or even larger disclosure risks than the corresponding cell suppressions since the agency now approximately provides the intruder the model.

Importantly, our 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. In fact, agencies can develop models that enforce (and do not enforce) the confidentiality rules as part of their disclosure checks. This will enable agencies to determine how much the protection relies on keeping the confidentiality rules secret. Finally, 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.

[Received November 2008. Revised October 2009.]

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. [565]

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State Space Models," *Biometrika*, 81, 541–553. [568,569]

Cohen, S., and Li, B. T. (2006), "A Comparison on Data Utility Between Publishing Fixed Intervals versus Traditional Cell Suppression on Tabular Employment Data," Statistical Survey Paper, U.S. Bureau of Labor Statistics, available at *http://www.bls.gov/ore/abstract/st/st060100.htm*. [565,566]

Cox, L. H. (1980), "Suppression Methodology and Statistical Disclosure Control," *Journal of the American Statistical Association*, 75, 377–385. [565,566]

——— (1995), "Network Models for Complementary Cell Suppression," *Journal of the American Statistical Association*, 90, 1453–1462. [565,566]

Dobra, A., Karr, A., and Sanil, A. (2003), "Preserving Confidentiality of High-Dimensional Tabulated Data: Statistical and Computational Issues," *Statistics and Computing*, 13, 363–370. [564]

Dobra, A., Karr, A., Sanil, A., and Fienberg, S. (2002), "Software Systems for Tabular Data Releases," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 529–544. [564]

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. [564]

Durbin, J., and Koopman, S. (2001), *Time Series Analysis by State Space Methods*, Oxford: Oxford University Press. [567]

Ferreira, M. A. R., and Lee, H. K. H. (2007), *Multiscale Modeling: A Bayesian Perspective*, New York: Springer. [565]

Ferreira, M. A. R., West, M., Lee, H. K. H., and Higdon, D. (2006), "Multiscale and Hidden Resolution Time Series Models," *Bayesian Analysis*, 1, 947–968. [565]

Fienberg, S. (2006), "Privacy and Confidentiality in an e-Commerce World: Data Mining, Data Warehousing, Matching and Disclosure Limitation," *Statistical Science*, 21, 143–154. [564]

Fischetti, M., and Salazar, J. J. (2001), "Solving the Cell Suppression Problem on Tabular Data With Linear Constraints," *Management Science*, 47, 1008–1026. [565,566]

Früwirth-Schnatter, S. (1994), "Data Augmentation and Dynamic Linear Models," *Journal of Time Series Analysis*, 15, 183–202. [568,569]

Gamerman, D., and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (2nd ed.), Boca Raton: Chapman & Hall/CRC. [564,567,569]

Gelfand, A., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [569]

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. [569]

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. [564]

Harvey, A. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge: Cambridge University Press. [567]

Jeffrey, R. C. (1992), *Probability and the Art of Judgement*, Cambridge: Cambridge University Press. [565]

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006), "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality," *The American Statistician*, 60, 224–232. [564]

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. [564]

Kelly, J. (1990), "Confidentiality Protection in Two and Three-Dimensional Tables," unpublished Ph.D. thesis, University of Maryland College Park. [566]

Little, R. J., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, New York: Wiley. [566,575]

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008), "Privacy: Theory Meets Practice on the Map," in *Proceedings of the 24th International Conference on Data Engineering*, Los Alamitos, CA: IEEE Computer Society, pp. 277-286. [564]

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, New York: Academic Press. [568]

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Theory*, New York: Wiley. [565,568]

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org*. [575]

Reiter, J. (2003), "Inference for Partially Synthetic, Public Use Microdata Sets," *Survey Methodology*, 29, 181–188. [576]

——— (2005), "Estimating Risks of Identification Disclosure for Microdata," *Journal of the American Statistical Association*, 100, 1103–1113. [564]

Reiter, J., and Raghunathan, T. (2007), "The Multiple Adaptations of Multiple Imputation," *Journal of the American Statistical Association*, 102, 1462–1471. [564]

Robert, C. P., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer-Verlag. [567]

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley. [566]

Schafer, J. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15. [567]

Schouten, B., and Cigrang, M. (2003), "Remote Access Systems for Statistical Analysis of Microdata," *Statistics and Computation*, 13, 381–389. [564]

Searle, S. R. (1982), *Matrix Algebra Useful for Statisticians*, New York: Wiley. [568]

Siotani, M., Hayakawa, T., and Fujikoshi, Y. (1985), *Modern Multivariate Statistical Analysis*, Columbus: American Sciences Press. [565,568]

Van Buren, S., and Oudshoorn, C. G. M. (1999), *Flexible Multivariate Imputation by MICE*, Leiden: TNO Preventie en Gezondheid. [575]

West, M., and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models* (2nd ed.), New York: Springer. [565,567]

Willenborg, L., and de Waal, T. (2001), *Elements of Statistical Disclosure Control*, New York: Springer. [564,567]