Probabilistic, Bayesian updating of IOTs: application to WIOD

Oleg Lugovoy[‡§] Andrey Polbin[§] Vladimir Potashnikov[§]

Conference paper
25th International Input-Output Conference
19-23 June 2017, Atlantic City, New Jersey, USA

Abstract

The paper summarizes the authors' efforts on developing and application of probabilistic method(s) for updating IO tables, preliminary presented and discussed on previous IIOA conferences. The core of the methodology is Bayesian framework which combines an information from observed data, additional believes (priors), and related uncertainties into posterior joint distribution of input-output table (IOT) coefficients. As we show in the paper, the framework can be applied to various IOT problems, including updating, disaggregation, evaluation of uncertainties in the data, and addressing incomplete/missing observations. The flexibility of the methodology is partially based on sampling techniques. We apply modern Monte Carlo Markov Chains (MCMC) methods to explore posterior distribution of IOT coefficients. We also compare results with mainstream methods of updating IOT to investigate its performance. Various indicators of performance and application to various data suggest different results. The overall performance of the method is similar or comparable with mainstream techniques. The main advantage the proposed methodology is an estimation of full profile of joint probability distribution of unknown IOT matrices. The method can be also combined with any other techniques through prior information.

[‡] Environmental Defense Fund, USA
[§] The Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia

Table of Contents

Introduction

The paper summarizes recent authors' efforts on developing and application of Bayesian methods of updating IO tables, preliminary presented by the authors on the 22th IIOA, and extends the methodology and results in extending the application[5] to the The World Input-Output Database (WIOD) (Timmer at al., 2015). We also update and extend estimates of IOT tables for Russia, and test the methodology on actual IOTs for US, considering it as a higher quality data.

The paper includes several parts. In the first part we discuss conceptual frameworks for updating, disaggregation, and balancing IO tables. In the second part we apply the methodology to WIOD data. We treat the last table for each country as unknown and estimate it with the Bayesian method using all previously available matrixes for constructing prior distribution. When specifying prior distribution we argue that Beta distribution for IO coefficients is more appropriate than Normal distribution and fit it for the each coefficient on previously available matrixes. We consider two point estimates of "unknown" IO table: posterior mode and posterior mean. To find posterior mode we use nonlinear optimization techniques, to explore posterior distribution we use modern MCMC methods. Posterior mode robustly outperforms competitive methods, popular in the literature, according to different closeness statistics. Posterior mean perform slightly worse than posterior mode. We conclude that point estimate of Bayesian method at least is compatible with the other methods on real data examples.

But the main contribution of our method is that it provide probabilistic estimate of IO coefficients consistent with all available data constraints. This property is very useful for analyzing uncertainty about IO coefficients and results of the models that calibrated to IO tables. After comparing point estimates of the Bayesian method of "unknown" IO table with alternative methods, we concentrate on the constructing creditable set for IO coefficients. We provide arguments that standard symmetric creditable interval for input-output coefficient is inappropriate and induce significant bias. We argue for using higher posterior credible set for characterization of the uncertainty. We construct credible sets for estimates of IRIOS tables and for the results of some simple IO models. We also perform Monte Carlo experiments were we show that posterior higher posterior credible set have better coverage properties.

In the third part of the paper, we upgrade and extend estimates of SUT tables for Russia. Russian statistical system is under transition for almost two decades from Soviet type Material Product System to the System (MPS) of National Accounts (SNA). The main transitional break in methodology took place in 2003-2004 when Russian statistical agency "Rosstat" started reporting

[5] In previous version of the paper we applied the methodology to IRIOS tables (van der Linden and Oosterhaven, 1995).

based on the new definition of economic sectors consistent with NACE, and stopped reporting using definition of activities inherited from the Soviet statistical system. This methodological break splits all industry level statistics into two periods with little consistency between each other. As a result, Rosstat stopped updating input-output tables (IOT) in 2003, based on the only benchmark survey conducted in 1995. The next survey is scheduled for 2011 with expected publication of results in 2015 or later. Official backward estimation is not expected. Therefore Russian statistics will miss IOT at least from 2004 to 2010. Also quality of officially updated IOT from 1996 to 2003 based on 1995 benchmark is questionable.

We apply Monte Carlo Markov Chains (MCMC) methods to disaggregate available in NACE classification SUTs (2006, 15 products by 15 activities) into larger 69 by 69 format. Since the 15x15 SUTs are published by Rosstat as preliminary estimates, they are not fully consistent with other available national accounts data, such as output and value added by industries. To take into account the data uncertainty, we introduce a measurement error for the aggregated io-coefficients. As result, we estimate posterior distribution of input-output coefficients for aggregated and disaggregated matrices, which are consistent with yearly national accounts information. Than we update the estimated 15x15 matrices for 2007-2012 period, using proposed sampling methods, and compare results with alternative approaches. Than we compared our estimation IO table of 2011 for Russia with issued by Russian statistical agency. Note that IO table of 2011 for Russia was issued in SNA 2008, while previous IO table was in SNA1993.

The paper includes three parts. In the first part, we discuss the conceptual framework of application of Bayesian techniques to probabilistic updating of IOTs, disaggregation, addressing measurement errors in data, missing observations, various specifications of priors, and computer implementation. In the second part, we test the methodology on actual data, World Input Output Database (WIOD), and compare its performance with other mainstream techniques of IOT updating. For estimation WIOD tables only based years was used, instead of previous paper.

In the third part we apply the methodology to build probabilistic IOTs for Russia. We compare our estimation for 2011 with IO table issued by Russian statistical agency, but in different SNA system. Based on information from national accounts for 70+ industries, and preliminary official IOT estimates for 15 main sectors, published by Rosstat for 2006, we are trying to reconstruct probabilistic IOTs for 77 sectors, using the Bayesian techniques for disaggregation and updating IOTs up to 2014.

In addition to another IOT updating technique, the main contribution and advantage of proposed methodology is a straightforward and practically achievable quantification of uncertainties in input-output tables, consistent with directly and indirectly linked with IOTs observed data, and any amount of additional information, which can be expressed by inequality constraints for IO coefficients and their linear combinations.

1. Conceptual framework

In this section we discuss an application of Bayesian framework and Monte Carlo Markov Chains method for updating, disaggregation, and balancing IOT.

1.1. Updating IOT with Bayesian methods

The basic problem of updating an IO matrix or more generally a SAM can be defined as finding of an unknown IO matrix with known sums of rows and columns, and known IO matrix for a previous year(s). Mathematically speaking, we need to find a matrix *A* with following restrictions:

$$\begin{aligned} &Y = AX, \\ &\sum\nolimits_i a_{i,j} = \bar{a}_j, \quad a_{i,j} \geq 0 \end{aligned} \tag{1}$$

where *Y*, *X* are known vectors and $\bar{a}_j$ are known sums of columns. Since there is no unique solution with the set of constrains on sum of rows and columns only, a known matrix A^0 (f.i. from previous year) is used as a starting point. The solution is usually reduced to finding such matrix *A*, which minimize some distance function from known matrix A^0 under a set of constrains (1).

The problem (1) can be also solved with Bayesian methods, which provide a natural and flexible way to incorporate any kind and amount of information either as a prior distribution or observable data. Moreover, Bayesian methods provide full density profile on estimated parameters with covariates. The information can be very valuable in evaluating quality of the estimates, magnitude with which each particular io-cell's estimate affects all others, the level of uncertainties and how they affect results of an analysis based on the estimated tables.

In Bayesian econometrics some prior information or beliefs about estimated parameter θ could be summarized by prior density function $p(\theta)$ according to Bayes theorem:

$$p(\theta \mid Y) = \frac{L(Y \mid \theta) p(\theta)}{\int L(Y \mid \theta) p(\theta) d\theta} \propto L(Y \mid \theta) p(\theta) \tag{2}$$

where $p(\theta \mid Y)$ is the posterior density and $L(Y \mid \theta)$ is the likelihood.

Bayesian inference is easy since the posterior density contain all the information one may need. The researcher could be interested in point estimate, credible set and correlation of parameters and construct it from posterior distribution. In Bayesian framework point parameter estimate is chosen to minimize expected loss function with expectation taken with respect to the posterior distribution. The most common loss function used for Bayesian estimation is the mean square error and the corresponding point parameter estimate is simply the mean of the posterior distribution.

Despite the attractiveness of this method, in the past, Bayesian inference was not so popular due to numerical integration needed in equation (2). In some cases when the prior on θ is conjugate with posterior on θ the posterior density can be obtained analytically. But in more general setup we know posterior density up to normalizing constant. Recently developed computer-intensive sampling methods such as Monte Carlo Markov Chain (MCMC) methods have revolutionized the application of Bayesian approach. MCMC methods are iterative sampling methods that allow sampling from posterior distribution $p(\theta | Y)$.

Heckelei *et al. (*2008) shortly discuss IOT update with Bayesian method and give an example on artificial data. Authors present a Bayesian alternative to the cross-entropy method for deriving solutions to econometric models represented by undetermined system of equation. In the context of balancing an IO matrix they formulate posterior distribution in the following way:

$$p(z | data) \propto I_{\Psi}(z) p(z) \qquad (3)$$

$$z = vec(A) \qquad (4)$$

Equation (4) means vectorization of matrix *A*. In equation (3) $p(z)$ is some prior distribution, $p(z | data)$ is the posterior distribution and $I_{\Psi}(z)$ is the indicator function that assigns weights of 1 if *z* satisfies the constraints (1) and 0 otherwise. Authors interpret the indicator function as the likelihood function. As estimate of *z* Heckelei *et al.* (2008) consider mode of posterior distribution which could be found with some optimization routine. And they illustrate proposed method balancing small 4x4 matrix with independent normal prior taking A^0 as prior mean.

However the proposed by Heckelei *et al.* (2008) method actually reduced to minimization yet another distance function from known matrix A^0. In this paper we concentrate on finding full density profile of posterior distribution with MCMC techniques and applying it to real data.

For convenience we consider equality and inequality constraints of the system of restriction (1) separately. Inequality constrains could be simply introduced in prior distribution by assigning 0 value of density in inadmissible domain. For example one could specify independent truncated normal distribution between 0 and 1 for each parameter of the matrix *A*. On the other hand if we have certain beliefs about some parameters we could introduce it as additional linear equality constraints. For example it is convenient to assign 0 values for elements of unknown matrix *A* if corresponding elements in the matrix A^0 are zeros.

At the next step let us consider linear equality constraints and rewrite it in the following form:

$$Bz = T \qquad (5)$$

where *B* is the known matrix, *T* is the known vector and $z = vec(A)$ is the unknown vector of estimated parameters. System (5) represents undetermined linear system of equations. And from linear algebra it is known that any solution of linear system (5) could be written in the form:

$$z = \tilde{z} + F^{(1)}\xi^{(1)} \tag{6}$$

where $\tilde{z}$ is the particular solution of the system (5) and $F^{(1)}$ is the fundamental matrix of solutions of homogeneous system $Bz = 0$. And any vector $\xi^{(1)}$ solves system (5). The particular solution and the fundamental matrix could be obtained by Gaussian elimination algorithm.

Columns of the fundamental matrix $F^{(1)} = [f_1^{(1)},.., f_k^{(1)}]$ represent basis of the Euclidean subspace. At the next step we could find the basis of the orthogonal complement of this subspace $F^{(2)} = [f_1^{(2)},.., f_{n-k}^{(2)}]$. Let us consider linear transformation of the original space:

$$\begin{bmatrix} \xi^{(1)} \\ \xi^{(2)} \end{bmatrix} = \begin{bmatrix} F^{(1)} \; F^{(2)} \end{bmatrix}^{-1} (z - \tilde{z}) \tag{7}$$

In the new system of coordinates prior density has the following form:

$$p_\xi(\xi) = \det\begin{bmatrix} F^{(1)} \; F^{(2)} \end{bmatrix} p_Z(\tilde{z} + F^{(1)}\xi^{(1)} + F^{(2)}\xi^{(2)}) \tag{8}$$

If we specify posterior distribution in the form (3) than posterior distribution will be the conditional distribution of random vector $\xi^{(1)}$ given the zero value of the random vector $\xi^{(2)}$:

$$p_\xi(\xi \mid data) = p_{\xi^{(1)}|\xi^{(2)}}(\xi^{(1)} \mid \xi^{(2)} = 0) \tag{9}$$

If prior distribution is multivariate normal distribution, posterior distribution of vector $\xi^{(1)}$ is also multivariate normal and we could compute posterior mean and covariance matrix analytically. But it doesn't guarantee nonnegative values of estimated matrix *A*. In general setup we use truncated prior distribution and know posterior density up to normalizing constant. To conduct inference about parameters we approximate posterior distribution (9) applying MCMC sampling methods. After generating the sample of vectors $\xi^{(1)}$ we could move to initial space using formula (6) and obtain the sample of vectors *z*, which represents elements of unknown matrix *A*.

To obtain sample from posterior distribution for examples in this paper we perform the Metropolis sampling algorithm, which is a special case of a broader class of Metropolis-Hasting algorithms, and apply a standard single-site updating scheme. As a proposal density for generating candidate parameter values we use normal distribution for each parameter of vector $\xi^{(1)}$. Standard deviations of the proposal density are iteratively selected during adaptive phase to guarantee acceptance rate for each parameter to be between 35 and 45 percent.

1.2. Computer implementation

As mention above, system (5) represents undetermined linear system of equations, with solution

$$z = \tilde{z} + F^{(1)} \cdot \xi^{(1)}$$

where $\tilde{z}$ is the particular solution of the system (5) and $F^{(1)}$ is the fundamental matrix, which consists of a system of linearly independent vectors form a basis in the subspace of solutions of (5).

The choice of the fundamental matrix for optimal MCMC is a nontrivial task. Let us look at a simple example. Suppose that the first two rows and first two columns fundamental matrix consist of zeros except for the first 2x2 elements, which are equal:

$$\begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix}$$

Assume that the density of the a priori distribution of the first two elements is $\left(N(0.1, 0.1), N(0.1, 0.01)\right)$. Obviously, with this configuration, you will need much more iterations on MCMC algorithm to obtain a qualitative assessment of the posterior distribution than in the case reconfigure the fundamental matrix with the first 2x2 elements:

$$\begin{vmatrix} 0 & -2 \\ 2 & 0 \end{vmatrix}$$

Use of the second embodiment reduces the number of required iterations and thus the time to more than a hundred times. If the vector prior distribution is $\left(N(0.1, 0.1), N(0.1, 0.001)\right)$, then the more than 10 thousand times.

Solutions for this simple case is obvious, but in a more complex case, not all may be so simple. The problem is that the effect change ξ on the density of the prior distribution is not obvious. To solve this problem, we transform the log prior density distribution using (6):

$$\log p(z) \sim -\sum_j \frac{\left(z_j - \mu_j\right)^2}{2\sigma_j^2} + const = -\sum_j \frac{\left(z_j^0 + \sum_i f_{ij} \cdot \xi_i - \mu_j\right)^2}{2\sigma_j^2} + const$$

Disclosure using parentheses and grouping can reduce the equation to the form

$$\log p(z) \sim \xi' \cdot H \cdot \xi + W \cdot \xi + Q + const$$

where H, W, Q known matrix. In particular, matrix H is equal

$$H = F^{(1)'} \cdot \Sigma \cdot F^{(1)}$$

where Σ is diagonal matrix with elements equal $-\frac{1}{2\sigma^2}$. Then, according to analytic geometry, there exists a coordinate system in the subspace $\xi^{(1)}$, and the corresponding transition matrix D, the matrix H can be reduced to the diagonal form. Arranging suitable ξ(1)0 can cause log prior distribution to the form:

$$p(\beta) \sim -\sum_j \frac{\left(\beta_j - \mu_j^\beta\right)^2}{2{\sigma^\beta}_j^2} + const$$

where μβ и σβ are known parameter from matrix D, W, Q and β is replace ξ(1) by equation:

$$\xi^{(1)} = D \cdot \beta + \xi^{(1)0}$$

and the corresponding fundamental matrix in the new coordinate system has the form:

$$F^{(1')} = F^{(1)} \cdot D$$

In this form, vector β have a clear interpretation — vector of independent multivariate normal distribution. In this case, we clearly understand the a priori distribution of β, and we can choose as a proposal density in the normal distribution with zero mean and σ equal to σβ.

In this approach, there is a clear geometrical interpretation. Isolines prior distribution are n2 — dimensional ellipsoid, centered at μ, and semiaxes proportional σ. The set of points satisfying the constraints in the form of equations are the hyperplane of dimension (n2 – p), where p — the number of linearly independent constraints. Projections of isolines on the hyperplane will be ellipsoids, with dimension equal to the dimension of the hyperplane, with center μβ and semiaxes proportional σβ. Basis vectors, column matrix $F^{(1')}$, parallel to the axes of the ellipsoids.

In this formulation the prior distribution, it is possible directly sample matrix, and check the final matrix in inequality constraints. Advantage of this method MCMC, compare with previous one is the outstanding performance of the resulting Markov chains, and disadvantage of this method is increasing share dropped matrices, and consequently time, together with the increasing dimension estimated matrix. As a result of this estimation econometrician can publish vector μ, σ, and the matrix D, from which the user can generate an arbitrarily large number of matrices required for its purposes.

Prior knowledge of the distribution density of β allows us to calculate the covariance matrix of the variables z, excluding inequality constraints. By definition covariance is:

$$cov_{ij} = \int\int (z_i - \mu_i)\left(z_j - \mu_j\right) dz_i dz_j$$

by use equation

$$z = \tilde{z} + F^{(1')} \cdot \beta$$

with distribution β equal:

$$\beta \sim N\left(\mu^\beta, \sigma^\beta\right)$$

then covariance is equal

$$cov_{ij} = \int\int \left(\sum_k f_{ik} \cdot \beta_k - \widetilde{z_\imath}\right)\left(\sum_k f_{jk} \cdot \beta_k - \widetilde{z_\jmath}\right) dz_\imath dz_j$$

Considering that, the covariance of two random independent variables is zero, then

$$cov_{ij} = \int \sum_k f_{ik} \cdot f_{jk} \cdot \beta_k \, dz_k = \sum_{\mathrm{k}} \mathrm{f_{ik} f_{jk}} \sigma_{\mathrm{k}}^2$$

or in the matrix form

$$\mathrm{COV} = F^{\left(1'\right)'} \cdot \Sigma^{\beta} \cdot F^{\left(1'\right)}$$

where sigma diagonal matrix with coefficients equal $-\frac{1}{2\sigma^{\beta^2}}$. As shown by experiments on real data R2 regression coefficients between the covariance obtained from the equation above, and covariance obtained from MCMC over .97, with a constant equal zero and interception in the range 0.9 — 1.1. In figure 8 shown comparison MCMC chains. Some срфшты in the base version of the algorithm, on the right hand, do not converge. Further increasing the number of iterations by several orders not improve the situation.

Figure 9 shows the distribution coefficients of the covariance between cells IO tables in 1998-2003, in the format OKONH, and 2004-2006 in the format of NACE. As can be seen most of the points located on the bisector and the regression coefficient, constructed between the coefficients close to unity, with the constant zero. R2 for a regression of more than 0.975.

1.3. Bayesian disaggregation of IO tables

The described above method of updating IO tables can be generalized and used for other purposes, including disaggregation. Let's consider the inverse problem to the disaggregation – the aggregation of an IO matrix $\tilde{A}$ of N industries into $\tilde{A}^*$ of dimension n, where N > n. Therefore matrix $\tilde{A}^*$ consist of rows and columns which are sums of rows and columns of matrix $\tilde{A}$. Let's matrix S with dimension $n \times N$ is responsible for the transformation. For example, if two first industries of $\tilde{A}$ should be aggregated into one industry of $\tilde{A}^*$, than the first row of S will have units in the first two elements, and zeros in others. In more general case:

$$S = \begin{cases} S_{i,j} = 1, \ i \in sector\ j \\ S_{i,j} = 0, i \notin sector\ j \end{cases} \tag{10}$$

Therefore, aggregation problem can be written:

$$\tilde{A}^* = S\tilde{A}S' \tag{11}$$

To come back to disaggregation one should find elements of unknown matrix $\tilde{A}$, consistent with (*). The equation () can be rewritten:

$$\mathrm{kron}\left(\mathrm{S},\mathrm{S}\right) * \mathrm{vec}\left(\tilde{A}\right) = \mathrm{vec}\left(\tilde{A}^*\right) \tag{12}$$

where $\text{kron}(\bullet)$ denotes Kronecker product of the matrices, $\text{vec}(\bullet)$ denotes a matrix' vectorization. Let's also assume that intermediate demand for an industry output does not exceed an output of the industry:

$$\sum_j a_{i,j} * x_j \leq x_i \tag{13}$$

The constrain can be presented similarly to (5):

$$D * z \leq X \tag{14}$$

where X is the final output.

1.4. Measurement errors in observed data

National accounts usually have several cycles of publication. First estimates are made on partially available data and usually considered as preliminary. As new data comes, the estimates are updating. Therefore the information for the same economic indicators published in various years may differ.

We faced the problem working on the disaggregation exercise on the real data. The aggregated version of "Use" matrix for 2006 was published earlier than the disaggregated production information for the same year. The data on output, value added, and intermediate consumption from the matrix is not consistent with the same but more detailed statistics. It is likely the information on production was updated, but the Use table was not.

To address the problem we introduce measurement errors to the observed data. We assume that the aggregated matrix, which was published earlier, is measured with an normally distributed error:

$$\left\{\text{kron}(\text{S},\text{S}) * \text{vec}\left(\tilde{A}\right) \text{-vec}\left(\tilde{A}^*\right)\right\} \sim N(0,\Sigma) \tag{15}$$

where

Σ — diagonal matrix with elements proportional to the square of $\text{vec}\left(\tilde{A}^*\right)$. Later we assume that standard deviation of the measurement error for each cell is equal to 10% of the value of the cell. Therefore for the density function of posterior distribution will be:

$$p(a \mid data) \propto p(a) L(data) I(a \mid data) \tag{16}$$

where

$p(a)$ — prior distribution density function,

$L(data)$ — likelihood function for the specified in (15) measurement error,

$I(a \mid data)$ — an indicator function which shows that all the io-coefficients satisfy the set of constrains.

1.5. Computer implementation

The MCMC sampling methodology is computationally intensive. Moreover, quality of results directly depends on number and length of chains. Initially developed algorithm with all the sequence of required operations of multiplications, summation, and comparison took 20 minutes to sample just one matrix. The time is not appropriate for large-scale calculations. For instance, to sample 15 million matrices (an experimentally found suggested minimum size of sample), the algorithm would require 5000 years. However, this straightforward algorithm has a lot of potential for time-efficiency.

First, the matrices are quite sparse. Standard procedures can be applied to improve the time performance. As result, number of elementary operation for 4761 elements decreased from 20 million to 370, with improved time to 0.1 seconds per matrix.

Second, the 370 operations can be paralleled. After reformulating the problem for standard graphical processor supporting CUDA technology, the time was improved to 0.006 seconds per one matrix. See table 1 for more details.

Table1. Time-performance of various sampling algorithms.

№	Algorithm	Software	Time of one matrix (69x69) sampling	Number of elementary operation of summation and multiplication	Comparison operations
1	MCMC	R	> 20 min	$(4761\text{-}N)^2+69^3>2e7$	4830
2	Optimized MCMC	R	~ 0.1 sec	70+N < 364	70+N < 364
3	Optimized MCMC	CUDA	~ 0.006 sec	70+N < 364	70+N < 364

Note: N is a number of linearly independent constrains.

2. Experimental updating of WIOD tables

Here we apply the proposed methodology to the WIOD (Timmer at al., 2015). For estimation WIOD tables only based years was used, instead of previous paper (see Appendex A). In the Bayesian framework we take previous based IO tables as a prior information, assuming initial independent truncated normal distributions for each IO coefficient. Standard deviations of prior distribution are estimated for each coefficient on the all based previously available tables. To compute posterior mean of coefficients we apply Markov chain Monte Carlo (MCMC) method with two chains and sampled length of 5,000,000 simulations. To compute posterior mode of coefficients we use nonlinear programming techniques.

The estimates include 24 experiments from WIOD database (the list of countries is presented in Appendix A). The accuracy of the Bayesian techniques is comparable with other methodologies, and outperforms most of them by major number of considered indicators. However, the best accuracy for the WIOD application keeps RAS method. It should be noted that WIOD tables were balanced by RAS-based methods (see Timmer at al., 2015) which might be the reason of highest RAS performance vs. other techniques (see table 1, more detailed results are in appendix).

One of the advantage of the Bayesian methodology is in its flexibility through using prior information. RAS estimates can be used as prior info as well. Table 2 presents comparative statistics between RAS and the Bayesian estimates, when results of RAS method are used to as prior for Bayesian estimation. The main difference here is that Bayesian mode will be equal to RAS estimation because it is fully consistent with the new data. Therefore, the first row is the table is empty.

Table 2– Estimation WIOD for based years, IO value previous as prior, compare with RAS.

Statistic	RMSE	MAE	MAPE	WAPE	SWAD	Psi	RSQ	AED
MODE is better RAS	13%	4%	0%	4%	8%	4%	13%	0%
MEAN is better RAS	4%	0%	4%	0%	4%	0%	8%	0%

Table 4– Estimation WIOD for based years, RAS as prior, compare with RAS.

Statistic	RMSE	MAE	MAPE	WAPE	SWAD	Psi	RSQ	AED
MODE is better than RAS	—	—	—	—	—	—	—	—
MEAN is better than RAS	33%	8%	0%	8%	29%	8%	33%	4%

As mentioned above, the main advantage and goal of the methodology however is in estimation of full profile distribution of unknown tables. Figure 1 presents an example – a part of an estimated IOT for Australia, and estimated distribution for each cell. The estimates of each cells are highly correlated because of linear relations between cells.

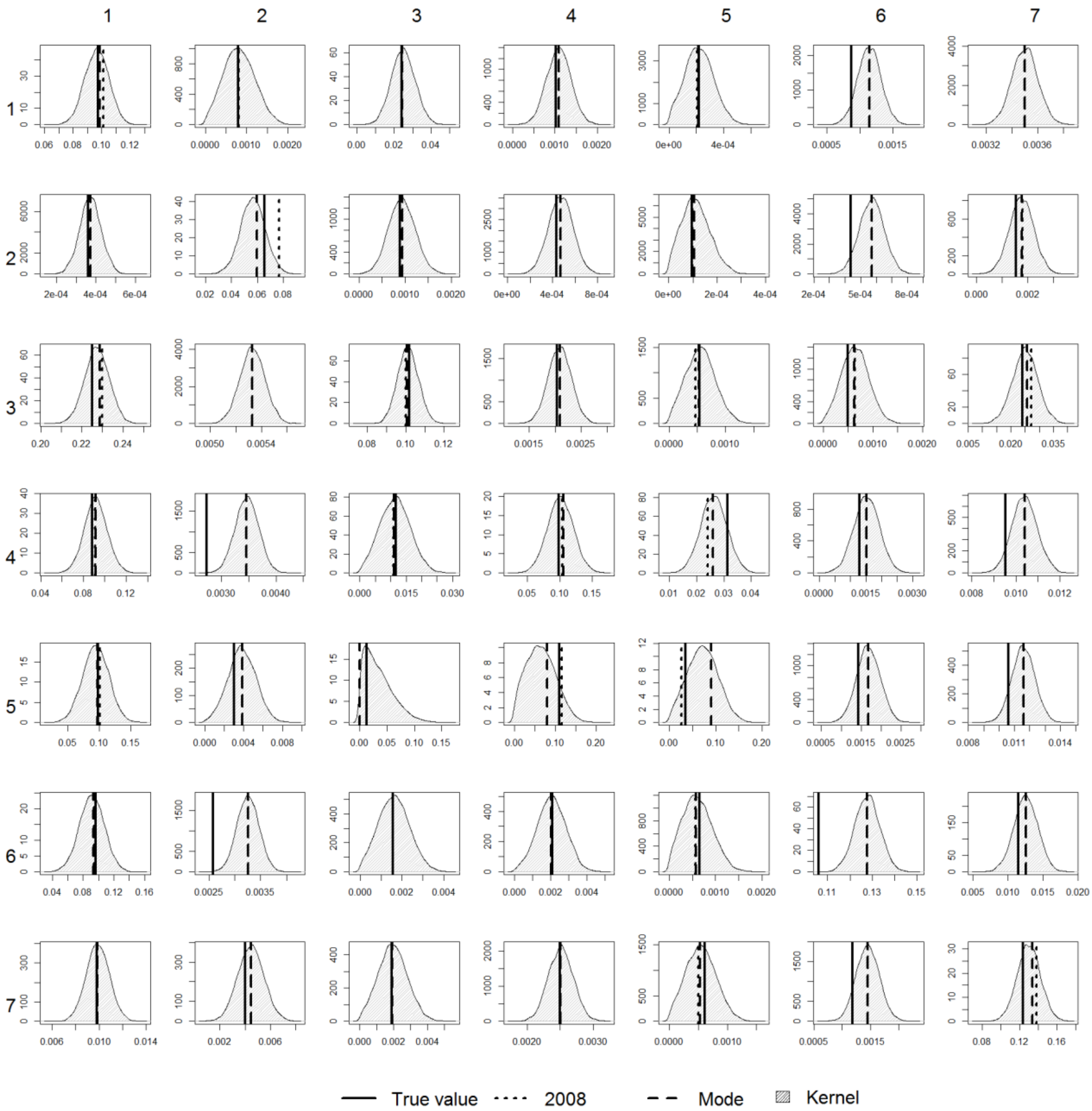

Figure 1. **Estimation WIOD for Australia 2008, IO value 2004 as prior.**

WIOD tables are highly processed, build on data from various sources, countries, and balanced with RAS-based methods. It should be cleared that the tables incorporate some uncertainties due multiple steps of data adjustment and balancing. Ideally the Bayesian methodology should be used on every step to accumulate uncertainties in the data and estimate confidence intervals for each cell, and full profile of the joint distribution. However, the application presented in the paper can be a second-best methodology to evaluate uncertainties in the tables, and/or update the tables for later years. As discussed in the conceptual framework part of the paper, updates are possible even when information of sum of rows and tables are not available or measured with errors. This will affect confidence intervals of (some of) the estimated cells, but fuzzy information will have lower impact on the prior tables.

With the goal to test the methodology for (relatively) less processed data (certainly IOTs are always a highly processed data because of the nature of their compilation), we apply the methodology to US IOTs, considering them as "less-processed data". The summary statistics of the estimates for 2002-2013 years, based on previous 5 years, are presented in the Appendix C.

The overall performance of the Bayesian method is also comparable with others (see Appendix B).

3. Disaggregation of 15 to 69 industries (OKVED) for Russia for 2006

Here we apply the developed MCMC procedure to disaggregate symmetric 15x15 Use table in the OKVED classification into 69x69 matrix, using data for output and intermediate consumption for the 69 industries. We had to add measurement error to the observed 15x15 matrix. The data on 69 industries was published in the later years and is not fully consistent with the 15x15 matrix. The parameters of the experiment with the main results are summarized in the Table .

As follows from the table, the quality of the estimates is notable lower. Some MCMC chains are experiencing convergence problem which shows Geweke statistics and high autocorrelation of the chains even with very large interval between saved samples (thin = 5000). Around 10% of the autocorrelation coefficients are higher than 0.43. Geweke statistics also reports success in convergence for around 87% of all cells, and more than 99.6% of cells have at least one converged MCMC chain.

The reason of the lower quality of estimates might be caused by the introduced measurement error to the each cell of the aggregated matrix to fit the data of larger dimension. The error increases possible ranges for each cell, as well as correlation between them, and may affect the convergence. It is likely that longer sampling and/or taking into account potential autocorrelation between the sampling values will improve convergence of MCMC chains, increase quality of the estimates. The problem will be addressed on the further steps of the research.

The resulting samples for the disaggregated cells were aggregated and their distributions are compared with priors on the Figure 1 in the appendix. As follows from the picture, posterior distributions (green and red lines on the figure) often displaced from initial priors, which are normally distributed mean value of observed 15x15 Use table for 2006, and standard deviation equal to 10% of the cell values. The main reason of displacement of the posterior distribution is likely the inconsistency of the newly observed disaggregated data and the initial aggregated table. The inconsistency results in the matrix rebalancing, which we observe as displacement of the posterior distribution from their priors.

It should be noted, that the estimates might be also improved if other data is taken into account. For example, certain estimate of intermediate demand can be recovered based on import, export, public spending and final consumption. Also more meaningful prior information can be assigned to some industries or cells in the matrix, based on the economic knowledge of the sectors.

4. Updating of “Use” table (OKVED) from 2006 to 2012

In this section we update the Use-2006 table to each following year up to 2012. The methodology is similar to the applied above disaggregation. The base year table is the observed Use2006 matrix, which is the same for the all years, presumably measured with errors. Similarly we use output and intermediate consumption data for 69 industries to update the table and disaggregate it for particular year.

As and earlier, there are two levels of priors in the model – for disaggregation and measurement errors. Uniform distribution (uninformative) priors were assigned for the disaggregation. Normal distribution priors were assigned to the measurement errors for each cell, with mean values equal to the base year matrix, and standard deviations equal to 10% of the cells value.

For sampling we applied Random Walk Metropolis Hasting algorithm, optimized for the particular task and parallelized for calculation on CUDA-enabled graphical processors. For each year we run two Markov chains with length of 15 million iterations, burning first 2/3 of the iterations and saving every 5000th observation. The overall process for one year took around 40 hours on a pretty standard computer with i7-2600K Intel processor and NVIDIA-560 graphical card. The resulting 69x69 matrices are too large for publishing (available on request). In the appendix we present aggregated version of the tables for 2007-2012 in comparison with prior information for each cell.

The results are pretty similar to the disaggregated 2006 table, with shift of some estimated parameters in comparison to the prior information. As and earlier, we assume that the main reason of the shifts caused by preliminary character of the published aggregated IO table for 2006. The later data disaggregated data is not consistent with the table, but the later was not updated by Rosstat. Also, changes in production structure could induce changes in the USE table as well. We will continue the detailed analysis of the estimated tables on industries level on the further step of research.

IO table of 2011 for Russia was issued by Russian statistical agency, but in different SNA system. In figure 2 shown compare between share of value added for different sector. The main changing sector K-O.

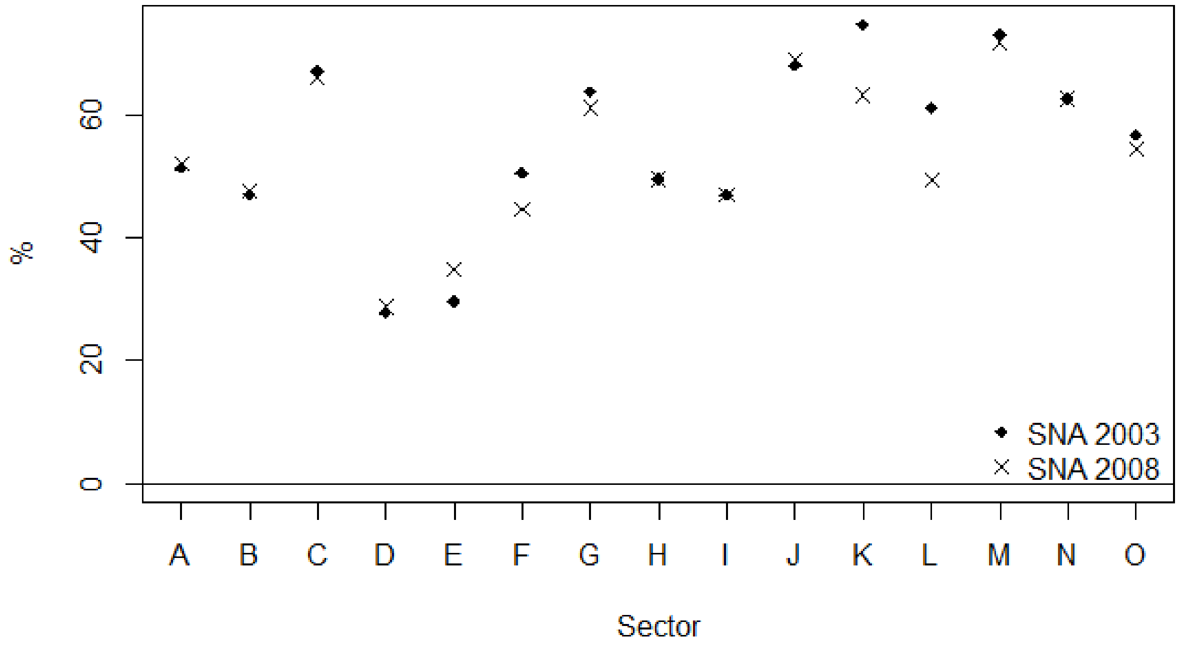

Figure 2. Share Value-Added in different SNA of 2011 for Russia.

In figure 3 shown comparison error estimation and changing io-cells. Note that column sums is fixed by estimation condition. So, sum of error for RAS and Bayes is equal. The more change is io-cells, the more estimate errors.

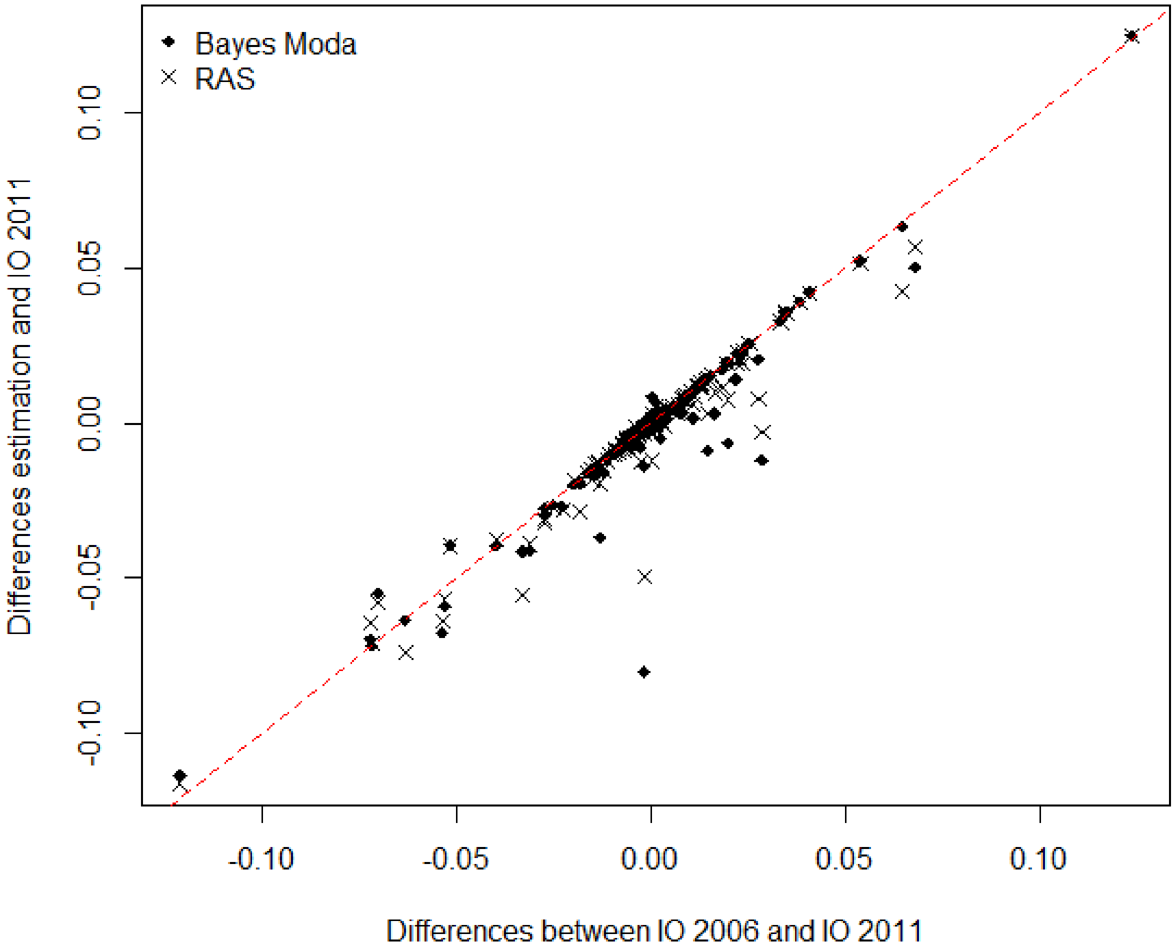

Figure 3. Compare error estimation and changing io-cells.

5. Concluding remarks

The presented methodology proposes sampling methods for updating, disaggregating, and balancing IOTs, and more largely national accounts. The main benefits of the methods is in natural incorporation of uncertainties into estimation process, flexibility in accommodation any kinds of data and information into estimation process, and full density profile for each of unknown parameters instead of point estimates.

In the paper we apply the proposed methodology to WIOD tables for 34 countries, US IOT for 8 recent years (based of 5 preceding years), and to sample Russian IOT, the most uncertain because of unavailable official statistics since 2003.

The overall performance of the methodology is comparable with other mainstream techniques. The precision of the estimates is normally higher than many other methods, as indicates the set of indicators. However, for real data application RAS method demonstrates usually better quality of estimates, which we interpret as broad application of RAS method and its derivatives for compilation of IOTs starting on national statistics data, to further IOTs processing and balancing. This findings doesn’t devaluate advantages of the Bayesian methodology, because of evaluation of uncertainties in the IOT coefficients. Moreover, the proposed methodology can be easily compared with any other preserved by a researched techniques using their estimates as a prior, and evaluating uncertainties using Bayesian methods.

The experimental updating, balancing and disaggregation of Russian IO table and updating US IOTs demonstrates a feasibility of application of sampling techniques for the large-scale problems with acceptable results. With developed algorithms, sampling of 15 million matrices of the 69x69 dimension can be performed in 40 hours on a modern consumer-class computer. Even with the achieved speed of calculation the methodology can be appropriately used. However, it is clear that the limit of performance is not reached yet. Further improvements of algorithms and involvement of professional computer clusters might improve the performance in hundreds and thousands of times.

The estimated WIOD tables will be provided for public with publication of the paper.

References

Eurostat (2008). European Manual of Supply, Use and Input-Output Tables. Methodologies and Working Papers. Luxembourg: Office for Official Publications of the European Communities.

Golan, A., Judge, G., Robinson, S. (1994). Recovering information from incomplete or partial multisectoral economic data. The Review of Economics and Statistics. 76(3), 541-549.

Golan, A., Judge, G., Miller, D. (1996). Maximum Entropy Econometrics. Chichester UK: Wiley.

Heckelei T., Mittelhammer R., Jansson T. (2008). A Bayesian alternative to generalized cross entropy solutions for undetermined econometric models. Institute for Food and Resource Economics Discussion Paper 2008: 2.

Hoff, P. (2009). A First Course in Bayesian Statistical Methods. Springer.

Kalantari, B., Lari, I., Ricca, F., Simeona, B. (2008). On the complexity of general matrix scaling and entropy minimization via the RAS algorithm. Mathematical Programming, Series A 112, 371–401.

Leon Y., Peeters, L., Quinqu, M., Surry, Y. (1999). The use of maximum entropy to estimate input-output coefficients from regional farm accounting data. Journal of Agricultural Economics 50, 425-439.

Miller, R. E., Blair, P. D. (2009). Input-Output Analysis: Foundations and Extensions. Cambridge: Cambridge University Press, 2nd edition.

Ntzoufras, I. (2008). Bayesian Modeling Using WinBUGS. Wiley

Robinson, S., Cattanbo, A., el-Said, M. (2000). Updating and estimating a social accounting matrix using cross entropy methods. Economic Systems Research 13, 47-67.

Stone, R. (1961). Input-Output and National Accounts, OECD, Paris.

Stone, R., Bates, J., Bacharach M. (1963). A program for growth 3, input-output relationships 1954-1966. Cambridge.

Temurshoev, U., Yamano, N., Webb, C. (2010). Projection of supply and use tables: methods and their empirical assessment. WIOD Working Paper Nr. 2.

van der Linden, J.A., Oosterhaven, J. (1995). European Community Intercountry Input–Output Relations: Construction Method and Main Results for 1965–85. Economic Systems Research, 7:3, 249-270.

Timmer, M. P., Dietzenbacher, E., Los, B., Stehrer, R. and de Vries, G. J. (2015), "An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production", Review of International Economics., 23: 575–605

Kuznetsov SY, Piontkovsky DI, Sokolov DD, Starchikova O.S. “An empirical comparison of mathematical methods for constructing dynamic series of the system of input-output tables”, HSE Economic Journal, 2016, vol. 20, No. 4, p. 711-730.

Appendix A. Comparison of Bayesian method performance with others for WIOD tables updating

In the experiment we updated IOTs for 24 various countries from WIOD database based only in based years. We used two previous based years as prior. Two previous based years was used as prior information for initial matrix and for estimation of standard deviation prior for each IOT cells. Based years for WIOD in table 1. The Bayesian estimation then have been compared with other methods, including cross entropy (CE), least squares (LS), non-linear least squares (NLS), weighted least squares (WLS), and RAS. Since for Bayesian method we obtain distribution instead of point estimate, we use two point-measures for comparison with other methods: mode (MODE) and mean (MEAN) values for each cell. MODE can be also calculated using non-linear programming techniques after specification of likelihood function. This approach has been used in this experiment; MEAN was estimated on the sample of matrices.

List of countries in the experiment:

Australia, Austria, Belgium, Bulgaria, Brazil, Canada, China, Germany, Czech Republic, Denmark, Spain, Estonia, Finland, France, United Kingdom, Greece, Hungary, Indonesia, India, Ireland, Italy, Japan, Korea, Lithuania, Luxembourg, Netherlands, Poland, Portugal, Romania, Russia, Slovak Republic, Slovenia, Taiwan, United States.

Table 1. Basic IO years used for WIOD construction (Kuznetsov at al., 2016).

	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Australia		+							+	+				+			
Brazil	+	+				+					+						
Bulgaria	+					+				+							
UK				+						+				+			
Hungry						+					+					+	
Germany						+					+						+
India				+					+					+			
Indonesia	+					+					+						
Ireland	+					+					+						
Spain	+					+					+						
Italy	+					+					+						
China			+					+					+				
Korea	+					+					+			+			
Luxemburg						+					+					+	
Netherland						+					+					+	
USA			+					+					+				
Czechia	+					+					+					+	
Sweden	+					+					+						
Japan	+					+					+						+

Table 2. Comparison of Bayesian MEAN vs others as % of cases when Bayesian method has higher or equal precision according to various statistics. Experiment with RAS used as a prior matrix.

	CE	LS	NLS	WLS	RAS
RMSE	67%	79%	75%	88%	33%
MAE	71%	92%	71%	100%	8%
MAPE	29%	100%	29%	100%	0%
WAPE	71%	92%	71%	100%	8%
SWAD	67%	83%	63%	96%	29%
Psi	71%	92%	71%	100%	8%
RSQ	75%	79%	75%	88%	33%
AED	71%	92%	71%	96%	4%

Table 3. Comparison of Bayesian MODE vs others as % of cases when Bayesian method has higher or equal precision according to various statistics. Experiment with RAS used as a prior matrix.

	CE	LS	NLS	WLS	RAS
RMSE	83%	100%	88%	100%	-
MAE	100%	100%	100%	100%	-
MAPE	96%	100%	96%	100%	-
WAPE	100%	100%	100%	100%	-
SWAD	88%	100%	92%	100%	-
Psi	96%	100%	96%	100%	-
RSQ	88%	100%	88%	100%	-
AED	96%	96%	100%	96%	-

Note: Missing values for RAS method means that Bayesian mode estimate is equal to RAS, i.e. the comparison is not applicable.

Table 4. Comparison of Bayesian MEAN vs others as % of cases when Bayesian method has higher or equal precision according to various statistics. Experiment where for previous based year matrix used as a prior.

	CE	LS	NLS	WLS	RAS
RMSE	21%	54%	21%	79%	13%
MAE	4%	75%	8%	100%	4%
MAPE	0%	100%	0%	100%	0%
WAPE	4%	75%	8%	100%	4%
SWAD	13%	63%	13%	79%	8%
Psi	8%	58%	8%	96%	4%
RSQ	21%	58%	21%	88%	13%
AED	4%	71%	4%	96%	0%

Table 5. Comparison of Bayesian MODE vs others as % of cases when Bayesian method has higher or equal precision according to various statistics. Experiment where for previous based year matrix used as a prior.

	CE	LS	NLS	WLS	RAS
RMSE	13%	46%	17%	75%	4%
MAE	0%	75%	0%	100%	0%
MAPE	13%	100%	13%	100%	4%
WAPE	0%	75%	0%	100%	0%
SWAD	8%	63%	8%	79%	4%
Psi	4%	63%	4%	100%	0%
RSQ	21%	58%	21%	83%	8%
AED	0%	88%	0%	96%	0%

Appendix B. Statistics for comparison of varios updating techniques.

Name	Formula	Description
RMSE	$\sqrt{\frac{1}{n^2}\sum_{ij}\left(x_{ij}^{true}-x_{ij}\right)^2}$	Root mean square error
MAE	$\frac{1}{n^2}\sum_{ij}\left\lvert x_{ij}^{true}-x_{ij}\right\rvert$	Mean absolute error
MAPE	$\frac{1}{n^2}\sum_{ij}\frac{\left\lvert x_{ij}^{true}-x_{ij}\right\rvert}{\left\lvert x_{ij}^{true}\right\rvert}\cdot 100$	Mean absolute percentage error
WAPE	$\sum_{ij}\left(\frac{\left\lvert x_{ij}^{true}\right\rvert}{\sum_{kl}x_{kl}^{true}}\right)\cdot\frac{\left\lvert x_{ij}-x_{ij}^{true}\right\rvert}{\left\lvert x_{ij}^{true}\right\rvert}\cdot 100$	Weighted absolute percentage error, which weights each percentage deviation of x_{ij} from x_{ij}^{true} by the relative size of the corresponding true element in the overall sum of the actual elements.
SWAD	$\frac{\sum_{ij}\left\lvert x_{ij}^{true}\right\rvert\cdot\left\lvert x_{ij}-x_{ij}^{true}\right\rvert}{\sum_{kl}\left(x_{ij}^{true}\right)^2}$	Standardized weighted absolute difference, which is effectively similar to WAPE with the difference being that absolute deviations are weighted by the size of the true transactions.
Ψ	$\frac{1}{\sum_{kl}x_{kl}^{true}}\cdot\sum_{ij}\left[\left\lvert x_{ij}^{true}\right\rvert\cdot\left\lvert\ln\left(\frac{x_{ij}^{true}}{s_{ij}}\right)\right\rvert+\left\lvert x_{ij}\right\rvert\cdot\left\lvert\ln\left(\frac{x_{ij}}{s_{ij}}\right)\right\rvert\right],$ $where\ s_{ij}=\frac{\left\lvert x_{ij}\right\rvert+\left\lvert x_{ij}^{true}\right\rvert}{2}$	Ψ is shows a linear relation between its value and the level of error.
RSQ	$cor\left(x_{ij}^{true},x_{ij}\right)^2$	(or coefficient of determination) – the square of the correlation coefficient between the elements of the actual and predicted matrices (x_{ij} and x_{ij}^{true}), when at least one of them is different from zero.
AED	$\sum_{ij}\left\lvert x_{ij}^{true}\cdot\log x_{ij}^{true}-x_{ij}\cdot\log x_{ij}\right\rvert$	Average entropy distance

Appendix C. Updating USA data tables

Estimates in the Industry Economic Accounts of the Bureau of Economic Analysis (BEA) are generally available at three levels of detail: sector (15 industry groups), summary (71 industry groups), and detail (389 industry groups). For most data products, estimates at the detail level are available only for estimate year 2007 (due to the availability of detailed data from the 2007 Economic Census); however, estimates of gross output at the detail level are also available annually. This table shows the relationship between these three levels of detail as well as how each level relates to the 2007 North American Industry Classification System (NAICS) code structure.

These statistics were prepared by the Industry Economic Accounts (IEAs) Directorate, Bureau of Economic Analysis (BEA), U.S. Department of Commerce. The statistics in these spreadsheets are not copyrighted.

Table 6– Estimation IO for USA from 2005 to 2013, IO previous year as prior, CI analysis, comparison with RAS.

Statistic	RMSE	MAE	MAPE	WAPE	SWAD	Psi	RSQ	AED
MOD is better than RAS	8%	8%	0%	8%	17%	8%	8%	0%
MEAN is better than RAS	17%	8%	0%	8%	17%	8%	17%	0%

Table 8 – Estimation IO for USA from 2005 to 2013, RAS estimation as prior, comparison with RAS.

Statistic	RMSE	MAE	MAPE	WAPE	SWAD	Psi	RSQ	AED
MOD is better than RAS	—	—	—	—	—	—	—	—
MEAN is better than RAS	50%	42%	0%	42%	50%	42%	50%	25%

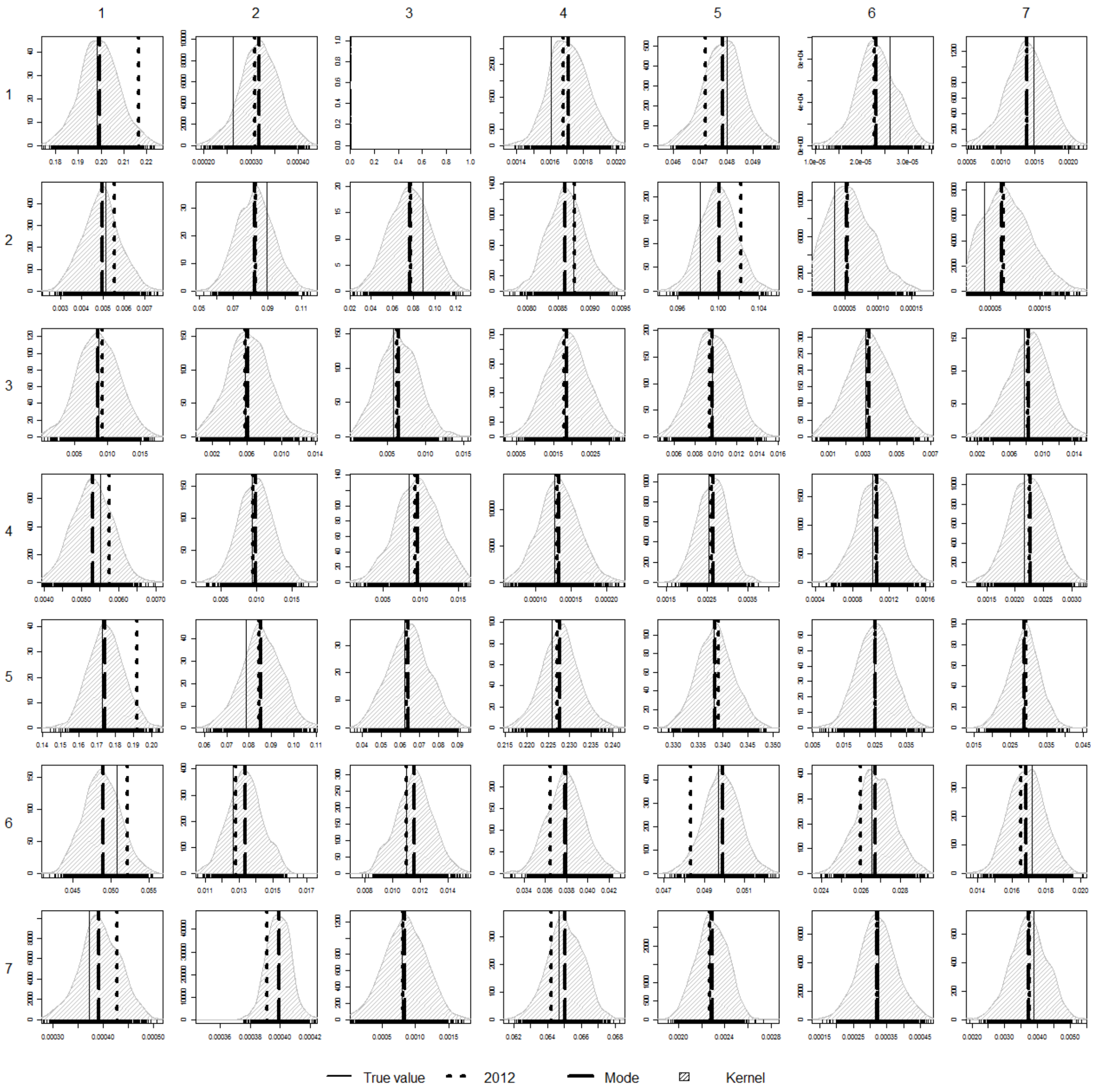

Figure 4. Estimation IO for USA for 2013, IO 2012 year as prior, first 7x7 sectors.

Appendix D. Estimation of Russian IOT for 2006-2012, selected output.

Table 1. Parameters and results of MCMC for 2006 year.

Parameter	*Value*
Number of iterations	4e6
Thin (step between saved observations)	5000
Burn (number of first dropped iterations)	1e5
success Geweke, %	87.8%
max ACF	0.996

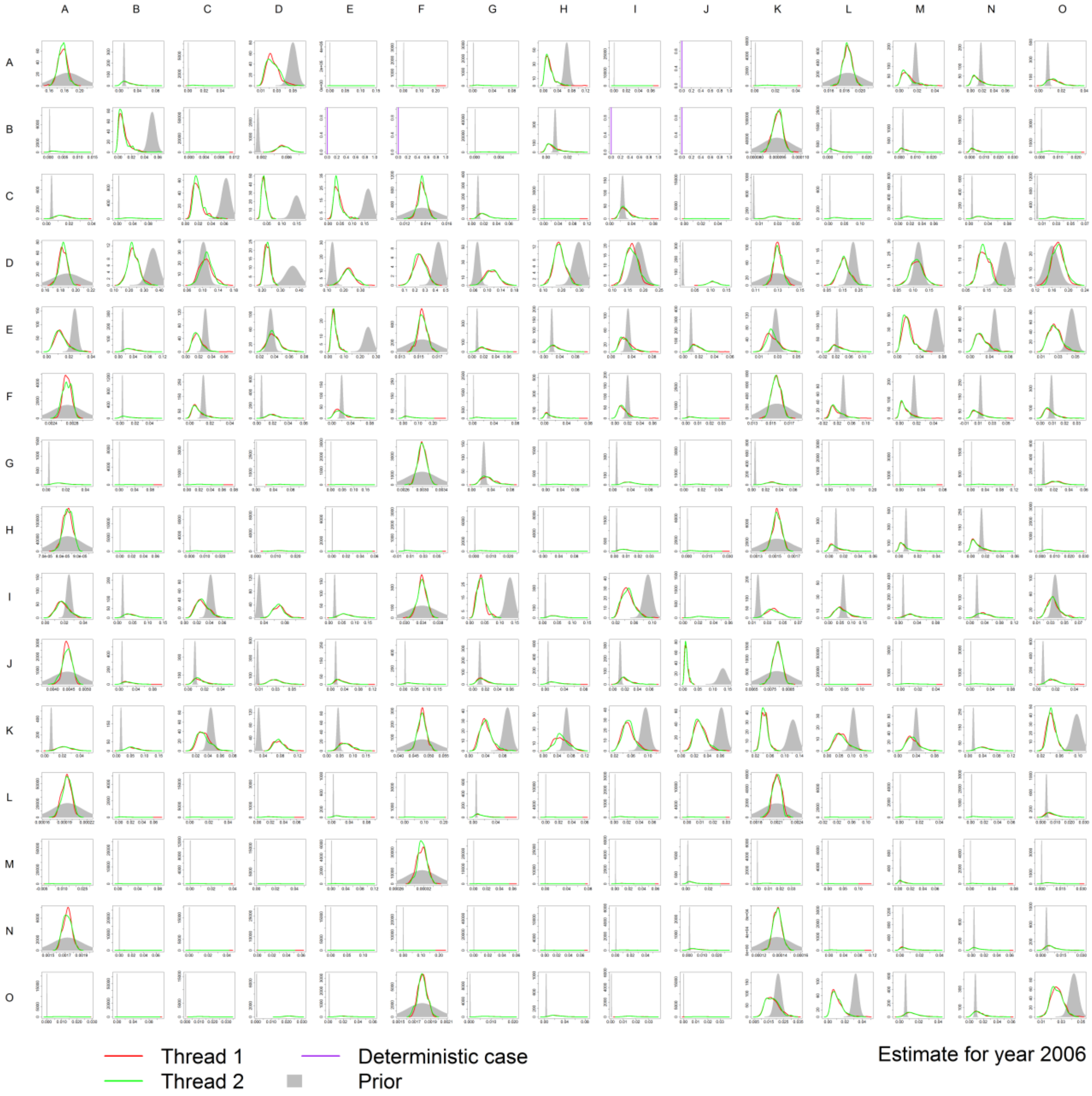

Figure 1. Kernel for aggregate matrix 15x15 from estimation for 69x69 for 2006.

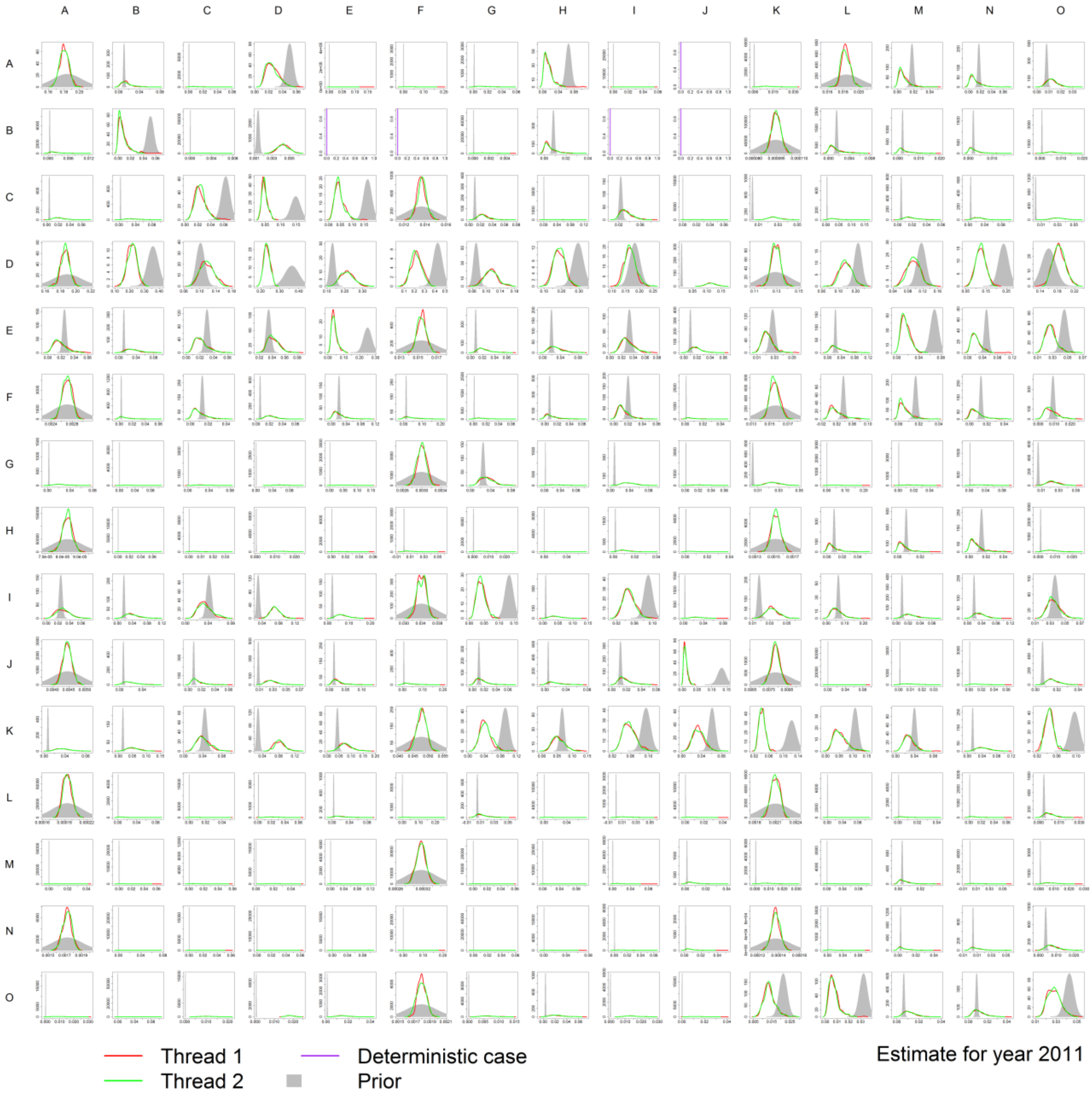

Figure 6. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2011.

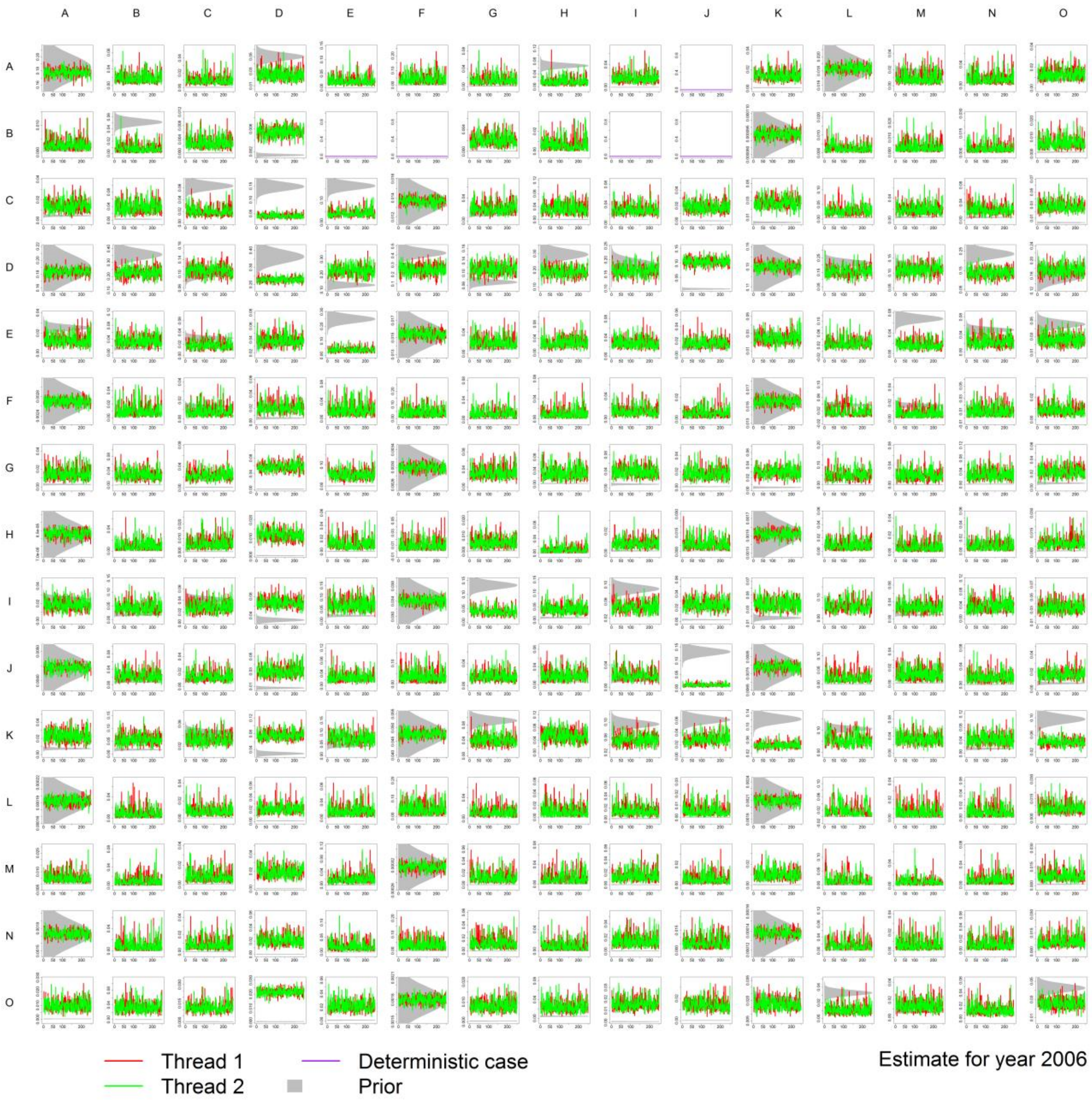

Figure 7. Prior distributions and MCMC chains (thread 1 & 2) for estimated 15x15 Use table for 2006.