Bayesian Updating of Input-Output Tables

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

Introduction

The paper continues efforts on developing Bayesian method of updating IO tables, presented by the authors on the 16th Annual Conference on Global Economic Analysis, and extends the methodology and results in several ways. In the current paper, we test our methodology on the "long" survey based IRIOS tables. We compare two point estimates of the Bayesian method of "unknown" IO table: posterior mode and posterior mean with estimates, which come from alternative methods popular in the literature. Than we discuss how to construct an appropriate creditable set for IO coefficients. We also upgrade and extend estimates of SUT tables for Russia.

This publication further provides a method for generating an arbitrarily large sample of IO tables, satisfying the constraints, and a priori information from the normal distribution of the multivariate normal distribution, which gives additional advantages compared with exact estimation of the most probable or sample matrices, as it allows the user to select the desired accuracy for analysis. The development of this approach is the rapid assessment of the

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

covariance matrix of the coefficients, and density parameters. However, a disadvantage of this method is poor characteristics for obtaining sample matrices with a large number of sectors.

The use of conjugate vectors to Hessian matrix of priori distribution, has greatly improved the parameters MCMC algorithm used to generate the sample most probable IO tables, and get rid of the problem of optimal choice of proposal density and poor convergence MCMC chains, at least for the case of normal prior distribution. In case of Beta prior distribution for calculation of the relevant parameters MCMC algorithm may be used their approximation by normal distribution.

The work consists of three parts. The first part is the core of the current paper, where we present our conceptual frameworks for updating, disaggregation, and balancing IO tables.

In the second part of the current paper, we test our methodology on the “long” survey based IRIOS tables (van der Linden and Oosterhaven, 1995). 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 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.

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:

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

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:

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

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

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:

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

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

Equation (4) means vectorization of matrix *A*. In equation (3) $p(z)$ is some prior distribution, $p(z \mid data)$ 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:

$$Bz = T \tag{5}$$

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 $\big(N(0.1, 0.1), N(0.1, 0.01)\big)$. 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 $\big(N(0.1, 0.1), N(0.1, 0.001)\big)$, 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

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

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:

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

by use equation

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

with distribution β equal:

$$\beta \sim N(\mu^{\beta}, \sigma^{\beta})$$

then covariance is equal

$$cov_{ij} = \int\int \left(\sum_k f_{ik} \cdot \beta_k - \tilde{z}_\imath\right)\left(\sum_k f_{jk} \cdot \beta_k - \tilde{z}_\jmath\right) dz_i dz_j$$

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

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

or in the matrix form

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

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:

$$S = \begin{cases} S_{i,j} = 1, \; i \in sector\ j \\ S_{i,j} = 0, i \notin sector\ j \end{cases} \qquad (10)$$

Therefore, aggregation problem can be written:

$$\tilde{A}^* = S\tilde{A}S' \qquad (11)$$

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

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

where $\text{kron}(\bullet)$ denotes Kronecker product of the matrices, vec(•) 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:

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

where

Σ — diagonal matrix with elements proportional to the square of $\text{vec}(\tilde{A}^*)$. 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,

$I(a\,|\,data)$ — 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 IRIOS tables

Here we apply our methodology on the “long” survey based IRIOS tables (van der Linden and Oosterhaven, 1995). We treat IO tables for 1985 for 5 countries: France, Netherlands, Italy, Belgium and Germany as unknown and try to estimate it with different methods and compare results. In the Bayesian framework we assume independent truncated normal distributions for each IO coefficient and use coefficients of 1980 IO table as prior mode. Below we will also try beta prior distribution. To specify standard deviations in prior distribution we estimate standard deviation for each coefficient on the all previously available tables. To compute posterior mean of coefficients we apply Markov chain Monte Carlo (MCMC) method with two chains and sampled length of 1 500 000 simulation. To compute posterior mode of coefficients we use nonlinear programming techniques.

We consider following alternative methods for updating input-output tables: RAS method and Cross-Entropy (CE), Least Squares (LS), Normalized Least Squares (NLS), Weighted Least Squares (WLS) distance minimization methods from previously available matrix. For detailed overview of different updating methods see (Temurshoev et al., 2010). Results of updating IO tables for considered countries are presented in tables 1-5.

As follows from the tables posterior mode robustly outperforms competitive methods, popular in the literature, according to different closeness statistics. Posterior mean perform slightly worse than posterior mode. For some cases (Belgium and Germany) performance of posterior mean is relatively bad. Bayesian method at least is compatible with the other methods on real data examples, but it is more appropriate to use posterior mode as a point estimate of unknown IO tables. Thus standard symmetric creditable interval for input-output coefficient is inappropriate and induce significant bias. Using higher posterior credible set for characterization of the uncertainty could potentially improve coverage properties. Results of experiments on constructing credible sets and experiments with beta prior distribution will be presented at the conference.

Table 1. Results of Updating IO tables for France

	Mean	Mode	CE	LS	NLS	WLS	RAS
RMSE	0,018	0,009	0,012	0,013	0,012	0,015	0,012
MAE	0,007	0,004	0,005	0,006	0,005	0,007	0,005
MAPE	0,884	0,722	0,602	5,982	0,571	13,723	0,692
WAPE	34,913	21,414	25,099	28,955	25,481	35,357	24,321
SWAD	0,194	0,104	0,133	0,144	0,137	0,157	0,122
Psi	0,333	0,204	0,242	0,264	0,246	0,303	0,235
RSQ	0,888	0,969	0,948	0,935	0,948	0,919	0,949

Table 2. Results of Updating IO tables for Netherlands

	Mean	Mode	CE	LS	NLS	WLS	RAS
RMSE	0,028	0,021	0,064	0,024	0,027	0,036	0,022
MAE	0,011	0,008	0,023	0,010	0,010	0,011	0,009
MAPE	1,682	1,545	0,939	3,107	1,627	4,844	1,397

WAPE	47,432	35,642	100,522	42,944	42,154	49,284	37,354
SWAD	0,292	0,164	1,000	0,238	0,228	0,255	0,194
Psi	0,435	0,323	0,379	0,388	0,382	0,404	0,341
RSQ	0,777	0,876	0,764	0,832	0,808	0,728	0,865

Table 3. Results of Updating IO tables for Italy

	Mean	Mode	CE	LS	NLS	WLS	RAS
RMSE	0,022	0,007	0,008	0,009	0,008	0,010	0,007
MAE	0,007	0,003	0,004	0,005	0,004	0,006	0,004
MAPE	0,932	0,749	0,792	0,931	0,789	1,485	0,739
WAPE	33,282	15,447	17,594	20,899	17,579	25,704	16,657
SWAD	0,175	0,063	0,065	0,092	0,065	0,104	0,069
Psi	0,316	0,152	0,173	0,196	0,173	0,224	0,164
RSQ	0,858	0,986	0,982	0,977	0,982	0,968	0,985

Table 4. Results of Updating IO tables for Belgium

	Mean	Mode	CE	LS	NLS	WLS	RAS
RMSE	0,043	0,005	0,005	0,008	0,006	0,010	0,005
MAE	0,015	0,002	0,002	0,003	0,002	0,004	0,001
MAPE	0,907	0,155	0,124	4,814	0,132	20,212	0,082
WAPE	67,714	9,452	8,454	14,484	8,754	18,979	6,800
SWAD	0,393	0,050	0,051	0,084	0,054	0,091	0,051
Psi	0,597	0,093	0,084	0,138	0,087	0,169	0,068
RSQ	0,627	0,993	0,992	0,984	0,991	0,972	0,994

Table 5. Results of Updating IO tables for Germany

	Mean	Mode	CE	LS	NLS	WLS	RAS
RMSE	0,033	0,008	0,008	0,010	0,009	0,012	0,010
MAE	0,012	0,004	0,004	0,005	0,004	0,006	0,005
MAPE	0,868	0,562	0,574	1,008	0,587	1,410	0,602
WAPE	53,596	16,119	17,282	21,895	18,479	25,963	20,973
SWAD	0,332	0,072	0,082	0,109	0,089	0,116	0,105
Psi	0,491	0,158	0,171	0,211	0,182	0,236	0,206
RSQ	0,710	0,982	0,979	0,967	0,976	0,953	0,968

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.

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.

4. 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 test our method on the “long” survey based IRIOS tables. Results of the experiments are in favor that point estimates from proposed Bayesian method are compatible with alternative methods for updating IO tables. We also provide framework to construct an appropriate creditable set for IO coefficients which has good coverage properties.

The experimental updating, balancing and disaggregation of Russian IO table 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.

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.

Appendix

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

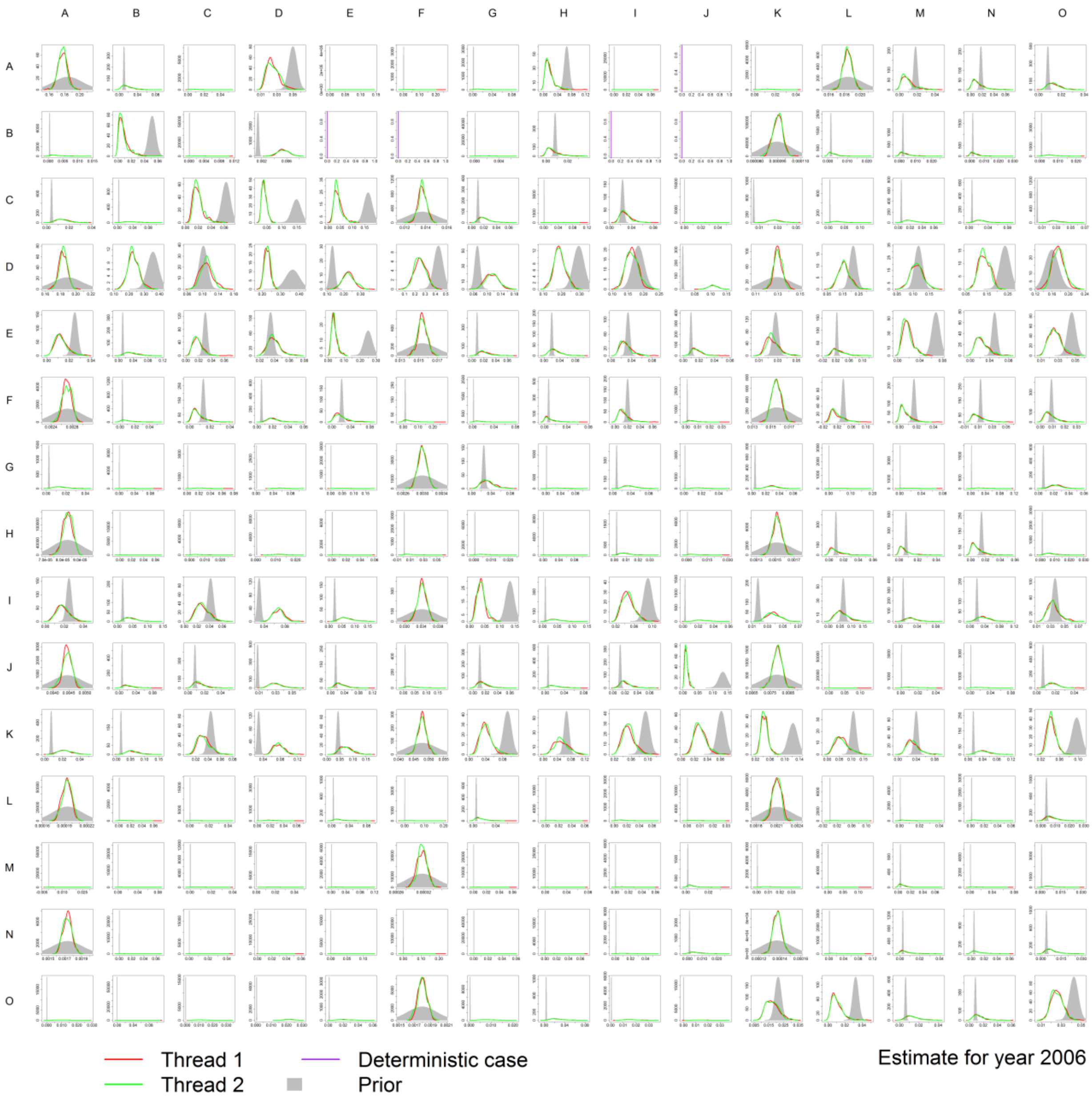

Figure 1. Kernel for aggregate matrix 15x15 from estimation for 69x69 for 2006.

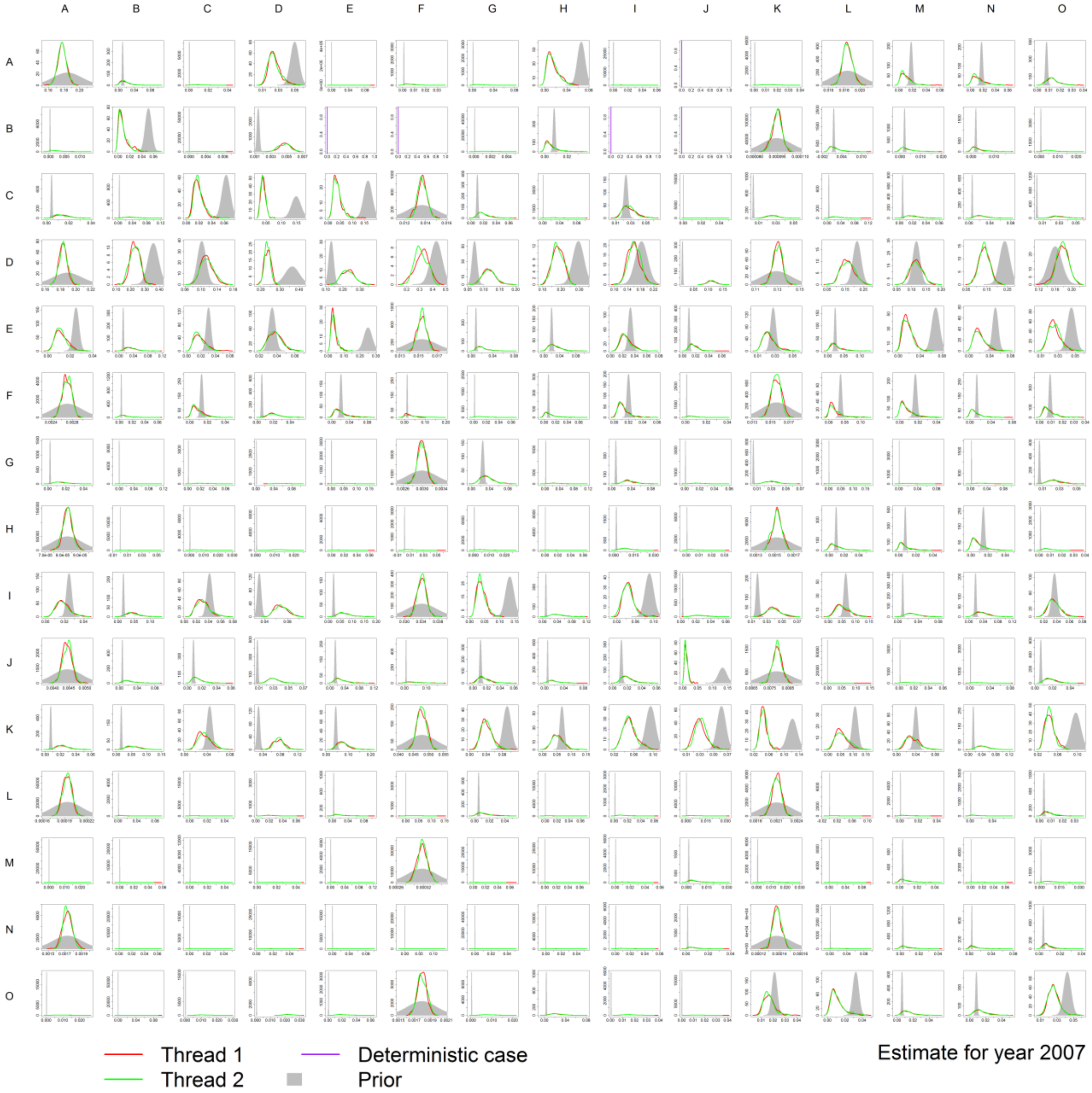

Figure 2. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2007.

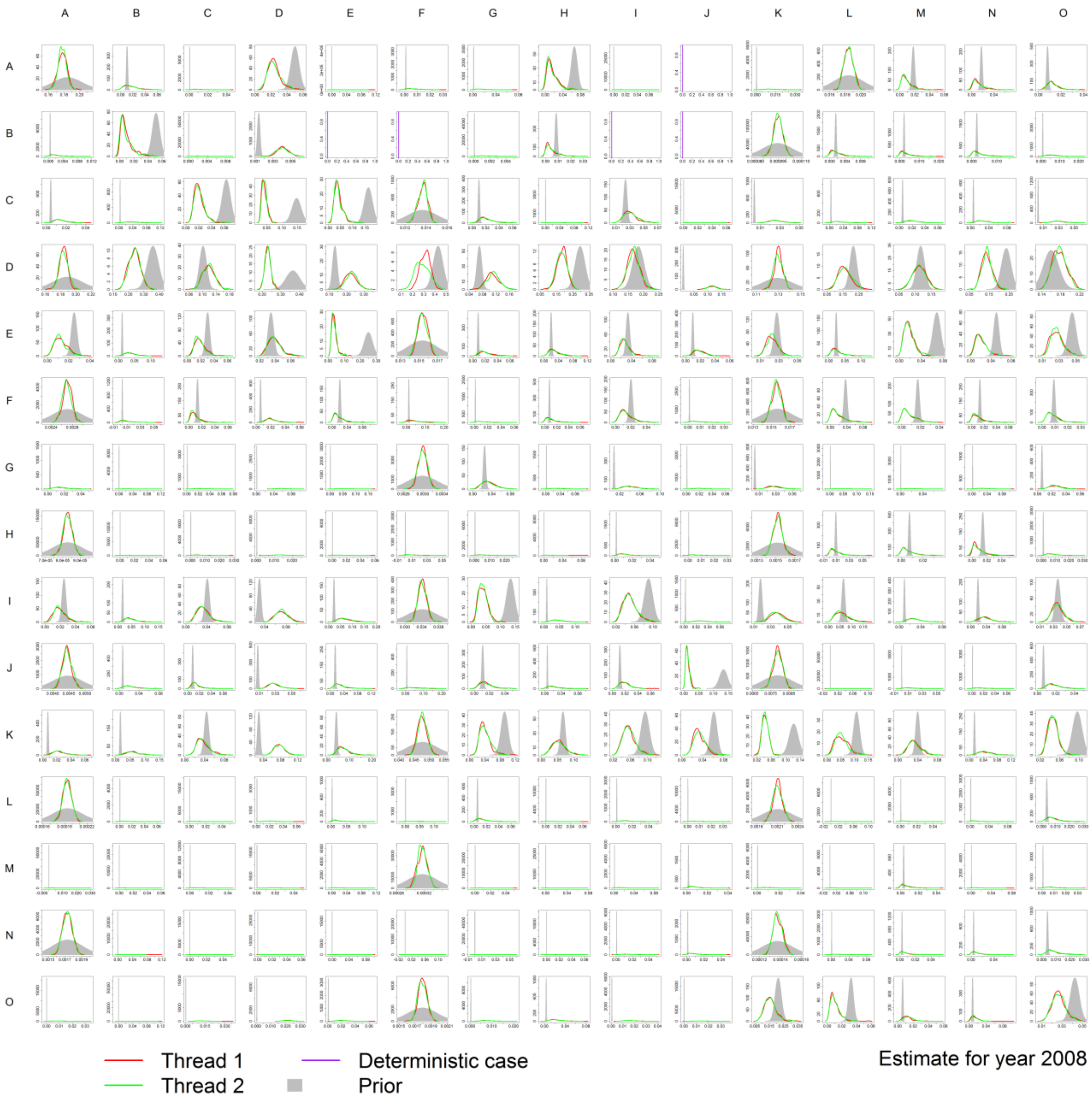

Figure 3. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2008.

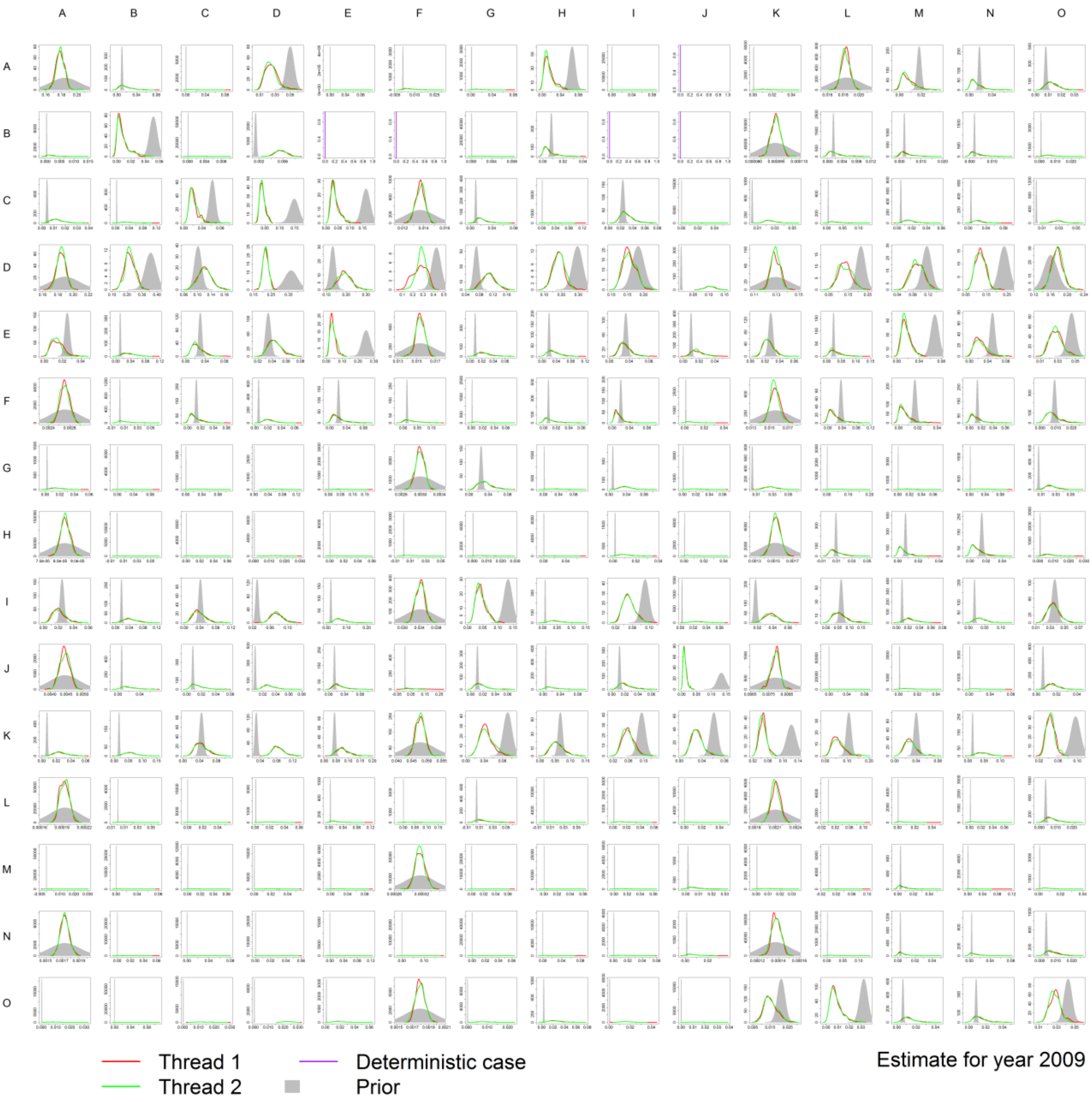

Figure 4. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2009.

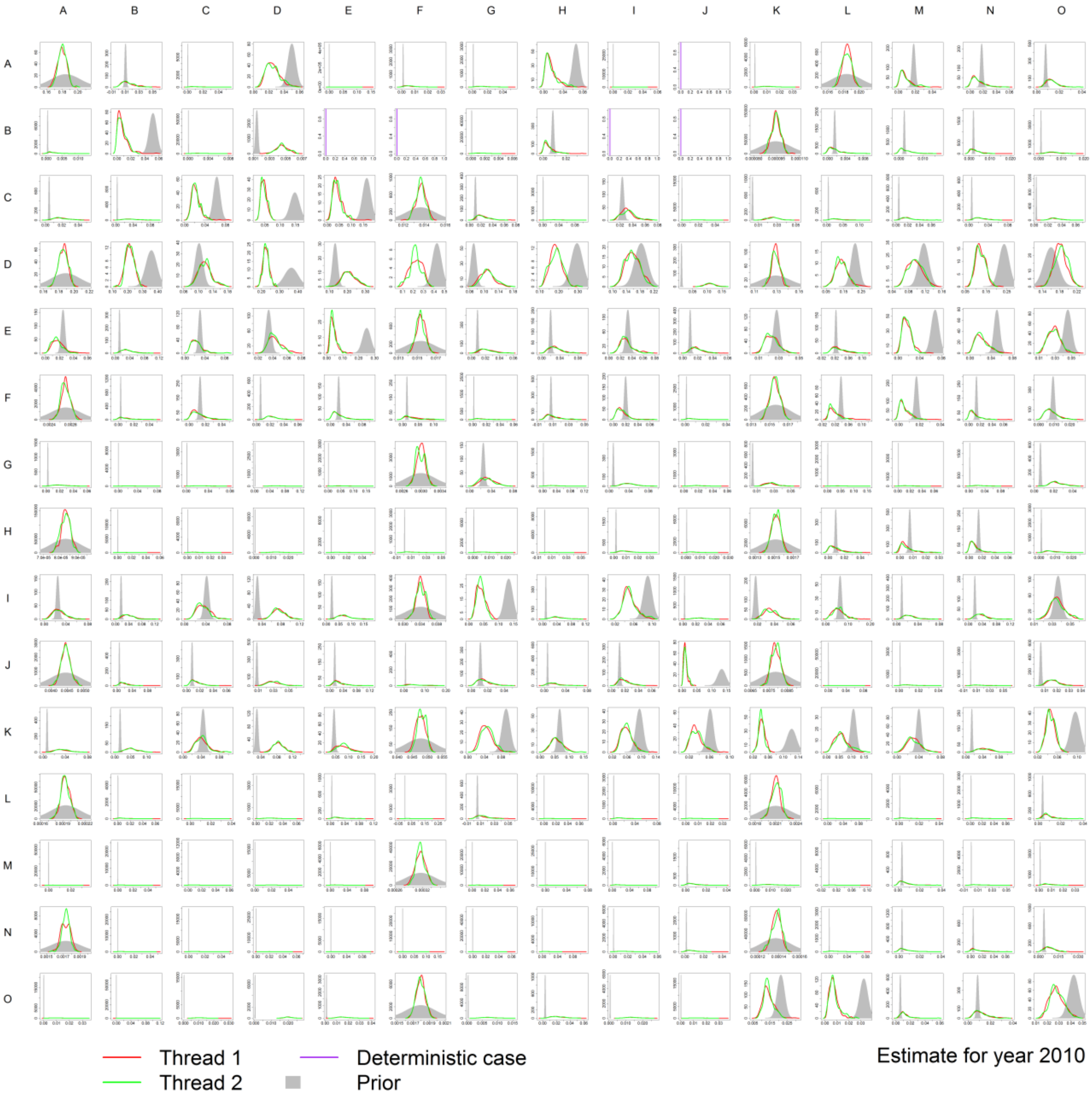

Figure 5. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2010.

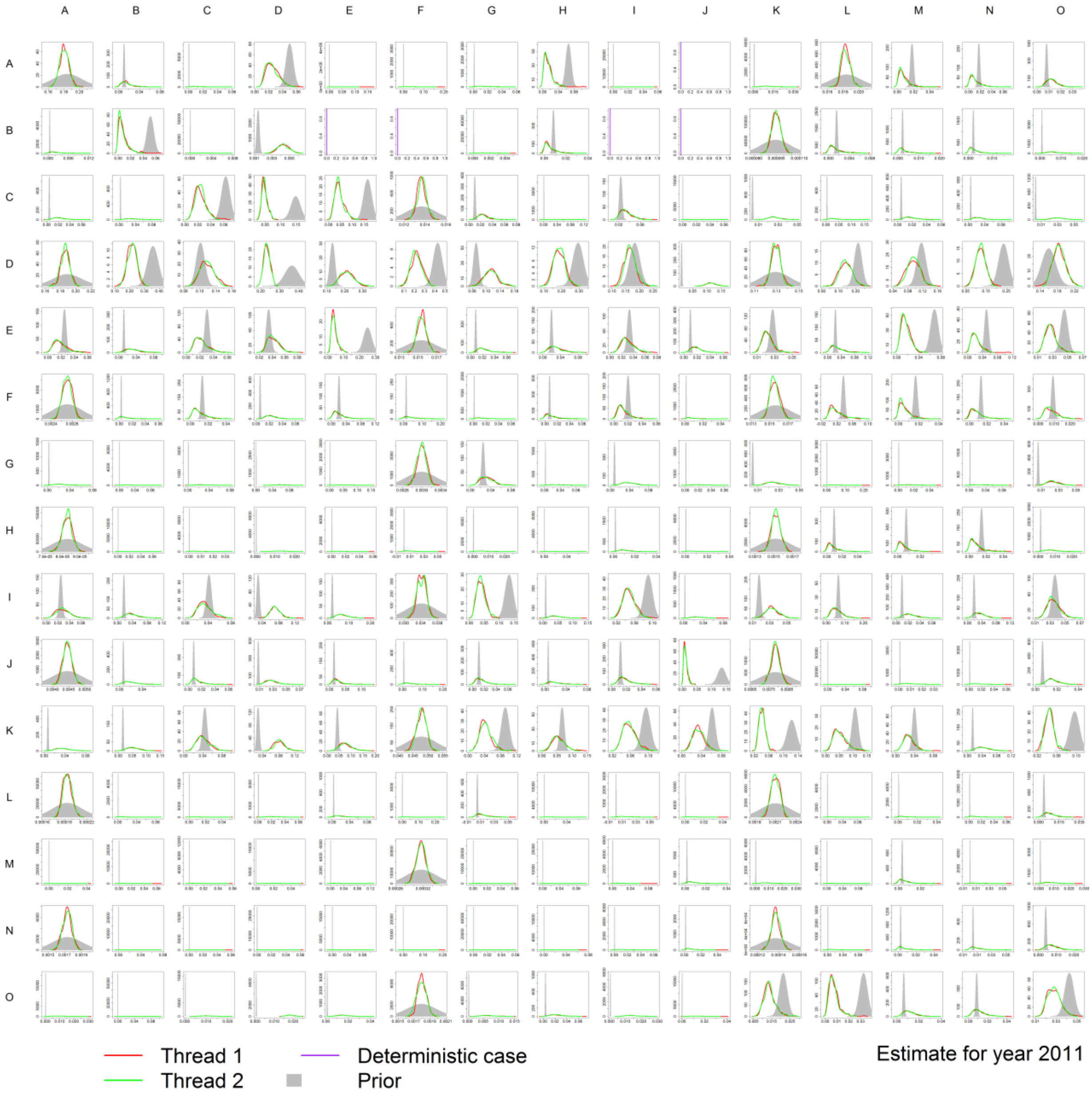

Figure 6. Prior and posterior distributions (thread 1 & 2) for estimated 15x15 Use table for 2011.

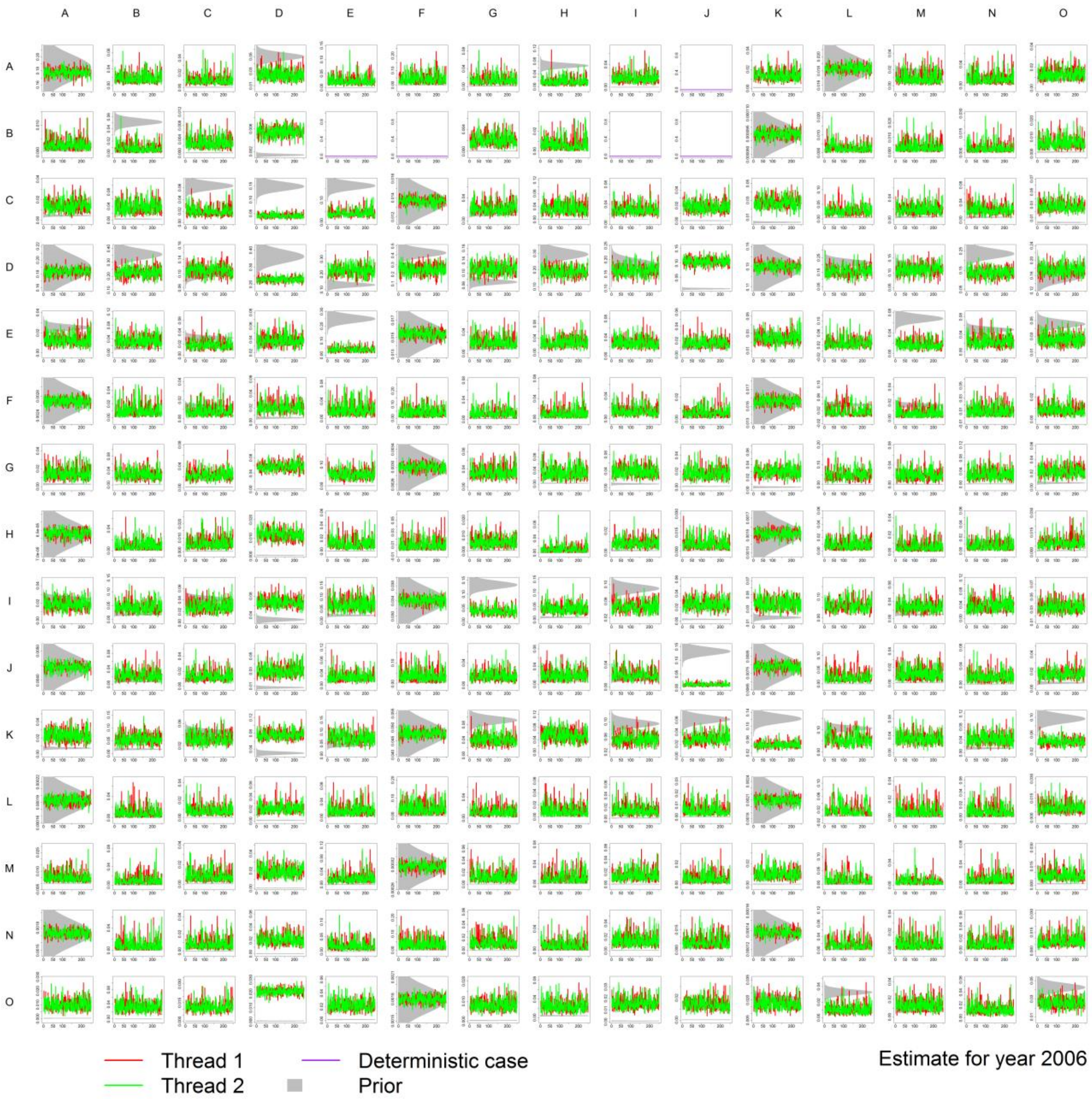

Figure 7. Prior distributions and MCMC chains (thread 1 & 2) for estimated 15x15 Use table for 2006.

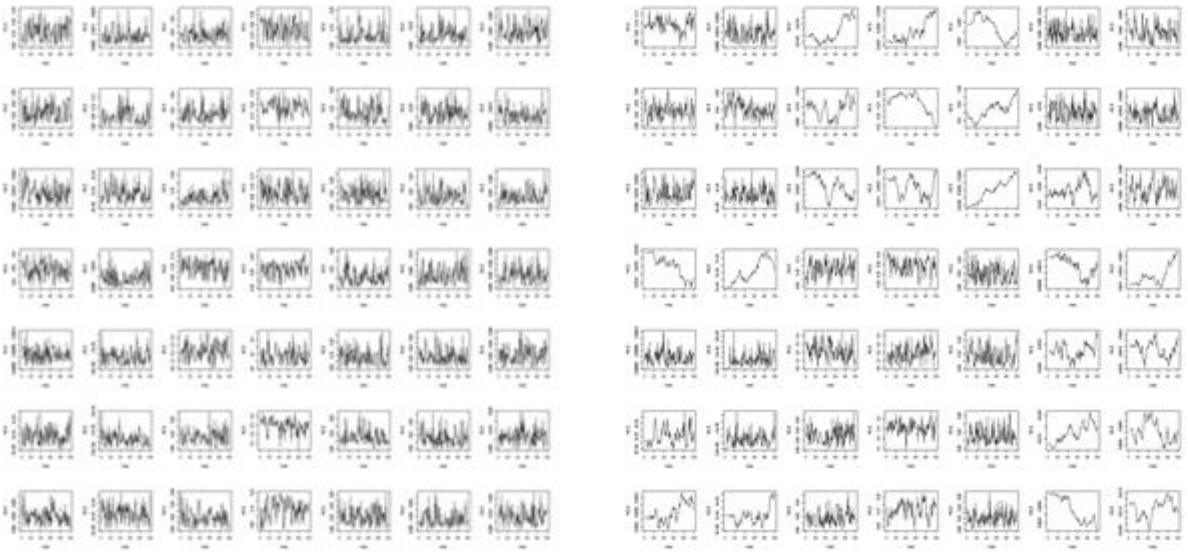

Figure 8. Compare MCMC chains: improvement (left), base version (right).

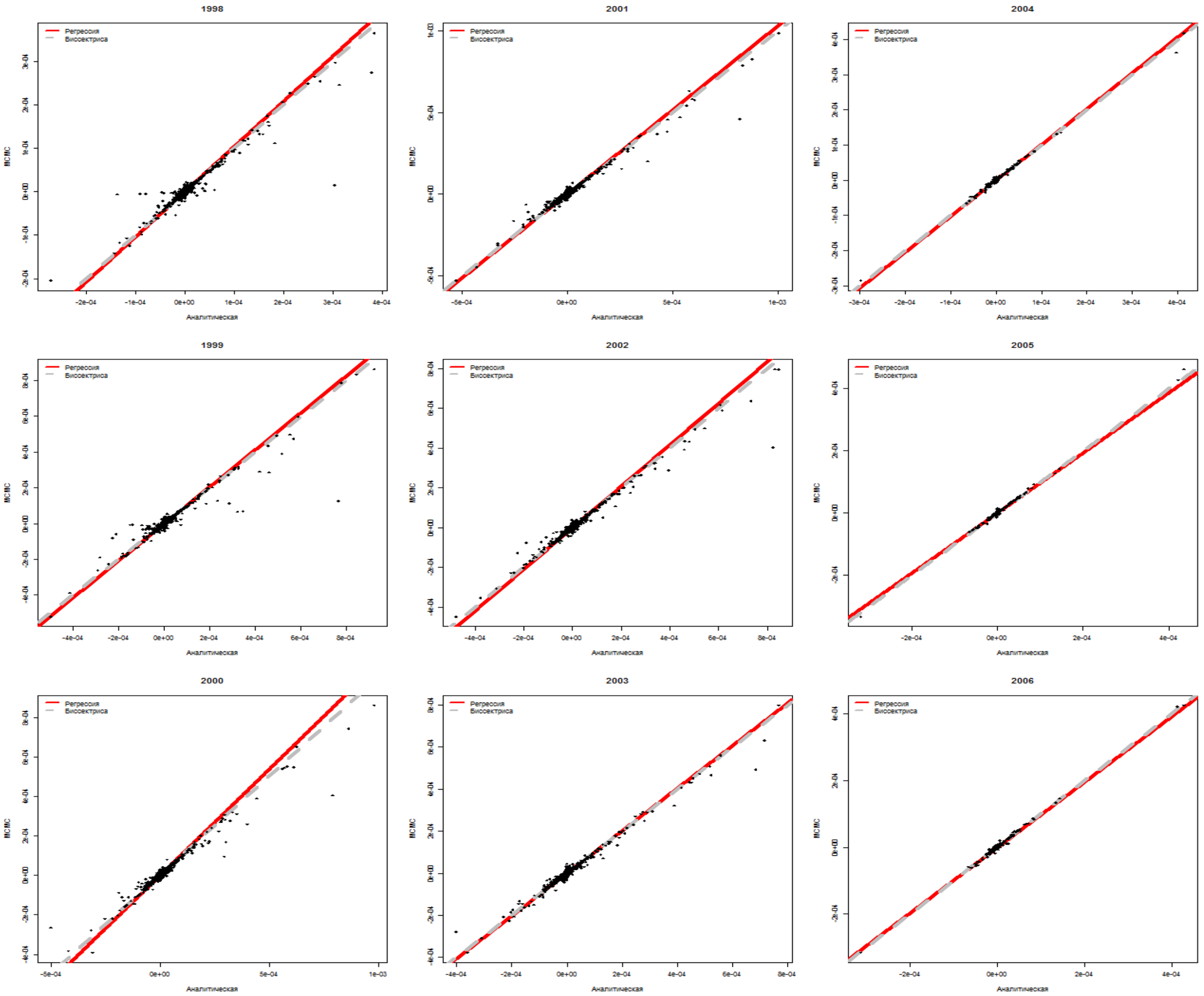

Figure 9. The distribution coefficients of the covariance between the cells in the IO table year OKONH 1998-2003, and 2004-2006 NACE.