Żądło, Tomasz

Article

On Misspecification of Spatial Weight Matrix for Small Area Estimation in Longitudinal Analysis

Comparative Economic Research. Central and Eastern Europe

Provided in Cooperation with:
Institute of Economics, University of Łódź

Suggested Citation: Żądło, Tomasz (2012) : On Misspecification of Spatial Weight Matrix for Small Area Estimation in Longitudinal Analysis, Comparative Economic Research. Central and Eastern Europe, ISSN 2082-6737, Łódź University Press, Łódź, Vol. 15, Iss. 4, pp. 305-318, https://doi.org/10.2478/v10103-012-0043-5

This Version is available at:
https://hdl.handle.net/10419/259135

10.2478/v10103-012-0043-5

TOMASZ ŻĄDŁO*

On Misspecification of Spatial Weight Matrix for Small Area Estimation in Longitudinal Analysis

Abstract

The problem of prediction of subpopulation (domain) total is studied as in Rao (2003). Considerations are based on spatially correlated longitudinal data. The domain of interest can be defined after sample selection what implies its random sample size. The special case of the General Linear Mixed Model is proposed where two random components obey assumptions of spatial and temporal moving average process respectively. Moreover, it is assumed that the population may change in time and elements' affiliations to subpopulation may change in time as well. The proposed model is a generalization of longitudinal models studied by e.g. Verbeke, Molenberghs (2000) and Hedeker, Gibbons (2006). The best linear unbiased predictor (BLUP) is derived. It may be used even if the sample size in the subpopulation of interest in the period of interest is zero. In the Monte Carlo simulation study the accuracy of the empirical version of the BLUP will be studied in the case of correct and incorrect specification of the spatial weight matrix. Two cases of model misspecification are studied. In the first case the misspecified spatial weight is used. In the second case independence of random components is assumed but the variable which is used to compute elements of spatial weight matrix in the correct case will be used as auxiliary variable in the model.

* Ph.D., University of Economics in Katowice

1. Introduction

We start from some general description of the problem. In the paper considerations are based on the model approach in survey sampling. In the survey sampling the main purpose is the estimation or prediction of characteristics of the whole population e.g. the population mean or the population total (i.e. sum of values of the variable of interest). In the practice of survey sampling typically it is not the only purpose of the survey – the estimation or prediction of subpopulation (domain) characteristics may be of interest of survey statistician as well. For example, from some population of people a sample is drawn. The key issue is to estimate the total amount of money spent for some type of goods in the whole population. Additional purpose of the survey is to estimate the total amount of money spent for the considered type of goods but not in the whole population but for inhabitants of some geographical region which (additionally) belong to the group of households consisting from 3 persons. If the division of the population due to geographical regions and household size is not taken into account in the sampling plan, the subpopulation size in the sample will be random. It means that it may be very small or even zero. What is more, if the problem will be considered in the case of longitudinal survey, we should take into account that the population may change in time, population elements may change their subpopulation's affiliation in time (that the household size, to which some person belongs to, may change in time) and that the temporal and spatial autocorrelation is observed.

2. Basic notations

Let us introduce some notation presented earlier by Żądło (2009). Longitudinal data for periods t=1,...,M are considered. In the period t the population of size N_t is denoted by Ω_t. The population in the period t is divided into D disjoint subpopulations (domains) Ω_{dt} each of size N_{dt}, where d=1,...,D. Let the set of population elements for which observations are available in the period t be denoted by s_t and its size by n_t. The set of subpopulation elements for which observations are available in the period t is denoted by s_{dt} and its size by n_{dt}. Let: $\Omega_{rdt} = \Omega_{dt} - s_{dt}$, $N_{rdt} = N_{dt} - n_{dt}$.

Let M_{id} denotes the number of periods when the i-th population element belongs to the d-th domain. Let us denote the number of periods when the i-th population element (which belongs to the d-th domain) is observed by m_{id}. Let

$m_{rid} = M_{id} - m_{id}$. It is assumed that the population may change in time and that one population element may change its domain affiliation in time (from technical point of view observations of some population element which change its domain affiliation are treated as observations of new population element). It means that i and t completely identify domain affiliation but additional subscript d will be needed as well. More about this assumptions will be written at the end of the next section.

The set of elements which belong at least in one of periods t=1,...,M to sets Ω_t is denoted by Ω and its size by N. Similarly, sets Ω_d, s, s_d, Ω_{rd} of sizes N_d, n, n_d, N_{rd} respectively are defined as sets of elements which belong at least in one of periods t=1,...,M to sets Ω_{dt}, s_t, s_{dt}, Ω_{rdt} respectively. The d*-th domain of interest in the period of interest t^* will be denoted by Ω_{d*t*}, and the set of elements which belong at least in one of periods t=1,...,M to sets Ω_{d*t*} will be denoted by Ω_{d*}.

Values of the variable of interest are realizations of random variables Y_{idj} for the i-th population element which belongs to the d-th domain in the period t_{ij}, where i=1,...,N, j=1,...,M_{id}, d=1,...,D. The vector of size $M_{id} \times 1$ of random variables Y_{idj} for the i-th population element which belongs to the d-th domain will be denoted by $\mathbf{Y_{id}} = \left[Y_{idj} \right]$, where $j = 1,...,M_{id}$. Let us consider values of the variables of interest $Y_{i'd'j'}$ for the i'-th population element which belongs to the d'-th domain observed in periods $t_{i'j'}$, where i'=1,...,n, j'=1,...,$m_{i'd'}$, d'=1,...,D. The vector of random variables $Y_{i'd'j'}$ (where i'=1,...,n, j'=1,...,$m_{i'd'}$, d'=1,...,D) of size $m_{i'd'} \times 1$ will be denoted by $\mathbf{Y_{s\,i'd'}} = \left[Y_{i'd'j'} \right]$, where j'=1,...,$m_{i'd'}$. The vector of random variables $Y_{i''d''j''}$ of size $m_{ri''d''} \times 1$ for the i''-th population element which belongs to the d''-th domain for observations which are not available in the sample is denoted by $\mathbf{Y_{r\,i''d''}} = \left[Y_{i''d''j''} \right]$, where j''=1,...,$m_{ri''d''}$.

The proposed approach may be used to predict the domain total for any (past, current and future) periods but under assumption that values of the auxiliary variables and the division of the population into subpopulations in the period of interest are known.

3. Superpopulation model

We consider superpopulation models used for longitudinal data (compare Verbeke, Molenberghs, 2000; Hedeker, Gibbons, 2006) which are – what is important for further considerations – special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). The following model is assumed:

$$\mathbf{Y_d} = \mathbf{X_d}\boldsymbol{\beta}_\mathbf{d} + \mathbf{Z_d}\mathbf{v_d} + \mathbf{e_d}, \tag{1}$$

where

$\mathbf{Y_d} = col_{1\le i\le N_d}(\mathbf{Y_{id}})$, where $\mathbf{Y_{id}}$ is a random vector of size $M_{id}\times 1$, $\mathbf{X_d} = col_{1\le i\le N_d}(\mathbf{X_{id}})$, where $\mathbf{X_{id}}$ is known matrix of size $M_{id}\times p$, $\mathbf{Z_d} = diag_{1\le i\le N_d}(\mathbf{Z_{id}})$, where $\mathbf{Z_{id}}$ is known vector of size $M_{id}\times 1$, $\mathbf{v_d} = col_{1\le i\le N_d}(v_{id})$, where v_{id} is a random component and $\mathbf{v_d}$ (d=1,2...,D) are assumed to be independent, $\mathbf{e_d} = col_{1\le i\le N_d}(\mathbf{e_{id}})$, where $\mathbf{e_{id}}$ is a random component vector of size $M_{id}\times 1$ and $\mathbf{e_{id}}$ (i=1,...,N; d=1,...,D) are assumed to be independent, $\mathbf{v_d}$ and $\mathbf{e_d}$ are assumed to be independent.

What is more, that vector of random components $\mathbf{v_d}$ obey assumptions of spatial moving average process, i.e.

$$\mathbf{v_d} = \lambda_{(sp)}\mathbf{W}_d\mathbf{u_d} + \mathbf{u_d}\ , \tag{2}$$

where $\mathbf{W}_d$ is the spatial weight matrix for profiles $\mathbf{Y_{id}}$, $\mathbf{u_d} \sim (\mathbf{0}, \sigma_u^2\mathbf{I_{N_d}})$. Hence,

$$\mathbf{v_d} \sim \left(\mathbf{0}, \mathbf{R_d}\right), \tag{3}$$

where $\mathbf{R_d} = \sigma_u^2\mathbf{H_d}$ and $\mathbf{H_d} = \mathbf{I_{N_d}} + \lambda_{(sp)}(\mathbf{W_d} + \mathbf{W_d^T}) + \lambda_{(sp)}^2\mathbf{W_d}\mathbf{W_d^T}$. Moreover, elements of $\mathbf{e_{id}}$ obey assumptions of MA(1) temporal process, i.e.

$$e_{idj} = \varepsilon_{idj} - \lambda_{(t)}\varepsilon_{idj-1}. \tag{4}$$

Hence,

$$\mathbf{e_{id}} \sim (\mathbf{0}, \mathbf{\Gamma_{id}}), \tag{5}$$

where elements of $\mathbf{\Gamma_{id}} = \sigma_\varepsilon^2 \begin{bmatrix} 1+\lambda_{(t)}^2 & -\lambda_{(t)} & 0 & ... & 0 \\ -\lambda_{(t)} & 1+\lambda_{(t)}^2 & -\lambda_{(t)} & ... & 0 \\ 0 & -\lambda_{(t)} & 1+\lambda_{(t)}^2 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1+\lambda_{(t)}^2 \end{bmatrix}$.

Let $\mathbf{Y} = col_{1\le d\le D}(\mathbf{Y_d})$, $\mathbf{V} = D_\xi^2(\mathbf{Y}) = diag_{1\le d\le D}(D_\xi^2(\mathbf{Y_d}))$ and $\mathbf{V_d} = D_\xi^2(\mathbf{Y_d})$.
Hence,

$$\mathbf{V} = diag_{1\le d\le D}(\mathbf{V_d}) = diag_{1\le d\le D}\left(\sigma_u^2 \mathbf{Z_d H_d Z_d^T} + diag_{1\le i\le N_d}(\mathbf{\Gamma_{id}})\right). \tag{6}$$

Let $\mathbf{Y_s} = col_{1\le d\le D}(\mathbf{Y_{sd}}) = col_{1\le d\le D}(col_{1\le i\le N_d}(\mathbf{Y_{sid}}))$,
$\mathbf{V_{ss}} = D_\xi^2(\mathbf{Y_s}) = diag_{1\le d\le D}(D_\xi^2(\mathbf{Y_{sd}}))$, $\mathbf{V_{ss\,d}} = D_\xi^2(\mathbf{Y_{sd}})$

Hence,

$$\mathbf{V_{ss}} = diag_{1\le d\le D}(\mathbf{V_{ss\,d}}) = diag_{1\le d\le D}\left(\sigma_u^2 \mathbf{Z_{sd} H_d Z_{sd}^T} + diag_{1\le i\le n_d}(\mathbf{\Gamma_{ss\,id}})\right), \tag{7}$$

where

$\mathbf{Z_{sd}} = diag_{1\le i\le n_d}(\mathbf{Z_{sid}})$, where $\mathbf{Z_{sid}}$ is known vector of size $m_{id}\times 1$, $\mathbf{\Sigma_{ss\,id}}$ is a submatrix obtained from $\mathbf{\Sigma_{id}}$ by deleting rows and columns for unsampled observations,

What is more,

$$\mathbf{V_{ss}^{-1}} = diag_{1\le d\le D}(\mathbf{V_{ss\,d}^{-1}}) = diag_{1\le d\le D}\left(\left(\sigma_u^2 \mathbf{Z_{sd} H_d Z_{sd}^T} + diag_{1\le i\le n_d}(\mathbf{\Gamma_{ss\,id}})\right)^{-1}\right). \tag{8}$$

Let $\mathbf{Y_r} = col_{1\le d\le D}(\mathbf{Y_{rd}}) = col_{1\le d\le D}(col_{1\le i\le N_{rd}}(\mathbf{Y_{rid}}))$,
$\mathbf{V_{rr}} = D_\xi^2(\mathbf{Y_r}) = diag_{1\le d\le D}(D_\xi^2(\mathbf{Y_{sd}}))$, $\mathbf{V_{rr\,d}} = D_\xi^2(\mathbf{Y_{rd}})$,
Hence,

$$\mathbf{V_{rr}} = diag_{1\le d\le D}(D_\xi^2(\mathbf{V_{rr\,d}})) = diag_{1\le d\le D}\left(\sigma_u^2\mathbf{Z_{rd}}\mathbf{H_d}\mathbf{Z_{rd}^T} + diag_{1\le i\le N_{rd}}(\mathbf{\Gamma_{rr\,id}})\right) \quad (9)$$

where $\mathbf{Z_{rd}} = diag_{1\le i\le N_{rd}}(\mathbf{Z_{rid}})$, where $\mathbf{Z_{rid}}$ is known vector of size $M_{rid}\times 1$, $\mathbf{\Gamma_{rr\,id}}$ is a submatrix obtained from $\mathbf{\Gamma_{id}}$ by deleting rows and columns for sampled observations. Let $\mathbf{V_{sr}} = Cov_\xi(\mathbf{Y_s}, \mathbf{Y_r}) = diag_{1\le d\le D}(Cov_\xi(\mathbf{Y_{sd}}, \mathbf{Y_{rd}}))$, $\mathbf{V_{sr\,d}} = Cov_\xi(\mathbf{Y_{sd}}, \mathbf{Y_{rd}})$. **Hence,**

$$\mathbf{V_{sr}} = diag_{1\le d\le D}(\mathbf{V_{sr\,d}}) = diag_{1\le d\le D}\left(\sigma_u^2\mathbf{Z_{sd}}\mathbf{H_d}\mathbf{Z_{rd}^T} + diag_{1\le i\le N_{rd}}(\mathbf{\Gamma_{sr\,id}})\right)$$

where $\mathbf{\Gamma_{sr\,id}}$ is a submatrix obtained from $\mathbf{\Gamma_{id}}$ by deleting rows for unsampled observations and column for sampled observations.

Similar model is considered by Żądło (2011) but instead of spatial and temporal MA models for vectors of random components considered in this paper (see assumptions (2) and (5)) he studied simultaneously spatial autoregressive (SAR) process and temporal AR(1) model. Model (1) (with assumptions (2) and (5)) similarly to the model proposed by Żądło (2011) may be used when the population changes in time and the domain affiliation of population elements changes in time. In this case observations of new element of the population or observations of the population element after the change of its domain affiliation form new profile $\mathbf{Y_{id}}$. It means that observations of new population element will be temporally correlated and spatially correlated with other population elements within the subpopulation. If the population element changes its domain affiliation its new observations will be temporally correlated (but temporally uncorrelated with old observations) and spatially correlated with other population elements within new subpopulation (but spatially uncorrelated with elements of the previous subpopulation).

In next sections three predictors of the total (of the sum of random variables) $\theta_{d^*t^*} = \sum_{i\in\Omega_{d^*t^*}} Y_{id^*t^*}$ in the domain of interest in the period of interest will be proposed.

4. First predictor – Spatial EBLUP

In the section, based on the Royall (1976) theorem, we derive the formula of the best linear unbiased predictor (BLUP) of the population total under the model (1). Let us introduce following notations:

$$\hat{\boldsymbol{\beta}}_{d^*} = \left(\mathbf{X}_{\mathbf{sd}^*}^{\mathbf{T}} \mathbf{V}_{\mathbf{ss\,d}^*}^{\mathbf{-1}} \mathbf{X}_{\mathbf{sd}^*} \right)^{-1} \mathbf{X}_{\mathbf{sd}^*}^{\mathbf{T}} \mathbf{V}_{\mathbf{ss\,d}^*}^{\mathbf{-1}} \mathbf{Y}_{\mathbf{sd}^*} \tag{10}$$

where

$$\mathbf{V}_{\mathbf{ss\,d}^*} = \sigma_u^2 \mathbf{Z}_{\mathbf{sd}^*} \mathbf{H}_{\mathbf{d}^*} \mathbf{Z}_{\mathbf{sd}^*}^T + diag_{1 \le i \le n_{d^*}} (\boldsymbol{\Gamma}_{\mathbf{ss\,id}^*}), \tag{11}$$

$\mathbf{X}_{\mathbf{sd}^*}$ is known $\sum_{i=1}^{n_{d^*}} m_{id^*} \times p$ matrix of auxiliary variables, $\mathbf{Y}_{\mathbf{sd}^*}$ is a $\sum_{i=1}^{n_{d^*}} m_{id^*} \times 1$ vector of random variables Y_{idj}.

The BLUP of the total in the domain of interest in the period of interest is given by:

$$\hat{\theta}_{(1)} = \sum_{i \in s_{d^*t^*}} Y_{id^*t^*} + \tilde{\mathbf{x}}_{\mathbf{rd^*t^*}} \hat{\boldsymbol{\beta}}_{\mathbf{d^*}} +$$

$$+ \boldsymbol{\gamma}_{\mathbf{rd^*}}^{\mathbf{T}} \left(\sigma_u^2 \mathbf{Z}_{\mathbf{rd^*}} \mathbf{H}_{\mathbf{d^*}} \mathbf{Z}_{\mathbf{sd^*}}^{\mathbf{T}} + diag_{1 \le i \le N_{rd^*}} (\boldsymbol{\Gamma}_{\mathbf{rs\,id}^*}) \right) \mathbf{V}_{\mathbf{ss\,d}^*}^{-1} \left(\mathbf{Y}_{\mathbf{sd}^*} - \mathbf{X}_{\mathbf{sd}^*} \hat{\boldsymbol{\beta}}_{\mathbf{d}^*} \right) \tag{12}$$

where

$\tilde{\mathbf{x}}_{\mathbf{rd^*t^*}}$ is a $1 \times p$ vector of totals of auxiliary variables in $\Omega_{rd^*t^*}$,

$\boldsymbol{\gamma}_{\mathbf{rd^*}}$ is a $\sum_{i=1}^{n_{d^*}} M_{rid^*} \times 1$ vector of one's for observations in period t^* (i.e. in $\Omega_{rd^*t^*}$) and zero otherwise.

If the unknown parameters σ_u^2, σ_ε^2, $\lambda_{(sp)}$, $\lambda_{(t)}$ in (12) will be replaced by some estimators we obtain the empirical best linear unbiased predictor (EBLUP) which remains unbiased under some weak assumptions (see Żądło (2004)). Because the spatial correlation is included in the assumed model, the EBLUP may be called spatial EBLUP (SEBLUP).

5. Second predictor – misspecified spatial weight matrix

In the previous section the BLUP and its empirical version were derived under the model (1) assuming that the structure of the spatial weight matrix $\mathbf{W}_\mathbf{d}$ (where d=1,...,D) is correct. In this case the new predictor is derived under the model given by formula (1) but the assumed structure of the spatial weight matrix is not correct. The misspecified spatial weight matrix will be denoted by

$\mathbf{W}_{\mathbf{d}\,(mis)}$ (where d=1,...,D). In this case we obtain some empirical predictor which is not EBLUP under (1) due to the misspecification of the spatial weight matrix – in the simulation study it will be denoted by SEBLUPmis.

6. Third predictor – independent random components

In this section we assume that population data obey assumptions of the model (1) but with $\lambda_{(sp)}=0$ and $\lambda_{(t)}=0$ what means that random components are assumed to be uncorrelated. For this model (i.e. under assumption that $\lambda_{(sp)}=0$ and $\lambda_{(t)}=0$) BLUP is given by:

$$\hat{\theta}_{(3)}=\sum_{i\in s_{d*t*}} Y_{id*t*}+\tilde{\mathbf{x}}_{\mathbf{rd*t*}}\hat{\boldsymbol{\beta}}_{\mathbf{d*}}+\sigma_u^2\boldsymbol{\gamma}_{\mathbf{rd*}}^{\mathbf{T}}\mathbf{Z}_{\mathbf{rd*}}\mathbf{Z}_{\mathbf{sd*}}^{\mathbf{T}}\mathbf{V}_{\mathbf{ss\,d*}}^{-1}\left(\mathbf{Y}_{\mathbf{sd*}}-\mathbf{X}_{\mathbf{sd*}}\hat{\boldsymbol{\beta}}_{\mathbf{d*}}\right) \tag{13}$$

where

$$\mathbf{V}_{\mathbf{ss\,d*}}=\sigma_u^2\mathbf{Z}_{\mathbf{sd*}}\mathbf{Z}_{\mathbf{sd*}}^{T}+\sigma_\varepsilon^2\,\mathbf{I}_{\sum_{i=1}^{n_{d*}} m_{id*}\times\sum_{i=1}^{n_{d*}} m_{id*}} \tag{14}$$

where $\mathbf{I}_{\sum_{i=1}^{n_{d*}} m_{id*}\times\sum_{i=1}^{n_{d*}} m_{id*}}$ is identity matrix of size $\sum_{i=1}^{n_{d*}} m_{id*}\times\sum_{i=1}^{n_{d*}} m_{id*}$, $\hat{\boldsymbol{\beta}}_{\mathbf{d*}}$ given by formula (10) where $\mathbf{V}_{\mathbf{ss\,d*}}$ given by (11) is repl aced by (14).

The predictor given by (13) is not BLUP under the model (1) due to the misspecification of the assumed model (because it is derived under assumption that $\lambda_{(sp)}=0$ and $\lambda_{(t)}=0$). In the simulation study the empirical version of the predictor (13) will be denoted by SEBLUPmis2.

7. Monte Carlo simulation study

The simulation study was conducted using R package (R Development Core Team (2012)). It is based on artificial longitudinal data from M=3 periods. The population size in each period equals N=200 elements which consists of D=20 domains (subpopulations) each of size 10 elements. The balanced panel sample is considered – in each period the same 40 elements are observed. The

sample sizes in D=20 domains are as follows {1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3}. Relatively small population size is due to very time-consuming computations. Relatively large sample size (comparing to the population size) is assumed to increase accuracy of SEBLUP due to the existence of spatial effect. The number of iterations is 2000.

In the simulation data are generated based on the model (1) assuming arbitrary chosen parameters $\sigma_u^2 = 1$, $\sigma_\varepsilon^2 = 1$, $\forall_d \beta_d = \beta = 100$, $\mathbf{X_{id}} = [1]_{M_{id} \times p}$ and different values of $\lambda_{(sp)}$ and $\lambda_{(t)}$. The spatial weight matrix is based not on geographical distances between profiles (in this case between elements) but based on the values of the auxiliary variable. In the simulation study the elements of spatial weight matrix (denoted by $\mathbf{W_d}$) were row-standardized inverses of absolute differences between sorted values of auxiliary variable generated from Beta(1,5) distribution. The Beta(1,5) distribution is distribution with positive asymmetry as many economic variables what means that considered distance between elements may be treated as a distance in some economic sense. Values of the variable were sorted to obtain the biggest values of the spatial weight matrix close to the diagonal. For this type of assumed as correct spatial weight matrix, assumption of row-standardized neighborhood matrix (where one element has two neighbours – one before and one after) as a misspecified spatial weight matrix (denoted by $\mathbf{W_{d\,(mis)}}$) is reasonable solution.

In the simulation three predictors are considered:

- The first predictor (denoted in the simulation by SEBLUP) is spatial EBLUP which is empirical version of (12), where parameters are estimated using restricted maximum likelihood method. The spatial weight matrix is given by $\mathbf{W_d}$ (described above) and $\mathbf{Z_{id}} = [1]_{M_{id} \times 1}$.
- The second predictor (denoted in the simulation by SEBLUPmis) is empirical version of (12) but where $\mathbf{Z_{id}} = [1]_{M_{id} \times 1}$ and $\mathbf{W_d}$ is replaced by row-standardized neighborhood matrix (denoted by $\mathbf{W_{d\,(mis)}}$) and parameters are estimated using restricted maximum likelihood method based on misspecified likelihood function (where $\mathbf{W_d}$ is replaced by $\mathbf{W_{d\,(mis)}}$)
- The third predictor (denoted in the simulation by SEBLUPmis2) is given by (13) where parameters are estimated using restricted maximum likelihood method based on misspecified likelihood function (assuming that $\lambda_{(sp)} = 0$ and $\lambda_{(t)} = 0$ and $\mathbf{Z_{id}}$ is a vector of auxiliary variable generated based on Beta(1,5) distribution which was used earlier (in the case of the first and the

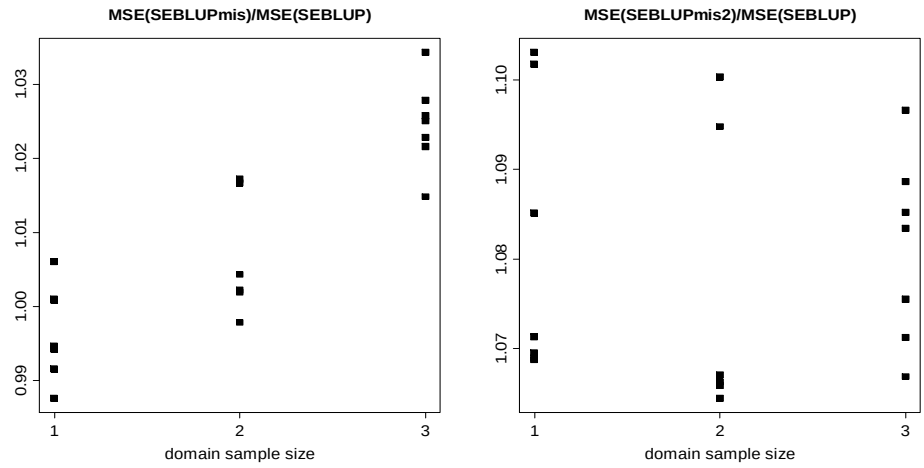
second predictor) to compute distances and then elements of spatial weight matrix.

In the simulation special cases of the predictors' equations are used for balanced panel sample and under assumption that $\forall_d \beta_d = \beta$.

Because we are mainly interested in the spatial effect, in 5 simulations we study different values of $\lambda_{(sp)}$ ($\lambda_{(sp)} = \{0,6; 0,7; 0,8; 0,9; 1\}$) and one value of $\lambda_{(t)} = 0,5$. To study maximum effect of both spatial and time effect in the last simulation it is assumed that $\lambda_{(sp)} = \lambda_{(t)} = 1$. Cases for $\lambda_{(sp)} = \{0,6; 0,7\}$ are not presented at graphs.

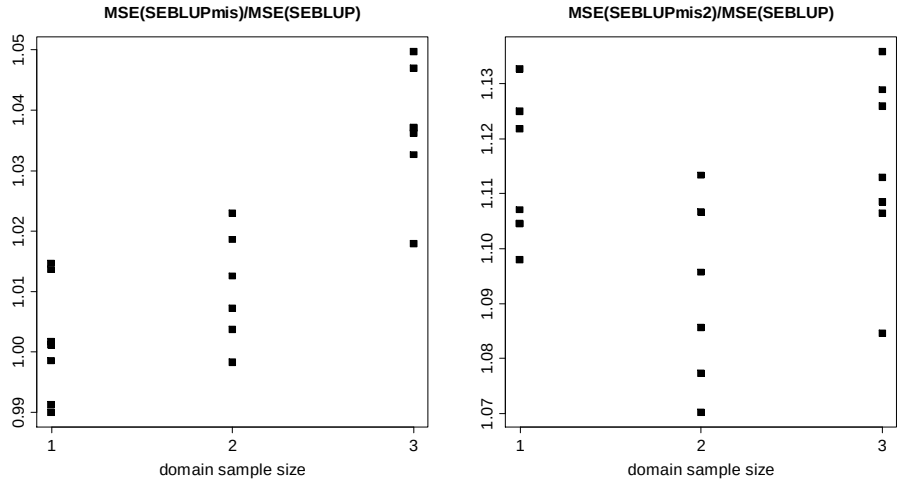
In the following graphs it is shown (on the right side) that the predictor SEBLUP may be more accurate comparing to BLUPind maximum for the considered cases from c.a. 9% to c.a. 17% for $\lambda_{(sp)} = 1$ and $\lambda_{(t)} = 0,5$.- see the right side of graph 3.

Graph 1. Simulation results for $\lambda_{(sp)} = 0,8$ and $\lambda_{(t)} = 0,5$



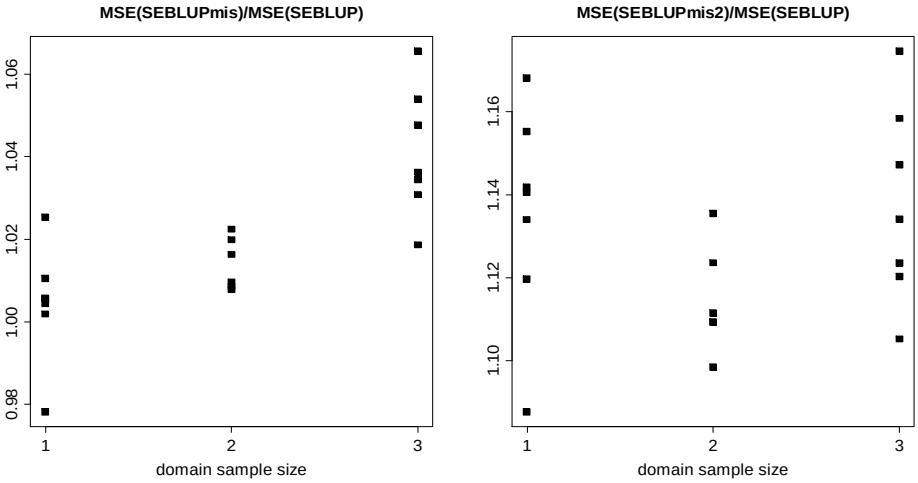
Source: Results based on own computations.

Graph 2. Simulation results for $\lambda_{(sp)} = 0{,}9$ and $\lambda_{(t)} = 0,5$



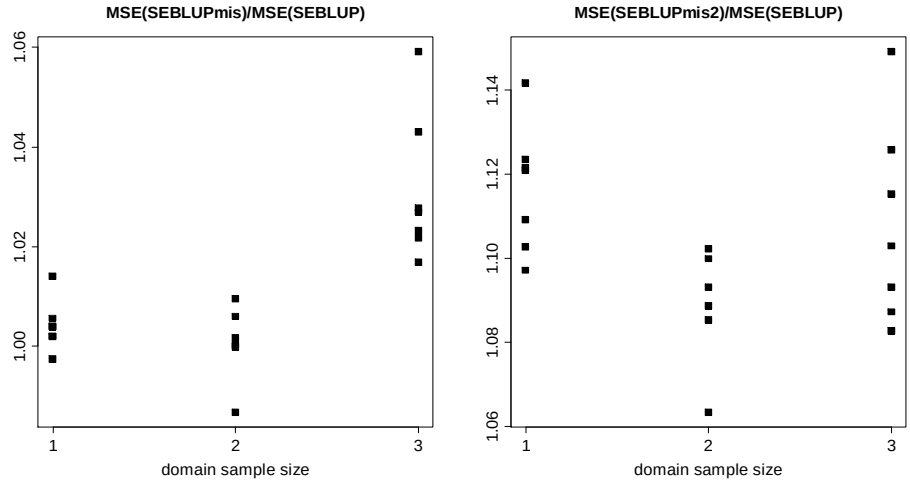
Source: results based on own computations.

Graph 3. Simulation results for $\lambda_{(sp)} = 1$ and $\lambda_{(t)} = 0,5$



Source: results based on own computations.

Graph 4. Simulation results for $\lambda_{(sp)} = 1$ and $\lambda_{(t)} = 1$



Source: results based on own computations.

The predictor SEBLUP may be more accurate comparing to SEBLUPmis maximum for the considered cases by c.a. 6% - see the left side of graph 3 and graph 4. What is very interesting, for all of the considered pairs of $\lambda_{(sp)}$ and $\lambda_{(t)}$ there are some cases when SEBLUPmis is better than SEBLUP. It results from the decrease of the accuracy of the first predictor due to the estimation of the model parameters (the decrease of the accuracy of spatial EBLUP comparing with spatial BLUP) – in some cases studied in the simulation study the maximum decrease was even greater than 5%.

Summarizing, the proposed predictor (the first predictor - SEBLUP) which takes the spatial and temporal autocorrelation into account may be better than the predictor derived under assumption of the lack of spatial and temporal autocorrelation (the third predictor – SEBLUPmis2) in the studied cases from c.a. 1% to c.a. 17%. The misspecification of the spatial weight matrix has small influence on the accuracy – the maximum decrease of the accuracy observed in the simulation was c.a. 6% but in many cases the predictor under assumption of the misspecified spatial weight matrix is better than the correct predictor what results from the decrease of the accuracy of SEBLUP due to the estimation of model parameters.

8. Conclusions

In the paper three predictors of the subpopulation total are proposed for longitudinal data and their accuracy is studied in the simulation study. It is shown that the considered misspecification of spatial weight matrix decreases the accuracy of the predictor only slightly.

References

Hedeker D., Gibbons R.D. (2006), Longitudinal Data Analysis, John Wiley, New Jersey

R Development Core Team (2011), *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna

Rao J.N.K (2003), *Small area estimation*, John Wiley and Sons, New Jersey

Royall R.M. (1976), The linear least squares prediction approach to two-stage Sampling, *Journal of the American Statistical Association*, 71, 657-473

Verbeke G., Molenberghs G. (2000*), Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New York

Żądło T. (2004), *On unbiasedness of some EBLU predictor*, [in:] J. Antoch (ed.), *Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg-New York, 2019-2026

Żądło T. (2009), On prediction of domain totals based on unbalanced longitudinal data, [in:] Wywiał J., Żądło T. (eds.) *Survey Sampling in Economic and Social Research*, University of Economic in Katowice, Katowice

Żądło T. (2011), *On accuracy of two predictors for spatially and temporally correlated longitudinal data*, submitted to publication

Streszczenie

O BŁĘDNEJ SPECYFIKACJI MACIERZY WAG PRZESTRZENNYCH W STATYSTYCE MAŁYCH OBSZARÓW W BADANIACH WIELOOKRESOWYCH

Rozważany jest problem predykcji wartości globalnej w podpopulacji (domenie) podobnie jak w Rao (2003). Zaproponowano przypadek szczególny Ogólnego Mieszanego Modelu Liniowego, gdzie dwa składniki losowe spełniają założenia

odpowiednio przestrzennego i czasowego procesu średniej ruchomej. Proponowany model jest uogólnieniem modeli wielookresowych rozważanych przez Verbeke, Molenberghs (2000) oraz Hedeker, Gibbons (2006). Wyprowadzona zostanie postać najlepszego liniowego nieobciążonego predyktora wartości globalnej w domenie. W badaniu symulacyjnym dokładność empirycznej wersji najlepszego liniowego nieobciążonego predyktora była analizowana zarówno w przypadkach prawidłowej jak i nieprawidłowej specyfikacji macierzy wag.