EURAREA

Instructions to SAS macro EBLUPGREG

Project Acronym	**EURAREA**
Project Full Title	**Enhancing Small Area Estimation Techniques to Meet European Needs**
Project/Contract No.	**IST-2000-26290**
Document Title	**EBLUPGREG User Instructions**
Version	**1.0**
Date	**20 August 2004**
Authors	**The EURAREA Consortium**
	Statistics Finland

1. Introduction

The SAS macro EBLUPGREG is a program for small area estimation using unit level data. It includes several estimators from the generalised regression estimator to the linear mixed models which may contain spatial and time correlation structures. The goal is to use sample data estimate totals or means of a continuous response variable and with the help of population data yield predicted values for the study variable over geographical regions.

2. System requirements

The EBLUPGREG –macro assumes the availability of SAS (version 8.0 or higher) including the IML –module. The macro was developed and tested in the following environment:

PC: IBM NetVista, one Intel Pentium 4 processor at 1.8 GHz, and 1.5 Gbyte RAM.
System: Windows XP, SP 1.
SAS: SAS for Windows, v. 8.2 .

The program is portable to other operating systems provided that the user takes care of the system-specific changes, e.g. in drives and paths. The program has also been tested on the trial version of SAS v. 9.1.

In the MS-Windows environment the minimum configuration with large data sets is dependent on the RAM memory. EBLUP-models can become heavy to run: especially when the sample data set is large, there are a lot of domains, the model contains many covariates (say more than five) and there are observations from various time points. Our suggestion is to have more than 512 Mbytes RAM although we have run tests successfully in less efficient environments, too (see appendix 2).

A normal execution time using a sample of 12,000 from 85 areas is 3-4 minutes for a GREG and an EBLUP model without spatial and time correlation structure. With spatial and time correlations the execution time can increase considerably: easily to 10 minutes with random time effects and even to 30 minutes with fixed time effects or time varying area effects (using the configuration above).

3. Theoretical background

Theoretical background for the estimators presented in this program are found in the EURAREA web-page http://www.statistics.gov.uk/eurarea under the work packages 2 and 3.

The main theory reports of the EURAREA project referred later in this paper are following:

Saei, A., and Chambers, R. 2003.Small Area Estimation: A Review of Methods Based on the Application of Mixed Models.

Saei, A., and Chambers, R. 2004. Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects.

These two papers introduce one broad class of estimators, the empirical best linear unbiased predictor (EBLUP) using different types of auxiliary information. However, the EBLUPGREG program includes and uses two other estimators as well: the Generalised linear regression model (GREG), and the Synthetic estimator (SYN).

Theoretical background of synthetic and GREG estimation methods were presented in other papers, and thus the reader is supposed to have access to sources listed below. Also some books containing chapters in small-area estimation would greatly help to understand the theoretical underpinnings and choices which were made during the process. Especially we can mention following sources:

Lehtonen, R., and Pahkinen, E. 2004. Practical Methods and Analysis of Complex Surveys, 2nd ed. Chichester: John Wiley & Sons. Ch. 6.

Lehtonen, R., and Veijanen, A. 1999. Domain estimation with logistic generalized regression and related estimators, in IASS Satellite Conference on Small Area Estimation, Proceedings. Riga: Latvian Council of Science, 121-128.

Lehtonen, R., Särndal, C-E., and Veijanen, A. 2003. The Effect of Model Choice in Estimation for Domains, Including Small Domains. Survey Methodology, Vol. 29, No. 1, 33-44.

Rao, J.N.K. 2003. Small Area Estimation. New York: John Wiley & Sons.

Särndal, C-E., Swensson, B., and Wretman, J. 1992. Model-Assisted Survey Sampling. New York: Springer-Verlag.

The classes of estimators in this program are:

GREG Generalised regression estimator,
GREG is based on the ordinary regression model fitted to the whole sample with sampling weights, see Särndal, Swensson, Wretman 1992, formula 6.4.13, p. 228.
Predicted values from the GREG estimator contains the prediction from the weighted XB and the normal weighted error correction mechanism.
Variance estimator as in Särndal et al., 1992, formula 10.5.12., p. 401.
If a domain is empty in the sample data set (i.e. sample size 0), the model predicts a value for the dependent variable, and correspondingly, the sampling variance is obtained from the synthetic estimator.

SYN Synthetic estimator,
synthetic estimator is based on the fixed part of a multilevel model fitted to the sample without weights. Predicted values from the SYN estimator are obtained from the weighted XB.
The MSE estimator as in Performance of Standard estimators - draft report by Patrick Heady, 2003, p. 16.

EBLUP Empirical linear unbiased predictor, ref. Saei and Chambers (2003, 2004).
Basic unit level EBLUP (neither spatial nor time correlation), see Saei and Chambers 2004, Ch. 4.1. All parameters are estimated using either restricted maximum likelihood (REML) or maximum likelihood (ML) method.
The predicted values contain weighted fixed and random effects.
The MSE estimators as specified in Saei and Chambers, 2004, Ch. 3.3. (see also p. 12).

EBLUP with spatial correlation structure
Besides the area-level random effects the model can be augmented with spatial correlation structure, see Chambers and Saei, 2004, Ch. 4.4. A natural choice is to provide the program with x-y coordinates of each area, calculated often to the mid-point. The program calculates the distance matrix by an exponential decay model or by a power model (see e.g. Littell et al., 1999, p. 305, or SAS v. 8 manual, 1999, p. 2138).
The predicted values contain weighted fixed and random effects.
The MSE estimator as in Saei and Chambers, 2004, Ch. 4.3. with the information matrix definition on p. 19.

EBLUP with autocorrelated time effect

Besides the area-level random effects the model can be augmented with the autocorrelated covariance structure. Only the first order autocorrelation structure (AR(1)) is provided. If the autocorrelation is set to zero, we get a special case: the time-independent model. These models assume that the observations are sampled from the same areas independently for various discrete time points, i.e. neither panel data nor assumptions on the length of the lags between the measurements. The models and the corresponding MSE estimators are described in Saei and Chambers, 2004, Ch. 4.2.

EBLUP with autocorrelated time effect, autocorrelation parameter fixed

Similar as above but the user sets the autocorrelation parameter to a prespecified value from the range [-0.999, 0.999]. The model is the same as above but in the MSE estimation the last row and column are removed from the information matrix.

EBLUP with fixed time trend

Sometimes data may not support the autocorrelated covariance structure. Then one can alternatively apply a fixed time trend. In this case time is considered as any other fixed (independent) covariate. Additionally the relation must be assumed to be linear. The model is described in Saei and Chambers 2004, Ch. 4.1., and the MSE estimator ibid, Ch. 3.3. (see also p. 12).

EBLUP with fixed categorical time effect

As above but here time is treated as a categorical effect (i.e. factor in the ANOVA terminology). Note, however, that the domains are here cross-classified of region and time. The model is described in Saei and Chambers 2004, Ch. 4.1., and the MSE estimator ibid, Ch. 3.3. (see also p. 12).

EBLUP with time varying area effects

All models above contain area random effects that are assumed invariant over time. However, this assumption may not always hold. One relaxation is to allow the area effects vary over time by introducing random time by area interaction effects. The model is an extension of the time autocorrelated covariance structure with the random area level variation. The model and the MSE estimator are described in Saei and Chambers 2004, Ch. 4.3.

The parameters of the EBLUP estimator include fixed and random effects, covariance matrices and ratios between the matrices, variance components, spatial correlations and autoregressive parameters of covariances. They can be estimated iteratively either by maximum likelihood (ML) or restricted (or residual) maximum likelihood method (REML). REML is the suggested estimation method but in the case of large variation in the data set the ML has sometimes performed better in simulation experiments.

Some practical adaptations were made in order to implement the theory into the program code. The most important conventions are presented in Appendix 1.

4. Program structure

The program structure is presented in graph 1. In principle three types of data are used: sample data set, population counts and sums of independent variables for the selected regions and time points, and geographic coordinates of the selected regions[1]. Only the sample data set is obligatory for the

[1] One should also keep in mind that the versatility of SAS program makes it possible to handle those data in many ways. The reader should consult the SAS/BASE manuals for data step programming in order to find the most suitable way for data input.

program but in order to obtain predicted values for the domains the user must provide the population data set as well. In addition, it is necessary to input a file containing geographical coordinates for spatially correlated models.

The choice of model type and estimation method will activate various parts in the program.
For models involving spatial correlation or longitudinal data consisting at least two samples from different time points from the same areas, the estimation method turns more complicated and can involve heavy iterative process.

A brief description of the parameter estimation is

A. cross-sectional models:

A regression model is fitted on sample data to obtain the fixed parameters.

GREG: the model is estimated over the whole sample population with sampling weights if available. Domain level predictions are obtained by applying the estimated parameters to the study variable with population data.

SYN: the model is fitted at the domain level in the sense that the random area effects from the mixed (or multilevel) model influence on the parameter estimates.
Domain level predictions are obtained by applying the estimated fixed parameters to the population data.

EBLUP: the model is fitted at the domain level and the domain level random terms are estimated in the iterative process (mixed model)
Domain level predictions are obtained by applying both the estimated fixed and domain level random parameters to the population data.

Models using spatial correlation:

GREG: as above

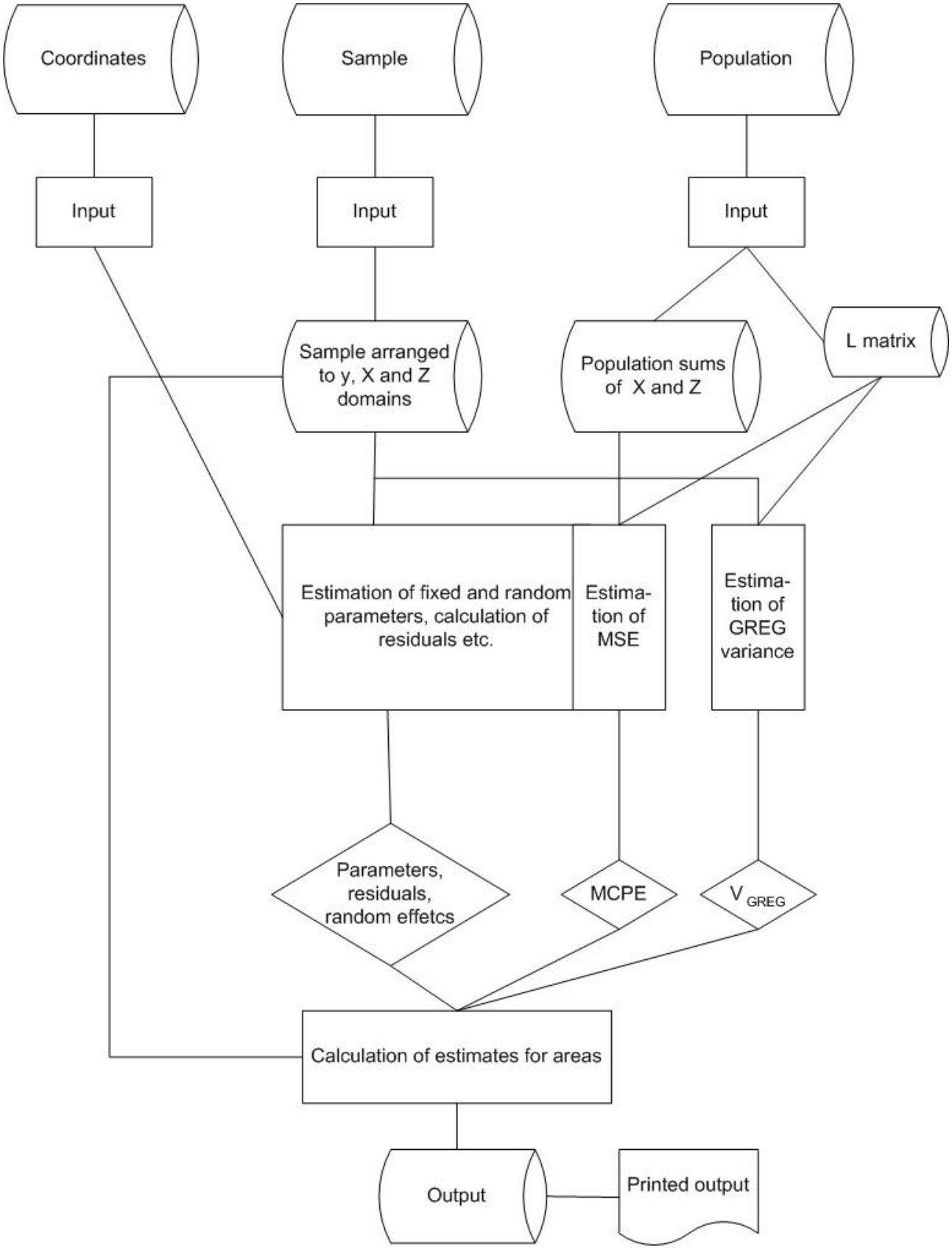
SYN and EBLUP: the spatial correlation based on the distance function affects the estimated covariance matrices and, thus, have effect on both fixed and random parameters.

Models using temporal correlation:

GREG: the model is estimated over the whole sample population (no weights) from all time points
Domain level predictions are obtained by applying the estimated parameters to the population data.

SYN and EBLUP: the first order autocorrelation of the covariance matrices affects the estimated covariance matrices and, thus, have effect on both fixed and random parameters. The time effect depends on the way time series data is taken into the model
(see different alternatives on p. 4-5)
Domain level predictions are obtained by applying both the estimated fixed and domain level random parameters to the population data.

The whole program consists of 16 sub-macros which call and combine procedures (9) and functions (23). The whole listing of the program is enclosed in the CD-ROM.

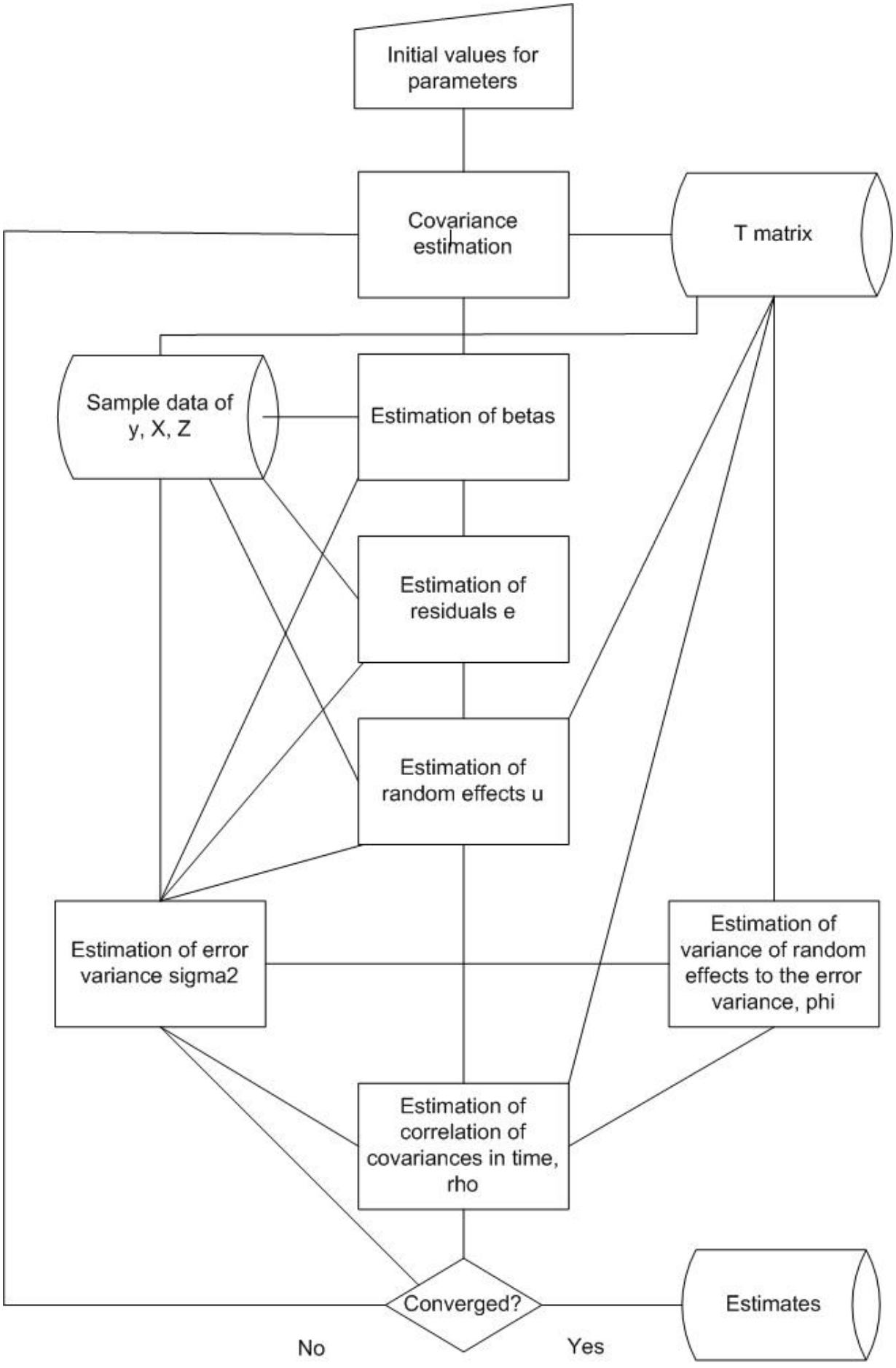

Coordinates
Sample
Population
Input
Input
Input
Sample arranged to y, X and Z domains
Population sums of X and Z
L matrix
Estimation of fixed and random parameters, calculation of residuals etc.
Estima-tion of MSE
Estima-tion of GREG variance
Parameters, residuals, random effetcs
MCPE
V GREG
Calculation of estimates for areas
Output
Printed output

Next diagram shows the estimation stage in a more detailed manner in the case of correlated time effects without spatial correlation structure.

The program calculates the naive standard errors for the fixed effect estimates and predicted space and time random effects. The estimation is basically the same as in SAS PROC MIXED (see Littell et al. 1999, 500-501). The more sophisticated standard error estimators, accounting for the estimation uncertainty of covariance parameters (Kackar and Harville, 1984, Prasad and Rao, 1990, Kenward and Roger, 1997, Rao 2003), are not applied here. However, the MSE estimators of the EBLUP domain estimators are calculated taking the estimation uncertainty of model parameters fully into account. This may yield considerable larger MSEs for domain estimates when compared to the naive MSE estimators.

For the EBLUP domain estimates, the program calculates the mean cross product error matrix, MCPE, see Saei and Chambers, 2004, Ch. 3.3. It consists of 4 components: $\boldsymbol{g_1}$ (general estimate of variation), $\boldsymbol{g_2}$ (uncertainty of estimating the β coefficients), $\boldsymbol{g_3}$ (uncertainty of estimating variance components σ^2's), and $\boldsymbol{g_4}$ (general uncertainty of the model). The MSE estimates are the diagonal elements of the estimated MCPE matrix. The EBLUPGREG program estimates all components and provides output in two ways: square root of the whole MSE and square root of all its components. Additionally, the program produces a MSE estimate without $\boldsymbol{g_3}$ component. In certain data sets the component $\boldsymbol{g_3}$ was found unstable and caused unexpected MSE estimates. In such a situation the three other components could be used instead because they still provide a good idea of the magnitude of the total error of the model. However, if the $\boldsymbol{g_3}$ component appears unstable there is a clear risk of incorrectly specified model.

The MCPE estimation was omitted from figure 2 because of its complexity.


Initial values for parameters
Covariance estimation
T matrix
Sample data of y, X, Z
Estimation of betas
Estimation of residuals e
Estimation of random effects u
Estimation of error variance sigma2
Estimation of variance of random effects to the error variance, phi
Estimation of correlation of covariances in time, rho
Converged?
No
Yes
Estimates

5. Data

The macro EBLUPGREG uses SAS v. 8 data sets[2] and all variables are referred to by their names. Three input files may exist: sample, population and coordinate files.
All data sets must satisfy the following requirements:

- only one dependent variable is allowed for each run
- variables must have the same names in different data sets in order to be correctly processed
- no missing values are allowed
- all domains (regions, time points) in the sample data set must be a true subset of the domains in the population data set. I.e. population data set must contain at least the same domains as the sample.
- all variables must be numeric. Character (or alpha-numeric) variables are not allowed.
- following data set names are reserved for the program:
  SSAMPLE12387423,
  SPOPULATIONSUMS12387423,
  TPOPULATIONSUMS12387423,
  TPOPULATION12387423,
  SUFILE12387423,
  TUFILE12387423,
  SCOORDINATES12387423

5.1. Sample data set

The sample data MUST contain following arguments used in the program:

Argument name	Description
y	response variable, always included.
xlist	names of quantitative x variables separated by blanks, no constant variables
regionIdentifier	name of the integer-valued region (or other domain) identifier, values>0, always included.

for models with correlated time

timeIdentifier:	name of the time period identifier (integer variable: 1,2,…,t; no breaks)

Additionally, if GREG estimates are required the user must provide:

weights	name of the sample weight variable.

Categorical variables cannot be used in the models. Instead, corresponding sets of indicator variables must be created and used as quantitative x variables. Sets of indicators must not be constant for all observations to avoid linear dependence with the constant (1), which is always included in the model. And as usual, one of indicator classes must be omitted to avoid singular matrices. For example, if respondent's gender is included the user can create a dummy variable either for male or female respondents.

[2] Note: The program code was written using long intrinsic variable names and therefore cannot be run with earlier SAS versions.

It is a good idea to rescale continuous variables with large values, i.e. by dividing them by some constant (say 100, 1000 etc.). If the scales of the covariates differ very much the estimation procedures can fail because of very large elements in the covariance matrices.

It is not necessary to sort the data sets in any way although sorting by time period and area index makes the analysis faster.

A cut from the test sample:

domain	time	y1	y2	y3	y4	x	xc	yc	weight
1	1	14.4326	1	3	0	14.9901	8.21243	5.29670	5
1	1	17.5515	6	1	0	16.8463	8.21243	5.29670	5
1	1	20.2743	10	1	1	18.7906	8.21243	5.29670	5
1	1	15.5744	1	2	0	14.2404	8.21243	5.29670	5
1	1	18.4442	4	3	0	16.3443	8.21243	5.29670	5
1	1	19.1186	3	1	1	17.0644	8.21243	5.29670	5
1	1	19.2151	10	1	1	17.7163	8.21243	5.29670	5
1	1	20.3930	15	3	0	19.4540	8.21243	5.29670	5
1	1	16.4125	2	1	0	12.4698	8.21243	5.29670	5
1	1	17.5037	9	1	0	17.0986	8.21243	5.29670	5
...	...	...	...	...	...	...	...	...	...
10	5	21.0379	24	1	0	23.0330	3.20587	6.77050	5
10	5	13.7426	0	3	0	9.6020	3.20587	6.77050	5
10	5	16.8099	1	3	0	15.0929	3.20587	6.77050	5
10	5	18.9447	4	3	0	15.5438	3.20587	6.77050	5
10	5	21.7510	24	1	1	26.4579	3.20587	6.77050	5
10	5	18.3771	12	1	0	18.8686	3.20587	6.77050	5
10	5	17.0208	6	3	1	14.9129	3.20587	6.77050	5
10	5	16.8416	10	2	0	19.1193	3.20587	6.77050	5
10	5	23.5388	21	3	1	21.9886	3.20587	6.77050	5
10	5	20.7281	3	3	0	17.1964	3.20587	6.77050	5
10	5	19.3926	17	1	0	20.8327	3.20587	6.77050	5

Note: If the sample data set is longitudinal but the aim is to run the estimates for only one time point using the cross-sectional data the user has to select the correct period data before invoking the program as there is no option for selecting subpopulations. (see the DATA phase statement before example 1 in the test program).

5.2. Population data set

The population data set is optional. However, it is advisable to include it whenever available because the total counts of auxiliary information at the domain level are necessary for predicting the study variable values for the very domains.
The population totals of auxiliary variables in regions are obtained either from data set "population" or "populationSums". In the case of large populations the "populationSums" is preferred to "population" data set.

The “populationSums” data set must contain the domain population size and domain sums of all covariates used in the model. The variables should correspond with the following arguments of the EBLUPGREG macro:

Argument name	Description
xlist	(non-empty) name list of quantitative x variables (covariates, no constant variables) separated by blanks, “populationSums” must contain sums of these over each domain
regionIdentifier	name of the integer-valued region identifier, always included.

regionsize	name of the variable which contains the domain (= region or region by time, depending on the model) population counts

Additionally for models with correlated time

timeIdentifier:	name of the time period identifier (integer variable: 1,2,…,*t*; no breaks). Normally the last time point *t* is regarded as the current (or latest) period but due to the model structure the order may be converted.

For example, consider a mixed model with time correlating structure (5 periods), the fixed part of the model defined as y = f(x), for regions=1 to 10. Now the auxiliary information can be given as the following file print-out shows:

time	domain	y1	y2	y3	y4	x	n
1	1	1282.24	555	112	20	1165.31	69
1	2	2566.18	2014	169	84	2671.44	120
1	3	1887.39	1154	210	49	1919.31	94
1	4	1889.59	1670	180	69	2054.49	86
1	5	1760.36	1187	125	47	1715.85	86
1	6	4721.08	4299	341	176	4997.97	204
1	7	835.74	382	86	13	745.90	46
1	8	1017.50	866	94	33	1088.94	47
1	9	898.09	840	88	33	997.20	40
1	10	3701.64	2875	424	111	3748.90	174
2	1	1330.24	695	107	24	1216.60	69
...	...	...	...	...	...	...	...
4	10	3499.46	2244	422	82	3649.93	174
5	1	1311.42	573	104	26	1211.46	69
5	2	2571.52	2079	166	79	2699.47	120
5	3	1926.45	1336	225	55	2040.77	94
5	4	1869.25	1628	190	65	2020.92	86
5	5	1710.26	1080	128	42	1659.49	86
5	6	4646.90	4128	347	160	4864.06	204
5	7	799.07	266	90	10	681.20	46
5	8	1044.62	950	95	39	1164.58	47
5	9	892.01	794	100	32	992.94	40
5	10	3570.38	2507	424	103	3599.26	174

Here we have sums for our study variable (y1), too, in order to test model performance. For the model regionIdentifier = domain, timeIdentifier = time and regionsize= n.

Data set "population" contains unit-level information about variables. The arguments are otherwise the same as above but “regionSize” is not needed because the data set will be cumulated over the domains. Thus the arguments look:

Argument name	Description
xlist	(non-empty) name list of quantitative x variables separated by blanks, no constant variables
regionIdentifier	name of the integer-valued region identifier, always included.

Additionally for models with correlated time

timeIdentifier:	name of the time period identifier (integer variable: 1,2,…,*t*; no breaks). Normally the last time point *t* is regarded as the current (or latest) period but due to the model structure the order may be converted.

Again, it is not necessary to sort the population or populationSums data set but sorting by time period and area index can improve performance.

5.3. Geographic coordinate file

A data set "coordinateFile" contains the x-y coordinates of the regions. This data set is necessary for the estimation of spatial correlations of area effects.

A coordinateFile data set must contain at least three variables which are read by two arguments:

Argument name	Description
regionIdentifier	name of the integer-valued region identifier
coordinates	names for x and y coordinates.

An example where regionIdentifier = domain, coordinates = xc yc:

domain	xc	yc
1	8.21243	5.29670
2	6.79939	1.36015
3	2.60367	7.03539
4	0.34268	3.01330
5	8.99029	1.76030
6	8.86374	4.78874
7	7.15563	6.21529
8	5.21768	6.64074
9	5.26464	3.19288
10	3.20587	6.77050

5.4. Files for testing purposes

It is possible to test the functioning of the model for checking, e.g. for validating the best alternative with respect to unbiasedness or RMSE.

Data set "ufileS" is used only in tests when the true area effects are known.
It has two variables defined by macro arguments

Argument name	Description
regionIdentifier	name of the integer-valued region identifier
trueU	name of the random effect variable.

Data set "ufileT" is used only in tests when the true time effects are known.
It has two variables specified by

Argument name	Description
timeIdentifier	name of the integer-valued time identifier
trueU	name of the random effect variable.

5.5. Output file(s)

An output data set is created by the macro when the user provides the program with a file name and population data set is read for prediction. It contains the following variables:

Name	Description
region	region (always)
time	time period, if timeIdentifier is given in macro call

EBLUP	EBLUP estimates, if eblup=1 in the macro call
GREG	GREG estimates, if greg=1 in the macro call
Synthetic	synthetic estimates, if synthetic=1 in the macro call
areaEffect	predicted area effect for each region
areaEffStdErr	standard error for the area effect
timeEffect	predicted time effect for each time period
timeEffStdErr	standard error for the time effect
sqrtMSE	square root of the estimated MSE of EBLUP
sqrtMSENoG3	square root of the estimated MSE of EBLUP, calculated without g3
sqrtg1	square root of the estimated g1 component of the MSE
sqrtg2	square root of the estimated g2 component of the MSE
sqrtg3	square root of the estimated g3 component of the MSE
sqrtg4	square root of the estimated g4 component of the MSE
sqrtMSEsyn	square root of the estimated MSE of the synthetic estimator
stdGREG	the estimated standard deviation of GREG

If a data set ufileS or ufileT is specified for testing, the estimated random effects are added to these data sets.

6. Invoking EBLUPGREG and the standard output

6.1. Invoking the macro

The macro contains the following arguments. Those marked are compulsory, others optional (depending on the model choice). In the macro call, it is possible to give initial values of some of the model parameters. If an argument is not needed, it can be omitted from the macro call. Some arguments have default values, which are used if the argument is not specified. The order of arguments is not relevant

Intrinsic name	compulsory	definition
%eblupgreg(		
sample=,	Y	name of sample data set
population=,		name of unit level population data set
populationSums=,		name of the data set containing population totals
regionSize=,		name of the variable defining regional population sizes if data set POPULATIONSUMS defined
y=,	Y	name of the dependent variable of the model
xlist=,	Y	name(s) of the independent variables in the model
regionIdentifier=,	Y	name of the region/domain identifier
timeIdentifier=,		name of the variable containing time periods compulsory if time effect model was chosen
test=**0**,		switch for the testing, values: 1 test, 0 no (default)
spatialCorrelations=**0**,		switch for spatial correlation model: 1 yes, 0 no (default)
spatialType='exponential',		type of the spatial model, if spatialCorrelations=1 value 'EXPONENTIAL' (default): specifies model with exponentially decaying autocorrelations: for regions distance d apart, correlation is exp(-alpha*d) value 'POWER': specifies a power model for regions distance d apart, correlation is d**alpha. Note: the program rescales the distances by dividing the original Euclidean distances by the mean distance.

spatialParameter=**0.5,**	initial value of the parameter alpha in the model for spatial autocorrelation, default is 0.5. In “EXPONENTIAL” model value >0, In “POWER” model value [0.0001, 0.9999]
timeParameter=**0.3**,	initial value of the parameter rho in the model for time effects autocorrelation, range = [-0.999, 0.999], default is 0.3
fixedTimeCorrelations=**0**,	switch to invoke time effects with fixed autocorrelation structure, values: 1 yes, 0 no (default) if 1 the TIMEPARAMETER must be given a value
timeCorrelations=**0**,	indicator of correlated time effects in the model, values: 1 correlated (AR1), 0 independent (default)
timeTrend=**0**,	indicator for a trend as fixed effects, values: 1 yes, 0 no (default)
timeIndicators=**0**,	switch to invoke categorical time effects in the model (instead of TIMETREND), 1 yes, 0 no (default)
timeVaryingAreaEffects=**0**,	switch to invoke a random effects model with time by area variation. 1 yes, 0 no (default)
coordinateFile=,	name of the file containing the x-y coordinates
coordinates=,	names of the coordinate variables in data file
weights=,	sampling weights (to be used in GREG only)
ufileS=,	data set containing the true area effects (used in tests) if TEST=1
ufileT=,	data set containing the true area effects (used in tests) if TEST=1 & timeParameter>**0**
trueU=,	name of the true random effects variable (used in test)
convergenceCrit=**1e-5**,	convergence criterion used to evaluate the max absolute change in the estimable parameter, default=**1e-5**
maxiterations=2**00**,	maximum number of iterations, default=200
initialPhiS=**1**,	initial value of the ratio of the variance of area effects to the error variance, default is 1
initialPhiT=**1**,	initial value of the ratio of the variance of time effects to the error variance, default is 1
initialSigma2=**1**,	initial value of the unit-level error variance, default is 1
modules=modules.eurarea,	name of the SAS library and catalog to contain the IML modules created by the macro, default name is modules.eurarea
estimateLastPeriod=**0**,	switch to order the program to predict the values for the last period only if any time effect is specified, values: 1 yes, 0 no (default) NOTE: the model is estimated using observations from all time periods.
parametersEstimatedBy='REML',	switch for the estimation method for the EBLUP: value ‘ML’ invokes the maximum likelihood est. value ‘REML’ (default) invokes the restricted maximum likelihood estimation method
eblup=**1**,	switch to invoke the EBLUP estimator, values: 1 yes (default), 0 no
greg=**0**,	switch to invoke the GREG estimator, values: 1 yes, 0 no (default)
synthetic=**0**,	switch to invoke the SYNTHETIC estimator, values: 1 yes, 0 no (default)

estimateMeans=**0**,		switch to invoke estimation of means instead of totals, values: 1 yes, 0 no (default)
stratified=**0**,		indicator to tell the program of a stratified sampling design: switch to invoke the GREG estimator, values: 1 yes, 0 no (default)
stratum=,		name of stratum if STRATIFIED=1
output=	Y	name of the output file
);		

The estimated parameter of the dependent variable is either a total or a mean at the required regional domains and time periods. The mixed model may contain both random effects associated with regions (area effects) and random effects associated with time (time effects).
When a time effects model is applied, one should be careful to choose the right combinations of various indicators:

Random time effect with the autocorrelation parameter rho to be estimated:
timeIdentifier= <name>,
timeCorrelations=**1,**

Random time effect with rho set by the user:
timeIdentifier= <name>,
timeCorrelations=1**,**
fixedTimeCorrelations=1,
timeParameter=<specified value [-1, 1]>,

Fixed time effect:
timeIdentifier= <name>,
timeCorrelations=0**,**
and one of the following:
timeTrend=1, (for linear trend) or
timeIndicator=1, (for categorical time effect – the program creates the indicators automatically)

Time varying area effects:
timeVaryingAreaEffects=1,
timeCorrelations=1,
timeIdentifier= <name>,
fixedTimeCorrelations=0,
spatialCorrelations=0

NOTE:

A lower bound for the ratio of random effect variance to the error variance is **1e-8**, and the correlation of time effects is restricted to [-0.999,0.999]. If the correlation coefficient reaches either boundary the procedure will stop trying to improve the AR1-parameter estimate because there is a risk to achieve very unstable information matrix. Instead the correlation will be fixed to the detached bound and the model is estimated using the fixed time correlation structure. In such a case the user is advised to finally try an alternative model.
There are many ways of introducing time, besides those listed above one can also create 0-1 - indicators to all but one time points which will lead to fixed effects of individual time points.

Convergence criteria in estimation algorithms have been written assuming that typical values of x and y variables are between 1 and 1000. If all responses are close to zero, estimation algorithms may converge too quickly; if the responses are too large, estimation takes more time. If the variables in the model have large differences in their scale the user is advised to transform them appropriately.

An example of a macro call:

Note: first the user will have to give the correct drive and path information for the modules (libname modules <DRIVE:\PATH>;) created by the macro on the first row of the main macro EBLUP_LINEAR.SAS, and save it. Thereafter one can use the macro by calling it as follows, in this case a model with spatial correlation structure:

```
libname a  <DRIVE:\PATH>;
filename EBLUPLIN  '<DRIVE:\PATH\> EBLUP_LINEAR.SAS';
%include EBLUPLIN;
%eblupgreg(sample=a.sample,
      populationSums=a.popsum,
      regionSize=popN,
      y=y1,
      xlist=x1 x2,
      regionIdentifier=domain,
      coordinateFile=a.coord,
      coordinates=xc yc,
      spatialType='exponential',
      spatialCorrelations=1,
      modules=modules.eurarea,
      eblup=1,
      greg=1,
      synthetic=1,
      output=a.blups);
```

6.2. Output

The standard output contains

A. Technical details of the estimation procedure
(1) Number of iterations (***ITERATIONS***)
(2) Information matrix for the model (***IMAT***)
(3) Inverse of the information matrix for the model (***INVI***)
(4) Smallest singular value of the information matrix (***MINSING***)
(5) Traces of the estimated MSE components g1-g4 of the EBLUP model (***TR1-TR4***).
(6) Estimated bounds for the alpha parameter for spatial correlation models (***MINALPHA***, ***MAXALPHA***)
(7) Correlation matrix of the estimated fixed effect parameters (***CORRB***)

B. Model parameters and diagnostics

(8) Estimates for the fixed beta parameters (***PARAMETERS***) and their respective standard errors
(9) Estimates for the variance ratio, ***PHI,*** for the area effect and time effects
(10) ***AREA EFFECT VARIANCE*** for models with random area effect (also time-varying)
(11) ***TIME EFFECT VARIANCE*** for models with random time effect

(12) Estimated 1st order autoregressive correlation parameter (***AR(1)***)
(13) Estimated ***ALPHA***, i.e. spatial correlation parameter calculated from the distance function
(14) Error variance ***SIGMA2***
(15) Coverage rates (***RATE***) and mean absolute relative errors (***MARE***) for the estimators if test=1.

References

Heady, P. 2003. Report on the performance of the “Standard” Estimators (Draft). London: Office for National Statistics.

Kackar, R.N., and Harville, D.A. 1984. Approximations for standard errors of estimators of fixed and random effects in mixed linear models, *Journal of the American Statistical Association,* Vol. 79, 853-862.

Kenward, M.G., and Roger, J.H. 1997. Small sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, Vol. 53, 983-997.

Lehtonen, R., and Pahkinen, E. 2004. *Practical Methods and Analysis of Complex Surveys*, 2nd ed. Chichester: John Wiley & Sons. Ch. 6.

Lehtonen, R., and Veijanen, A. 1999. Domain estimation with logistic generalized regression and related estimators, in *IASS Satellite Conference on Small Area Estimation, Proceedings*. Riga: Latvian Council of Science, 121-128.

Lehtonen, R., Särndal, C-E., and Veijanen, A. 2003. The Effect of Model Choice in Estimation for Domains, Including Small Domains. *Survey Methodology*, Vol. 29, No. 1, 33-44.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. 1999. *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

Prasad, N.G.N., and Rao, J.N.K. 1990. The Estimation of the Mean Squared Error of Small-Area Estimators, *Journal of the American Statistical Association,* Vol. 85, 163-171.

Rao, J.N.K. 2003. *Small Area Estimation*. New York: John Wiley & Sons.

Saei, A., and Chambers, R. 2003. Small Area Estimation: A Review of Methods Based on the Application of Mixed Models.

Saei, A., and Chambers, R. 2004. Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects.

SAS Institute. 1999. *SAS/STAT User’s Guide Version 8*. Cary NC: SAS Institute Inc.

Särndal, C-E., Swensson, B., and Wretman, J. 1992. *Model-Assisted Survey Sampling*. New York: Springer-Verlag.

Practical adaptation of the theory for the program code

Ari Veijanen 12.8.2004

Theory report by Saei & Chambers (2004): Small Area Estimation Under Linear and Generalized Linear Mixed Models With Time and Area Effects.

(1) **Transformations of domain sums**

In (1.4) of Saei & Chambers (2004), the domain sums of the response are defined by

$$\theta = a_s y_s + a_r y_r .$$

In our algorithm, the matrix a is not used. Domain sums of variables are used without referring to the matrix a. A matrix L is used to transform the vector of domain totals. When our goal is to estimate domain totals, L is an identity matrix. For domain means, L contains the inverses of domain population sizes.

(2) **Check of linear independence**

If the columns of X are linearly independent, then the rank of X equals the number of columns in X. Since rank(X'X) = rank(X), this condition can be checked by comparing rank(X'X) with the number of columns in X'X. The rank of X'X (denoted by XtX) is evaluated by IML as follows:

rankXtX=round(trace(ginv(XtX)*XtX)).

This equation is recommended in SAS/IML documentation (SAS OnlineDoc, version 8; IML function rank). If the calculated rank is smaller than the number of columns in X'X, the procedure checkXrank

prints a note and model fitting will not proceed. The result may be inaccurate. If the calculated rank is larger than the true rank, the linear dependencies in X could be overlooked, but this error is probably rare in practice. If the calculated rank is smaller than the number of columns, it is always best to modify the X matrix in order to avoid numerical instability in the model fitting.

(3) **Inversion of large covariance matrices**

To avoid the inversion of the large matrix

$$\Sigma_s = W_s + Z_s \Omega Z_s'$$

we apply equation (a) in theorem on p. 4 of Saei & Chambers (2004):

$$\Sigma_s^{-1} = W_s^{-1} - W_s^{-1} Z_s T_s^* Z_s' W_s^{-1},$$

where

$$T_s^* = (\Omega^{-1} + Z_s' W_s^{-1} Z_s)^{-1}.$$

In our application, $W_s = I$. Therefore,

$$\Sigma_s^{-1} = I - Z_s T_s^* Z_s'.$$

This is applied in several cases. Matrix

$$X_s' \Sigma_s^{-1} X_s$$

is calculated as

$$X_s' X_s - X_s' Z_s T_s^* Z_s' X_s$$

in G_2 (p. 5 in Saei & Chambers (2004); procedure calculateMCPE) and in

$$T_s = T_s^* + T_s^* Z_s' X_s (X_s' \Sigma_s^{-1} X_s)^{-1} X_s' Z_s T_s^*$$

(eq. 7 on p. 8 of Saei & Chambers (2004); matrix Tphi in procedures estimationT, calculateParameters and estimationS).

By equation (2.2) of Saei & Chambers 2004, p. 4,

$$\hat{\beta} = \left(X_s' \Sigma_s^{-1} X_s\right)^{-1} X_s' \Sigma_s^{-1} y_s .$$

In IML function betaEst we replace Σ_s^{-1} by $I - Z_s T_s^* Z_s'$ and apply

$$\hat{\beta} = \left(X_s' X_s - X_s' Z_s T_s^* Z_s' X_s\right)^{-1} (X_s' y_s - X_s' Z_s T_s^* Z_s' y_s).$$

(4) **Random effect estimation**

By equation (2.2) in Saei & Chambers (2004), p. 4, the random effects are estimated by

$$\hat{u} = \Omega Z_s' \Sigma_s^{-1} (y_s - X_s \hat{\beta}).$$

This equation is simplified by applying equation (b) in theorem on p. 4 of Saei & Chambers (2004). When we replace $Z_s' \Sigma_s^{-1}$ by $\Omega^{-1} T_s^* Z_s'$, we obtain

$$\hat{u} = T_s^* Z_s' (y_s - X_s \hat{\beta}).$$

This is applied in procedures calculateParameters, estimationT and estimationS.

(5) **Derivative of a matrix with respect to a ϕ-parameter**

The derivative of any inverse matrix A^{-1} with respect to a parameter ϕ is calculated as

$$\frac{\partial}{\partial \phi} A^{-1} = -A^{-1} \left(\frac{\partial}{\partial \phi} A \right) A^{-1}.$$

This is applied in procedures informationMatrix and MCPES.

(6) **Scaling the distances in spatial correlation models**

An IML function distances calculates the Euclidean distances between region centres and scales them by dividing each distance by the average distance between regions. The scaling may reduce numerical estimation problems, when the average distance is large.

(7) **Spatial correlation derivatives**

An IML function corrDva calculates the derivative of the spatial correlation matrix with respect to the alpha parameter. When the correlations are exponentially decaying, the element (i,j) in the correlation matrix is

$$\exp(-\alpha d_{ij})$$

and its derivative with respect to α is

$$-d_{ij} \exp(-\alpha d_{ij}).$$

In the power model, the corresponding derivative is

$$d_{ij} \alpha^{d_{ij}-1}.$$

(8) **Information matrix in the case of fixed time effect in the AR(1) model**

When the AR(1) correlation parameter ρ is fixed, the last row and column of the information matrix are removed.

(9) **Calculation of final estimates**

After the iterative estimation algorithm (in estimationT and estimationS) has converged, the final estimates of residuals and random effects are calculated using the same equations as in the iterations. Moreover, if the sample contains empty regions, we recalculate certain matrices for all the regions in population.

(10) **Covariance matrix of fixed parameter and random effect estimates used for standard errors**

The covariance matrix of beta parameters and the random effects is

$$C = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^-$$

(SAS online documentation for SAS/STAT, MIXED procedure, chapter 41, section 23). In our procedure calculateCovBU, R=I and G is the inverse of the covariance matrix of random effects.

(11) **GREG estimator methodology**

In procedure gregEstimator, the sampling designs permit simple equations for the estimates of variances. The domain variances are defined in eq. 10.5.12 in Särndal et al. 1992, p. 401 by

$$\hat{V}_d = \sum_{k,l} \breve{\Delta}_{kl} \frac{g_{dks} e_{ks}}{\pi_k} \frac{g_{dls} e_{ls}}{\pi_l},$$

where

$$\breve{\Delta}_{kl} = 1 - \frac{\pi_k \pi_l}{\pi_{kl}} \text{ for } k \neq l \text{ and } \breve{\Delta}_{kk} = 1 - \pi_k .$$

The inclusion probabilities π_k are calculated as inverses of sampling weights.

In SRSWOR,

$$\breve{\Delta}_{kl} = 1 - \frac{\pi_k \pi_l}{\pi_{kl}} = 1 - \frac{\pi_k \pi_l}{\pi_k (n-1)/(N-1)} = 1 - \frac{n/N}{(n-1)/(N-1)} \quad (k \neq l).$$

This is independent of k and l and therefore it is denoted by $\breve{\Delta}$ (variable DeltaG in the algorithm).

We also denote

$$\tilde{e}_k = \frac{g_{dks} e_{ks}}{\pi_k} \text{ (variables gres in the algorithm).}$$

Then the domain variance under SRSWOR is

$$\begin{aligned}
\hat{V}_d &= \sum_{k,l} \breve{\Delta}_{kl} \frac{g_{dks} e_{ks}}{\pi_k} \frac{g_{dls} e_{ls}}{\pi_l} \\
&= \sum_k \sum_{l \neq k} \breve{\Delta} \tilde{e}_k \tilde{e}_l + \sum_k (1-\pi_k) \tilde{e}_k^2 \\
&= \breve{\Delta} \sum_k \tilde{e}_k \left(\sum_l \tilde{e}_l - \tilde{e}_k \right) + \sum_k (1-\pi_k) \tilde{e}_k^2 \\
&= \breve{\Delta} \left(\sum_k \tilde{e}_k \right)^2 - \breve{\Delta} \sum_k \tilde{e}_k^2 + \sum_k (1-\pi_k) \tilde{e}_k^2 \\
&= \sum_k (1-\pi_k) \tilde{e}_k^2 + \breve{\Delta} \left[\left(\sum_k \tilde{e}_k \right)^2 - \sum_k \tilde{e}_k^2 \right].
\end{aligned}$$

This is denoted in the algorithm by sum1 + DeltaG*(sumge**2 - sumge2).

In stratified sampling with constant weights within strata,

$$\breve{\Delta}_{kl} = 1 - \frac{\pi_k \pi_l}{\pi_{kl}} = 1 - \frac{\pi_k \pi_l}{\pi_k (n_h - 1)/(N_h - 1)} = 1 - \frac{n_h / N_h}{(n_h - 1)/(N_h - 1)} \quad (k \neq l \text{, both in same stratum } h).$$

For $k \neq l$ in different strata, $\breve{\Delta}_{kl} = 0$. The algorithm is similar to the algorithm under SRSWOR with small modifications.

Appendix 2. Test runs with two different computer set-ups.

A test data set with following options were run:

a) cross-sectional data, 1000 observations and 12 areas,
   GREG, SYN, and EBLUP estimators
b) same data, GREG and SYN, estimators; EBLUP estimator with spatial correlation structure
c) longitudinal data, 5 years each containing 1000 observations and 12 areas,
   GREG and SYN estimators; EBLUP estimator with independent time correlation structure
d) same data as in c) but time varying area effects in EBLUP.

PC	IBM NetVista,		IBM Aptiva 2158-282
Processor	Intel Pentium 4 at 1.8 GHz,		AMD K6 at 400 MHz
RAM	768 Mbyte		128 Mbyte
Hard disk	40 Gbyte		8 Gbyte HD
System	Windows XP, SP 1.		Windows 98
SAS	SAS for Windows v. 8.2	SAS for Windows v. 9.1	SAS for Windows v. 8.2
a)	2 sec	1 sec	8 sec
b)	4 sec	3 sec	12 sec
c)	30 sec	21 sec	1 min 38 sec
d)	47 sec	38 sec	2 min 57 sec
Total	1 min 23 sec	1 min 3 sec	4 min 55 sec

Appendix 3. Test runs for different models

3.1. Cross-sectional data, random area effects – neither spatial nor time series correlation structure. Estimates for the domain means requested.

Program call:

```
%eblupgreg(sample=sample5,
    populationSums=popsum5,
    regionSize=n,
    weights=weight,
    y=y1,
    xlist=x,
    regionIdentifier=domain,
    modules=modules.eurarea,
    estimatemeans=1,
    eblup=1,
    greg=1,
    synthetic=1,
    test=1,
    output=test1);

    run;
```

Here we have also requested a test of the coverage of estimators for y.

```
                                ITERATIONS

                                        17

                                      IMAT

                    information matrix 16.996453 10.480823
                                       10.480823 141.72845

                                           INVI

                  inverse of information matrix  0.061647 -0.004559
                                                -0.004559 0.0073929


                                                          MINSING

               smallest singular value of information matrix 16.121915

                         traces of g1, g2, g3 and g4 in MCPE


                             TR1       TR2       TR3       TR4

                        0.6124989 0.0481327 0.0745334 0.2569319

                                                        RATE

                 EBLUP confidence interval coverage rate        100 %

                                                          RATE2

              EBLUP confidence interval coverage rate (no g3)       100 %

                                                        RATE

                   GREG confidence interval coverage rate        100 %
```

NAME	MARE
EBLUP	0.0108732

NAME	MARE
GREG	0.0149202

NAME	MARE
synthetic	0.0113959

CORRB

estimated correlation matrix of beta:

1	-0.918287
-0.918287	1

PARAMETERS

parameter	in mixed model	std. error	in GREG
intercept	10.130942566	0.5203481669	9.9180215623
X	0.5027544621	0.0220598283	0.5151476947
phi(area)	0.1168438513		
area effect variance	0.2776913381		
sigma2	2.3766020631		

The output data set contains the predicted values and their respective root-mean square errors, and the area effect estimates:

Obs	region	EBLUP	GREG	synthetic	area Effect	effect StdErr	sqrtMSE	sqrt MSENoG3
1	1	19.4174	19.5959	18.9580	0.40866	0.34964	0.34564	0.32000
2	2	21.5492	21.5630	21.4407	0.10130	0.31173	0.27303	0.25444
3	3	20.0845	19.7582	21.0459	-0.88118	0.32625	0.30075	0.27878
4	4	22.1860	22.2897	21.9452	0.21899	0.33435	0.31740	0.29378
5	5	20.0832	20.1736	19.8323	0.22822	0.33635	0.31570	0.29193
6	6	22.7853	22.8856	22.1183	0.64010	0.28275	0.21340	0.20200
7	7	17.4493	17.3702	17.5761	-0.10687	0.39028	0.40377	0.37328
8	8	22.4724	22.3927	22.5883	-0.09805	0.37617	0.38338	0.35443
9	9	22.3293	22.1436	22.6111	-0.23217	0.39235	0.41628	0.38529
10	10	20.2379	20.1741	20.5306	-0.27901	0.28639	0.23005	0.21672

Obs	sqrtg1	sqrtg2	sqrtg3	sqrtg4	sqrt MSEsyn	stdGREG	true Value
1	0.25872	0.08947	0.09238	0.16570	0.22502	0.32024	19.0060
2	0.21614	0.04668	0.07001	0.12587	0.20684	0.31315	21.4293
3	0.23431	0.05145	0.07978	0.14203	0.20602	0.39752	20.4941
4	0.24466	0.06539	0.08497	0.14890	0.20997	0.34696	21.7355
5	0.24466	0.05652	0.08497	0.14890	0.21251	0.34480	19.8867
6	0.17498	0.02963	0.04867	0.09648	0.21157	0.19039	22.7789
7	0.29592	0.10104	0.10885	0.20385	0.25552	0.55545	17.3710
8	0.28172	0.08030	0.10334	0.19952	0.21719	0.31592	22.2259
9	0.30308	0.09518	0.11145	0.21802	0.21751	0.53876	22.3004
10	0.18660	0.03518	0.05456	0.10446	0.20713	0.25949	20.5194

For comparison results from similar model estimated by SAS/PROC MIXED:

```
proc mixed data=sample5;
class domain;
model y1=x /s ;
random intercept/s sub=domain g;
run;
```

The Mixed Procedure

Model Information

Data Set	WORK.SAMPLE5
Dependent Variable	y1
Covariance Structure	Variance Components
Subject Effect	domain
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
domain	10	1 2 3 4 5 6 7 8 9 10

Dimensions

Covariance Parameters	2
Columns in X	2
Columns in Z Per Subject	1
Subjects	10
Max Obs Per Subject	41

Number of Observations

Number of Observations Read	194
Number of Observations Used	194
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	745.49671651	
1	3	734.93285108	0.00057128
2	1	734.80759787	0.00004175
3	1	734.79921459	0.00000028
4	1	734.79916017	0.00000000

Convergence criteria met.

Estimated G Matrix

Row	Effect	domain	Col1
1	Intercept	1	0.2777

The SAS System

The Mixed Procedure

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	domain	0.2777
Residual		2.3766

Fit Statistics

-2 Res Log Likelihood	734.8
AIC (smaller is better)	738.8
AICC (smaller is better)	738.9
BIC (smaller is better)	739.4

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	10.1309	0.5203	9	19.47	<.0001
x	0.5028	0.02206	183	22.79	<.0001

Solution for Random Effects

Effect	domain	Estimate	Std Err Pred	DF	t Value	Pr > |t|
Intercept	1	0.4086	0.3496	183	1.17	0.2440
Intercept	2	0.1013	0.3117	183	0.32	0.7456
Intercept	3	-0.8811	0.3262	183	-2.70	0.0076
Intercept	4	0.2190	0.3343	183	0.65	0.5133
Intercept	5	0.2282	0.3363	183	0.68	0.4983
Intercept	6	0.6401	0.2827	183	2.26	0.0248
Intercept	7	-0.1069	0.3903	183	-0.27	0.7845
Intercept	8	-0.09805	0.3762	183	-0.26	0.7947
Intercept	9	-0.2322	0.3923	183	-0.59	0.5548
Intercept	10	-0.2790	0.2864	183	-0.97	0.3312

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
x	1	183	519.41	<.0001