A COMPARATIVE ANALYSIS OF ALTERNATIVE ECONOMETRIC PACKAGES FOR THE UNBALANCED TWO-WAY ERROR COMPONENT MODEL

by Giuseppe Bruno[1]

Notwithstanding it was originally proposed to estimate Error Component Models (ECM) using balanced panels, most applications use unbalanced panels. When unbalanced panels include both time and individual random effects, special computational problems arise. Wansbeek and Kapteyn (1989) analyze algorithms for estimating unbalanced two-way ECM; Baltagi et al. (2002) compares three classes of estimators for the unbalanced two-way error model. Here I show some differences between theoretical findings and empirical applications by investigating how various proposed algorithms are implemented in the most widely used econometric packages and by providing a comparative appraisal of the different methods the various packages used to estimate the unbalanced two-way ECM. An illustration examines the determinants of the bank deposit yield in Italy, comparing the outcomes of six popular econometric packages available for the analysis of panel data: E-Views, LIMDEP, RATS, SAS, STATA, and TSP. The packages give strikingly different numerical results. While the relevant documentation is often elusive as to algorithmic details, my findings suggest one reason for the differences lies in the means used for computing the variance of the two idiosyncratic error component terms. Finally, I examine the small-sample properties of the various algorithms by means of Monte Carlo experiments. According to preliminary results, these algorithms show approximately the same performance in terms of bias and variability.

Keywords: Unbalanced panel data, Variance components, two-way error
JEL classification: C15, C23, C87

Prepared for the 9th International Conference of Computing in Economics and Finance. July 11-13, 2003, Seattle, Washington.

[1] Bank of Italy, Research Department, Rome.

1. *Introduction*

The last twenty-five years have witnessed a steady increase of interest for the analysis of panel data (i.e., data characterized by both a time and a cross-section dimension).

The simplest possible case is one in which the number of time observations is the same for each individual observation unit (*balanced panels*). A partial list of relevant references includes Wallace and Hussain (1969), Nerlove (1971), Swamy and Arora (1972) and Fuller and Battese (1974). These papers propose different methods to estimate Error Component Models (ECM); while the various approaches share the same asymptotic properties, they differ as to the computational details. Indeed, previous empirical work has shown that different econometric packages often deliver different numerical results. As to the sources of those differences, no firm conclusion may be easily reached, partly because essential algorithmic details are often omitted in the package documentation.

A more realistic occurrence is one in which some observations are missing (*unbalanced panels)*. This case is relatively simple to deal with when the model only includes either individual or time effects (the small sample properties of a wide range of different estimation methods for the case of unbalanced panels with one-way error are investigated in Baltagi and Chang (1994)). If the panel is unbalanced and the model includes random two-way effects, then special computational problems arise. A theoretical analysis of the algorithms that have been proposed in the literature to estimate ECM with unbalanced panel data when the error has a two-way structure may be found in Wansbeek and Kapteyn (1989) (WK). Building on their work, Baltagi *et al*. (2001, 2002) have carried out a comparison of three different "families" of alternatives estimators for the unbalanced two-way error model. Davis (2002) extends the WK algorithms to unbalanced data with multi-way error components. As to the packages that are available to deal with unbalanced panel models, some of them simply do not allow estimation of two-way error models with unbalanced panel samples. As to the rest of the available packages, they have been found to provide remarkably different numerical results. Once again, it is hard to ascertain the sources of those differences, because of insufficient algorithmic documentation.

This paper aims at filling the gap between theoretical findings and empirical applications. To this end, we investigate how the algorithms that have been proposed in the literature have been implemented in a few of the most widely used econometric packages. Our

goal is to provide a comparative appraisal of the different estimation methods that the various packages use to deal with unbalanced panel data when the error is modeled with a two-way structure (note that, although some algorithms share the same label, their actual functioning is in fact rather different across packages).

Starting from a purposely modified version of the well known Grunfeld dataset (see Boot and De Wit 1960), we first compute a numerical benchmark for the estimation of an unbalanced panel data model with a two-way error panel. We then investigate the determinants of bank deposit yields. The available data set is very large, in that it includes more than 10,000 observations for the years 1990-2000. The dependent variable is the interest rate on current account deposits while the pre-determined variables of our model are the Herfindahl index, the number of banks in each province, the rate of growth of bank deposits, the ratio of banking costs to total assets.

The core of the paper focuses on the comparison of the numerical outcomes produced, for the model sketchily described above, by six of the most popular econometric packages available for the analysis of panel data (E-Views, LIMDEP, RATS, SAS, STATA, and TSP). Not all of these packages can deal with the case of unbalanced panel data models with a two-way error structure. For those that can, our results show that they deliver strikingly different numerical results. While the package documentation is often elusive as to the actual algorithmic details, our findings suggest that one reason for those differences rests with the method followed to compute the variance of the two idiosyncratic error component terms.

The small sample properties of the random-effects algorithms included in those packages are investigated by means of Monte Carlo experiments. The preliminary results of these simulations, with reference to the bias and the Monte Carlo variability of the different estimation procedures, seem to show an equivalent behavior implying therefore a basic analogy among these procedures.

2. The model and the algorithms proposed in the literature

In this paper we consider the following panel data regression model:

$$y_{i,t} = x'_{i,t}\,\beta + u_{i,t} \qquad t = 1,2,\ldots,T; \qquad i = 1,2,\ldots,N_t; \qquad (2.1$$

where:

$y_{i,t}$ is the observation on the dependent variable for the *i*th individual at the *t*th time period,

$x_{i,t}$ is *i,t*th observation on a vector of *k* nonstochastic regressors

β is a *kx1* vector of regression coefficients including the intercept.

The panel is unbalanced and we observe only N_t individuals in period *t*. The disturbances of (3.1) are assumed to follow the two-way error structure:

$u_{i,t} = \mu_i + \lambda_t + \nu_{i,t}$ where $\mu_i \propto N(0, \sigma_\mu^2), \lambda_t \propto N(0, \sigma_\lambda^2), \nu_{i,t} \propto N(0, \sigma_\nu^2)$

The literature proposes essentially two classes of algorithms:

a) two-step feasible GLS
b) Maximum Likelihood.

All the examined packages adopt the first method mainly because of its computational simplicity.

The first step of the feasible GLS is the estimation of the three variance components $\hat{\sigma}_\nu^2$, $\hat{\sigma}_\mu^2$, $\hat{\sigma}_\lambda^2$. The second step, starting from an unbiased estimate of $\hat{\Omega}$ the variance-covariance matrix of the disturbances, consists in a data transformation using the Cholesky factorization of $\hat{\Omega}^{-1} : y^* = \hat{\Omega}^{-1/2} y$, $x_k^* = \hat{\Omega}^{-1/2} x_k$ and finally an OLS regression on the transformed data. A straightforward spectral decomposition is available for $\hat{\Omega}$ when the panel is balanced but when the panel is unbalanced there is no simple representation for the inverse. This case is explored in detail in Wansbeek and Kapteyn (1989) and Baltagi et ali. (2002). The first paper derives the true projection matrix onto the null-space of the time and individual dummies and uses this matrix for deriving an unbiased estimator for the three variance components.

3. The empirical application

The application presented here examines the determinants of the interest rates paid on bank deposits in 95 Italian provinces between 1990 and 1999. The data on interest rates have been taken from the Central Credit Register and refer to deposits of 10,000 € or more with a sample of banks.

The data set is arranged using the individual code of each bank and the province in which it is active. The unit of observation is a subset of the Cartesian product of the sets of banks and provinces (for example, bank *i* in province *j* gives one unit of observation, i.e. the couple *i,j*). The total number of individuals is obtained by summing the number of banks over the provinces. This number equals 1,298 in our sample. Summing the individuals over the available years we get a total of 10,561 observations after removal of few outliers.

The dependent variable in the regressions is the average rate on the total deposits of

each bank. There has been an extensive debate in the United States on the determinants of interest rates, with two opposing points of view. According to the first, market concentration is still an important factor influencing bank interest rates. According to the second, the link between traditional measures of competition and banking prices disappeared with the deregulation of the banking system.

In this paper we measure concentration using the Herfindahl index for deposits. Different control variables are used in the literature to check for the effect of concentration on interest rates. We use per-capita income in each province, which is an indicator of the degree of economic development, and the growth rate of deposits in each province, which may influence the return offered on deposits. Other control variables include banks' costs, which may influence deposit rates on the assumption that banks consider their total costs when fixing deposit rates. Here we consider two aggregates: the ratio between costs and total assets (COSRAT) and average staff costs per employee (COSPER).

4. A comparison among the packages

All the considered packages allow some kind of panel data estimation with built-in commands; in the following we emphasize the most relevant differences arisen.

4.1 E-Views

We have used the version 4.0 of the package in the windows environment because a UNIX version is not available.

The basic panel data features are essentially the one-way fixed and random effects model. It deals with both balanced and unbalanced panel. When dealing with unbalanced panel data missing values must be explicitly coded to pad time ranges. The one-way random effects models are not available when the panel has over 750 different cross sections. The two-way error specification is not a built-in feature and has to be programmed by the user.

4.2 LIMDEP

We have had access to version 7.0 of the package under the UNIX operating system. This package supplies a comprehensive menu for both one-way and two way error component models. The estimation algorithms deals automatically with balanced and unbalanced panel.

No special preprocessing is required for missing values.

4.3 RATS

We have worked with version 5.01 of RATS under the UNIX operating system. This version of RATS provides both a one-way and two-way error modeling for both balanced and unbalanced datasets. Analogously to E-Views panel data always have to have the same number of entries per individual. Missing values are used to pad time ranges. It takes about half an hour for carrying out the estimation on the whole dataset (10500 observations) when both the individual and time period effects are taken into account in the estimation. It seems like the algorithm spends a lot of time in building the set of dummy variables. When they are built by the user the estimation is much faster.

4.4 SAS

Version 8 of SAS has been used in our experiments. This is the only package that follows the assumptions proposed in WK, therefore the **TSCSREG** proc requires a dataset with at least two observations for each individual. From the original dataset composed by 10561 observations, application of this rule removed only 22 observations.

4.5 Stata

We have been working with Stata verson 8.0. This version of STATA provides one-way fixed and random effects modeling for both balanced and unbalanced datasets. The estimation algorithms take care of the unbalanced nature of the dataset. The two-way error specification is left to the programmer.

4.6 TSP

Version 4.5 of TSP has been used in our experiments. This version of TSP provides just a one-way error modeling for both balanced and unbalanced datasets. Two way error specifications are left to the user. As well as in LIMDEP, SAS and Stata, the estimation algorithms deals automatically with unbalanced panel requiring no preprocessing for missing values.

4.7 Comparison of the Estimation results

In this section we present some numerical results for the estimation of one specification severally for each package.

In following tables we present some results for the three packages that provide built in estimation command for two-way unbalanced panel. Here we introduce an *ad-hoc* benchmark built using the Grunfeld investment dataset and showing both the coefficients and the estimated variance components

Balanced panel	FIRMVAL	CSTOCK	CONSTANT
LIMDEP	.11212	.33135	-66.782
RATS	.11058	.29469	-54.999
SAS	.11076	.31675	-61.279

Two-way R.E. (balanced)	σ_v^2	σ_μ^2	σ_λ^2
LIMDEP	2675.4	7584.2	5081.9
RATS	2644.1	270.53	8294.7
SAS	2675.4	8119.7	112.38

Unbalanced panel	FIRMVAL	CSTOCK	CONSTANT
LIMDEP	.09018	.22535	-17.079
RATS	.11545	.12341	-15.593
SAS	.09136	.22878	-19.842

Two-way R.E. (unbalanced)	σ_v^2	σ_μ^2	σ_λ^2
LIMDEP	2054.0	7292.7	205.26
RATS	2026.6	35.275	7559.4
SAS	2054.0	7111.9	0

Aside from the constant, the coefficient estimates are very close to each other in the balanced case. In the unbalanced case differences are larger. As far as the variance components are concerned only the uncorrelated error shows some similarities whereas the two idiosyncratic terms are radically different from each other.

In our application we tested all the panel estimation methods available in the six

examined packages and the results are summarized in the following tables:

Table n. 1.

One-way R.E.	Herfin	Cosper	Cosrat
E-Views	**NA**	**NA**	**NA**
LIMDEP	**.049427**	**-5.94800**	**.920840**
RATS	**.054999**	**-5.91761**	**.699054**
SAS	**.055264**	**-5.91611**	**.687921**
Stata	**.055784**	**-5.91316**	**.665925**
TSP	**.056172**	**-5.91094**	**.649390**

Table n. 2.

Two-way F.E.	Herfin	Cosper	Cosrat
E-Views	**-.22802**	**-6.4892**	**4.05737**
LIMDEP	**-.22802**	**-6.4892**	**4.05740**
RATS	**-.22802**	**-6.4892**	**4.05737**
SAS	**-.22802**	**-6.4892**	**4.05737**
Stata	**-.22802**	**-6.4892**	**4.05737**
TSP	**-.22802**	**-6.4892**	**4.05737**

Table n. 3.

Two-way R.E.	Herfin	Cosper	Cosrat
E-Views	**NA**	**NA**	**NA**
LIMDEP	**.67739**	**-45.1170**	**42.0240**
RATS	**-.16214**	**-7.62244**	**5.08594**
SAS	**-.00758**	**-6.21705**	**2.41711**
Stata	**NA**	**NA**	**NA**
TSP	**NA**	**NA**	**NA**

As can be seen from the three tables the one-way Random-effects and the two-way fixed effects algorithms produce essentially the same numerical results across the different estimation package. When we go to the two-way Random-effects the situation is completely different. Three of the examined packages do not provide built-in estimation methods for this model. Albeit the other three packages do offer built-in two-way Random-effects algorithms the numerical results show no similarity whatsoever. In order to provide a clear picture for the random-effect algorithm the following table show the values of the variances components provided by the three packages:

Two-way R.E.	σ_v^2	σ_μ^2	σ_λ^2

LIMDEP	**.3069**	**1.1574**	**3.8697**
RATS	**.3068**	**4.8722**	**.21288**
SAS	**.3069**	**0.5244**	**5.2269**

There is only an agreement on the computation of the basic uncorrelated disturbance, the two idiosyncratic terms are substantially different. Of course these variance components are relevant determinants for the regression coefficients.

It is also relevant to mention some algorithmic details of the various packages.

LIMDEP adopts an ANOVA estimator for the variance components, using the two-way fixed-effects residuals for estimating the variance of the remainder disturbance. The other two idiosyncratic disturbances are estimated by means of the between groups and between time periods regressions. After the estimation the three variance components, random-effects estimates of the coefficients are achieved by running OLS on the variables transformed according to following:

$$z^{*}_{i,t} = z^{*}_{i,t} - \vartheta_{1i}\bar{z}_{i.} - \vartheta_{2i}\bar{z}_{.t} + \vartheta_{3i}\bar{z}_{..}.$$

Also RATS uses an ANOVA estimator for the variance components using the fixed-effect residuals. Unfortunately the documentation does not go very deep into details of the computation for the variance components for the two idiosyncratic components and the treatment for the unbalanced case.

SAS adopts the method described by Fuller and Battese (1973) for the variance component estimation. When the panel is unbalanced the estimation method is modified according to Wansbeek and Kapteyn (1989).

The experiments carried out show that a second reason for these substantive differences in the estimates are linked with the presence of the missing observations.

5. Monte Carlo Analysis

We have carried out some Monte Carlo experiments using the packages that provide a two-way Random-effect estimation algorithm.

The estimators' efficiency has been studied by running a parametric bootstrap procedure using the OLS residuals for estimating the three variance components.

The simulation is based on the generation of multiple draws for the dependent variable with a fixed set of explanatory variables.

The parametric bootstrap consists in the following steps It consists in a Monte Carlo exercise made up of seven steps:

1) Obtain the residuals for estimating the three variances

2) run a control random-effect estimation to compute the "true" estimated model; this task is carried out separately for each package and provides a *true value* (a benchmark) for the dependent variable and for the structural coefficients;

3) generate three normal random samples with zero mean and the standard deviations computed in step 1;

4) add the three simulated residuals to the true value of the dependent variable computed in step 3, thus simulating a new draw for the dependent variable;

6) perform a two-way random-effects estimation using this new dataset;

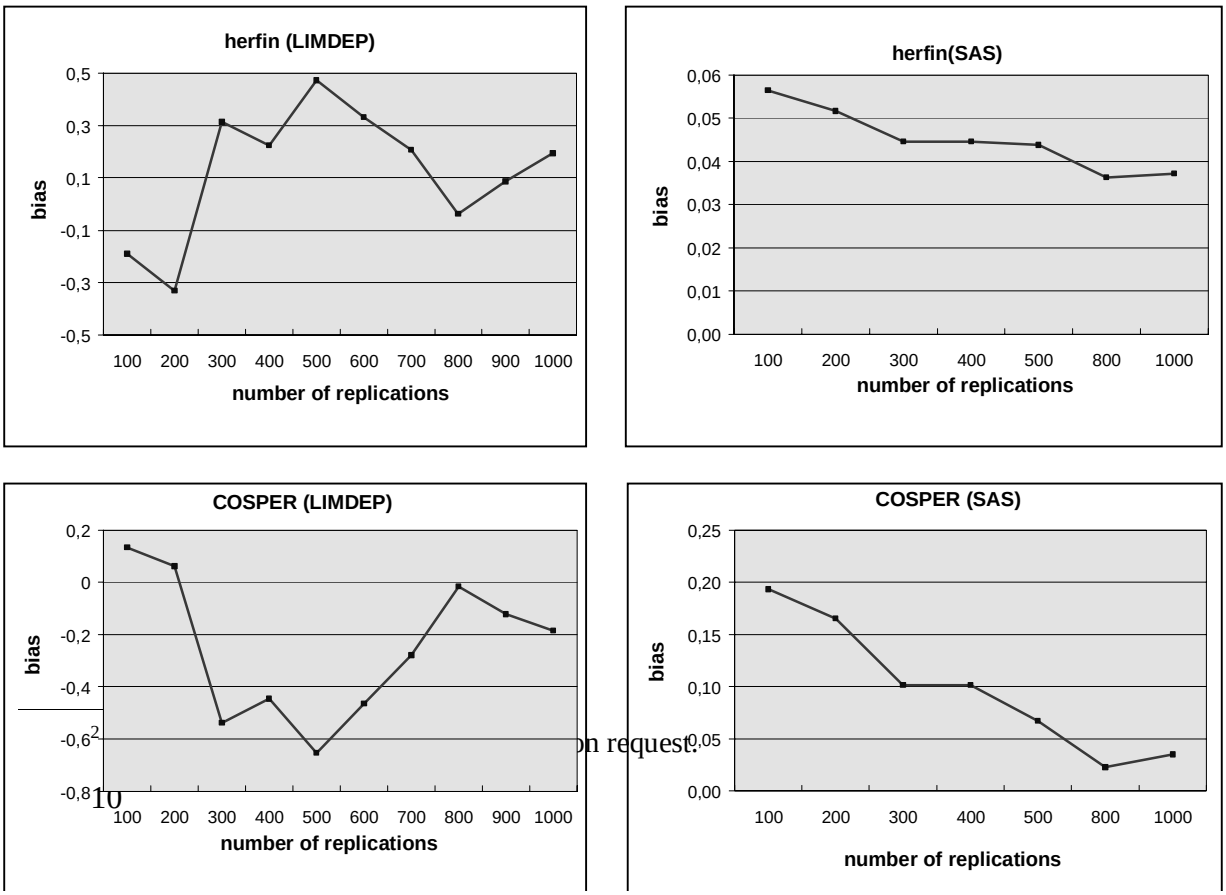
7) compute the mean and the standard deviation of the replications.

From the performances standpoint, LIMDEP is the fastest software. We have already mentioned the fact the RATS takes about half an hour for just one two-way R.E. estimation therefore we gauged the results of only three Monte Carlo experiments.

With the aim of comparing also the programming complexity, the programs for the Monte Carlo experiments have been written using the looping features made available by each package[2].

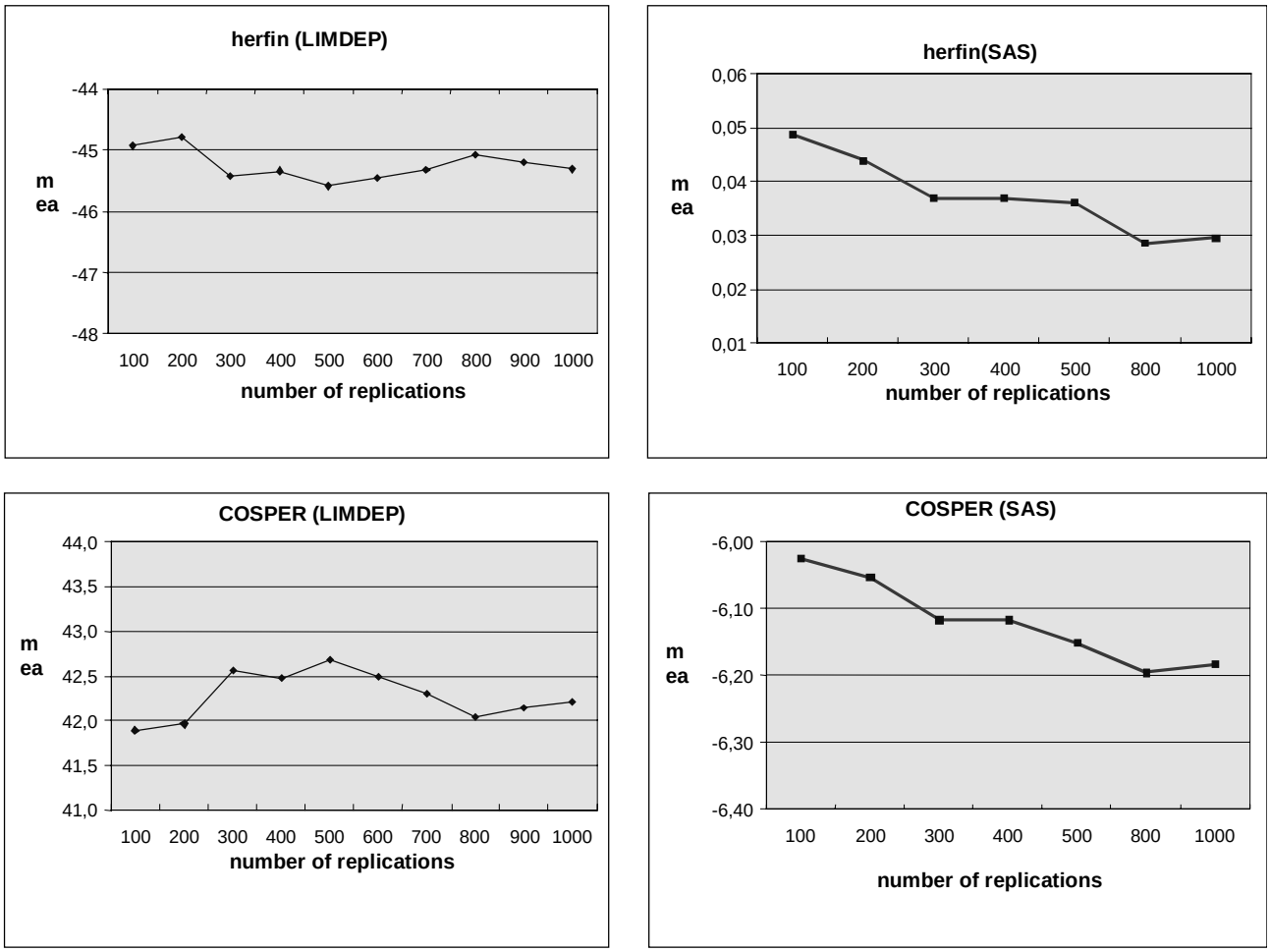
A second and more important insight is the evaluation of the bias and standard deviation for the Monte Carlo replications for each of the packages examined. Figures n.1 show the behavior of the bias and the standard deviation around the Monte Carlo mean computed for the coefficients of the Herfindahl concentration index and the normalized cost of personnel.

Fig. 1



[2] on request.

Fig. 2



Before drawing conclusions, we note some details of the Monte Carlo experiment.

1) The data-generation process is given by the fitted value of the random-effects estimate for each of the three packages.

2) The total panel length was kept fixed over the different replications.

Taking into account these experimental conditions, the observation of the graphs allows the following considerations:

a) The estimated coefficients do remain stable over replications therefore the estimators are robust with respect to the noise injection;

b) The bias seems decreasing and approaching the zero (for RATS we did not collect enough information)

From the programming standpoint this experiment provide us with some insight on the

programming flexibility and the performances of different packages.

6. Concluding remarks

We have shown that when two-way random effects modeling is adopted, some econometric packages fail to provide built-in procedures. Even more worst the packages that do provide this sort of algorithms disagree, in a substantial way, in the estimates of the coefficients. The differences persist even on Monte Carlo experiments. One reason is the way in which the variance components are computed, the other one is related to the difference with which the packages deal with panels heavily unbalanced. Finally is worth remarking that the documentation of some packages often lacks important details required for a thorough understanding of the algorithms. When accuracy is at stake this lack means a costly software reverse engineering.

References

Baltagi, B. H., S.H. Song and B.C. Jung (2001), *The unbalanced nested error component regression model*, Journal of Econometrics, n. 101, pp. 357-381.
Baltagi, B. H., S.H. Song and B.C. Jung (2002), *A comparative study of alternative estimators for the unbalanced two-way error component regression model*, Econometrics Journal, Vol. 5, pp.480-493.
Baltagi, B. H. and Y-J. Chang (1994), *Incomplete panels. A comparative study of alternative estimators for the unbalanced one-way error component regression model*, Journal of Econometrics, n. 62, pp. 67-89.
Boot, J. and G. de Wit (1960), *Investment demand: an empirical contribution to the aggregation problem*, International Economic Review, n. 1, pp. 3-30.
Bruno, G., and R. De Bonis (2003), *A comparative study of alternative econometric packages with an application to Italian deposit interests rates*, Journal of Economic and Social Measurement, forthcoming.
Davis, P. (2002), *Estimating multi-way error components models with unbalanced data structures*, Journal of Econometrics vol. 106 pp. 67-95.
Fuller, W.A., G.E. Battese (1974), *Estimation of Linear Models with Crossed-Error Structure*, Journal of Econometrics vol. 2 pp. 67-78.
Nerlove, M. (1971), *A note on error component models*, Econometrica vol. 39 pp. 383-396.
Wallace, T.D., A. Hussain (1969), The Use of Error Components Models in Combining Cross-Section with Time Series Data, Econometrica vol. 37 pp. 55-72.
Swamy, P.A., A. A. Arora (1972), The exact finite sample properties of the estimators of coefficients in the error components regression models, Econometrica vol. 40 pp. 261-275.
Wansbeek, T., A. Kapteyen (1989), *Estimation of the error-components model with incomplete panels*, Journal of Econometrics vol. 41 pp. 341-361