EURAREA: An overview of the project and its findings

Patrick Heady and Martin Ralphs, Office for National Statistics, UK.

1. A short introduction

One of the purposes of this conference is to present the findings of the EURAREA project, and consider the steps that follow from it. So we need to start by looking at what EURAREA was, what it was attended to achieve, and what was its relation to other applied and theoretical research. In this paper, we would like to consider EURAREA's contribution under three headings:

1. Empirical evaluation of SAE methods
2. Making SAE "NSI-friendly"
3. Creation of an environment for future empirical research

2. Empirical evaluations and their implications

In the research proposal that we submitted to the European Commission we presented small area estimation as a promising methodology which so far had mostly been applied on the other side of the Atlantic. We proposed to investigate:

1. the potential effectiveness of these methods in the context of European official statistics
2. the scope for using recent theoretical innovations (such as methods involving spatial and temporal autocorrelation) to enhance their effectiveness
3. to make recommendations for their application.

Thus, though the project provided some scope for theoretical innovation (some of which has been published in journals as well as in the EURAREA Report, for example Dehnel et al., 2004 and Zhang and Chambers, 2004), its main focus was on the application and evaluation of existing methods, and of methods that were already being developed elsewhere. And, fortunately, our evaluation of these methods has generally confirmed previously positive results.

2.1. A summary of the conclusions from the EURAREA evaluation

The main conclusions from the evaluation are summarised below.

1. Model-based estimation methods substantially outperform design-based methods for very small areas (NUTS4 / 5), and achieve comparable or slightly better levels of precision for medium-sized areas. However, this finding does not always extend to the performance of confidence intervals calculated using model-based methods. Though in some instances they performed well, in others coverage rates were substantially below face value.

2. Model misspecification is a potential source of error. If models are fitted using unit-level covariate data alone, the fixed effect component of the estimators is liable to severe bias as a result of the 'ecological' effect. Additionally, misspecification of the distribution of random terms may underlie some of the problems with confidence intervals.

3. Making use of data from earlier time periods for the area concerned, via either the random or fixed part of the model, substantially enhances the precision of estimates for individual small areas. Interestingly, allowing for the spatial auto-correlation of random area effects was less effective in our simulations. It is possible that greater improvements might be achieved with different spatial autocorrelation structures or distance metrics, but in general we saw a more pronounced gain from incorporating time series data.

4. The enhanced log-linear methodology proved effective in estimating change-since-last-census for cross-classified data, with the use of a generalized linear structural mixed model achieving the best results in most cases. The associated confidence intervals tended to be underestimated for SPREE and GLSM estimators, but were generally too conservative in the case of the GLSMM estimator in our experiments.

5. The standard deviation of the set of estimated area means generated from a single sample tends to either underestimate (in the model-based case) or overestimate (in the design-based case) the standard deviation of the set of actual area means. In principle, model-based estimators can be adjusted to reduce this problem. Such adjustments are not possible with design-based estimators.

6. Effective model-based estimation requires that sample data can be matched to area-level covariates with high explanatory power. If possible, unclustered sample designs are also favourable and increase the success of the estimation models.

2.2. Our results in the European context

Although these results support theoretical expectations and are in that sense unsurprising, they are interesting and new from the point of view of European statistical policy because they show the specific effect of these general findings for the choice of estimators for the kinds of subject matter and spatial unit that are important to European policy makers.

Policy implication 1: Useful estimates for very small areas

A key finding from EURAREA is that useful estimates can be made for *very* small areas (NUTS4/5) using small area estimation techniques and model-based approaches in particular. The typical gain achieved is illustrated below in Figures 1 and 2.

Here, we show Mean Squared Error (MSE) performance for key estimators expressed as a proportion of the MSE that would arise if the National Sample Mean was used as the estimator for each local area. Of course, the National Sample Mean is not actually a sensible small area estimator. But the results do tell us the amount of error that would be incurred by making the false assumption that all areas had the same mean value, which would be the natural default assumption in the absence of any form of small area estimation. They therefore provide a useful benchmark against which to assess the performance of the other techniques.

In Figure 1, we see that at NUTS3 level all of the estimators perform substantially better than the national sample mean, but that model-based estimators are usually (except in the case of income) as good or better than their design-based counterparts.

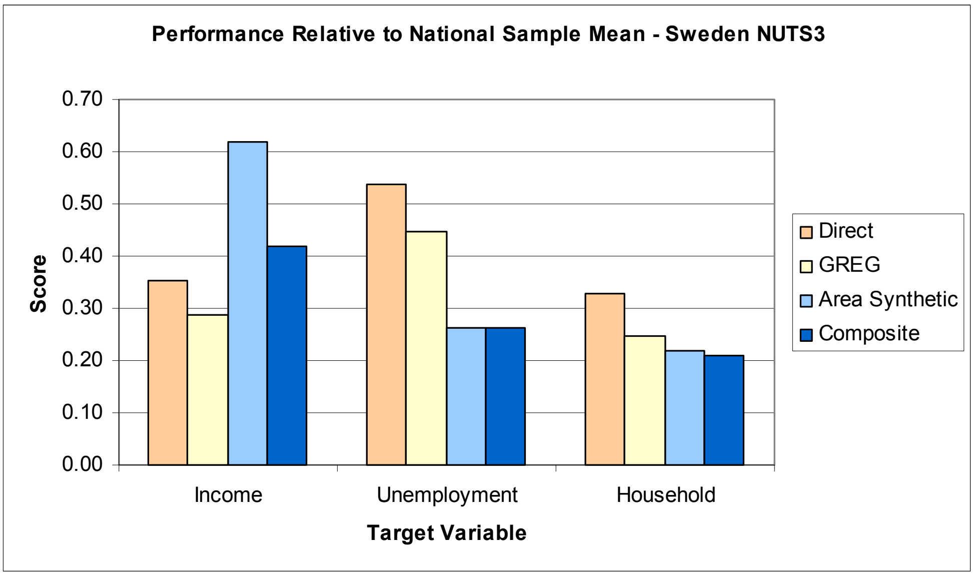

Figure 1 –MSE performance relative to the MSE of the National Sample Mean for three target variables in Sweden at NUTS3 (NSM = 1.0).

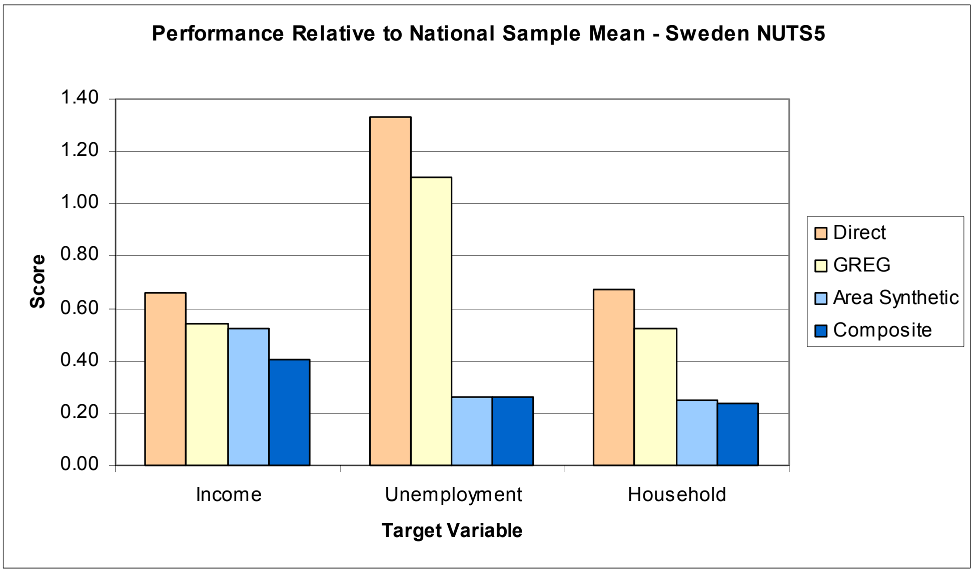

Figure 2 –MSE performance relative to the MSE of the National Sample Mean for three target variables in Sweden at NUTS5 (NSM = 1.0).

In Figure 2, the results are much more clear cut. Direct and GREG estimators actually perform worse than the national sample mean in the case of ILO Unemployment and are always less successful than their model-based counterparts. The composite estimator is usually the best performer.

Policy implication 2: Estimating the distribution of area values: problems of over-shrinkage

The performance of estimators for particular areas is important when resource allocation occurs on an area-specific basis, but other policy applications require estimates that robustly reflect the distribution of area values across the country. This is important if a government wishes to assess the extent of geographic inequality or if applications for funding by some higher-level institution (such as the European Community) are dependent on the number of areas in a country which fall below some specified threshold. From this point of view, a reasonably good set of estimates might be one for which the empirical standard deviation of the true area values was close to the empirical standard deviation of the estimated area values.

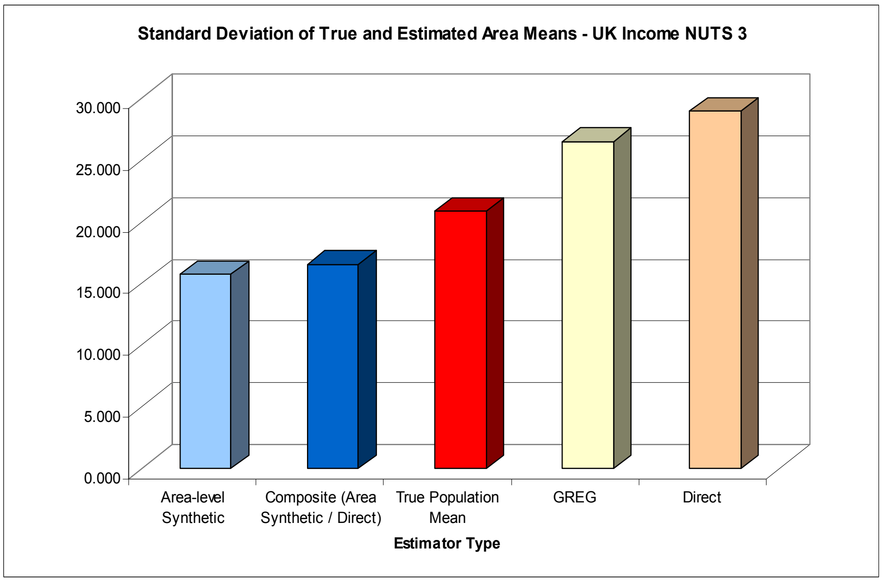

Figure 3 – Comparing the true standard deviation of area means with that produced by different estimation strategies for Income at NUTS3 in Northwest England and North Wales.

In Figure 3, we compare the true standard deviation of area means for NUTS3 areas in the United Kingdom with the standard deviations of estimates of these means produced using the Direct, GREG, area synthetic and composite methods described above. The direct estimator tends to overestimate extremes in the distribution, and as a result the standard deviation of area values is over-inflated. The area level synthetic estimator has the opposite effect, and tends to "shrink" the estimates towards the centre of the distribution. The result is understatement of extreme values, often referred to as “over-shrinkage” in this context, which is equally problematic when our goal is the description of the overall distribution.

There are a number of proposed methods for dealing with over-shrinkage (for example see Spjøtvoll and Thomsen, 1987, Rao, 2003 and Zhang, 2004) and this is an area where further empirical work, perhaps using EURAREA datasets, could be valuable.

Policy implication 3: EURAREA findings are consistent

It is important to emphasise that the specific conclusions from the evaluation programme are very much the same for all the European countries in EURAREA despite widely different socio-economic systems and statistical infrastructure.

2.3. Towards the practical implementation of SAE

The findings of the project also point to the remaining work that needs to be done to make SAE operational in European national and EU contexts:

- In all countries, the current design of major national surveys was adequate to support SAE methods that were close in effectiveness to the theoretical optimum;
- The main adaptations that were needed were availability (at least within the NSI) of precisely geo-coded survey data;
- The improved availability of powerful covariates would substantially increase the predictive power of SAE techniques;
- The practical evaluation of alternative approaches to dealing with over-shrinkage was an important area for applied research, particularly in the context of resource allocation within the EU.

Although the methods considered in EURAREA are certainly not exhaustive, the results that have emerged are sufficient to show that, given the political and administrative will to implement them, small area estimation techniques already have the capability to play a major role in resource allocation problems.

3. Making SAE "NSI-friendly"

Important as these findings are, there was more to EURAREA than that.
Its wider significance is related to a paradox: the fact that, although European researchers have been prominent in the development and application of SAE and related methods - names such as Särndal, Holt, Goldstein, Pfeffermann and Kordos spring to mind - European statistical offices have been much slower to adopt these methods, and when they have done so, have often applied them in a rather hesitant and marginal way. This is particularly striking when one reflects that most of the key papers on which SAE applications are based are by now anything from 10 to 25 years old. One has to ask whether the statistical offices have simply been waiting for a thorough evaluative study, or whether there are deeper obstacles to the adoption of SAE methods.

We would like to suggest that there are deeper obstacles, and that a second major contribution of the EURAREA project may be the extent to which it helps staff in NSIs to overcome these obstacles. These obstacles can be summed up as follows:

1. The methods are felt to be intellectually inaccessible. The statistical theory that underpins them is quite complex, and the practitioner must grapple with an additional layer of theory to do with computational algorithms in order to implement them efficiently. This becomes increasingly critical as the volume of data increases. The situation is further complicated because the way in which the theory is presented and published means that it is mainly available at researcher rather than practitioner level.

2. The methods are felt to be practically inaccessible, because software requirements, particularly in the case of more advanced models, do not usually fit with extant NSI statistical software systems (in particular the facilities offered by modules such as SAS Proc MIXED or SPSS are rather limited).

NSIs could of course adopt “black-box” solutions: buying in a package that enabled one to specify estimators without fully mastering the underlying theory or the way in which it is implemented. In some ways this makes pragmatic sense, but there is a fundamental problem. NSIs are supposed to be authoritative organisations, taking responsibility for the figures they produce – and this role is hard to reconcile with a “black box” approach.

The EURAREA team started to tackle this problem when we decided to program all our estimators ourselves. It was given in the contract that we would have to write some programs - for those estimators that were not yet implemented in standard packages. However, in the event we resolved, without a great deal of discussion, to program all our estimators ourselves. We believe that it was the wish to fully understand all aspects of the methods that was responsible for this collective decision. The result certainly proved educational for us: there is no better way of testing your understanding of a piece of theory than trying to write an implementation program that actually works!

Of course, if EURAREA is to have a lasting impact on NSI understandings and attitudes, the value of this education must be extended beyond the members of the EURAREA team itself. We have tried to provide for this in two ways: firstly by writing the programs in open code, so that colleagues can play around with them, and so partly replicate our own learning experiences. Secondly, we have tried to structure the EURAREA report in a way that will make the connection between theoretical and implementation issues transparent to readers: whenever possible linking texts on objectives, theory, implementation and actual effectiveness closely together.

Before moving on, we hasten to say that the programming work done by the different EURAREA teams was far from being purely educational. Table 1 lists the set of program tools that were developed by the project team, together with the estimators they implement and the groups responsible for developing them. These programs both extend the range of estimators that can be implemented via SAS and greatly improve on the efficiency and speed of some existing programs.

Program	Implements	Authors
Standard Estimators (SAS v8)	Direct, GREG, Unit-level synthetic, area-level synthetic, composite estimators	SNTL Consulting Office for National Statistics UK
EBLUP_TS (SAS v8)	Composite estimator with area-level time effect	University of Southampton UK Office for National Statistics
EBLUPGREG (SAS v8 / SAS v9)	Unit-level composite estimator with time or spatial effects GREG estimator Synthetic estimator	Statistics Finland University of Jyväskylä University of Southampton
EBLUP_SPACE (SAS v8)	Unit-level composite estimator with spatial effects	ISTAT, Italy University Roma III University of Southampton
FISHERSCORMIX FISHERSCORMIX2 (SAS v8 / C++)	Synthetic estimator with sample weights	INE, Spain University of Miguel Hernandez, Spain
SPREE / GLSM / GLSMM (SAS v8)	Cross-classification estimators for two and three way tables	ISTAT, Italy Statistics Norway

Table 1 – EURAREA programs and functionality

4. Creation of an environment for future empirical research

An equally important part of EURAREA was the research environment that made our evaluation study possible. The considerable investment in datasets and programming that we undertook may make it easier to pursue related research in future.

The simulation strategy that we chose required a considerable investment in database construction and in the construction of programs to run the simulations and implement our chosen performance criteria. The basic simulation setup is illustrated in Figure 4, while the datasets developed for the project are listed in Table 2.

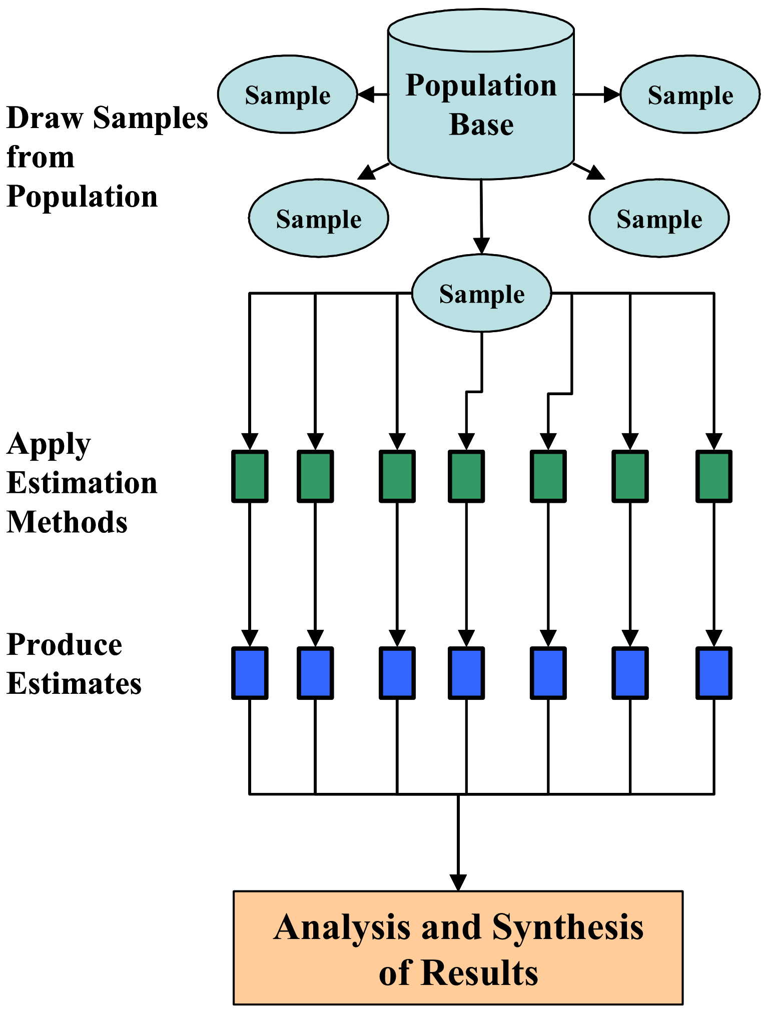

Figure 4 - The EURAREA simulation process. Repeated samples are drawn from a population base and a range of estimation methods are applied to each sample.

Country	Simulation Universe	Total Population	Total Households	NUTS 3 Areas	NUTS 4/5 Areas
Finland	100% of Finland	4.12 million (over 16 years old)	2.25 million	20	85
Italy	25% of Italy			18	151
Poland	5% of population	2.06 million	0.5 million	44	373
Spain	5 Autonomous Communities (Regions)	15.4 million (over 16 years old)	5.92 million	18	215
Sweden	100% of Sweden	5.5 million (16-64 years old)	3.47 million	24	289
United Kingdom	25% of England and Wales	13.9 million	4.9 million	13	2275

Table 2 - Summary information about project databases

One of the main findings of EURAREA is that it is technically feasible to simulate statistical procedures and explore their performance on large population databases. This model of experimental design can be taken forward and applied outside the EURAREA project, since the databases and simulation programs remain in existence and can be used to evaluate other statistical techniques. The experience of ONS in making wider use of its EURAREA data resources can serve as an example. Since the completion of the EURAREA research programme, the databases configured for the project have been used for a range of different application, and others are planned for the future:

a) Evaluating further small area estimators. The small area estimation project team at ONS is evaluating a range of small area estimation methods using the EURAREA datasets, focussing in particular on optimal model selection and fitting, deriving consistent estimates for different geographical levels and estimation of change over time.

b) Evaluating area-construction algorithms. ONS has adopted optimisation procedures in the construction of reporting geographies (known as "Super Output Areas") for the 2001 census in order to produce area units of uniform population size and homogeneity in terms of key properties such as tenure mix. We have used the EURAREA population bases to compare the properties of these new, optimised geographical units with existing administrative hierarchies. We also plan to use EURAREA data to evaluate maintenance requirements for the new geography across the inter-censal period.

c) Testing data-recasting methodology and associated confidence intervals. Changing geographical boundaries and consequent problems for temporal comparison between small areas are a particular problem in Britain. ONS has been developing methods to move data between overlapping boundary systems. Again, EURAREA datasets have been used to provide a basis for empirical comparison between different data recasting methods.

The experience of working with these data and simulation systems also contributes to the research environment, by bringing certain issues into a clearer focus. Two examples will make the point.

1. Under the hierarchical modelling approach which EURAREA shares with most work on model-based SAE, the areas appear as distinct units with no internal spatial differentiation, and linked at most by spatial auto-correlation of expected area values. Once you start working with databases with individual address data, and use the same databases to construct artificial boundaries within what would otherwise be continuous urban sprawl, it quickly becomes apparent that the usual hierarchical SAE set-up is by no means the only way of posing the estimation problem.

2. A problem that members of the EURAREA team discussed amongst ourselves was the relation between our simulation exercise, based as it was on repeated sample selections from a set of given populations, and the theoretical under-pinning of model-based estimation, based (at least in its non-Bayesian versions) on the notion that the model describes the random processes underlying the generation of the observed populations. Some of us felt that this meant that repeated simulations on given populations were not fair tests of the performance of model-based estimators – while others of us felt that contact with the richness of real data, and some acquaintance with the actual processes of community development and boundary construction, exposed the

models as merely analytically convenient fictions. We do not want to take sides here! Our point is simply that the process of constructing a shared simulation methodology brought different viewpoints into focus and made possible a meaningful debate in which both theoreticians and practitioners could join.

The final demonstration that EURAREA has succeeded in creating an environment for continuing research in spatial estimation is the fact that many members of the EURAREA team are presenting papers at this conference, based on work that they have continued to do after the formal end of the EURAREA project itself. The ultimate test of the project's value is that it has helped put more European researchers, particularly researchers linked to NSIs, into a position to contribute to, and learn from, wider developments in the field of spatial estimation and modelling.

5. Acknowledgements and Notes

We would like to thank the European Union for funding the EURAREA project, and all of our project partners for working with us in such a constructive manner. We would also like to thank Eurostat for providing support and organising constructive project reviews.

The estimators referred to in this paper are defined fully in the EURAREA Project Reference Volume. This, and all of the SAS programs developed in EURAREA, are available for download from http://www.statistics.gov.uk/eurarea.

6. References

Dehnel, G., Golata, E. and Klimanek, T., 2004, Consideration on optimal sample design for small area estimation, *Statistics in Transition*, Vol. 6, No. 5, p. 725 – 754.

Rao, J.N.K., 2003, *Small area estimation*, Wiley, New York.

Spjøtvoll E. and Thomsen, I. 1987, Application of some Empirical Bayes methods to small area statistics, *Bulletin of the International Statistical Institute*, Vol. 4, p. 435 – 450.

Zhang, L., 2004, Simultaneous estimation under nested error regression model, *Statistics in Transition*, Vol. 6, No. 5, p. 655 – 666.

Zhang, L. and Chambers, R.L., 2004, Small area estimates for cross-classifications, *Journal of the Royal Statistical Society, Series B*, Vol. 66, Issue 2, p. 479 – 496.