Second draft – Tuesday 15th May

Royal Statistical Society, Discussion Meeting, 9th May 2018

'From start to finish: A framework for the production of small area official statistics'

Discussion on the paper by Tzavidis, Zhang, Luna, Schmid and Rojas-Perilla

Ian R. Gordon (London School of Economics) – I.R.Gordon@lse.ac.uk

I very much welcome this ambitious and timely paper, and appreciate the opportunity to introduce discussion of some of the issues it raises. It sets itself a commendably broad task, in terms both of the holism of its first-to-last perspective and the generality of domains and situations which its guidance should frame. Practically it moves us towards these goals by fulfilling two very different kinds of function.

First and foremost, it documents an exemplary, rigorous and *experimental* exercise in generating small area income estimates in a situation where direct estimates (of varying precision) are available for about half the cases and lacking for the other half, and where an inequality measure (the Gini coefficient) is required as well as one of average conditions. But crucially, it also opens up key questions about credible best practice approaches in a burgeoning area of activity, emphasising the importance of the user/statistician interface at stages from need recognition to actual use of the constructed estimates.

There are evident tensions between these aspects – with more definitive achievement on the former; but a lot of grist for productive debate on the latter – notably about how users/ statisticians can effectively collaborate – to be found between the lines of the paper as well as in its direct treatment of the issue. A tight experimental design (including withholding of the best predictor variable from the modelling stage for use in the subsequent evaluation) reflects a research context in which users were scarcely involved. But it encourages conjecture (for this discussant) as to some positive differences that user engagement might have made, both to this specific case and to the design of a framework.

Those I would like to raise reflect my own professional position, as an artisanal statistician/ semi-pro user, and disciplinary affiliation, as a geographer. These lead me to a couple of substantial reservations about the purely technical framework guiding the case exercise. The first relates to the very cursory treatment of the choice/definition of independent variables deployed to generate the modelled estimates. As a specific example, I am struck by the fact that despite repeated emphasis on the nonlinear basis of the Gini coefficient as a source of problems, any real difficulties with this lie in a need to ensure that predictor variables (or transformations of these) can effectively target the two tails of the income distribution, which Figure 5's plots of residuals show not to have been achieved at either end. Any useful practice guidance manual on small area estimation needs to point up the importance of addressing such substantive specification issues for particular situations, as well as offering more generalised ('text book' type) suggestions for suitably robust estimation and evaluation techniques.

Second draft – Tuesday 15th May

The other issue of substance I would raise about the ‘small area’ estimation (and evaluation) process reported here involves an almost complete lack of attention to geography – give or take some maps suggesting a need for it to be attended to. How far geography actually matters in any case, and how effectively it may already have been picked up by other independent variables, cannot be prejudged. But, in principle, it is clear that local income levels (or other statistics) are liable to be substantially affected by: location, spatial dependence, settlement/agglomeration size and density, and ecological influences from the mix of population/economic activities – as for example the fact that statisticians living in upper middle class districts of London are likely to have substantially higher incomes than those in working class suburbs ! All are passed over in the paper, including the very strong likelihood that municipalities lacking direct estimates because there were no locals in a survey sample will differ significantly in ways (e.g. settlement size) requiring attention to sample selection bias in the small area estimation procedure. The substantial (positive) bias in mean income estimates for out of sample municipalities (*averaging* 11-12% according to Table 4) suggests that this is a real factor in this case – presumably because places with fewer residents are liable both to be uncovered and to have lower incomes (cet. par.).

In spelling out this pair of reservations, my intention is not to devalue the real strengths and technical sophistication of this paper, nor the impetus it gives toward developing a practice-related framework (or good practice manual) for both users and suppliers of small area statistics. Indeed, I think the current burgeoning of work, encouraged by Big Data sources, GIS technology and better geo-data referencing – with the potential for Gresham’s Law effects as the range of producers and outlets extends – rather urgently requires an initiative of that kind. But the kind of generality it aims for should (in my judgement) prioritise adequacy to context and substance, rather than specifying universally applicable techniques.

In working towards this, we might usefully debate a couple of general issues which the paper raises (for this reader). One is about what the real value is of the procedural *‘parsimony’* that the paper advocates. For me it lies very largely in the kinds of transparency that both statisticians and users require in order to generate and adopt suitable small area measures/indicators – which deserves to be addressed more directly. The other is about the helpfulness or otherwise of a current assumption (by international organisations as well as in this paper) that area estimates should have *micro-data foundations*. For me this seems rather a distraction in a task where causality is not an obvious issue – unless there are some reasons to expect that there might be important interaction effects (e.g. perhaps for predicting numbers in the tails of the income distribution?) – particularly since ecological effects are entirely apposite in this context.

For this timely stimulus, as well as its professionalism, I am very pleased to move the vote of thanks. (895 words).