Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2003: Challenges in Survey Taking for the Next Decade

2003

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2003
Challenges in Survey Taking for the Next Decade

FUTURE CHALLENGES FOR METHODOLOGY FOR OFFICIAL STATISTICS

D Holt[1]

ABSTRACT

Methodology plays an important role in assuring the quality of individual statistical outputs. It also plays a wider strategic role within the National Statistical Office (NSO) by promoting the professional reputation of the organisation, by contributing to the development of international standards and by promoting public confidence in the NSO and its outputs. Hence when considering future directions for methodological research and development it is vital that the strategic needs of the NSO itself are taken into account. This paper attempts to illustrate the link between the strategic needs of the NSO and the methodological needs that this generates.

KEYWORDS: Business Strategy; Methodological Priorities; Model-Based Estimation; Performance Monitoring; Small-Area Estimation.

1. INTRODUCTION

The question of identifying the future challenges for methodological development for Official Statistics is not easy. Short term methodological developments often flow naturally as continuations of existing development programmes but the same is not always true of longer term developments. By the time new statistical needs are properly articulated it is often rather late for the timely development of supporting methodology. Furthermore any attempt to identify future methodological needs can appear to be a shopping list with relatively little structure or sense of overall direction. This paper will not be a comprehensive review of future lines of development. It will instead try to set out a process that attempts to link the strategic needs of the National Statistical Office with the methodological development programme.

What is Methodology?

The mission of any National Statistical Office (NSO) is to produce statistics that are relevant to the needs of the user community, are of high – and defensible- quality (in respect of the main uses to which the statistical results are put) and command public confidence in the outputs. This latter requirement is fundamental since if statistics do not command public confidence this undermines their uses for:

- developing and monitoring public policy,
- managing public services,
- supporting government and business people in decision making,
- supporting other uses such as for educational purposes and
- providing the public with information to monitor government itself as a cornerstone of the democratic process.

To many the core of methodological development to the present has centred on sampling theory (including properties of sampling frames), statistical estimation, survey methodology and statistical methods related to data capture and processing (such as editing and imputation and outlier detection). More recently the role of statistical

[1] D. Holt, University of Southampton. The author was formerly Director of the Office for National Statistics (ONS) and Head of the Government Statistical Service for the United Kingdom.

modelling has played a greater role in addressing issues such as small-area estimation and the impact of measurement error on estimates. Of course, statistical models (either explicit or implicit) have a much longer provenance in Official Statistics and topics such as adjusting for non-response, time-series analysis and seasonal adjustment, all of which can be viewed as having an implicit or explicit model-based foundation, have been widely used for a long period.

In reality a much wider definition of methodology is needed. The whole framework of National Accounts, the development of satellite accounts for specific areas (e.g environment, labour or health accounts) and topics such as input-output analysis to support the National Accounts balancing process are areas for methodological development.

More generally the framework for producing Official Statistics rests on identifying the concepts that we seek to measure and from these developing the definitions and thence statistical measures that will support each statistical output. Developing and establishing this framework has a methodological component even though many aspects will engage others within the NSO and external users. Hence the framework of definitions, classifications and best practices that together make up international standards are part of my broad definition of methodology.

As methodologists we strive to establish principles and practices, algorithms and general theory that can be applied to a variety of situations and hence provide a general framework within which systems for Official Statistics are established. Survey design, post-stratification, generalised regression estimation systems (GREG), automatic edit and imputation and seasonal adjustment are all examples of generalised methodologies. But equally important is the professional expertise that ensures that these generalised approached are applied sensitively and sensibly to each specific application and that this is then properly evaluated.

The impact of Methodology

Strong methodology plays a vital role in ensuring that the NSO fulfils all of the elements of its fundamental mission: relevance, quality and public confidence. It does so, not simply by underpinning each individual statistical output, but more generally.

The basic role of good methodology as the cornerstone of each individual statistical output is well established. At this level the users rely on it: the need for high quality, defensible methods and known and published margins of error so that users may interpret the results is well understood and will not be developed further here.

But the effect of good methodology goes beyond its impact on a particular output and supports the NSO's mission both in terms of the way that the organisation functions and on public confidence in the organisation and its outputs more generally. Strong methodology supports three general aspects:

- it promotes an NSO-wide culture of systematic evaluation and review that is central to the maintenance of quality standards,
- this wider culture is one of the ways in which users build confidence in the professionalism of the NSO and as a consequence in its individual outputs, and
- it supports the research and development needed to ensure that new statistics are developed, or established statistics reviewed and renewed to ensure that the outputs of the NSO continue to be relevant to users' future needs.

Given the central role of methodology to the mission of the NSO it follows that one cannot consider the question of future challenges for methodology except in the context of the future strategic needs of the NSO itself. As such the priorities may vary in different countries although, given the shared global issues that NSO's face, there will be a strong element of commonality. Specifically the strategic directions in which methodology should develop cannot be generated solely by an internal assessment of future directions within the methodological group itself.

2. THE STRATEGIC CONTEXT

In this section a brief picture of some of the strategic pressures facing NSO's is painted. These have a strong influence on the future challenges for methodological development.

Maintaining the Relevance and Quality of Current Outputs

Perhaps the most immediate set of challenges that NSO's face is to maintain the relevance and quality of existing outputs. As the economic and social conditions in society change there is a continuing need to review and revise statistical outputs. In the main this will require updating the underpinning methodologies for the existing outputs but can also require the development of new statistical measures to capture emerging aspects of society. The areas in which change has occurred and continues to occur are well known. For example:

- Economic globalisation: the increasing economic dependence of countries and economies on each other; the organisation and delivery of economic activities crossing national boundaries continues to represent a greater share of world economic activity. It is not unusual for companies to be based in one country, having design capability in another, to be raising capital for investment in a third and establishing manufacturing or service units in others. The basic factors of production including labour may be spread across the globe. This is not new but the steady increase in such activities places additional strain on Official Statisticians to measure reliably the activities within National Accounts.
- In large part, globalisation is fuelled by the impact of new information and communication technologies (ICT) and the way that these have allowed business activities to be restructured to take advantage of economic benefits. The underpinning feature is the growth and extensive penetration of ICT into almost every aspect of the economy and of society more generally. The access to information and its active use is affecting every aspect of industry and commerce, public services such as Health and Education and Government more generally. It is also changing the way in which citizens live their lives, the ways in which they access and use information for consumption, for interacting with the economy and public services, for personal development and for leisure purposes. Thus there are changes affecting the supply and demand sides of the economy and changes that go beyond the production boundary into every aspect of society more generally. Very many countries have policies to promote the development of and public access to ICT (the so-called Information Society) in order to release the potential for future economic growth and competitiveness. Official Statisticians must track the changes taking place in society and monitor the effectiveness of public policies designed to support these.

The Geographical Dimension

There are increasing pressures to satisfy user requirements at the supra-national and sub-national levels. The first is mainly driven by the need for internationally comparable statistics both for national and international purposes. They are used by national governments to monitor the country's performance against comparators; to ensure that economic competitiveness is maintained or enhanced; to monitor economic and social developments in other countries and the outcome of alternative economic or social policies that other states may adopt.

Increasingly in some regions they are used for national participation in international decision-making and resource allocation. For these purposes internationally comparable statistics are essential.

They are required too by international agencies to monitor national performance and to make comparisons. The World Bank, IMF and bilateral funding agencies depend heavily on Official Statistics to monitor the impact of policies and technical assistance programmes. For example a review of UN Summits and major conferences during the 1990's identified over 280 statistical indicators needed to monitor UN policies made through conference decisions (UNSC, 2002).

Whilst one may think of economic statistics as the area in which the need for internationally comparable statistics is strongest there are clear regional and wider international interests in areas of the environment, health, education and culture.

At the same time there is increasing demands for sub-national statistics for various policy purposes. There is an increasing tendency in a number of countries towards devolution of some policy and decision making responsibilities from federal to some form of regional government. Below this there is a growing demand for statistics at much smaller areas (so called neighbourhood statistics) to monitor the impact of social and economic policies on small communities. This is often linked to concerns about social disadvantage and multiple deprivation.

Related, but additional to these developments there is growing interest in possible alternatives to the traditional Census either through the use of administrative registers as pioneered in Nordic countries or by use of large scale continuous surveys (the so called rolling Census approach)

Administrative Sources

Sub-national statistics increase the pressure to use administrative sources (perhaps in conjunction with survey sources) to improve estimates. They also re-emphasise the existing work on small area estimation. As data sources are disaggregated further there are questions about confidentiality and disclosure. The pressure to use administrative sources more fully is also driven by the extensive, more accessible data bases that ICT supports and the potential to reduce the burden on respondents by utilising existing information. One may expect that as the use of ICT transforms the administrative systems used by governments to manage large scale public services such as tax, social protection, health and education, so the opportunities and requirements to use these sources for statistical purposes will increase.

Neither the upward pressure for internationally comparable statistics nor the downward pressure for sub-national statistics is new but the pressure is sustained and reflects the fact that there is still much for NSO's to do to meet the needs of users.

Performance Measurement

There is an increasing demand in a number of countries for statistics to monitor the performance of public services at the level of identifiable delivery units. In part this is driven by a desire to improve public accountability for the performance of public services and in part by a growing emphasis on users of public services as consumers who need information to allow them to access public services effectively. Hence there is a growing trend to monitor and report on such things as school educational attainment rates, hospital performance measures, police district clear-up rates for crimes, even the patient survival rates for individual heart surgeons.

Underpinning all of these trends and developments there is the requirement that the NSO must produce a relevant, coherent, authoritative picture of the economic and social situation and so justify continued public confidence in the outputs.

In the remainder of this paper an attempt is made to link some of the strategic needs listed above with the implications for future methodological developments. Inevitably any implications will be partial and there will be other developmental needs that could be identified.

3. SAMPLING THEORY FRAMEWORK

The basic sampling theory on which most of Official Statistics rests is familiar and is derived from Neyman (1934). We assume a fixed finite population of size N with unit values $Y_i, i = 1,..., N$. From this is selected a sample, *s,* with values $Y_i, i \in s$ based upon a random sampling scheme p_s. Inference is based on the randomisation distribution generated by the sample design.

The great advantage of this approach is that Official Statisticians can defend the estimates as unbiased because they are assumption free. Assuming a complete frame and complete response, then no matter what the choice of, for example, stratified, multi-stage design (including the choice of stratification and auxiliary variables, stratum boundaries and sample allocation) the randomisation distribution will ensure that the design consistent estimator will be unbiased (or effectively so if we are concerned with ratios or regression estimators and reasonably large

samples). If the design is good, and the choice of stratification variables, cluster sizes, auxiliary variables and sample allocations is sensible the estimates will be essentially design unbiased and of reasonable efficiency. If the choices are poor, the estimates will still be design unbiased but with poorer efficiency.

This is a public confidence justification and there is much to be said for it. We note that the basic defence for the statistical output is not that Official Statisticians are highly competent and professional but that we use methods that specifically exclude professional judgements that could result in bias for the resulting estimates. Wherever possible we have clung to this in a very determined way: for example creating a general framework for essentially design consistent estimation including GREG estimates. This approach has served Official Statisticians well, has produced defensible statistical outputs and hence promoted public confidence in the NSO and its work.

Of course at the margin, it has always been something of an illusion: methods for adjusting for non-response, imputing missing values and specific procedures such as the seasonal adjustment of time series have always had either an explicit or implicit set of underlying assumptions for justification.

However it has been clear for some time that there are some circumstances (e.g. small area estimation) in which the price for design unbiasedness is too high and explicitly model based estimators are growing in use.

I will suggest in due course that there are situations where model based estimates are not just unavoidable but integral to the objective and that this raises a number of explicit issues for methodologists in NSO's. But first I comment on one or two areas within small area estimation where the level of methodological development still appears to fall short of the strategic needs of the NSO.

4. SMALL AREA ESTIMATION

There is a long history of methodological work on this topic resulting in a gradual transition from so called synthetic estimation (ad hoc methods within the general framework of ratio and regression estimation) into model based inference that seeks to reflect the essential characteristics of the population structure and to provide appropriate estimates of parameters, measures of uncertainty etc. There is now an established body of results that draw on general statistical theory and methods for mixed models (for example using Empirical Best Linear Unbiased Prediction, Hierarchical Bayes and other methods). A number of models have been given particular prominence:

Fay-Herriot: $\theta_i = g\left(\overline{Y}_i\right) = z_i^T \beta + v_i, \qquad i = 1,\ldots, m$

Random intercept: $y_{ij} = x_{ij}^T \beta + v_i + e_{ij}$

Two Level:

$y_{ij} = x_{ij}^T \beta_i + e_{ij} \qquad j = 1,\ldots, N_i \qquad i = 1,\ldots, m$

$\beta_i = z_i \ \gamma + v_i$

The Fay-Herriot model (Fay and Herriot, 1979) models the small area means (or functions of them) in a direct way using area level covariates z_i to account for some of the between area variation.

The random intercept model uses unit-level auxiliary variables x_{ij} and allows for further between area variation by allowing the intercept term in the regression to be a random co-efficient. After model estimation the small area means are essentially derived by aggregating the predicted values within each area and adding in the predicted value of the random intercept for the particular small area.

The two-level model takes these ideas one stage further by using unit level auxiliary variables x_{ij} and allowing any or all of the regression parameters (not just the intercept) to be random and allowing some of the between area

variation to be explained by area-level auxiliary variables $z_{i.}$ The random terms at level 2 (the area) may be correlated.

These models may be made more general, the precise error structures need to be specified and can include heteroscedastic variance and covariance structures where appropriate. Moreover the models can be generalized (for example to model discrete binary variables). Rao (2003) gives an excellent account of the models and related estimation approaches.

The attempt to link this approach with the strategic needs of the NSO suggests a number of issues that will require further methodological development:

- The use of a model based approach means that the historic defense for the resulting estimates: that they are essentially assumption free and therefore unbiased with respect to the randomization distribution no longer holds. The choice of the model formulation, error structures, auxiliary variables will all affect the resulting estimates and are based on the judgment of the Official Statistician responsible for the analysis. Hence the retention of public confidence in the estimates and the defense against external criticism of the results will depend on the professionalism of the analyst responsible and more broadly on the regard in which the organization is held. This raises extensive questions about what would be contained in the methodological report supporting the estimates and how quality would be demonstrated and reported.
- The very richness of the choice of models within the mixed-model framework calls for more research on model selection and model fit diagnostics. This will include outliers at both the small area and individual data point levels. Often there is an interplay between the way that the structural part of the model is specified and the error components which means that there is a need for model selection procedures and diagnostics that take account of this interplay. Additionally model fit is often concerned with overall fit to the data as a whole but given that the end product will be estimates for each small area, the need for diagnostics to identify areas that do not appear to conform to the overall model is clear.
- Since much Official Statistics data is discrete there is still more work needed on generalized mixed linear models.
- In practice the small areas often vary in size so that some would be supported by direct estimates based on the survey data (e.g. a direct estimate of Ontario may be adequate but not for Prince Edward Island). The acceptability to users of hybrid estimation needs further exploration.
- Also the same general thrust that has increased the demand for small area estimates also requires in them being used to make inter-area comparisons. The statistical properties of such comparisons when the estimates are based on mixed model procedures needs further development.

In many applications there is a growing body of opinion that the Fay-Heriot model or the random intercept model captures the population structure adequately and is therefore recommended. This is probably based on the fact that small area means are essentially derived from predictions at the small-area mean of any auxiliary variables and consequentially variation in the small area slopes does not contribute to the prediction. Consider a simple two-level model with only one auxiliary variable:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + v_{0i} + v_{1i} x_{ij} + e_{ij}$$

$$y_{ij} = \beta_0^* + \beta_1 \left(x_{ij} - \overline{X}_i\right) + v_{0i} + v_{1i} \left(x_{ij} - \overline{X}_i\right) + e_{ij}$$

$$\hat{\overline{Y}}_i = \hat{\beta}_0^* + \hat{\beta}_1 \left(\overline{x}_i - \overline{X}_i\right) + \hat{v}_{0i} + \hat{v}_{1i} \left(\overline{x}_i - \overline{X}_i\right) \cong \hat{\beta}_0^* + \hat{v}_{0i}$$

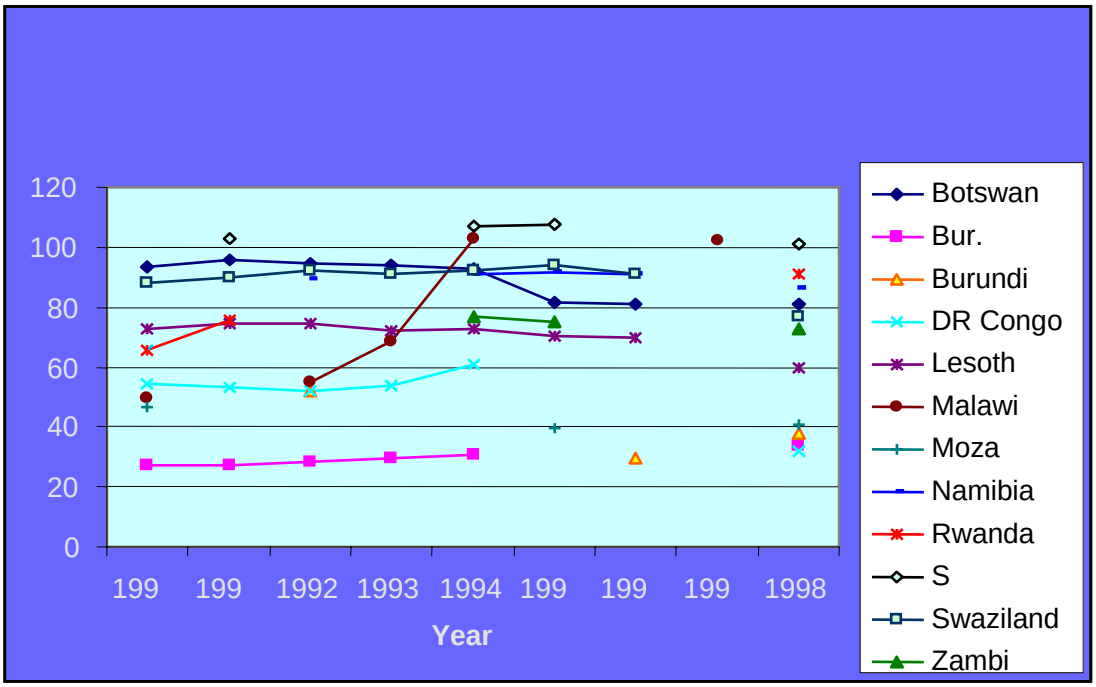
The random slopes do not contribute to the small area predicted mean since the prediction is made for $\overline{x}_i = \overline{X}_i$.

However one of the problems in small area estimation is to find auxiliary variables that provide sufficient information for the prediction. One obvious source is to use time series data from small areas so that time is an integral part of the formulation. In many practical situations the prediction will be for the most recent time period and the growing emphasis on ‘rolling population surveys’ or so called ‘rolling Censuses’ will reinforce this need for

'latest period' estimates. In this case the mean of the auxiliary variables will be at the mean period over which data is available while the prediction will be made for the latest period. In this case the variable small-area slopes will make a direct contribution to the prediction and the case for Fay-Heriot or the random intercept model is not so clear.

Monitoring International Development Indicators

The same issue arises in a very different context concerning the monitoring of the Millennium Development Goals for all countries of the world (UNSD, 2004). In this case the ideal data set for each statistical indicator would be a time series for each country with indicator values for every year. In practice many countries do not measure each indicator each year and in fact the country time series are often very incomplete. Figure 3.1 illustrates the case for a selection of Sub-Saharan African countries for the indicator measuring the enrolment rate in primary education.



Data source: UNESCO

Figure 3.1: Enrolment Rates for Primary Education in Selected African Countries

There are few relatively complete time series and some countries have only one or two observations. The countries are analogous to the small areas in the previous discussion and the requirement is to predict the country value for a specific period (often the most recent) and hence to make an estimate of the indicator for the region as a whole. Holt (2003) describes the issues in more detail.

Neighbourhood Statistics

The growing demand for 'neighbourhood statistics' in a number of countries may be regarded as a further example of the growing demand for small area statistics to which NSO's must respond. There are however a number of additional methodological questions that need to be addressed. The way that NSO's have addressed this demand is to assemble statistical information on each small area (neighbourhood) from a variety of different sources. These will include the Census(s), various administrative sources and, in some cases survey based estimates or combinations of survey and administrative data. The resulting 'data set' is an amalgam of a range of 'statistical objects' that may include a series of cross-tabulations (perhaps derived from the same or different sources). From the perspective of the statistical producer, the methodology now needs to cover not just making a small-area

estimate but making multidimensional estimates (e.g. a cross-tabulation) for each small area. The methodological needs listed earlier must be extended to this case.

However it is valid to ask whether the NSO's responsibility is simply focussed on making a series of related or unrelated estimates for each small area (the 'statistical objects') or whether it goes beyond that. For many users the 'statistical objects' are likely to be an intermediate product that will be fed into further analyses that meet the users real need. For example governments may feed the values into funding formulae or into models that drive or develop social programmes. Governments and local authorities may want to monitor measures of social deprivation over time to asses the impact of social programmes. Others may need 'statistical objects' (e.g. cross-tabulations or measures of association between variables) which are unavailable in the neighbourhood statistics data set. These uses raise a whole set of methodological questions about the statistical properties of these main uses of the data set that go beyond simply producing the properties of each 'statistical object'. What support should the NSO provide to such uses. Is it to investigate the methodological properties of the main uses (which can often only be done within the NSO because of access to the underlying data) and to provide suitable information, guidance and advice?

For example consider the relatively simple case of wishing to estimate the covariance between two variables for which no information on joint distribution is provided in the neighbourhood statistics data base. In effect the information on the two variables comes from separate sources. This gives rise to the ecological fallacy problem (Steel, Holt and Tranmer, 1996) and we consider a simple components of variance model model for each of the two variables:

$$y_{pij} = \mu_p + v_{pi} + \varepsilon_{pij},\ p = 1, 2;\ i = 1,..., m;\ j = 1,..., N_i$$

$$V(v_{pi}) = \sigma^2_{vp}; \qquad V(\varepsilon_{pij}) = \sigma^2_{\varepsilon p}$$

$$Cov(v_{1i}, v_{2i'}) = \sigma_{v12}, i = i'$$

$$Cov(\varepsilon_{1ij}, \varepsilon_{2i'j'}) = \sigma_{\varepsilon 12}, i = i',\ j = j'$$

$$Cov(y_{1ij}, y_{2i'j'}) = \begin{cases} \sigma_{v12} + \sigma_{\varepsilon 12} = \sigma_{T12} & i = i',\ j = j' \\ \sigma_{v12} & i = i' \\ 0 & \text{else} \end{cases}$$

If sample variances and covariances based on the joint observations of the two variables are available then essentially unbiased estimates can be obtained. However if the estimates for the two variables come from different sources then the usual estimates are unavailable. In particular if only small area means of the two variables are available one might use:

$$\bar{s}_{12} = \frac{1}{m}\sum_i w_i (\bar{y}_{1i} - \bar{y}_1)(\bar{y}_{2i} - \bar{y}_2) \qquad w_i = n_{1i}\ n_{2i} / n_{12i}\ .$$

It may be shown that:

$$E(\bar{s}_{12}) = \sigma_{T12} + (\bar{w}^* - 1)\sigma_{v12}, \qquad \bar{w}^* = \bar{w}\left(1 - \frac{C_w^2}{m}\right).$$

Hence the estimate is biased for σ_{T12} and the bias depends on the level of overlap between the samples that the two separate means are based upon. In the extreme case when there is no overlap the estimate is based on the between

area component of covariance and does not include the between unit, within area component. As the overlap diminishes the bias becomes more severe.

A common approach to overcoming this (which is implicit in so-called statistical matching) is to take account of the joint distribution between each of the two variables and some auxiliary variables that are available for each. In effect if both variables are presented within the neighbourhood data set as cross-tabulations against, for example, age and sex groups, then these may be used as auxiliary variables. We assume a linear model linking each variable of interest with the auxiliaries:

$$y_{pij} = \mu_{p.z} + z_{ij}^T \beta_{pz} + \nu_{pi} + \varepsilon_{pij} .$$

Steel et al (1996) show that:

$$E(\bar{s}_{12} \mid z) = \sigma_{T12} + \beta_{1z}^T (\bar{s}_{z12} - \Sigma_{zz}) \beta_{2z} + (\overline{w}_i^* - 1) \sigma_{12.z}$$

and hence

$$\hat{\sigma}_{T12} = \bar{s}_{12} - \beta_{1z}^T (\bar{s}_{12z} - \Sigma_{zz}) \beta_{2z}$$

with bias $(\overline{w}^* - 1)\sigma_{v12.z}$.

The objective is to find the auxiliary variables **Z** that allow the estimate to be adjusted to minimize the residual bias. This reflects the residual component of covariance after allowing for the auxiliaries. This is essentially the approach of data fusion/data matching and identifies. The problem is that the residual bias will be unknown (unless the individual level data is available to the NSO but not published because of confidentiality constraints). Where there is small levels of overlap between the sources of the two variables even a small residual covariance component will be inflated by the multiplier based on the weights and could be very large. If one of the sources is based upon a census (or administrative source that is effectively full coverage) and the other is based upon a survey then the overlap between the two sources will be based on the sample size for the survey source in each area since $n_{1i} n_{2i} / n_{12i} = n_{1i}$ if this is the survey sample size.

Note that to provide this adjusted estimator users would need an estimate of the covariance matrix between the auxiliary variables Σ_{ZZ} and the NSO would need to provide this as an additional piece of information if it could not be derived from the neighbourhood statistics data set.

The point of this section is not to advocate some particular solution to a specific problem but to illustrate that the statistical properties of the uses to which the data set might be put may depend on access to the underlying data that only the NSO may have and may require the NSO to provide additional information. It certainly requires methodological investigation of the uses to which the data are put that takes the role of the NSO beyond simply producing the small area statistics and making them available.

5. THE USE OF ADMINISTRATIVE DATA

Another thematic strategic need for NSO's is the growing use of administrative data for statistical purposes. In part this is driven by the need for small area statistics (for which large administrative data sources may support disaggregation to small areas in a way that surveys cannot). In part it is also driven by the need to reduce costs and the data compliance burden on respondents. In part it may also be driven by the fact that some administrative sources may provide more complete coverage than surveys that are subject to unknown levels of non-response bias. The usual countervailing arguments are well known:

- that often the administrative source does not measure precisely the required concept,

- that often the coverage of the administrative source (and the level of maintenance for births and deaths for example) is not of as high quality as one would wish,
- that the administrative source is neither designed nor operated for statistical purposes and this can affect the way in which the data can be used, and
- that the source does not contain such a rich supply of related variables as a survey questionnaire would and this affects the depth of any possible analysis and the coherence of the estimates with those derived from other sources.

The administrative sources may be used alone or better still, some of the difficulties listed above may be reduced if the source is used in conjunction with survey data. The administrative source may be used as a sampling frame, or as a source of auxiliary variables in ratio or regression estimation. The precise forms of use will depend on whether the individual records within the two sources can be matched or not.

The use of administrative data for what one might term standard uses as a sampling frame or as the source of an auxiliary variable for estimation are well established within the usual quality assessment framework for Official Statistics. There are issues about coverage and quality (and in particular that many administrative sources have deficiencies that are more strongly reflected in some sections of society). For example tax records are often systematically deficient when one considers the low paid and this will have a differential effect on the use of the source as a frame, as an auxiliary variable or as a primary data source for statistical compilation.

However I shall briefly make the case that there is a much less well-developed quality framework for administrative data than there is for survey data and that this is a significant deficiency. Official Statisticians have spent more then 40 years researching and understanding the interaction between data collectors and the respondents as well as such matters as the effect of question wording and questionnaire design on the responses provided. More recently computer assisted data collection has provided additional information as a by-product of the data collection process to help our understanding of the process itself.

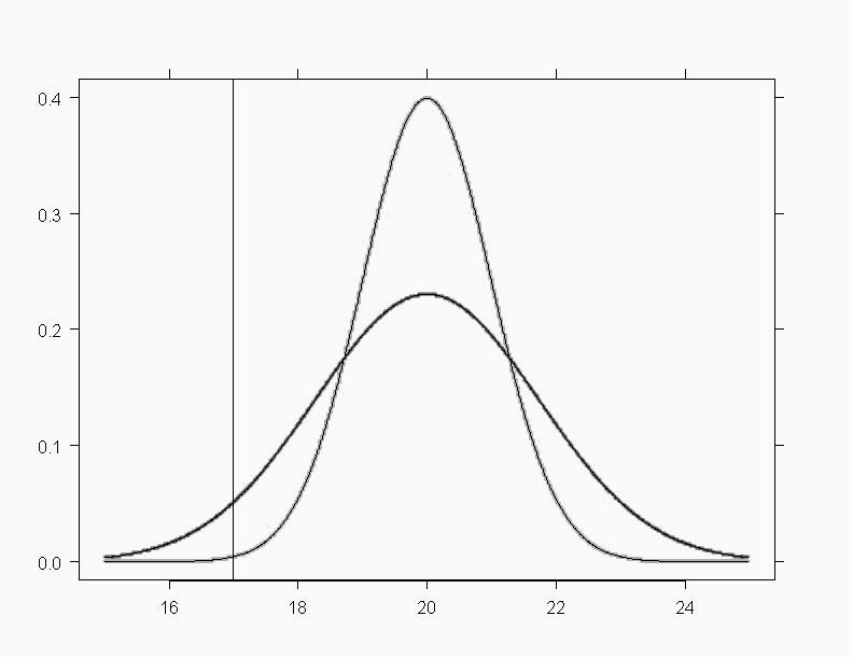
The information exchange between members of the public and any administrative system (and which will be the source for later statistical use) is also a form of interaction. Unlike the survey process, where we attempt to measure responses without affecting the respondents' real-world behaviour, the administrative systems are often designed to do just that. For example the rules on eligibility for a social benefit are often designed to encourage certain kinds of behaviour in the applicants. Whereas official statisticians attempt to design questions that are neutral on behaviour the administrative system may be designed with the opposite intention. If official statisticians are to use such sources for statistical purposes then I suggest that a framework of data quality, akin to that developed through survey methodology, may prove to be beneficial. At the very least it would provide a framework for assessing statistical quality and at best it may help to provide understanding of how to reconcile the administrative and statistical needs. If we are to present statistics derived from administrative sources for users we need a better developed quality framework in which to present the results.

Measurement Error

One quality issue which affects both survey and administrative data and has not been given enough attention is the question of measurement error. Fuller (1995) draws attention to the fact that a commonly used model assumption for measurement error is that the variable observed x is related to the true value X through an error term u which is assumed to be unbiased:

$$x = X + u \qquad\qquad E(u) = 0 \text{ and } V(u) = \sigma_u^2.$$

Fuller points out that for estimating means and variances (for example using simple random samples) the usual estimators remain unbiased for the observed variable and that one might therefore conclude that measurement error is not a problem. He goes on to observe, however, that even for estimating the proportion of the population above or below any particular threshold the estimates are not generally unbiased. Figure 4.1 presents a simple schematic diagram to illustrate the point:



The narrower curve represents the distribution of the true values and the wider curve that of the observations (with measurement error). The area under the wider curve is greater for the proportion of the distribution below any particular threshold. Hence the measurement error creates a bias in the estimate of the proportion below the threshold.

Normally this bias would not be readily apparent but Skinner et al (2002) provide an example based on the analysis of administrative data where this distortion is directly relevant to the purpose. In the late 1990's the UK Government introduced a minimum wage which applied to the large majority of the labour force (there are some fairly small exceptions). The case concerns estimating the proportion of people earning below the minimum wage from the UK Labour Force Survey. Clearly these estimates are important for monitoring the policy.

The sources of hourly pay that have been available for statistical use are a direct measure of hourly pay rate from administrative sources and a measure derived from the LFS by dividing income by usual hours worked. The former source is known to have coverage problems particularly in terms of the low paid and the latter is subject to measurement error in part because the denominator is a measure of usual hours rather than the specific hours in a period to which the numerator (pay) relates. An additional direct measure of hourly pay was derived from a subset of LFS respondents by adding suitable questions to the questionnaire. It is the comparison of the derived and direct LFS measures that we focus on.

Figure 4.2 (from Skinner et al, 2002) shows the joint distribution of both the direct and derived measures.

Figure 4.2 shows clearly the effect of the measurement error on the derived (x) variable where the distribution freely crosses the National Minimum Wage figure and a significant number of respondents are shown earning an hourly rate below this figure. In comparison, the distribution of the direct measure (y) has a much sharper cut-off at the minimum wage value.

The example illustrates a case where the effect of the measurement error is directly relevant to the policy interest since the derived measure (with substantial measurement error) implies that significant numbers of people are earning below the minimum wage and hence that the intention of the policy is thwarted. The direct measure substantially corrects this impression and creates a better picture of the true situation. Other uses of data that is subject to measurement error may be just as severely affected but the impact is not as apparent because there is no external basis for validity checking.

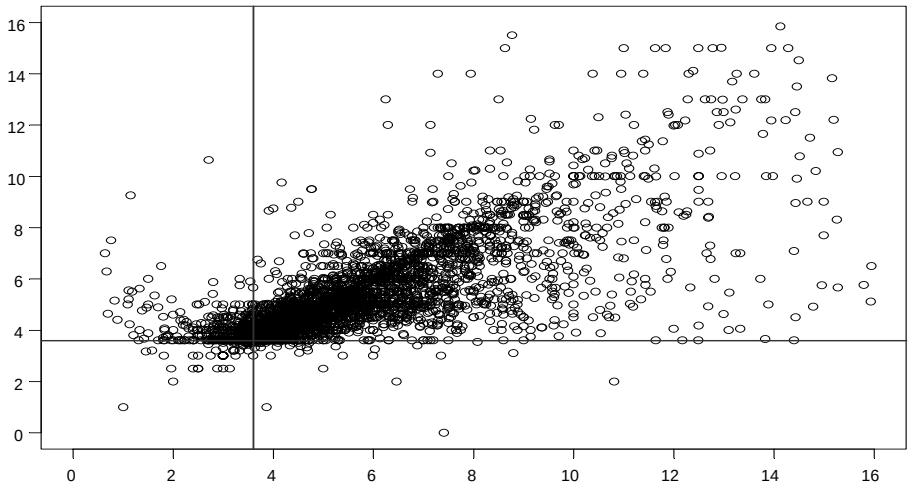


Figure 4.2: Joint distribution of Direct (y) and Derived (x) Measures of Hourly Pay (£ per hour) from UK Labour Force Survey. (Value of National Minimum Wage shown by Vertical and Horizontal Lines) source: Skinner et al (2002)

6. MEASURING PERFORMANCE

The final main section is concerned with a comparatively recent demand for statistical measures of performance at a disaggregated level; usually at the level of a delivery unit for a public service (see Bird et al, 2004). For example, the proportion of school-leavers exceeding a specified level of educational attainment (as measured by examination successes) for individual schools; the length of waiting lists for elective surgery for individual hospitals or the clear-up rate for crimes for individual police authorities. There are some examples at lower levels (such as the mortality rates for surgical teams or individual surgeons). Often the indicator values are presented as 'league tables' or banded performance measures (e.g. 3*, 2* and 1* hospitals).

The motivation for this development is the need to provide measures of unit performance for two purposes: to help the public (as consumers of the public service) make choices about which unit to use and secondly to increase the level of public accountability of Ministers and the staff and management of identifiable public service providing units. In terms of public accountability it is often the case that specific targets of performance are established against which the public may judge performance (e.g. 75% of high school students will achieve a specific level of educational attainment). In practice there are three elements to the requirement: a decision about the concept to be monitored, the development and monitoring of an appropriate statistical measure and finally the setting of a performance target based on the measure chosen.

However these three distinct aspects are often confounded and politicians or policy officials may couch the concept in terms of a specific measure and set a target without any reference to the statistical advisability or even feasibility of the measure chosen.

This use of statistical indicators raises a number of methodological issues that are outside of the usual framework for the methodology of official statistics. These either require further development or, where the developments have taken place in other settings, the methods need to be embedded within the culture of NSO's.

Statistical Design

The choice of statistical indicator has two important aspects for the NSO. First it is important that the concept chosen for accountability purposes has wider acceptance than simply being determined by those responsible for the public service. If the concept is seen to be 'politically chosen' then there is a risk that the NSO will be perceived as

supporting a politically tainted measure and this may call into question the statistical independence and integrity of the NSO. Second, even if the concept is sound, there is the inevitable tension between what is conceptually desirable and what is statistically feasible to measure. Third, the choice of a performance target is a political one although statistical analysis may well inform the choice.

There are questions about design that official statisticians are used to dealing with: whether it is better to collect information from delivery units on a 100% basis or whether sampling of cases/subjects would be more efficient. For example in the case of carrying out random mandatory tests among prisoners for the detection of illegal drugs in the blood stream the data collection process implies that costs are largely proportional to the number of samples obtained at each institution whereas in the case of educational attainment information this can be collected from school records for all pupils in a school for essentially no additional cost.

In normal cases one would look at the size of a feasible effect of a policy and calculate the sample size required to detect this effect with required power at the reporting unit level. However when the target has been chosen by policy officials or Ministers it may be aspirational in nature and may not reflect the level of effect that would be important to measure or realistic to achieve. Choosing sample sizes to give sufficient power to detect meaningful change or trend may go beyond the choice of target for aspirational purposes.

More importantly there may be small reporting units that simply do not generate the number of observations required, even on a 100% sampling basis, to give sufficient power to detect the size of change expected (see below).

Data Quality and Audit

The way in which the statistical indicators are used may have important consequences for those being monitored. Unsatisfactory measures may result in personal or organisational sanctions and satisfactory measures may result in personal rewards for managers (performance related pay) and for the organisation (greater levels of financial or managerial autonomy). In some cases unsatisfactory measures may result in imposed management or organisational change and may also result in the service unit experiencing a decline in staff morale and finding it harder to recruit staff of the highest calibre. Hence there are significant pressures on staff in the service units to ensure that the statistical indicator is as satisfactory as possible. Insofar as these pressures lead to genuine change to the operational effectiveness of the unit this is part of the rationale behind the policy but they may lead to perverse behaviours that may alter the statistical indicator without affecting the underlying performance.

In the extreme, perverse behaviour may lead to data falsification or, if there is any element of judgement in codifying the data, to recording practices which affect the indicator value. For example an indicator measuring prisoner-on-prisoner violence may require that each apparent incident be investigated by prison staff to determine whether such an incident has occurred. The more rigorous the investigation of every possible incident the more likely it is that the level of incidents recorded will be increased. Hence prison staff may be inclined to dismiss some incidents as ‘general unruly behaviour’ rather than assess them as potential cases of prisoner-on-prisoner violence.

The methodological implication for design is that data definitions and classification criteria need to be precise enough to avoid the effects of perverse behaviour and data quality audit needs to be alert to possible effects.

A second example of perverse behaviour that does not result in data falsification but can act against the service improvement that performance monitoring is intended to foster is given by Dranove et al (2002). The performance of named cardiac surgeons in New York, as measured by patient outcome, was published. An evaluation showed that subsequent risk averse behaviour by the surgeons resulted in reluctance to operate on the most severely ill (and hence those with the poorest prognosis) and a countervailing increase in medical intervention for less severe patients. The effect was a deterioration in patient outcomes overall.

Analysis and Interpretation

There are important methodological issues surrounding the analysis and interpretation of data used for performance monitoring. These surround how to control for case mix and context between service units and also how to interpret the measures in the presence of the inherent variation in the data.

Controlling for case mix.

The fundamental purpose of performance monitoring is to monitor the performance of the delivery unit. Consider, for example, monitoring school performance on the basis of the proportion of school leavers achieving a given standard. One might reasonably think that a school in a poor inner-city location with children drawn from communities with a variety of languages in daily use would face much greater educational challenges than another school in a more favourable location. Or if the standard of educational attainment on entry to each school was measured through some national assessment then one would expect the schools with the higher ability intake subsequently to produce the higher ability leavers.

This suggests that the schools performance measure should be 'controlled' for variations that will affect comparability and hence lead to so-called 'value-added' performance measures. The same thing arises in the medical context where controlling for the case mix of patients (for example a measure of the severity of the illness at the start of treatment) is essential if the treatment effect is to be isolated from the severity of the cases.

Note that it is not always easy to agree on which sources of variation should be adjusted for. For example when comparing schools should the final educational attainment be adjusted for the truancy level in the school. The argument in favour is that if there are high levels of truancy then the schools level of achievement is expected to be affected (adjusting for an exogenous variable). The contrary view is that the level of truancy reflects the success of the school in fostering enthusiasm and interest in education and hence reflects something that the school should influence (no control for an endogenous variable).

Note also that even when controlling for inherent variation is the right thing to do it is not always easy to measure the controlling variable accurately. The assessment of the severity of an illness at the time when medical treatment starts may not be straightforward.

Assuming that these problems can be faced and an estimation of 'value-added' performance made there are additional methodological problems for the official statistician. One of the common ways of approaching the problem is by using a mixed model as described in section 4:

$$y_{ij} = x_{ij}^T \beta_i + e_{ij} \qquad j = 1,...N_i \qquad i = 1,..., m$$

$$\beta_i = z_i \ \gamma + v_i$$

In the case of school performance for example i may index schools and j individual students. The variables $\boldsymbol{x}$ represent individual level covariates and $\boldsymbol{z}$ school, or neighbourhood, covariates.

Most commonly a random intercept model is chosen so that there is only one random term v associated with the intercept term in the regression model.

Whereas in section 4 when such models have been proposed for small area estimation, a design consistent alternative has been available even if it is inefficient. For performance monitoring the model-based formulation is integral to the objective. The value-added measure of performance is the predicted value of the random term v since the objective is to measure the performance of the unit conditional on the covariates. Hence the model formulation is not simply a convenient way of estimating a small area population parameter but determines the object of inference. The value of the performance measure will depend on the choice of covariates at the individual and school level together with the precise way in which the terms are included and the formulation of the multi-level random components.

To produce value-added performance measures for public consumption is a large step from the sort of statistics that NSO's usually produce. Public confidence and acceptance will depend on the professional reputation of the NSO, the quality of the underlying analysis, the demonstration through methodological reports of the soundness of the model selection procedures and the use of diagnostic measures.

Note that when there is only one school level random term, the random intercept, the interpretation of the value-added performance is clear. However if there is a need to include additional components as school-level random slopes for the covariates then the interpretation of school performance becomes more difficult to convey. Suppose for example that there is a random term associated with the student's gender. This would imply that the relative performance of schools will be different for male and female students. In general the policy maker's wish is to have a simple, easily conveyed measure of school performance. Describing the results if there are school differences of this kind will add to the presentational challenges.

Inherent Variation

Another methodological issue is how to present results taking account of the inherent variation in the data. Too often interpretations are made, league tables are presented, delivery units are categorised into classes of success or failure on the basis of the latest data value without paying due regard to the inherent uncertainty in the estimates. The data (e.g. the crime clear-up rate or the school success rate) is treated as a precise measure of the underlying performance. Schools, surgeons or police authorities can be branded as succeeding or failing on the basis of the latest data.

In practice the uncertainty associated with a mixed-model prediction of one unit's performance can be substantial and in some situations there is no remedy. If the unit of analysis is a school that has only 100 school leavers in any year then this is all the data available (unless the school's performance is judge by pooling data over several years and this may defeat the management intention).

The limitations of published league tables that rank delivery units on the basis of the performance measure are well known. Goldstein and Spiegelhalter (1996) give a good account of the imprecision.

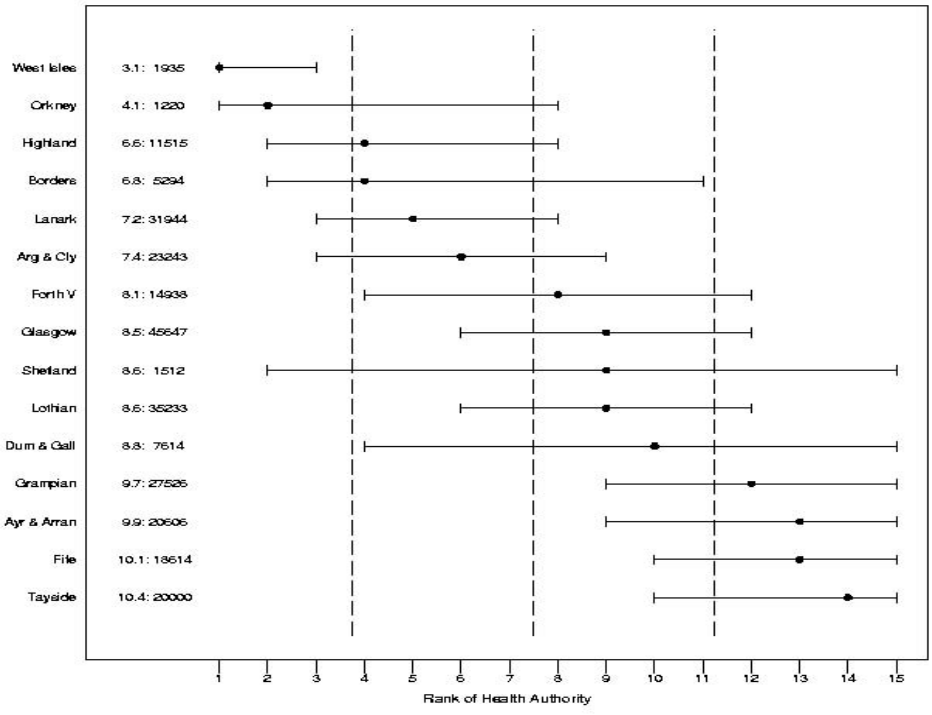


Figure 4.3: Median and 95% Confidence Intervals for Ranks of Scottish Health Authorities with Respect to Teenage Conception Rates, 1990-92.
Source: Bird et al (2004), from Goldstein and Spiegelhalter (1996).

Figure 4.3 is derived from Goldstein and Spiegelhalter (1996) and shows the median and 95% confidence intervals for Scottish Health Authorities with respect to teenage conception rates. The plot demonstrates that only the authority with the lowest rate can safely be assigned to the first quartile and the four authorities with the highest

rates to the upper half of the ranked order. Apart from these the uncertainty is too great to draw any definitive conclusions about the rank order of the authorities.

In practice the change in ranks from one year to the next (as a measure of change in performance) would be even more uncertain.

Too often league tables or star bands (3*, 2* and 1*) are produced and interpreted with a level of assumed precision which is simply not supported by the level of uncertainty for the inference. There is a major task to help users to understand the limitations of the tables and to help them guard against over-interpretation.

A related point is that often performance targets are cascaded without a clear understanding of the consequences. Suppose for example the target is set that 75% of school leavers should attain a particular educational standard. At the national level in the UK (with about 600,000 students in each year), to be 99% certain of meeting the target the system would have to operate with a slightly higher proportion than 75% achieving the target in order to allow for any random fluctuations that might occur in any year. For a teacher with a class of 30 children to be 99% certain of meeting the target the achievement rate would have to be close to 90%. If the teacher and the nation were both to operate with 80% success rate, the national performance measure would almost always exceed the target whereas the teacher would run a substantial risk of failing the target in any year. We also note that the confidence interval for the individual teacher cannot be narrowed since 100% of the students in the class are tested. The uncertainty is inherent in the data properties and the inference.

There are thus various methodological issues concerning the analysis and presentation of performance measures. Control for case mix is often essential but appropriate covariates are not always available; the precise model formulation which is a professional judgement will affect the performance measures; the parameter of interest is integral to the model formulation and there are important issues surrounding the inherent variation in the measures and how these are interpreted. National Statistical Offices will have to address these if they are drawn into producing performance measures.

7. SUMMARY AND CONCLUSION

In summary this paper advances the view that methodological developments matter, not simply as a way of ensuring the quality and acceptability of individual statistical outputs, but beyond this. In a systemic manner they establish the professional reputation of the NSO, they contribute to the development of international standards and they hence help to promote public confidence in the NSO and its outputs.

Given the strategic role of methodology within the NSO, it is argued that planning future methodological developments must take account of the strategic needs of the organisation. This is not easy but the paper tries to make connections between strategic developments and the methodological implications. A process for connecting these is essential if the future methodological developments are to reinforce the strategic needs of the NSO.

REFERENCES

Bird S M, Cox D R, Farewell V, Goldstein H, Holt D, Smith P C (2004) Performance Indicators: Good, Bad and Ugly. J Roy. Statist. Soc (to appear)

Dranove D, Kessler D, McClellan M, Satterthwaite M (2002). Is more information better? The effects of report cards on Health Care Providers. National Bureau of Economic Research, working paper No w8697, Cambridge Ma.

Fay R E and Herriot R A (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. J. Amer. Statist. Assoc. 74, 269-277.

Fuller W A (1995). Estimation in the Presence of Measurement Error. International Statistical Review, 63, 121-141.

Goldstein H and Spiegelhalter D J (1996). League Tables and their Limitations: Statistical Issues in the Comparison of Statistical Performance (with Discussion). J. Roy. Statist. Soc. A, 159, 3, 385-443.

Holt D (2003). Methodological Issues in the Development and Use of Statistical Indicators for International Comparisons. Survey Methodology, 29, 1, 5-18.

Neyman J (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *J. Roy. Statist. Soc., 97, 558-625.*

Rao J N K (2003). Small Area Estimation. John Wiley, New Jersey.

Skinner C J, Stuttard N, Beissel-Durrant G and Jenkins J (2002). The Measurement of Low Pay in the UK Labour Force Survey. Oxford Bulletin of Economics and Statistics, 64, Supplement, 653-676.

Steel D G, Holt D and Tranmer M (1996). Making Unit-Level Inferences from Aggregated Data. Survey Methodology, 22, 1, 3-15.

UNSC (2002) An Assessment of the Statistical Indicators Derived from United Nations Summit Meetings. UN Statistical Commission, New York.

UNSD (2004). Millennium Indicators Data-base. http://unstats.un.org/unsd/mi/mi_goals.asp