

Combining Surveys and Administrative Sources in data production

Integrazione tra indagini e fonti amministrative nella produzione di dati.

Ignazio Drudi

Dipartimento di Scienze Statistiche, Università degli Studi di Bologna
Via Belle Arti 42 8 – 40125 Bologna, e-mail: drudi@stat.unibo.it

Riassunto: Dati amministrativi e indagini dirette sono da sempre l'ossatura dei principali sistemi statistici nazionali, tuttavia, queste due fonti di dati hanno rivestito un ruolo complementare, nel senso che le prime erano chiamate ad operare in quegli ambiti fenomenici in cui le fonti amministrative erano carenti o non utilizzabili. La tesi sostenuta in questo intervento è che le nuove tecnologie informatiche di raccolta e gestione delle informazioni determinano una nuova esigenza di integrazione tra le due strategie di produzione dei dati, soprattutto in ambito economico. La necessità di nuove sinergie è individuabile in molte fasi del processo di produzione dei dati: dal processo di selezione delle popolazioni di interesse, alla costruzione di un sistema di controllo della qualità, dalla definizione di strumenti di rilevazione alla fase di revisione.

Keywords: Administrative data, Surveys, Predictive models, Economic Statistics, Selection bias, Data quality.

1. Introduction

Both direct surveys (DS) and administrative sources (AS) have been used since years in many countries in the production of statistics. Nevertheless, the objectives of statistical offices are formulated in terms of statistical information needs, while administrative accomplishments leads to individual data, which often are collected using a form recorded during a specific and issue-oriented work. The interaction between the required statistical information and the raw register (or register-like) data can be ideally represented as a fork having numerous prongs: one prong ignores entirely the administrative data and plans autonomous target oriented projects (i.e. conventional surveys); the opposite prong represents statistical production process that rely solely on data collected by others (i.e. register-based estimates of population); while a third option includes hybrid combinations of varying degrees of the two former options, ranging in the continuum between the above two extremes.

The choice among these options depends on the trade-off between interests and needs of statistical and administrative agencies, expressed as a set of attributes: population covered, type of data-collected, time reference, frequency and timing of updating, maintenance and reliability of the produced files, their accessibility and their content flexibility. But, first of all, a fundamental role is played by the available technology.

Mainly due to difficulties in accessing and managing large amount of data, for a long time AS have been confined to an ancillary role with respect to DS, they have been used to identify few useful auxiliary variables or to set-up frames of populations to investigate, and so on, in a perspective very close to the first prong described above. In

many cases, this is the best choice and the interaction between AS and DS will follow the above way for a long time. DS are the irreplaceable base of the knowledge about social phenomenon, personal opinion, attitudes and expectations. But technological changes have historically led to organizational change so that the growing possibility to access and manage large administrative data-bases has driven central statistical agencies to making an increasing use of such data, especially for economic statistics. This practice is quickly spreading in almost all developed countries, as these sources become available in computer readable format and offer some great potentials to improve or even to replace direct surveys. In other words, at least for business survey, we are moving towards the second prong of the fork, nevertheless, although gains in using administrative data can be considerable, they still presents substantial drawbacks, but even if, nowadays, a complete substitution is merely an utopia, DS and AS are changing their role in statistical data production.

The aim of this paper is to show how changes in technologies and in statistical use of AS creates new needs of integration with DS and, more precisely, that DS can and must play an irreplaceable role in assessing quality of data stemming from AS.

In this paper it is obviously impossible to face all problems concerning data-quality in AS, thus we choice to concentrate our attention on the framework in which, in our opinion, integration is more urgent and important: the ex-post treatments of three typical problems of Administrative data: undercoverage, self-selection and timeliness.

The rationale of the choice is twofold: topics considered can be tracked back to the same origin: the statistician has limited or no control on data generation process of an AS and hardly ever the needs of the statistical and administrative organizations, coincide perfectly. As a consequence, and this is the second point, using AS in order to produce statistical knowledge requires additional work. As we shall show, this work often consists in defining unusual estimation strategies. (Drudi, 1997; Drudi Filippucci, 2000).

In section 2, we describe the common thread linking undercoverage, self selection and timeliness, showing that they can be faced using different estimation tools referring to a general inference strategy. Such tools will be sketched in section 3, concluding that all of them show the same need for statistical evidence about their base hypotheses. The way DS can provide for such evidences will be shown in section 4. Section 5 reports some concluding remarks and tries to outline an hypothetical integrated system.

2. Undercoverage, self-selection and timeliness in AS

A fundamental problem to face when we want exploit an AS for statistical purposes is the correspondence of the target populations. Usually they differ because a set of factors.

Extent of the administrative acts

Legal accomplishment generating administrative records usually concern only some kind of individual or organizations, which are compelled to fill up some administrative form on the basis of their judicial status or their activity or if they presents some peculiar characteristic. Nowadays, at least in Italy, there is a strong propensity to simplify administrative burden, so that almost all laws provide for exemptions.

For example only stock companies have to draw out and register a complete (and uniform) balance sheet, V.A.T. registers are compulsory only for some professional

activities, small firms can adopt a simplified version of accounting, employees and self-employees have different accomplishment for pensions and so on.

As a consequence, each AS has its own target population, usually a subset of the entire population, which, on the contrary, is the main target of statistical agencies.

Different definition of Statistical units

Only juridical persons are legal subjects, and hence when statistical unit of interest has not a judicial status it has low importance for AS. This is the case of establishment and firms. While the interest for the most of economic statistics is on the former, the latter are the units having legal relevance, and hence are considered in administrative databases. The problem is, obviously, more heavy for multi-product and pluri-located firms. As a consequence, for example we could not compile regional accounts relying solely on balance sheet as they are drawn up by firms.

Different definition of Variables

Especially in economic field administrative acts provides for peculiar definitions of principal measures of activity, for example some small firms pay their taxes on an *a-forfeit* basis. Another case is the definition of depreciation allowance used in fiscal sheet that is quite different from the statistical concept of capital consumption.

Missing units and missing items

Missing units and missing items are two types of non-response that may be present in AS. Although providing answers to administrative data collection is usually mandatory, some non-response can still exist. They may be caused by voluntary non-compliance or by processing problems (computer problems, non-readable fields, corrupted files, etc.). As a matter of fact, at least in Italy, this problem cannot be under evaluated.

Processing errors

Several processing errors can occur when dealing with administrative data files. This is because the check phase when exists has been planned in order to achieve administrative goals, and some problems, crucial for statistical use, like keying errors, consistency problems, coding errors, etc., may be of no interest.

Lack of classification

A correct classification of economic activity is vital for the production of economic statistics. The problem must be carefully considered because usually administrative offices have limited interest in coding firm activity and ex post coding is a costly process.

Timeliness

While deadlines for administrative accomplishment are compulsory, data recording and checking often take place with some delay. Nowadays different ways to accomplish administrative declarations are allowed, like electronic data interchange or the use of Internet connections. Obviously in such cases information are immediately available, but subjects are free to choice the way they prefer, so that available data concerns only a well defined subset of units. In this sense timeliness produces a sort of “temporary undercoverage” introducing a serious selection bias in the early estimates.

All the above problems produce a main effect: they introduce a selection bias in the units (or in the variables) we can observe and use for statistical purposes. So, we are compelled to consider the process of estimation of the variables referred to the population of interest on the basis of AS. In other words, we need a strategy of inference. The question is clear when we consider surveys sampling but is not well discussed when we deal with AS.

In current literature concerning inference, some contributions on this topic (Deming, 1950; Sarndal, 1978; Smith, 1994) can be retrieved, albeit not explicitly referred to statistical use of AS, and they jointly show the necessity to use predictive inference. At least two statements can be recalled in order to clarify the central role of predictive approach. First of all, several authors clarify that the kind of inference needed depends on relationship between observed units and target population. More precisely, Smith (1999, p. 2) stress that: *“if the target (population) is the fixed finite population from which the sample was drawn then the inferences are descriptive. If the target is some other population, so that even a perfect census leaves inferential uncertainty, then the inference is analytic.... For analytic inference models are a necessity, ... (while) ... for descriptive inference there is a choice between model-based predictive inference and randomisation inference.”*

Secondly, considering that the statistician has no chance to control the selection mechanism, the inference must operate “after the selection”. Following Royall we can stress that: *“After the sample is selected, observations on the (observed) units ..must be used to make inferences, or predictions, about unobserved y values. The strength of the probability sampling approach, its freedom from assumption about the y's, now becomes a weakness..... Inference from sample to non-sample units requires a stronger relationship than a simple random design. Prediction models ...link the two sets of units and provide a basis for inference.”* (Royall, 1988, page 411).

As is well known, (Royall, Pfeffermann, 1982; Royall, Cumberland, 1981; Royall 1988, Cassel, Sarndal, Wretman, 1977) the key concept of predictive inference is the idea of superpopulation. Roughly speaking, in this context we suppose that the observed population is a realization of a stochastic variable containing all possible populations, hence the elements in the population are not fixed, as they are in the traditional approach, but rather they are treated as random variables. The basic idea is that a model links the actual population to a potentially realized superpopulation. The core of the inference is the correct specification of model, under a correct model, the estimates are design-robust, in other words we can face the problem of selection-bias.

In order to clarify the role of DS in this approach, we need to describe the estimation process in more detail.

3. Predictive model

In superpopulation models, given a population of N units, we assume that a variable y (a vector containing the y_i 's of each unit) be the realization of stochastic vector $\mathbf{Y}$. Besides, albeit not necessary, let us each individual in the population have known auxiliary variables $\mathbf{X}$ (a matrix $N \times P$). In such a situation a simple model describing the superpopulation could be:

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

$$V(\mathbf{Y}) = \mathbf{V}\boldsymbol{\sigma}^2$$

Where $\mathbf{V}$ is the (true) variance-covariance matrix of the superpopulation, conditional to $\mathbf{X}$. Now, given a subset of s observed units and of r unobserved ones, matrix $\mathbf{V}$ can be decomposed in four sub matrices: V_{ss} , V_{rr} , V_{rs} and V_{rr} . In matrix notation the model become:

$$E(\mathbf{Y}) = E \begin{bmatrix} \mathbf{Y}_s \\ \mathbf{Y}_r \end{bmatrix} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix} \beta \quad \text{and} \quad \mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$$

It is well known (Royall, 1988) that a BLU estimator of a linear combination of y_i , like $\mathbf{I}'\mathbf{Y} = \mathbf{I}'_s\mathbf{Y}_s + \mathbf{I}'_r\mathbf{Y}_r$, for example means and total, is given by:

$\mathbf{I}'\mathbf{Y} = \mathbf{I}'_s\mathbf{Y}_s + \mathbf{I}'_r \left\{ \mathbf{X}_r \hat{\beta} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} (\mathbf{Y}_s - \mathbf{X}_s \hat{\beta}) \right\}$ where $\hat{\beta} = (\mathbf{X}'_s \mathbf{V}_{ss}^{-1} \mathbf{X}_s)^{-1} (\mathbf{X}'_s \mathbf{V}_{ss}^{-1} \mathbf{Y}_s)$ is the usual GLS estimator for observed data. In other words, if the superpopulation model is correct, BLU estimation is obtained summing up observed data, regression estimates for unobserved ones and an error component obtained from observed standardized residuals, weighted by the covariance between observed and unobserved data, which play the role of correcting for selection bias.

Superpopulation model offers a general framework encompassing several empirical estimation strategies (Bolfarine, Zacks, 1992). Such strategies differ with respect to the specified model, the estimation techniques, the definition of selection process, the choice of appropriate specification depends on the available data and the statistical interest.

As an example, let us refer to a real AS. National Institute for Social Security (INPS) collects data about number of workers from all firms with employees. The most of them use paper forms to communicate data, but, since few years, it is possible to provide data using electronic forms. A growing number of firms tend to choose this new way. As a consequence, referring to an year of interest, while the latter are available in computer readable form after 3-4 months, the former are released after about 1 year. Obviously the choice between traditional and new forms is influenced by different factors, the most of them concerning the type of firm, like its size, organization, complexity, skill, and so on.

As a result the pattern of available information at a given time point (say at the end of the first quarter of year $t+1$) concerning the previous year t , can be summarized as follows:

Table 1 – Availability of data for INPS archives

Status of unit		Description	Data Availability	
t-1	t		t-1	T
I	I	Always electronic form	O	O
NI	I	Electronic form adopted in year t	O	O
I	NI	No more electronic form	O	NO
NI	NI	Always paper form	O	NO
I	A	Deaths	O	A
NI	A	Deaths	O	NO
A	I	Births	A	O
A	NI	Births	A	NO

Where I stands for “firms using electronic forms”, NI for “firms using paper forms”, A for “non existing firms”, O for observed, NO for non observed. It should be evident that a selection bias due to timeliness is in hand.

Let us suppose we are interested in estimating the total amount of employees on the basis of such data. Note that total occupation is the key variable for the most of economic statistics and it is the first variable to estimate. This situation has the drawback that no auxiliary variable is available.

In such a situation the simplest model we can implement is.

$$O_{it} = \beta O_{i,t-1} + \varepsilon_{it}, \text{ with variance-covariance matrix } \mathbf{V}$$

Where O_{it} stands for number of employees in firm i at time t , and ε is a residual.

Following Royall, we could estimate β from observed firms and then correct for unobserved data. But even a such simple model implies a large amount of uncertainty.

First of all, we have to expect a complex structure of variability: clearly residual will be auto correlated and unobserved heterogeneity must be controlled. Recalling that for the years preceding t , all firms are observed, we can try to improve our estimation using a longer period of time, that is a time series of cross-section. This perspective move us towards models for panel data, which requires not only the specification of a general model, but the structure of time dependence too. Besides, if we consider the whole estimation procedure we are compelled to identify also the form of sectional correlation (the $\mathbf{V}_{rs}$ matrix), so that the whole variance-covariance matrix is a mixture of time and selection dependence.

The model must be modified in order to fit this new conceptual scheme, more precisely we need a mixed model which contains both fixed and random effects parameters, and allows to specify a wide variety of covariance structure (Henderson, 1990; Searle, Casella, McCulloch, 1992; Littell et al. 1996; Verbeke and Molenberghs, 1997)

The mixed model extends the general linear model by allowing a more flexible specification of the covariance matrix of ε . In other words, it allows for both correlation and heterogeneous variances.

The mixed model is usually written as:

$$O_{it} = \beta' O_{i,t-1} + \gamma' \mathbf{Z} + \varepsilon_{it}$$

where everything is the same as in the above general linear model except for the addition of the known design matrix, $\mathbf{Z}$, and the vector of unknown *random-effects parameters*, γ . The matrix $\mathbf{Z}$ can contain either continuous or dummy variables, just like $\mathbf{X}$. In our example, $\mathbf{Z}$ will consist of an individual dummy and a set of variables describing the characteristic of the firms (sector of activity, size at a certain time, complexity, skill etc.). The key assumption on residual is much less restrictive than in general linear models, namely they are:

$$E \begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad Var \begin{bmatrix} \gamma \\ \varepsilon \end{bmatrix} = \mathbf{V} = \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}$$

Where $\mathbf{G}$ contains intra and inter individual heterogeneity, and $\mathbf{R}$ collect the time dependence of the phenomenon. The variance of $\mathbf{O}$ is, therefore, $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, and we can model $\mathbf{V}$ by setting up the random-effects design matrix $\mathbf{Z}$ and by specifying covariance structures for $\mathbf{G}$ and $\mathbf{R}$. As can easily seen, this a quit general formulation of the model, for example simple random effects are a special case of the general specification with $\mathbf{Z}$ containing only dummy variables, $\mathbf{G}$ containing variance components in a diagonal structure, and $\mathbf{R} = \sigma^2 \mathbf{I}$. The general linear model is a further special case with $\mathbf{Z} = \mathbf{0}$ and $\mathbf{R} = \sigma^2 \mathbf{I}$

Estimation is more difficult in the mixed model than in the general linear model. Not only we have β as in the general linear model, but there are unknown parameters in γ , $\mathbf{G}$, and $\mathbf{R}$ as well. Least squares is no longer the best method. Generalized least squares (GLS) is more appropriate. However, it requires knowledge of matrix $\mathbf{V}$ and, therefore, knowledge of $\mathbf{G}$ and $\mathbf{R}$. Lacking such information, one approach is to use *estimated*

GLS, in which some suitable estimation of $\mathbf{V}$ have to be inserted. The goal thus becomes finding a reasonable estimate of $\mathbf{G}$ and $\mathbf{R}$. In the most of cases, the best (often the only) approach is to use *likelihood-based* methods, exploiting some assumption on γ and ε distribution (Hartley, Rao, 1967; Patterson, Thompson 1971; Harville 1977; Laird, Ware, 1982; Jennrich, Schluchter, 1986). In addition, like Verbeke and Molenberghs (1997) clearly state, no satisfactory solution is available for mixed models estimation in presence of selection bias.

At the beginning of the present section we stated that, in predictive inference, the main condition for the correctness of the estimates is the correctness of the specified model together with the correct estimation of variance-covariance matrix. In the light of the description of an empirical example of estimation it is clear that a “well specified model” requires a composite set of assumptions about the statistical nature of the phenomenon we are interested in. The concern is not simply to state a relation linking some variables, but it is strictly necessary to provide hypothesis about time dependence, kind of intra and inter individual heterogeneity, a-priori distribution of random-effects parameters and about error term, ignorability of selection process.

On the other hand, as is the case, if the selection is informative neither validation of the model nor useful information about variability structure can be obtained by observed data

4. The needs for integration

As described above, the first need for integration stems from the correct specification of the model. A second instance to be considered depends on the particular nature of economic data we deal in. It is well known that economic activity is characterized by a considerable turn-over: new firms are created, some of the existing ones die, others change sector of activity or legal status, and so on. Such changes hardly can be included in an autoregressive model, and we have to devote special attention to this topic. Besides, in our example we must expect changes in the way the firms provide their data, so that the border between observed and unobserved units changes during the time.

The more natural answer to these questions, is a direct observation of the behaviour of units uncovered by AS, selected using an ignorable selection mechanism, that is to say a DS based on a sampling scheme. At a first glance, this statement appears to be a paradox. If we want to replace DS, we still need more DS. The scenario seems now to be: we use AS for some population subset and DS for the rest, where, referring to the example, “the rest” should probably be composed by small and less structured firms, difficult to observe in any case.

The paradox can be bypassed considering carefully the differences between traditional DS and the ones required to supplement the use of AS. Main differences concerns organization of the survey, sample size and design, estimation procedure. A sketch of such differences is given below.

- a) Lack of timeliness as we stated before, determines a selection bias. In order to overcome this problem, we need data about unobserved units. It is to be stressed that they do provide the data in the same period of the observed ones, but in non computer readable format and the amount of data requires a long time for recording information. Hence there is no real necessity to interview firms. An

adequate sample of paper forms can be selected and registered, so that cost are highly reduced and respondent burden is null. Note that such a strategy can provide estimation for births and deaths too.

- b) the unique solution only for two distinct problems: population undercoverage and variables not considered in administrative forms. Nevertheless, an accurate recognition of all available AS and political agreement with administrative offices in order to modify or add forms, could probably reduce the amount of necessary DS.
- c) An irreplaceable role for usual DS is, on the contrary, in measuring and controlling quality of the AS data. Tools like reinterviews aimed at evaluating response bias, simple and correlated response variance could become a *routine*, exploiting resources nowadays devoted to implement large surveys.
- d) Primary interest in order to validate our models for inference, as we have stressed, is on individual changes in time, hence DS have to adopt some kind of panel design. In such survey there is a great opportunity to improve timeliness and data quality and to reduce respondent burden using new technologies (Filippucci 2000).
- e) The main reason justifying DS rely on the estimation and validation of the parameters of the model, not on population totals or means, which will be estimated using the model. As is well known, regression coefficients and covariance estimators are, in general, more efficient than estimators of means and totals, because they exploit the correlation between the two (or more) variables linked by the model. Hence, *coeteris paribus*, the same sample size gives better estimates.
- f) We have a set of robust hypotheses based on complete (past) and partial (contemporary) observations, and we are looking for evidences confirming their validity in the present time and for unobserved units. This implies that the only task required to sampling is to ensure ignorability of the selection mechanism. In this sense selection is driven by the estimation process through, for example, tests and indices for goodness of fit or for model selection. As a consequence, it is possible to consider sampling schemes without fixed size, like sequential samples or similar.
- g) Past experience can be exploited in order to identify the selection bias of available data. As a consequence we should be able to estimate the selection bias and to identify the units with the higher probability to cause such a bias on the basis of influential analysis, for instance. These units could be included with probability 1 in the sample.
- h) Dealing with recording of paper administrative forms, sample size must be chosen considering different factors than in usual surveys. The marginal cost of observation is negligible, while time become the crucial constraint.

5. Conclusion

In this paper we presented a general strategy to obtain estimates from AS including the main approaches to model estimation especially when we deal with longitudinal information. The methodological framework we outlined authorises a large utilisation of the AS under two conditions: being able to identify the mechanism of units selection

and having or collecting appropriate supplementary information to evaluate and correct the bias. As we have shown that a complete substitution of DS with AS is merely an utopia, but massive utilization of administrative requires DS, although with a different role. First of all surveys should provide data in fields where we do not have other information and those information necessary to specify and check models to be used in making inference. Secondly, DS could become a strong check for administrative data quality. To great extend official statistics still rely on independent surveys, but in the context of a national statistical system information should stem first of all from statistical utilisation of AS and survey sampling should be finalised to acquire information to integrate, update and checking. National statistical system will more and more be characterised by utilisation of many data bases, more or less complete, we need to integrate and to check in view of ensuring appropriate data quality.

The impact on costs and timeliness of the outlined integration must be carefully evaluated and controlled experiment have to planned, nevertheless, considering the opportunity to sample directly paper forms, it is reasonable to presume a relevant reduction required resources. Main problems will arise in a different context. In fact, this kind of survey, are rather different from the current ones, the main goal is confirmatory analysis of given models, so that efficient sampling design can be found not in the field of usual sampling theory, but much more in the field of model assisted sample survey and estimators to be used belong to the “design and model unbiased” set. Statistical agencies and users have limited experience in this field and probably a lot of work is required in order to change the current point of view on data production process.

References

- Bolfarine H., Zacks S. (1992), *Prediction theory for finite populations*, Springer-Verlag.
- Cassel C. M., Sarndal C.E., Wretman J. H. (1977), *Foundations of Inference in Survey Sampling*, Wiley.
- Deming W.E. (1950), *Some theory of sampling*, Wiley.
- Drudi I. (1997), L'impiego dei dati fiscali per l'analisi delle piccole imprese, in: *Atti del convegno SIS La statistica per le imprese*, Tirrenia Stampatori, Torino.
- Drudi I., Filippucci C. (1994), Poststratified estimates using repeated observations, the case study of farm accounting data network, in: *Scritti di Statistica economica*, Quintano C. (a cura di), Curto editore.
- Drudi I., Filippucci C., (1996), Inference from longitudinal non-random surveys: a case study, *Rivista di statistica applicata*, n.1.
- Drudi I., Filippucci C., (1998), *Inference from non-random samples with partial overlap*, Proceedings of NTTS '98, Sorrento.
- Filippucci C. (1996), *A Predictive Approach to Estimating Agricultural Production from Farm Accounting Data in Italy*, Proceedings 5° Independent Conference of International Association of Official Statistics, Reykjavik.
- Filippucci C. (2000), Strategie e tecniche di formazione dei dati. Un quadro delle problematiche e delle prospettive, in: *Tecnologie informatiche e fonti amministrative nella produzione dei dati*, Filippucci C. (editor), Franco Angeli, Milano.
- Hartley H.O., Rao, J.N.K. (1967), Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model, *Biometrika*, 54, 93–108.

- Harville D.A. (1977), Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems, *Journal of the American Statistical Association*, 72, 320–338.
- Henderson, C.R. (1990), Statistical Method in Animal Improvement: Historical Overview, in: *Advances in Statistical Methods for Genetic Improvement of Livestock*, Springer-Verlag, 1–14.
- Jennrich, R.I., Schluchter, M.D. (1986), Unbalanced Repeated-Measures Models with Structured Covariance Matrices, *Biometrics*, 42, 805–820.
- Laird, N.M. and Ware, J.H. (1982), Random-Effects Models for Longitudinal Data, *Biometrics*, 38, 963–974.
- Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996), *SAS System for Mixed Models*, SAS Institute Inc.
- Patterson, H.D. and Thompson, R. (1971), Recovery of Inter-Block Information when Block Sizes are Unequal, *Biometrika*, 58, 545–554.
- Royall R.M. (1970), On finite population sampling theory under certain linear regression models, *Biometrika*, 52.
- Royall R.M. (1988), The prediction approach to sampling theory, in, *Handbook of Statistics*, Krishnaiah, Rao (editors), vol. 6, North-Holland.
- Royall R.M., Cumberland W.G. (1981), An empirical study of the ratio estimator and estimators of his variance, *Journal of American Statistical Association*, n.76.
- Royall R.M., Herson J.H. (1973), Robust estimation in finite populations, *Journal of American Statistical Association*, n.68.
- Royall R.M., Pfeffermann D. (1982), Balanced samples and robust Bayesian inference in finite population sampling, *Biometrika*, n.69.
- Sarndal C., (1978), Design Based and Model Based Inference in Survey Sampling, *Scandinavian Journal of Statistics*, n.5, pp. 27-52.
- Sarndal C., Swensson B., Wrettmann J. (1992), *Model assisted Survey sampling*, Springer-Verlag.
- Searle, S. R., Casella, G., and McCulloch, C.E. (1992), *Variance Components*, Wiley.
- Smith T.M.F. (1983), On the validity of inferences from non-random sample, *Journal of Royal Statistical Society, A*, vol. 146.
- Smith T.M.F. (1991), Post-stratification, *The Statistician*, n.40.
- Smith T.M.F. (1994), Sample survey 1975-1990, an age of reconciliation?, *International Statistical Review*, vol.62, n.1.
- Smith T.M.F. (1999), *Recent Developments in Sample Survey Theory and their impact on Official Statistics*, Proceedings of 52th ISI session, Helsinki.
- Smith T.M.F., Sudgen R.A. (1988), Sampling and assignment mechanism in experiments, surveys and observational studies, *International Statistical Review*, 56.
- Verbeke, G. and Molenberghs, G., eds. (1997), *Linear Mixed Models in Practice: A SAS-Oriented Approach*, Springer.