Collecting data through surveys only when all else fails? The case of surveys in the ECB

Sebastian Pérez-Duarte[1]

European Central Bank, Frankfurt, Germany sebastien.perez-duarte@ecb.int

When met with user requirements involving the provision of new data, statisticians have to determine the "best" source for such information, through the reuse of existing information, the adaptation of existing sources, or the creation of a new collection exercise. The solution is determined by matching the costs of the different possibilities and their benefits – in essence, finding the solution that minimizes costs (with a broad definition thereof) for a given utility of the data. We sketch a general model of data collection and data use, taking into account two main abstract dimensions of data, namely the quantity of information (depth, e.g. number of dimensions or variables) and the precision of the measurement (e.g. quality, bias, standard error). Administrative data and survey data, as well as existing or new data, can be all fitted in this model. We apply this theoretical framework to recent data collection exercises in the statistical department of the ECB.

Key Words: administrative data, utility, cost, quality, census, survey

1. Introduction

It is trivial to write that data are collected or produced and (hopefully) used. What data producers prepare, and what data users utilise are not only two sides of the same coin, but are best designed one with the other. In this paper we attempt to model on the one hand the production process of data in terms of the costs of production given the characteristics of the data, and on the other hand the use of the data, which we summarise with the utility drawn from the data and its associated characteristics. Good data are mostly likely more costly to produce, but may allow greater value to be extracted from them. The central question in this paper is how the user can ultimately choose which data source best fit his needs for the minimum cost.

The first caveat we must raise is that our model will be a gross simplification of reality: data quality is the result of a large number of decisions, and cannot be summarised in a couple of dimensions. Likewise, the utility provided by a particular dataset is an abstract concept, which, although intuitively understandable, is difficult to pin down.

There is no obvious measure for quality. In the case of a sample survey, bias and variance are perhaps the most spontaneous measures, but they apply to one variable at a time. The particular variable to use would of course depend on the context: all datasets contain several variables, and the quality in this multivariate setup is harder to pin down.

Brackstone (1999) describes the management of data quality in the context of official statistics along six dimensions: relevance, accuracy, timeliness, accessibility, interpretability, and coherence. We consider in this paper two

[1] All views expressed are those of the authors alone and do not necessarily reflect those of the ECB or the Eurosystem.

main dimensions, into which we can subsume most of Brackstone's dimensions: *depth* and *precision*. We call "*depth*" a measure of the amount of information that is available in the data, taking into account its relevance, interpretability and coherence. This could be for example based on the number of variables, the number of categories, or the frequency.

"*Precision*" is a measure of accuracy, and would take into account the different sources of error, either systematic or random (sampling, non-response, etc.) in the spirit of Groves and Lyberg (2010) and Lyberg (2012). Precision allows a phenomenon to be measured and decisions to be taken on this basis. Bias, variance, rounding, and bunching are all elements of this concept, which we will not attempt to specify more clearly here.

At one corner of this depth and precision space we could place focus groups, where extremely detailed information on thought processes is collected on a very limited number of participants; at the other extreme we could place a census, which collects a limited set of variables on the whole population.

In this paper we outline administrative and survey data, then sketch a model of data production and use, before applying it to recent ECB developments.

2. Administrative and survey data

The main dichotomy we will consider in the rest of the paper is between two general families of data, administrative on one hand and survey on the other. In reality the frontier between administrative data and survey data is blurrier, but in debates over the source of new data, the choice whether to consider a new data collection or tap into administrative data is often one of the first decisions, and we will focus our exposition on these two options.

Administrative data have been the source of statistics for many years. Nordbotten (2008) describes enumeration of resources in Babylon and by Egyptian pharaohs. Administrative data would seem to allow high precision, and in many cases contain much information, and it is thus no wonder that such data sources, where available, have been integrated into statistical outputs. Although administrative data have many qualities, they are not the be-all and end-all of statistics (see Groen 2012 for the analysis of sources of errors in survey and administrative data). One of the central questions in administrative data is the trade-off between the quality of statistical output and the respondent burden (Daas et al. 2008; Kavaliauskiene 2010). Administrative data may be of poorer quality than comparable survey data (Blank et al. 2009). Of course, some information cannot be collected from administrative data. We attempt to answer the question of administrative *vs.* survey data in a simple model of cost and utility.

Cost and utility in the context of data

Producing a dataset with a particular depth and a particular precision has a cost. We assume that there is either one or a set of data production functions, and that for any combination of depth and precision the cheapest alternative can be found.

A cost function for survey data is not impossible to imagine. Contacting persons, households, firms, and conducting a survey has obvious and tangible

costs, which depend on the complexity of the questionnaire and the number of questions – which we consider as the depth – as well as on the sample size and other design elements – which we take to be closely related to the precision.

For administrative data the costs are less evident – after all, the data have already been collected when the user plans on using it. Nevertheless, such data still need to be cleaned and processed for statistical use, and even though they may be collected for other purposes, they still impose a response burden on respondents, and have thus a cost which is also increasing in the depth and the number of respondents (a crude measure of precision).

Once a dataset has been produced, it may be used, and found to be useful, or not. There is no obvious measure of the utility provided by a dataset. Spencer (1985) introduces a risk function to proxy for this, but in this paper we only consider an abstract utility function, which will positively depend on depth and precision, with a trade-off between the two.

3. A model of data production and data use

We consider thus the two main dimensions of data, namely depth (d) and precision (p).

The cost $C(d,p)$ of producing data with given depth and precision is increasing in depth and precision. This implies that the frontier of this data production function is non-increasing. Intuitively, increasing depth while keeping precision fixed gives higher costs of collecting but makes the data more useful. Administrative data may not be directly available in various depths and precisions, but from a given dataset one can construct for the same price less precise and shallower versions of it.

We will assume that the user extracts a utility U from the dataset, which depends on the same depth and precision. The higher the precision, the higher the utility; the deeper the data, the higher the utility. Decreasing returns to precision and depth, probably. At any point, users might be tempted to trade off more precision with less depth, so a constant U curve in (d,p) space is downward sloping and convex. The slope is related to the relative preference for precision and depth.[2]

The dataset selected is the result of maximising the utility for a given cost, or similarly minimising cost for a given utility. Such a maximum, as well as the cost and utility functions, are displayed in Figure 1. A change in the utility, for example by an increased preference for precision over depth, will make the utility curve steeper, and lead to data with higher precision, at the expense of lower depth.

A smooth production frontier, as shown by the isocost curve in Figure 1, is more likely the result of a smooth process, e.g. the trade-off within one survey

[2] Without entering into specifics, it is possible to provide a more explicit formulation to the model before. Precision can be taken as the inverse of the width of confidence intervals, and is thus proportional to the square root of the sample size, which implies that costs are then related to the square of the precision. Depth is then only the number of questions. By taking into account fixed costs for the setup of a survey, costs can be specified as $C(p,d) = a_1 + a_2 p^2 + a_3 d + a_4 p^2 d$ while utility can be written as $U(p,d) = p^\theta d^{1-\theta}$.

between the length of the questionnaire and the sample size. Introducing an existing administrative dataset, available for the same cost, will change the production frontier as displayed in Figure 2. In this example, the administrative dataset has much higher precision but somewhat rather low depth. Depending on the relative appetence for precision, the administrative dataset is likely to be chosen.

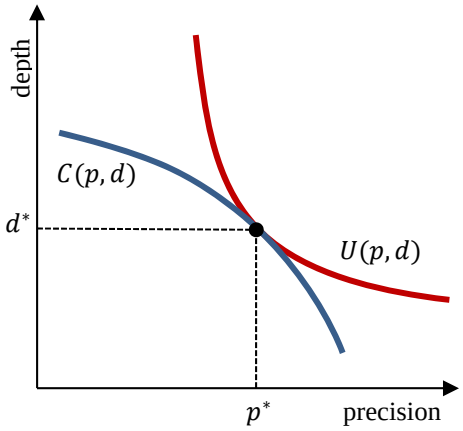

Figure 1: Depth-precision space, cost and utility functions

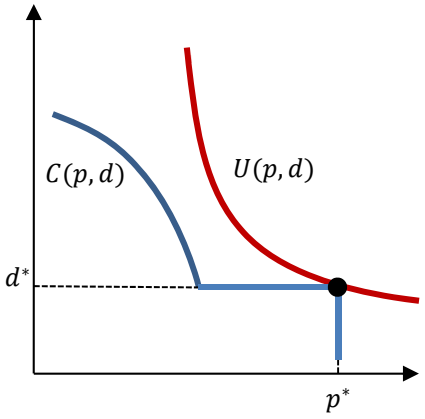

Figure 2: Mix of survey and administrative data

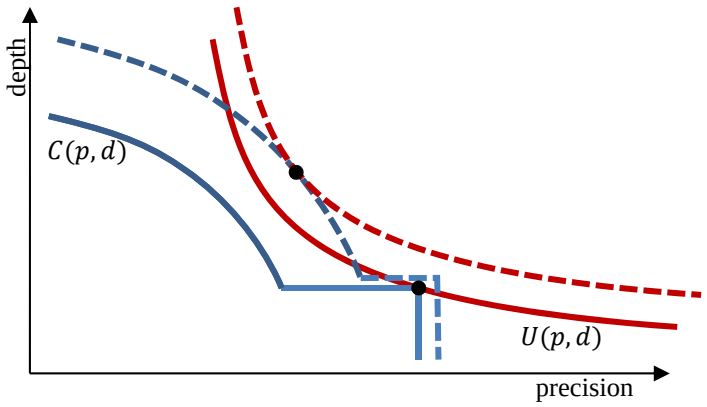

Figure 3: Varying the total budget available in a mixed survey

While the descriptions so far have taken the total cost as constant, in applications the cost is also a decision factor, in the sense that if substantial increases in the variance can be achieved with limited increase in the cost, different parameters for the optimisation can be set. In this respect, the gradient of the production function is the crucial element. We argue that the gradient is much steeper for administrative data – if not specifically the monetary cost, the process and the length of time required to change an administrative collection make such changes much costlier. Two isocost curves are shown in Figure 3, where the dotted lines show the outcome of an increase in the budget available, and the corresponding shift from administrative to survey data. The opposite is of course also possible – an increased pressure on costs leads a survey to be abandoned and the comparable administrative survey to be developed.

4. Application to ECB data collection exercises

The model above can be applied to recent and less recent developments at the ECB, with the on-going work on the EC/ECB survey on the access to finance of SMEs (SAFE), the Eurosystem Household Finance and Consumption Surveys (HFCS), and the ECB Monetary and Financial Institutions Interest Rates (MIR).

"Pure" surveys: household survey, enterprise survey

Information on the financial situation of firms, available through the yearly financial statements of firms, is somewhat detailed and comparable (at least within each country). However, it suffers from a lack of timeliness, and is not always available for the smallest firms. Finally, comparability across countries is not high due to the different accounting regimes. In terms of the model of the previous section, this did not place administrative data sufficiently to the upper left-hand corner of the (p, d) space, and the SAFE survey of SMEs, a joint endeavour of the European Commission and the ECB, was developed: it is a survey of 7,500 to 15,000 firms in the European Union, assessing the financing conditions of firms bi-annually, through a set of qualitative questions.

Information on household assets and liabilities was so far mostly available through the sector accounts of the System of National Accounts. These accounts offer a comprehensive and consistent view of the balance sheet of the economy as a whole and the household sector within it. Nevertheless, these accounts do not provide any insight into the distribution of these assets and liabilities, nor to their characteristics. In most of the euro area, the Eurosystem HFCS was developed from the beginning as a survey, due to the lack of administrative data.[3]

Cutting of the tail and selection of the largest: the case of interest rates

The ESCB collects regularly statistics on interest rates (MIR) provided by Monetary and Financial Institutions (MFIs) in the European Union. The regulation providing the basis for these statistics[4] allows reporting countries to collect either from all MFIs (census) or from a sample of MFIs, in this later case either by purposive selection of the largest MFIs after suitable stratification of the MFIs, or by random sampling (possibly with probability proportional to size). The selection between one of the different options was left to the National Central Banks, and resulted in varying methods to stratify and sample MFIs. Although this is out of the scope of this paper, the model presented above would be helpful in illuminating both the setup of the MIR in the regulation and its implementation by member states – in the case of interest rates, the variation between institutions is usually lower than the variation within, which allows a much higher precision for a given sample size. Huerga et al. (2013) study quality indicators for the MIR that apply to purposive sampling, and link the precision of the MIR to the variation and the potential bias in the interest rates.

---

[3] Exceptions are the Finnish and Danish register-based data.
[4] Regulation (EC) No 63/2002.

5. Conclusion

Central banks are used to conducting both surveys (e.g. for business sentiment) and censuses (e.g. of financial institutions). This paper proposes a simple model of data selection through the utility and the cost of each data source, which differ in their precision and the depth of information available. Mixing administrative and survey data causes the data production frontier to be non-convex, which leads to potential important changes in the preferred options when either preferences or budgets change. Surveys developed over the past 10 years by the Eurosystem show the usefulness of this data collection mode.

References

Andersen, A.L., A.M. Christensen, N.F. Nielsen, S.A. Koob, and M. Oksbjerg (2012), "The Wealth and Debt of Danish Families" *Danmarks Nationalbank Monetary Review*, 2nd Quarter 2012 part 2.

Blank, Rebecca M., Kerwin Kofi Charles, and James M. Sallee (2009), "A Cautionary Tale About the Use of Administrative Data: Evidence from Age of Marriage Laws", *American Economic Journal: Applied Economics*, Vol. 1, No. 2 (April), pp. 128-149.

Brackstone, Gordon (1999), "Managing Data Quality in a Statistical Agency", *Survey Methodology*, Vol. 25, No. 2 (December), pp. 139-149.

Daas, Piet J.H., Judit Arends-Tóth, Barry Schouten, and Léander Kuijvenhoven (2008), "Quality framework for the evaluation of administrative data", *Proceedings of Q2008 European Conference on Quality in Official Statistics.*

Groen, Jeffrey A.(2012), "Sources of Error in Survey and Administrative Data: The Importance of Reporting Procedures", *Journal of Official Statistics*, Vol. 28, No. 2, pp. 173–198.

Groves, Robert M. and Lars Lyberg (2010), "Total survey error: past, present, and future", *Public Opinion Quarterly*, Vol. 74, No. 5, pp. 849-879.

Ho, Frederick W. H. (2005), "Survey as a Source of Statistics and Factors Affecting the Quality of Survey Statistics", *International Statistical Review / Revue Internationale de Statistique*, Vol. 73, No. 2, (Aug.), pp. 245-254.

Hodges, James S. (1987), "Uncertainty, policy analysis and statistics", *Statistical Science*, Vol. 2, No. 3, August, pp. 269-291.

Hoffmann, Eivind (1995), "We must use administrative data for official statistics–but how should we use them?", *Statistical Journal Of The United Nations Economic Commission For Europe*, Volume 12.

Huerga, Javier, S. Pérez-Duarte and J.M. Puigvert (2013), "Quality measures in non-statistical sampling: MFI Interest Rates Statistics (MIR)", *Proceedings of the 59th World Statistics Congress*, Hong Kong.

Kapteyn, Arie, and Jelmer Y. Ypma (2007), "Measurement Error and Misclassification: A Comparison of Survey and Administrative Data", *Journal of Labor Economics*, Vol. 25, No. 3 (July), pp. 513-551.

Kavaliauskiene, Daliute (2010), "Use of Administrative Data. Efforts to Find Balance between Simplification for Respondents and Quality of Statistical Output", *Simply 2010: Conference on Administrative Simplification in Official Statistics*.

Lyberg, Lars (2012), "Survey quality", *Survey Methodology*, Vol. 38, No. 2, pp. 107-130, December.

Nordbotten, Svein (2008), "The Use of Administrative Data in Official Statistics – Past, Present, and Future – With Special Reference to the Nordic Countries", *Official Statistics in Honour of Daniel Thorburn*, pp. 205–223.

Spencer , Bruce D. (1985), "Optimal Data Quality", *Journal of the American Statistical Association* , Vol. 80, No. 391 (Sep.), pp. 564-573.