Impact and benefits of micro-databases' integration on the statistics of the *Banco de Portugal*

Paula Menezes[1], Luís D'Aguiar[2]
[1] Statistics Department, Banco de Portugal, PORTUGAL, e-mail: pamenezes@bportugal.pt
[2] Statistics Department, Banco de Portugal, PORTUGAL, e-mail: laguiar@bportugal.pt

Abstract

Data are critically important to good decision-making. In an increasingly complex economy, conventional data collecting schemes are no longer sufficient. To deal with the challenge of maintaining its statistics relevant to the users in an ever-shifting environment, the *Banco de Portugal* decided to explore the largely unused statistical potential of the available micro-databases and to integrate the existing administrative and survey data, thus enhancing the basic information infrastructure while protecting confidentiality. This presentation will address the benefits and problems to be dealt with when two or more data-sources are to be integrated.

Key Words: micro-data, infrastructure, integration, knowledge

Introduction

Economies are constantly faced with new challenges. To remain relevant, official statistics have to keep up with the rapid changes of modern times, which typically require the availability of commensurate statistical data that users may exploit in an accurate and reliable way. Policy-makers, financial supervisors and regulators, just to name a few, require as much rich and timely information as possible to take appropriate decisions.

The *Banco de Portugal* (hereinafter referred as "the Bank") – or any other major producer of official statistics, for that matter – has to ensure that the statistics for which the Bank is accountable retain relevancy over time and are able to cope with the speed and the scope of the main stakeholders' ever-increasing demand for comprehensive, detailed and high-quality information.

However, the process of continuously adapting the statistical output to new phenomena has a number of serious limitations. Conventional data collecting systems cannot simply keep on expanding indefinitely to cope with the ever-increasing need to fill the information gaps perceived by the users or in anticipation to their possible future data requirements. Amongst the possible motives for not pursuing recurrently this approach one could point out, *inter alia*, the following:

– The resulting overburdening of respondents goes against well-established best practices.

– The related initial and maintenance costs are far from being negligible, both to the agency that collects the data and to the respondents.

– New statistical datasets (or significant enhancements to existing ones) require lengthy preparation time (years, rather than months) and, once launched, are supposed to remain in operation for a prolonged period of time (typically around five years, in the case of the Eurosystem statistical reporting systems). This time-lag could even be further extended, should the revision result from a major methodological change, as it is often the case.

– *Ad hoc* surveys are, in general, too time-consuming and expensive, not to mention reliant upon the willingness to participate on the part of the target population.

In fact, the response given by conventional data collecting systems to new statistical demands – stemming from, *e.g.*, the need to conduct macro-prudential analyses or to accommodate new data requirements related to the Bank's participation in the European System of Central Banks (ESCB) – is problematic, costly and could possibly turn out to be counterproductive, which helps to understand why more and more central banks have been reusing the available micro-data, thus recognising that such information is both useful and necessary to respond to the data requirements of the complex world we live in, and to better address new issues and challenges as they arise.

Throughout this paper the term "micro-data" will be used to refer data about individual persons, households, businesses or other entities; it may be data directly collected by the Bank or obtained from other sources, such as administrative sources.

Changing the statistical paradigm

Data are critically important in making well-informed decisions. Poor quality data or, *a fortiori*, lack of data can lead to an inefficient allocation of resources and imposes high costs on the society. In ever more complex economies, the traditional approach to the compilation of official statistics – *i.e.*, producing standard statistical tables that can only address a set of predefined questions – is becoming increasingly insufficient and ineffective.

The Bank's strategy to deal with the challenge of maintaining its statistics relevant to the users in a shifting and more demanding environment, while attending to the need to keep the reporting burden on respondents at an acceptable level, was to enhance the overall efficiency of the statistical framework by further exploring the largely unused statistical potential of already existing data sources. In fact, statistically edited micro-data, which include *e.g.* data from administrative sources not originally intended for statistical purposes or even data related to the Bank's prudential supervision function, offer an unusual array of interesting features, *inter alia*:

– *Very good coverage* of the population in most of the cases.

– *Relatively low reporting costs*, thus helping to mitigate the constraints imposed by the response burden of the reporting agents.

– *Increased flexibility and agility* as regards the compilation of new statistics, *e.g.* related to financial and other structural innovations.

– *More rapid response to* ad hoc *data requirements* from the users – in many cases, almost in real time.

Moreover, the evolution in network and communication protocols, database systems and multidimensional analytical systems has somewhat removed the potential disadvantages of having to deal with the huge amounts of data normally associated with the handling of micro-databases. (Aguiar *et al.*, 2011)

Best practices in compiling official statistics advocate that all data should be collected only once: any form of double reporting or redundant collection should be avoided and, if existing, be terminated. Accordingly, data already available – due to whatever reasons – should be reused, if found useful, for statistical purposes. Obvious candidates are data from existing Central Credit Registers, as well as data from Central Balance Sheet Offices databases and information collected within the framework of the Bank's prudential supervision function. The experience of the Bank in this area has shown that the use of such information for statistical purposes can lead to a significant reduction of the response burden, higher data quality and lower costs.

On the national level, a formal exchange of administrative data with institutions outside the central bank, like the national statistical institute (NSI) or the tax authorities, would also help to reduce the reporting costs. An important precondition would be the maintenance of common company registers with the NSI. Extending this idea across national borders, one could think of common international databases – *e.g.*, exchanging micro-data on significant cross-border mergers and acquisitions that need

to be recorded symmetrically in the respective statistics of both affected countries. (Liebscher *et al.*, 2008)

Micro-databases managed by the *Banco de Portugal*

For the last 10 years the Bank has been developing and maintaining several micro-databases based on item-by-item reporting and has been exploring the statistical potential of these complementary sources of information with significant positive impacts on the overall quality of its statistical output.

The databases managed by the Bank's Statistics Department include:

– The *Securities Statistics Integrated System* (SSIS) database, a security-by-security and an investor-by-investor database that provides, in a single repository, data on the securities issues and holdings required by the different statistical domains (*e.g.*, monetary and financial statistics, external statistics, securities statistics and financial accounts), thus replacing the separate and distinctive data storing systems that were previously in place.

– The *Central Credit Register* (CCR), an administrative database that stores credit-related information supplied by all the resident credit-granting financial institutions.

– The *Central Balance Sheet Database* (CBSD), which stores granular information on virtually all the resident corporations, collected through the so-called *Informação Empresarial Simplificada* (IES), a joint effort of four distinct Portuguese public entities – the Ministry of Finance, the Ministry of Justice, *Instituto Nacional de Estatística* (the Portuguese NSI) and the *Banco de Portugal* – consisting of yearly submissions of information by corporations, in a single, paper-free, electronic form, to fulfil reporting obligations of accounting, fiscal and statistical nature.

Besides complementing and helping to cross-check the information gathered through the conventional channels, these micro-data have proved to be of great importance to the understanding of the developments in the Portuguese financial system, especially in the wake of the recent global financial crisis.

So far, this approach has permitted, *inter alia*:

– *Improving the responsiveness to new users' requirements*, particularly those arising from *ad hoc* information requests, with proven results in reducing or eliminating data gaps and in monitoring and assessing the evolution of the Portuguese financial system.

– *Curtailing the follow-up procedures as regards data collecting schemes*, whereby respondents are re-contacted after the initial submission of data, to obtain missing information and/or to verify and, if necessary, to correct questionable data.

– *Enhancing the quality control procedures* (*e.g.*, by cross-checking elementary/raw data from different statistical domains), thus increasing the efficiency of the production process and improving the quality of end-products.

– *Avoiding data redundancy*, while at the same time expanding significantly the range of statistics available.

As an example, the use of the available micro-databases for the compilation of the Portuguese flow-of-funds within the national financial accounts has been extremely helpful, as it allows for a much better understanding of the interlinks within the resident economy and *vis-à-vis* the rest-of-the-world.

Deepening data integration

In keeping with such course of action, the Bank has been developing an approach that, once completed, will allow for a higher level of integration of the available administrative and survey data. The goal is to achieve a significant enhancement of the basic data infrastructure without jeopardizing the provisions in the legislation, codes of practice and protocols that protect data confidentiality. In addition, a reduction in

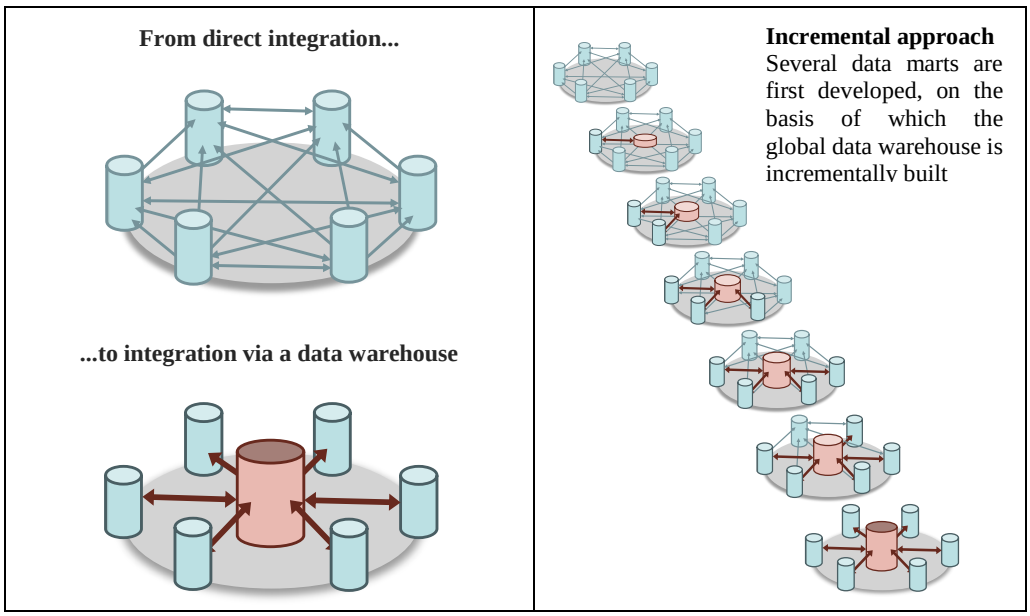
respondent burden and an increase in the breadth and depth of the information available to policy-makers and researchers are expected.

An architecture based on business intelligence – broadly defined as a category of applications (*e.g.*, decision support systems, query and reporting, online analytical processing, statistical analysis, forecasting, and data mining) and technologies for gathering, storing, analyzing, and providing access to data, to help a variety of users to make better business decisions (Terzić, 2008) – could significantly contribute to meeting the Bank's concerns in this area. With this in mind, the Bank set off a study in 2008 aiming at defining a business intelligence framework to be used as a reference in all future information technology developments in the statistical realm. This framework will be built upon three pillars (Aguiar *et al.*, 2011):

– A *common technological infrastructure* across the various information systems, to facilitate the integration and re-usage of components and to promote data access efficiency and transparency to final users.

– A *centralised reference database*, to provide common reference data (*e.g.*, identification criteria for the relevant entities that are observed, characterization of variables and classifications) and to enable the linkage of information across different sources and systems.

– A *data warehouse approach*, to guarantee a central access point to all statistical data, independently of the input source or the production process. This implies, *inter alia*, a data structure specified on the basis of common criteria, valid across the different sets of data.



At the moment, the Bank's statistical information subsystems are in the process of being reformulated according to this model: the SSIS and the balance of payment and international investment position statistics, on one hand; the CBSD and the CCR, on the other hand.

Data integration is concerned with integrating unit record data from different administrative and/or survey sources to compile new official statistics which can then be released in their own right. Integration of micro-data is a powerful approach to enrich the already available information – *e.g.*, by allowing efficient cross-data comparisons and quality checks among the different statistical domains. Surely, this is easier said than done – it is a rather complex process and, pending on the degree of integration to be achieved, it can be characterized by different features.

The prevailing, stand-alone, islands of information may be very diverse, making it technically difficult to create homogeneous information systems. In addition, there are

many different *practical and methodological problems that must be previously addressed* when two or more sources are to be integrated, *inter alia* (Di Zio, 1998):

– Harmonising populations – *e.g.* determining the group of entities that belong to a given institutional sector (financial and non-financial corporations, general government, households and non-profit institutions serving households) –, identification criteria, reference periods (annually, quarterly,...), variables and classifications.

– Adjusting for measurement errors (accuracy) and for missing data.

– Deriving variables.

However, such shortcomings may very well be offset by the possible *benefits of integrated data sets*. The latter include, according to UNECE (2009):

– Compiling new or enhanced statistics.

– Producing more disaggregated information for measures where some information currently exists.

– Carrying out research using composite micro-data that cover a wider range of variables for a larger number of units than available from any single data source.

– Potentially improving or validating existing data sources.

– Possibly reducing respondent burden.

These benefits could be illustrated by the following case, extracted from the Bank's own statistics: a given corporation, providing annual accounting data under its IES reporting obligations, might also be answering to the Bank's ISII survey (*Inquérito sobre Investimento Internacional*) and, at the same time, having its securities issues and holdings recorded in the SSIS database; in an integrated system, it would be possible to ensure the compatibility of these data at a micro-level, thus providing a powerful tool for the compilation of financial accounts (which require that total uses equal total resources in the domestic economy). Nonetheless, as referred above, having a partial integration, *e.g.* one that allows for a unified view on two different sub-systems like the CBSD and the CCR, clearly enriches analytical data awareness. In fact, the idea of pooling together all the data on financial or non-financial corporations available at the *Banco de Portugal* is rather appealing; it would allow us to have a more specific and detailed view on this particular institutional sector. In the case of the financial corporations, the advantages would be even stronger should we take into account the information of yet another important subsystem: the data used for supervisory purposes.

Concluding remarks

The implementation of an architecture framework such as the one summarized above will contribute to the construction of a coherent and integrated statistical system as opposed to having multiple systems that coexist but are not connected in an efficient way.

Such approach has only been possible because of the possibilities brought in by the information technology (IT) revolution. But even though IT has enabled the statistical community to carry out the current procedures for collecting, compiling and disseminating statistics more efficiently, albeit at a non-negligible cost, it is important to reflect on how such revolution can be used to introduce new and more effective procedures.

Benefits are evident but there are also problems, challenges and cautions with the use of integrated micro-data, particularly those related to confidentiality issues. As said before, data already available should be reused if found useful for (other) statistical purposes; that being the case, it is necessary to strictly safeguard their confidentiality and to ensure that the sharing is legally allowed or explicitly agreed by the reporting agents. However, because of legal constraints, confidentiality makes the access to some useful data sources problematic and disclosure is a constant problem when we need to release data.

A data integration process is complex and can be characterized by different steps. One of these steps is adopting a unified view on the existing micro-data data sources creating a customized view on a sub-set of data (*e.g.* the financial or non-financial sectors).

Integrated micro-data have the potential to support, if need be, the drilling down of the most summarized levels of data to the most detailed ones, which may help to confirm (or to disprove) trends and developments conveyed by macroeconomic statistics and, concomitantly, to explore and/or to elucidate their possible implications for *e.g.* financial stability analysis and systemic risk assessment. (D'Aguiar *et al.*, 2011)

References

Aguiar, M. & Martins, C. (2011). "Adding business intelligence to statistical systems – The experience of Banco de Portugal". *Eurostat Conference on "New Techniques and Technologies for Statistics".* Brussels, February 2011.

D'Aguiar, L., De Almeida, A., & Casimiro, P. (2011) "Promoting enhanced responsiveness to users' data needs: the experience of the Banco de Portugal in exploring the statistical potential of micro-databases". *Proceedings of the 58th ISI Session. Dublin, August 2011.*

D'Aguiar, L. & Lima, F. (2009) "Credit risk transfer – Dealing with the information gap". *The IFC's contribution to the 57th ISI Session, Durban, August 2009.* IFC Bulletin No 33, August 2010.

Di Zio, M. (1998) "Integration of micro-data: benefits and challenges". *ESCAP Workshop on Census and Survey Microdata Dissemination: Benefits and Challenges.* Bangkok, 18-20 June 2008.

Koch, C. (2001) "Data integration against multiple evolving autonomous schemata". *PhD thesis in Computer Science at TU Wien, Austria.* May 2001.

Liebscher, K. & Schubert, A. (2008) "Torn between new data needs and respondents' fatigue – Are efficiency gains the philosopher's stone?" *4th ECB Conference on Statistics.* Frankfurt am Main, 24-25 April 2008.

Terzić, A. (2008) "Business Intelligence Solutions – Cognos BI 8". Comter Systems, Inc., Fairfax, August 2008.

UNECE (1992) "Fundamental principles of official statistics in the UNECE region".

UNECE (2009) "Principles and guidelines on confidentiality aspects of data integration undertaken for statistical or related research purposes".