BANCA D'ITALIA

EUROSISTEMA

Questioni di Economia e Finanza

(Occasional papers)

Remote processing of firm microdata at the Bank of Italy

by Giuseppe Bruno, Leandro D'Aurizio and Raffaele Tartaglia-Polcini

December 2008

Number 36

The series Occasional Papers *presents studies and documents on issues pertaining to the institutional tasks of the Bank of Italy and the Eurosystem. The* Occasional Papers *appear alongside the* Working Papers *series which are specifically aimed at providing original contributions to economic research.*

The Occasional Papers *include studies conducted within the Bank of Italy, sometimes in cooperation with the Eurosystem or other institutions. The views expressed in the studies are those of the authors and do not involve the responsibility of the institutions to which they belong.*

The series is available online at www.bancaditalia.it.

REMOTE PROCESSING OF FIRM MICRODATA AT THE BANK OF ITALY

by Giuseppe Bruno*, Leandro D'Aurizio*, Raffaele Tartaglia-Polcini *

Abstract

Providing the possibility to run personalised econometric/statistical analyses on the appropriate data sets by remote processing allows greater flexibility in the production of economic information. Binding confidentiality requirements are required with business survey data. The Bank of Italy's infrastructure allows its business survey data to be exploited, while preserving anonymity of individual data. The system is based on the LISSY platform and has been already adopted by the Luxembourg Income Study (LIS) and other research centres. Firms' privacy is safeguarded by forbidding potentially confidentiality-breaking programme statements and by denying the visualisation of individual data. Data confidentiality is protected by removing key identifiers from the database and by trimming data in the right tail of the distribution. The platform provides its services through plain-text e-mails. The authorised user sends an e-mail containing an identifying header followed by a statistical programme to a predetermined address. The system checks the validity of the header, strips out the code and submits it in a batch to one of the econometric/statistical packages available (SAS and Stata). The outputs are mailed back to the user after passing an array of automatic and manual checks.

JEL Classification: C81, C88.
Keywords: microdata, confidentiality, remote access.

Contents

---

* Bank of Italy, Economic Research and International Relations. Corresponding author: Raffaele Tartaglia-Polcini Via Nazionale 91, 00184 Rome, Italy (email: raffaele.tartagliapolcini@bancaditalia.it). The authors wish to thank Giovanni D'Alessio for his precious comments. The opinions expressed in this paper are the authors' only and do not necessarily reflect those of the Bank of Italy. Sections 1 and 2 should be attributed to Raffaele Tartaglia-Polcini, Sections 3 and 5 to Leandro D'Aurizio and Section 4 to Giuseppe Bruno.

1. Introduction: from the surveys to the statistical research services

The Bank of Italy started regular sample surveys in 1965, when the first wave of the Survey of Household Income and Wealth took place. In 1974 the interviews for a survey of manufacturing firms were conducted for the first time by the Bank's branches. Since then, resulting microdata have been extensively used by economists at the Research Department of the Bank of Italy for policy use as well as for economic research.

The goal has never been to compete with official aggregate statistics, rather to take advantage of direct access to respondents and the possibility of laying out questionnaires tailored to the Bank's research needs[1]. Moreover, microdata availability made possible to exploit the panel dimension of the phenomena of interest. Aggregate statistical tables for both surveys have been regularly published on the Bank's Annual Report; more detailed results of the household survey, with standard tabulations and aggregations, have been published on a regular basis. On the contrary, detailed results from the business surveys were circulated only internally and, since 1997, also to the firms taking part in the interviews. Only from 2003 were detailed results of the business surveys published in form of a dedicated report.

Data from the household surveys have been made available to interested researchers, on request, since the early nineties, initially on tape, then on floppy disks and later on cd-roms. Data from the business surveys had never been made available outside the Bank until 2008.

Empirical research in economics has for a long time suffered from the unavailability of microdata. When this opportunity came to light, different solutions were adopted. The goal of a wider but confidentiality-compatible access to microdata as a strategy to foster economic research on Italian economic data is part of the Bank's commitment to transparency and accountability. This paper gives a synthetic account of the latest accomplishments at the Bank of Italy on external access to microdata, with special reference to sample business data. The new system BIRD (Bank of Italy Remote access to micro Data) is presented and motivated.

A reference survey of existing systems is found, for example, in Rowland (2003). A prototype of a remote access system is given in Keller-McNulty and Hunger (1998). A justification of the architecture chosen by the Bank of Italy for its own system is found in Schouten and Cigrang (2003).

Section 2 accounts for the confidentiality rules adopted in managing external microdata access at the Bank of Italy. Section 3 describes the databases currently available on BIRD. Section 4 outlines the technical features of the platform; Section 5 concludes.

2. The confidentiality code for the Bank of Italy surveys

Protecting confidentiality in microdata has a twofold motivation. It is required and sanctioned by the law; but even if it were not, it would be expected by the respondents to ensure reliability of the given answers. Yet, access to microdata for scientific purposes is to be advocated if scientific research is to be advanced. A trade-off between security and accessibility of databases emerges.

Whether the very availability of datasets for processing is an intrinsic threat to data confidentiality – independent of the means adopted – is still an open question. The

[1] It must be noted that the Bank of Italy is not part of the Italian Official Statistical System (SISTAN), hence participation in the surveys described here is not mandatory under the law.

anonymisation of microdata aimed at public release comes up to the mind as the first viable option: if data variability is not too high and outliers are ruled out, eliminating identification variables and collapsing classifications are viable solutions. This *public use file* (PUF) solution can safely be adopted for household data, where sensitive information like figures on wealth and income cannot be easily identified. If anonymisation is performed in such a way that makes disclosure by merging with other datasets unlikely, the PUF can safely be distributed.

This is the solution chosen by the Bank of Italy for microdata from the Survey of Household Income and Wealth (SHIW). Here, even the original data are formally anonymous, since the dataset provided by the company in charge of the interviews is already stripped of trivial identification variables. Prior to public release, individual records are further stripped of sensitive information (like those regarding morbidity) and other fine geographical classifications which may ease disclosure. In the past, release used to materialise on a compact disc mailed to whomever requested it and is today accomplished by making the data available for anonymous download on the web, together with the relevant metadata.

In some cases anonymisation is unsafe, due to the high number of outliers in the dataset. This happens customarily with business data. Here, access to databases can be monitored by a number of means. In the traditional “data lab”, the researcher has to show up in person at the place where data are stored: here she can login to the desired dataset while her processing is carefully scrutinised. Such a solution has been considered for years the only secure device for data access from external researchers. During the last decade, also thanks to widespread use of Internet services, the possibility of remote processing came into light and seemed from the very beginning a neat improvement with respect to previous solutions. The underlying idea is to let researchers access the lab remotely via some secure device. The initial implementation of this new device enabled the user to produce tables according to some predefined patterns. Issues here are about the minimum size of cells preventing disclosure of individual information and how to prevent singling out individual information through repeated queries. A more flexible access to the database, allowing also regression models to be run, requires that legitimacy checks be implemented through some form of filters applied to the strings submitted. Although this solution is not completely leakproof, as it will be illustrated below, yet it seems the most advanced, able to reconcile confidentiality and external usability of the datasets. This is called RADL (Remote Access Data Lab, using Trewin’s taxonomy, 2003) and is the solution adopted at the Bank of Italy for business data, known under the acronym of BIRD (Bank of Italy’s Remote access to micro Data).

The measures taken to safeguard data confidentiality of the remotely available datasets are first of all the usual anonymisation measures adopted in the Public Use Files.

For business data access, the Bank of Italy’s confidentiality code is based on principles like:

1. multiple selection criteria for eligibility, with some redundancy;
2. automatic legitimacy checks performed on the commands used to access data;
3. automatic and manual checks performed on the log, the output and the logic of submissions;
4. checking effort increasing with disclosure risk.

Let us briefly discuss each of these points.

1. The users' eligibility is based on the following criteria: a) personal identification via a valid id document, b) proven affiliation to a body (to rule out applications from "maverick" researchers), but not necessarily academic (for undergraduate students, a presentation letter from a professor may be requested); c) submission of a consistent (even if synthetic) research project; d) formal agreement with the Privacy law and the Deontological code (signed form). This openness policy is tempered by any perceived irregularity in elements a) to d) which may lead to application rejection[2].
2. In BIRD, access to dataset is possible exclusively through submission of a set of commands in one of the packages supported (to date, SAS and Stata). Before passing the batch programme on to the package parser, a legitimacy check is performed. In principle, for each of the supported packages a subset of keywords or sequences of keywords is deemed as "forbidden" and blacklisted if those commands are potentially able to disclose individual information. This is trivially the case of commands like *list* in Stata and *PROC PRINT* in SAS. If a batch programme contains one of those keywords or sequences, execution is blocked and the user is notified. There is however a limitation to this approach: to pin down every possible suspicious keywords or sequence in a statistical package can be a Sisyphean task, so many controls are left to manual examination.
3. Such drawbacks can be attenuated if a "graylist" is created, by listing suspicious keywords which, when found in the input source programme, or the output, or both will trigger closer inspection. This examination can be made automatically[3] or manually. For repeated submissions, the person in charge of the check must examine the logic of the programmes and assess the risk of complementary disclosure. For the moment BIRD is in a learning mode: all checks on the output are performed manually in order to gather enough experience.
4. The principle of graduality applies to the BIRD system in all cases where the researcher needs to access a dataset other than the standard version. Two typical cases are: i) access is requested to the "complete" version of the dataset (i.e. the anonymised but not winsorised data); ii) merging is requested between an internal and an external dataset. Such cases entail a much tougher security checking process. Specific authorisation profiles would be set up and ex-post checks would be exclusively manual. In case ii), also eligibility would be restricted, motivation scope narrowed and the whole project specifically supervised.

It is quite clear at this point that the effort undertaken to setup a system that is at the same time very open and reasonably secure is based on a series of manual interventions, at least on the eligible individuals and on the results that are to be released. BIRD as it is offered today is apt to serve a small number of motivated researchers. It has to be recalled that BIRD services are not part of the institutional goals of the Bank and are therefore formally offered on a "best effort" basis, even if the users can ask queries about their submissions if they have not got an answer within two working days[4]. The emergence of a much larger number of researchers and/or a larger number of submissions pro-capita would require a revision of the checking system, of the eligibility criteria and a consequent reassessment of costs.

[2] No application has been rejected to date; all of them came from academics or university students (both graduate and undergraduate).
[3] The option of automating some of the checks on the output will be considered in the future and the "graylist" automatic detection, on request of the Bank of Italy, will be implemented in the coming release of the LISSY platform.
[4] In the first months of operation of the BIRD system on a limited number of users, the time waiting for an output to be released has rarely exceeded two hours (within office hours).

3. The structure of the available datasets

3.1 An overview of the available data

Data from the yearly survey on Industrial and Service firms are available. The survey has been uninterruptedly carried out since the beginning of the seventies, but data are available in BIRD only since 1984, after they became to be collected into structured electronic archives.

The sample is a representative panel of the population of Italian firms. Until 1998, only firms with 50 employees and more were covered, belonging to the industrial processing segment and the sample size hovered around one thousand units. Starting from 1999 the target population was progressively enlarged and from 2002 the panel sample covers the population of firms belonging to the industrial and service sectors with 20 employees and more (the service sector do not include the financial institutions, for which the Bank of Italy collects census data).

The panel features physiological attrition rates, with "deaths" representing companies either no longer active, or fallen below the dimensional threshold of 20 units or no longer willing to participate in the survey waves. These units are replaced by newly enrolled ones, selected so as to be as close as possible to those exiting the panel, in terms of size, geographical location of administrative headquarters and sector of economic activity.

The table A in appendix shows the evolution of both the sample and the reference population (see Bank of Italy, 2005 and 2007 for further details on the survey design).

The target population is a subset of the whole population of Italian firms, since units with 1-19 employees are left out, but it represents a sizeable chunk of the total turnover, investment and number of employees of the whole population (respectively 70.0, 81.4 and 77.7 per cent of the totals for employees, turnover and investment for the whole industrial sector: Istat, 2005)[5].

3.2 The data structure

The elementary record represents all the information collected for a single firm in a survey year. The total record number is therefore the sum of the yearly sample sizes.

A peculiar feature of the survey, faithfully mirrored in the archive, is that variations between two adjacent calendar years can be computed within the same survey year, since, for the main variables concerning employment, turnover and investment levels, the firm reports values for the survey year and the previous one, together with a forecast for the following year[6].

The same firm within the dataset is univocally identified by the same identification key (the IDENT field, that is artificially generated and totally unrelated with any firm's characteristics). A certain IDENT can in some cases mark a firm that has been changing its structure during the years through mergers and acquisitions processes (also recorded in the database), but the fact that the IDENT value stays the same guarantees that changes have not been so big as to alter the basic firm structure.

An expansion weight, summing up to the population totals in terms of number of firms is also provided, derived from the original survey weight after being corrected to

[5] For the service sectors covered by the survey the coverage percentages are smaller, albeit still significant (58.5, 50.1 and 56.6 respectively).

[6] If we indicate with *t-1* and t the previous year and the survey year, *t+1* refers to the year for which a forecast is available. Interviews are carried out in the first months of year *t+1*.

take into account of the total missing responses and the post-stratification adjustments (Kalton and Flores-Cervantes, 2003).

It is preferred not to calibrate weights to reproduce totals for other variables (such as the number of employees) because this procedure would add too much variability to the weight distribution and furthermore weights could no longer be constrained to be greater than the unit value (Deville *et al*., 1993).

Strata are represented by combinations of sectors of economic activities and size classes, whereas post-strata are formed by combinations of macro-areas and very aggregate classifiers of economic activities and class–size.

Whenever the user wants to analyse variables collected without a scale factor[7], we recommend to weight results by a factor obtained as the product of the weight times a scale variable (like turnover or number of employees), so that the frequencies take into account the different firm sizes.

Many quantitative variables are imputed so that they can be used at the same time without the hindrance of dealing with different partial response rates. An indicator variable is always available to the analyst to leave imputed valued out, should he/she deem it reasonable.

Monetary values for investment and turnover are also available at constant prices and in such a case their value refers to the most recent year present in the database. Deflators are internal to the survey, since they are individually collected. Before being used to deflate monetary values, they are first averaged by sector of economic activity (extreme values are discarded). See D'Aurizio and Tartaglia-Polcini, 2008 for a discussion about the empirical properties of the various deflation techniques.

3.3 Data – level confidentiality safeguards

Besides the usual preliminary actions of stripping individual firm records of direct identifier and collapsing of classification cells, an additional measure we take is not to release field containing free-format text, which quite obviously could make security breaches likelier. This practice is also implemented within the remote access system used by Statistics Canada (Rowland, 2003).

A final tool to safeguard the confidentiality of individual data is a preventive treatment of all the quantitative variables, for each of them a cut-off upper value is defined and, for a small number of big companies, the corresponding values are set equal it, plus a disturbance term that prevents the data variability from decreasing too much. This solution does still provide valid inputs for all kinds of analyses except for those where estimates of totals of monetary variables are required: should this be the case, the researcher can ask to use the database with the original values[8], but the outputs will take more time to be returned.

We have decided not to use data-perturbation techniques (currently available on specialised software packages like Argus (Nordholt, 1999), since their utilisation might make results too much influenced by the perturbation techniques used and the selection of the best perturbation is still widely debated in the statistical literature (Reiter, 2005).

Our strategy quite closely follows the best international practices of the most advanced statistical institute (see Söderberg, 2005 for a review of the solutions of

[7] This happens, for instance, with the rate of change of capacity utilisation or of percentages (such as the percentage of hours worked overtime).

[8] By using weights the researcher can produce statistics like ratios and variation rates closely aligned with those released in the Bank of Italy's official publications, which do not contain estimates of totals (Bank of Italy, 2007). He/she might decide not use weights for very specific analyses (see also Ritchie, 2005).

Statistics Sweden and Capobianchi, 2006 for a general overview covering the systems of the major statistical public bodies).

3.4 The short-term additional features: expectations and plans available to the external user

Starting from the end of 2008, another set of data will be available with data coming from the short-term outlook survey carried out every year between September and October on the same panel sample of the major survey.

These data are grouped in separate archives, one for each survey year. In order to make long-term analyses an historical archive has also been created.

An additional feature of these dataset is the availability of weights summing up to the population of employees. The analyst has therefore the double option either to produce frequencies not directly related to the firm dimensions (by the traditional weights) or, by using these new weights, to compute frequencies that directly take into account the different firm sizes. The utilisation of these latter weights is strongly recommended whenever employment and wage dynamics are dealt with.

The analyst can match exactly the units in the two datasets, since the firms are identified with the same IDENT, so that the patterns of association between plans/expectations and subsequent realisations clearly show up.

4. The IT architecture: user interface and security rules

The technical infrastructure hosting the system providing remote processing service for the firm survey micro-data is based on a couple of standard PC. The machines are equipped with exactly the same software and work in an active cold stand-by configuration. The two PC are interconnected through a switched 100 Mbit/s Local Area Network (LAN).

The software application offering the remote processing capability (henceforth LISSY, Cigrang and Coder, 2003) is based on a standard java front-end application relying on a relational Data Base Management System back-end. As understandable from the premises the LISSY software can run on a variety of operating system ranging from Windows to different UNIX and LINUX breeds. For compatibility with the rest of the research department our current choice has fallen on Windows XP.

Among the different web technologies available for remote processing the wish to provide the users with the maximum flexibility of a statistical package has led to the employment of simple text mail for communicating from and to the external users.

LISSY system is based on two software components:

a) the PostOffice

b) the Batch machine.

The general structure of the software architecture is depicted in figure A in the Appendix.

4.1 The PostOffice

The PostOffice is the software front end. It periodically goes out to check the system e-mailbox. The PostOffice behaves like a multithread server programme. It carries out the two never ending loops detailed in the following.

The first is based on the following steps:

1. Is there an-email in the system mailbox?

1.1. No. Hold on x second and go back to step 1.
1.2. Yes. Do the following
2. split the mail in the authentication and user's code section
2.1. performs users' authentication
2.2. if unsuccessful reply to the sender mailbox and go back to step 1
3. checks the code for the presence of keywords breaching the security rules:
3.1. if there are security breaches an explanation is sent back the sender
3.2. otherwise the user's programme is inserted into the queue of the job to be executed
4. Is there a job in the completed jobs queue?
4.1. No. Go back to step 1.
4.2. Yes. Prepare the output log and mail it to the sending user
4.3. Go back to step 4

The PostOffice is configured only to accept plain text e-mails refusing the HTML based mails that may contain malicious code. In any case, no kind mail attachments is accepted. A job breaching this rule is refused, the sender receives an email back to explain why the job has been refused.

4.2 The System Database

The database systems employed in the LISSY system are any JDBC compliant database engine providing transactional security. At the Bank, the Oracle DBMS accomplishes the task of storing all the information about system features, users' profile, security rules and statistical package configuration. A certain number of tables are used to manage the system. Some of them, the USER tables contain all information provided by when someone registers to access the database. Some other tables, the JOB tables, gives exact information about all the jobs received. All incoming jobs are logged as well as all system performance parameters. Both these sets of tables among some others allow getting very precise statistics about usage and performance of the system.

All the tables are generated and maintained by the PostOffice component of the system. An existing LISSY database system can easily be integrated into it. All it needs is a single, dedicated user account that has resource permission granted in order to create and maintain it's own tables.

4.3 The Batch Machine

The Batch Machine constitutes the second pillar of the software application composing the LISSY system. The Batch Machines are able to run one or more of the statistical packages offered by the project. Each Batch Machine acts independently without any knowledge of the number, function, or status of other batch processors that may or may not exist on the system. This independence allows the addition or removal of a Batch Machine at any time for balancing workload purposes.

The Batch Machine continually checks the queue of the jobs to be executed and, when a new programme is written in the queue of jobs to be executed, it is dispatches the job to the chosen statistical package and waits for the output log. Upon job completion the Batch Machine carries out the final security checks and eventually writes the log into the completed jobs queue.

4.4 The Data Server

The data server is simply the repository for all of the datasets available for access. Separate directories are maintained for each one of the file formats required by the statistical packages available.

The centralisation of datasets guarantees that all requests are accessing the same data.

The directories containing these data files are write protected so that a user could not accidentally change them as part of its job submission. For obvious security reasons there is no direct link between this data and the mail server.

4.5 The Statistical Packages

The current version of the LISSY system is compatible with three statistical packages: SAS, SPSS and Stata. The need to design a careful set of security rules warranting the highest protection of the micro-data confidentiality has driven the choice of the statistical packages towards application already well known inside the research area. Therefore the actual system built at the Bank of Italy provides only SAS and Stata. The system's users do not access the package directly but only through the mediation of the PostOffice and the Batch machine. These two components performs all the relevant security checks on the input programme and the output log.

4.6 The Security Rules

The checks aimed at safeguarding the confidentiality of the firms micro-data are carried out at all the different phases starting from the production of the datasets stored on the data server. As a first measure, data are preventively treated before their release to external users (sub – section 3.3). The second control is based on allowing TCP/IP traffic on predefined port. The third check verify the plain text nature of the incoming e-mail and the absence of any attachment. Another policy rule is based on the specification of a set of words / sequence of words, belonging to the particular statistical package: they are forbidden since they could increase the risk of individual firm disclosure, particularly for observations in the distribution tails. Upon completion of this syntactical analysis the job is promoted to the executable state. Here the Batch machine will perform other checks preventing the exceeding of thresholds on the size of the output log and the maximum processing time.

All these technical controls are reviewed periodically and before any major upgrade of any critical software component.

5. Conclusions and further developments

The effort of building a complex remote access system for Bank of Italy's business survey micro-data was basically driven by the aim to pursue the accountability, that has recently become engrained into central banks' behaviour in their nature of public bodies, also in the domain of data dissemination.

Central Banks collect a wide range of micro-data that are inputs to reports and analyses which support the decisional processes at many levels. Such data can also foster economic research within Central Bank's research departments without overlapping with the core activities of monetary policy and banking supervision.

A public institution therefore fulfils its accountability obligation by making available to the general public as many data as possible among those it collects; a policy of

openness towards external users for micro-data represents perhaps the most thorough accomplishment of this task.

The sector of society that benefits most from such a policy is the research community, that has been constantly lobbying for more openness from data-collecting public institutions. Starting from the beginning of the nineties of the 20th century, micro-economists and policy makers have been able to enrich their statistical analyses through wider micro-data access (Abowd and Lane, 2003), as they became increasingly less satisfied with the standard statistical outputs released in official publications. Statistical surveys provide many information about the workings of modern societies and data collected and organised by professional statistical agencies have become a major support to analyse social and economic behaviour and produce solidly grounded empirical research and policy evaluation.

National Statistical Institutes have been the first to implement systems of remote micro-data access in order to face this new demand. This practice has taken hold even in Northern European countries such as Denmark and Sweden, with very restrictive legal frameworks shielding individual privacy (Borchsenius, 2005; Söderberg , 2005).

As far as data stored by Central Banks are concerned, the main thrust of this pressure has been the desire to exploit a unique depth of detail in some sectors of the economy.

The opening to a less restricted public of these new set of data will on one hand enable the scientific community to produce new analyses and at the same time will produce the desirable effect of making results obtained by researchers working within the institution “owning” the data verifiable and replicable from other scholars, in accordance with the best practices of any scientific field.

Public bodies must however face two inescapable constraints when they grant wider access to data:

The first one pertains data confidentiality, that ought to be guaranteed, either because it is often mandated by law, or because, although not legally binding, it is a commitment towards the sample units, in order to encourage survey participation. The need to guarantee the respondent’s privacy even more than what would be required by law is particularly felt in business surveys, in which firms’ representatives accept to disclose some details pertaining the firms’ strategy under the assumption that their competitors could never get to these information, since data will be disclosed only in aggregate form.

The second constraint concerns the human and technological resources that must be specifically devoted to data dissemination processes. They are not negligible, since the organisational structure required to set up a system of data access must be currently maintained: both the technical infrastructure and the necessary documentation need ongoing updating and potential users must be informed of the system existence, of its main features and of the subsequent updating.

We have striven to consider all these aspects in the design of our system, which is basically addressed to econometricians willing to apply advanced techniques to business survey data.

The system has currently been working for six months and has successfully satisfied the first users’ requests. Besides the imminent release of new datasets on the system (see sub-chapter 3.4), we are planning future developments in the next two or three years:

- new versions of the system will offer submission of programmes by accessing a web link, as an alternative to traditional e-mail submissions. This enhancement will provide safer identifications procedures;
- the statistical open source package R will become available;
- special projects could be created, which would enable selected users to merge other datasets with those offered by the system.

References

Abowd, J. M. and J. I. Lane (2003), “Synthetic Data and Confidentiality Protection”, Technical paper No. TP-2003-10, U.S. Census Bureau, LEHD Program.

Bank of Italy (2005), Supplements to the Statistical Bulletin - Sample Surveys - Survey of Industrial and Service Firms -Year 2003, New series, Volume XV, Number 55 - 20 October 2005.

Bank of Italy (2007), Supplements to the Statistical Bulletin - Sample Surveys - Survey of Industrial and Service Firms -Year 2006, New series, Volume XVII, Number 41 - 12 July 2007.

Borchsenius, L. (2005), “New Developments In The Danish System For Access To Micro Data”, paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005).

Capobianchi, A. (2006), “Review dei sistemi di accesso remoto: schematizzazione e analisi comparativa” (in Italian), Documenti Istat.

Cigrang, M. and J. Coder (2003), “LISSY Remote Access System”, paper presented at the Joint ECE/Eurostat work session on statistical data confidentiality (Luxembourg, 7-9 April 2003).

Deville, J.-C., C.-E. Särndal and O. Sautory (1993), “Generalized Raking Procedures in Survey Sampling”, *Journal of the American Statistical Association,* vol. 88, n. 423, Theory and Methods.

D’Aurizio, L. and R. Tartaglia-Polcini (2008), “Use of Deflators in Business Surveys: An Analysis based on Italian Micro data”, *Journal of Official Statistics*, vol. 24, n. 2, p. 277-300.

Kalton, G. and I. Flores-Cervantes (2003), “Weighting Methods”, *Journal of Official Statistics*, vol. 19, n. 2, p. 81-97.

Keller-McNulty, S. and E.A. Hunger (1998), “A Database System Prototype for Remote Access to Information Based on Confidential Data”, *Journal of Official Statistics*, Vol. 14, No. 4, pp. 347-360.

Nordholt, E. S. (1999), ”Statistical Disclosure Control of the Statistics Netherlands Employment and Earnings Data”, paper presented at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality (Thessaloniki, Greece, 8-10 March 1999).

Reiter, J. (2005), “Estimating Risks of Identification Disclosure in Microdata”, *Journal of the American Statistical Association,* vol. 100, n.472, p. 1103–1112.

Ritchie, F. (2005), “Access to Business Microdata in the UK: Dealing with the Irreducible Risks”, paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005).

Rowland S. (2003), “An Examination of Monitored, Remote Microdata Access Systems”, paper presented at the National Academy of Sciences Workshop on “Access to Research Data: Assessing Risks and Opportunities”, October 16-17, 2003.

Schouten, B. and M. Cigrang (2003), “Remote access systems for statistical analysis of microdata”, *Statistics and Computing,* n. 13, p. 381–389.

Söderberg, L.-J. (2005), “Mona-Microdata On-Line Access at Statistics Sweden”, paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality (Geneva, Switzerland, 9-11 November 2005).

Trewin, D.J. (2003), Access to microdata – issues, organisation and approaches, paper presented a the Conference of European statisticians (Geneva, Switzerland, 10-12 June 2003).

Appendix

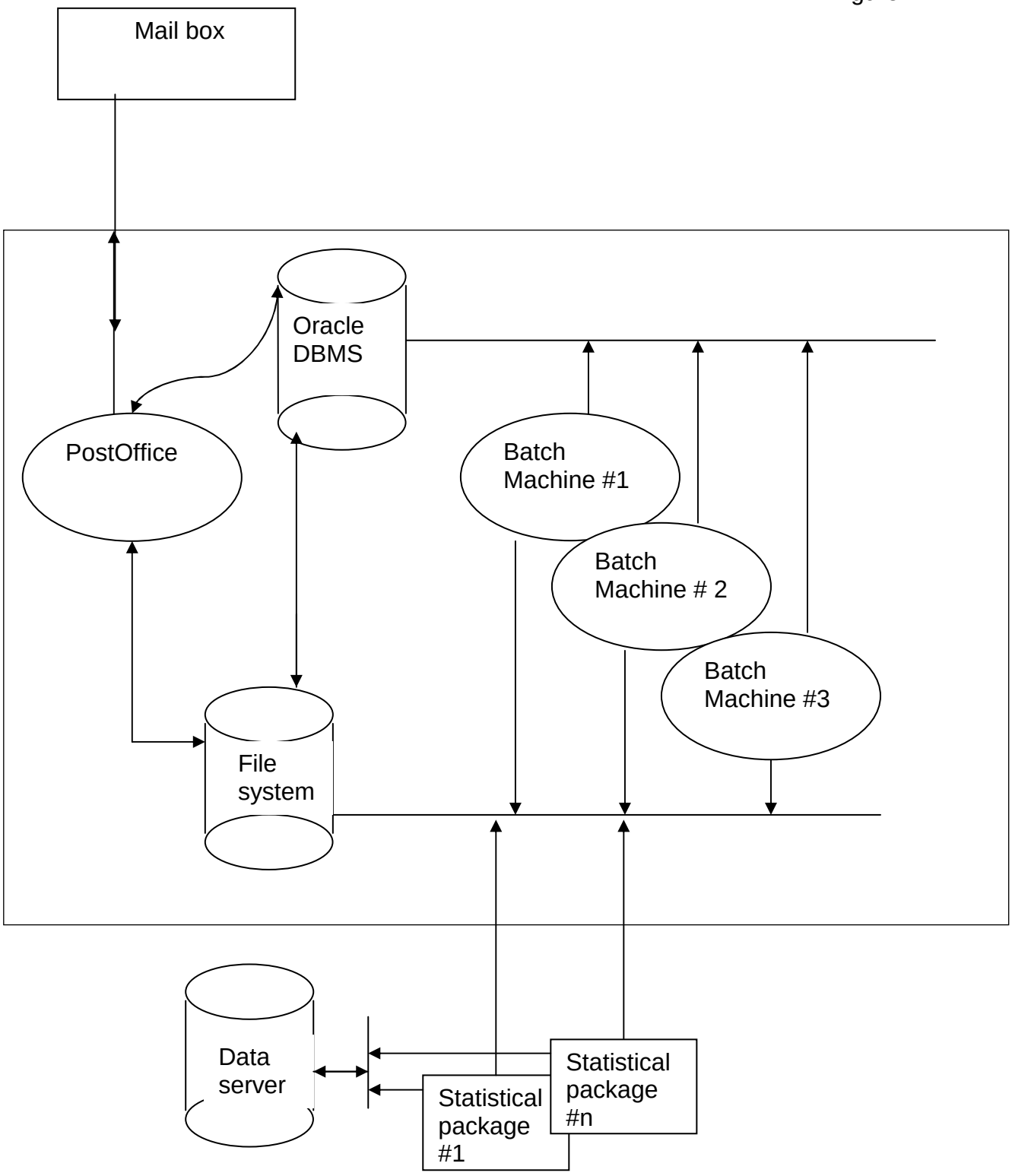
Table A

SAMPLE AND POPULATION SIZES OF FIRMS, 1984-2007[1]

Survey year	Total industry excluding construction				of which: manufacturing				Services			
	20-49		50 or more		20-49		50 or more		20-49		50 or more	
1984	-	-	-	-	-	-	865	12,601	-	-	-	-
1985	-	-	-	-	-	-	877	12,457	-	-	-	-
1986	-	-	-	-	-	-	874	12,313	-	-	-	-
1987	-	-	-	-	-	-	1,069	11,917	-	-	-	-
1988	-	-	-	-	-	-	1,039	12,025	-	-	-	-
1989	-	-	-	-	-	-	1,053	11,883	-	-	-	-
1990	-	-	-	-	-	-	1,071	11,739	-	-	-	-
1991	-	-	-	-	-	-	1,027	12,041	-	-	-	-
1992	-	-	-	-	-	-	994	11,658	-	-	-	-
1993	-	-	-	-	-	-	995	11,185	-	-	-	-
1994	-	-	-	-	-	-	953	11,037	-	-	-	-
1995	-	-	-	-	-	-	996	10,880	-	-	-	-
1996	-	-	-	-	-	-	1,060	11,411	-	-	-	-
1997	-	-	-	-	-	-	1,002	11,792	-	-	-	-
1998	-	-	-	-	-	-	998	11,609	-	-	-	-
1999	-	-	1,135	11,712	-	-	1,107	11,502	-	-	-	-
2000	-	-	1,474	12,029	-	-	1,428	11,798	-	-	-	-
2001	1,022	27,516	1,764	12,629	1,000	27,075	1,713	12,389	-	-	-	-
2002	1,194	26,235	1,852	12,095	1,162	25,755	1,797	11,828	367	17,509	556	7,956
2003	1,236	26,173	1,905	12,254	1,200	25,713	1,848	11,978	374	18,339	620	8,338
2004	1,234	25,685	1,916	11,954	1,204	25,230	1,861	11,677	410	18,915	666	8,650
2005	1,277	24,999	1,950	11,796	1,243	24,552	1,890	11,516	444	19,440	715	9,042
2006	1,209	-	1,900	-	1,172	-	1,838	-	428	-	715	-
2007	1,128	-	1,852	-	1,093	-	1,785	-	397	-	686	-

(1) Source: Istat (2005) for the population sizes. The numbers shown refer only to the population of firms covered by the survey.

Figure A


Mail box
Oracle
DBMS
PostOffice
Batch
Machine #1
Batch
Machine # 2
Batch
Machine #3
File
system
Data
server
Statistical
package
#n
Statistical
package
#1

RECENTLY PUBLISHED "OCCASIONAL PAPERS" (*)

N. 12 – *L'organizzazione dell'attività creditizia e l'utilizzo di tecniche di scoring nel sistema bancario italiano: risultati di un'indagine campionaria*, by Giorgio Albareto, Michele Benvenuti, Sauro Mocetti, Marcello Pagnini and Paola Rossi (April 2008).

N. 13 – *L'offerta di mutui alle famiglie: caratteristiche, evoluzione e differenze territoriali. I risultati di un'indagine campionaria*, by Paola Rossi (June 2008).

N. 14 – *I divari territoriali nella preparazione degli studenti italiani: evidenze dalle indagini nazionali e internazionali*, by Pasqualino Montanaro (June 2008).

N. 15 – *I conti pubblici nel decennio 1998-2007: fattori temporanei, tendenze di medio periodo, misure discrezionali*, by Maria Rosaria Marino, Sandro Momigliano and Pietro Rizza (July 2008).

N. 16 – *How to interpret the CPIS data on the distribution of foreign portfolio assets in the presence of sizeable cross-border positions in mutual funds. Evidence for Italy and the main euro-area countries*, by Alberto Felettigh and Paola Monti (August 2008).

N. 17 – *Prices of residential property in Italy: Constructing a new indicator*, by Salvatore Muzzicato, Roberto Sabbatini and Francesco Zollino (August 2008).

N. 18 – *La riforma della regolamentazione dei servizi pubblici locali in Italia: linee generali e insegnamenti per il futuro*, by Magda Bianco and Paolo Sestito (September 2008).

N. 19 – *I servizi pubblici locali tra mercato e regolazione*, by Daniele Sabbatini (September 2008).

N. 20 – *Regolamentazione ed efficienza del trasporto pubblico locale: i divari regionali*, by Chiara Bentivogli, Roberto Cullino and Diana Marina Del Colle (September 2008).

N. 21 – *La distribuzione di gas naturale in Italia: l'attuazione della riforma e i suoi effetti*, by Silvia Giacomelli (September 2008).

N. 22 – *Il settore dei rifiuti urbani a 11 anni dal decreto Ronchi*, by Paolo Chiades and Roberto Torrini (September 2008).

N. 23 – *Il servizio idrico in Italia: stato di attuazione della legge Galli ed efficienza delle gestioni*, by Michele Benvenuti and Elena Gennari (September 2008).

N. 24 – *Il servizio di taxi e di noleggio con conducente dopo la riforma Bersani: un'indagine sulle principali città italiane*, by Chiara Bentivogli (September 2008).

N. 25 – *Il project finance nei servizi pubblici locali: poca finanza e poco progetto?*, by Chiara Bentivogli, Eugenia Panicara and Alfredo Tidu (September 2008).

N. 26 – *Le grandi imprese italiane dei servizi pubblici locali: vincoli, opportunità e strategie di crescita*, by Magda Bianco, Daniela Mele and Paolo Sestito (September 2008).

N. 27 – *Domanda e offerta di servizi ospedalieri. Tendenze internazionali*, by Giovanni Iuzzolino (September 2008).

N. 28 – *L'assistenza ospedaliera in Italia*, by Maurizio Lozzi (September 2008).

N. 29 – *L'efficienza tecnica degli ospedali pubblici italiani*, by Alessandro Schiavone (September 2008).

N. 30 – *Il difficile accesso ai servizi di istruzione per la prima infanzia in Italia: i fattori di offerta e di domanda*, by Francesco Zollino (September 2008).

N. 31 – *Il debito pubblico italiano dall'Unità ad oggi. Una ricostruzione della serie storica*, by Maura Francese and Angelo Pace (October 2008).

N. 32 – *Il rischio dei mutui alle famiglie in Italia: evidenza da un milione di contratti*, by Emilia Bonaccorsi di Patti and Roberto Felici (October 2008).

N. 33 – *New policy challenges from financial integration and deepening in the emerging areas of Asia and Central and Eastern Europe*, by Valeria Rolli (October 2008).

N. 34 – *La banda larga in Italia*, by Emanuela Ciapanna and Daniele Sabbatini (October 2008).

N. 35 – *Emerging market spreads in the recent financial turmoil*, by Alessio Ciarlone, Paolo Piselli and Giorgio Trebeschi (November 2008).

* Copies of the QEF are available on the Bank of Italy's website www.bancaditalia.it.