Remote Acces Systems for Statistical Analysis of Microdata

Discussion paper 03004

Barry Schouten and Marc Cigrang

The views expressed in this paper are those of the author
and do not necessarily reflect the policies of Statistics Netherlands.
The authors like to thank Leon Willenborg for his advice.

Statistics Netherlands

Voorburg/Heerlen, July 2003

Explanation of symbols

.	= data not available
*	= provisional figure
x	= publication prohibited (confidential figure)
–	= nil or less than half of unit concerned
–	= (between two figures) inclusive
0 (0,0)	= less than half of unit concerned
blank	= not applicable
2002–2003	= 2002 to 2003 inclusive
2002/2003	= average of 2002 up to and including 2003
2002/'03	= crop year, financial year, school year etc. beginning in 2002 and ending in 2003

Due to rounding, some totals may not correspond with the sum of the separate figures.

Publisher
Statistics Netherlands
Prinses Beatrixlaan 428
2273 XZ Voorburg
The Netherlands

Printed by
Statistics Netherlands - Facility Services

Cover design
WAT ontwerpers, Utrecht

Information
E-mail: infoservice@cbs.nl

Where to order
E-mail: verkoop@cbs.nl

Internet
http://www.cbs.nl

ISSN: 1572-0314
Key figure: X-10
Production code: 6008303004

Statistics Netherlands

REMOTE ACCESS SYSTEMS FOR STATISTICAL ANALYSIS OF MICRODATA

Summary: A remote access system is a facility where users can submit queries for statistical information from their own computer. These queries are handled by the statistical agency and the generated, possibly confidentialised, output is returned to the user. This way the agency still keeps control over its own data while the user does not need to make frequent visits to the agency.

For some years, the Luxembourg Income Study (LIS) and Luxembourg Employment Study (LES) have made use of an advanced remote access system. At Statistics Netherlands and at other statistical institutes recently the need for a similar system has been expressed. In this paper, we discuss the characteristics, limitations and desired properties of a remote access system. We illustrate the discussion by the system used at LIS/LES.

Keywords: Microdata, Remote access, Remote execution, Confidentiality

1. Introduction

Governmental organisations need information from individual persons, companies and institutions in order to test the effectiveness of their policy and to set new policies. To this end, in many countries legislation permits special statistical agencies to obtain individual and possibly sensitive data. These agencies have the task to gather, manage and disseminate huge amounts of statistical information. However, although legislations are different from country to country, usually agencies are only permitted to disseminate information that cannot be reduced to information about individuals.

To achieve transparency of governmental policies and to facilitate research, statistical agencies publish information that is generally accessible. Still, governmental organisations and researchers often require more specific information than the information that is disseminated periodically. Many agencies have therefore set up facilities to handle specific, ad hoc requests.

Clearly, the protection of the confidentiality of microdata and the dissemination of statistical information are two conflicting objectives. The one extreme, the absence of any publications, is completely safe, while the other extreme, the full provision of all gathered microdata, is completely unsafe. Statistical agencies have to take a stand in between those two extremes based on the legislations. For a discussion, see Duncan and Pearson (1991) and Fienberg (1994). This stand, however, largely determines the type of access that is granted to microdata.

There are basically three different types of facilities that are used in order to deal with specific requests for statistical information: on-site, cd-rom and remote access.

On-site facilities are special locations, usually within the office of the statistical agency, where researchers can perform statistical analyses on microdata directly. The generated output of on-site analysis sessions is checked before it is released to the researcher. In most cases the researcher has to agree to sign a contract or to become a sworn-in employee for the duration of the research. Examples of on-site access are the Research Data Centers set up by Statistics Canada and the National Center of Health Statistics (NCHS) in the United States. Statistics Netherlands also has two on-site locations.

The microdata can also be physically distributed on cd-rom or any other device, so that a researcher can analyse the microdata behind his own workstation. Direct identifiers are always removed and the microdata are often perturbed and recoded to make disclosure improbable. The Australian Bureau of Statistics, for instance, puts Confidentialised Unit Record Files (CURF's) at the disposal of researchers under contract. Statistics Netherlands also supplies selected researchers with cd-roms under contract. These cd-roms contain the confidentialised microdata from a number of household surveys.

The third type of facilities is a remote access system, which will be the focus of this paper. A remote access system offers researchers the possibility to analyse microdata from a remote location, usually their own place of work, via a web connection. The microdata remain physically at the office of the statistical agency, and the user of the remote access system submits statistical queries. These queries are executed by the agency and the possibly confidentialised output is returned. For this reason such systems are sometimes also called remote execution systems.

Remote access systems, thus, combine the virtues of on-site facilities and cd-roms. The microdata are not physically distributed and hence remain under the control of the agency, while the users of the system can perform analyses at their own place of work. However, the trade-off between a user-friendly system and a safe system is evidently present for remote access systems. Disclosure control will always limit the detail of the output.

An extensive amount of literature has been written about the security and confidentiality of statistical databases. See Adam and Wortmann (1989), Keller-McNulty and Unger (1993) and Doyle et al. (2001) for an overview. Relatively little attention has been devoted, however, specifically to remote access systems of statistical microdata. Keller-McNulty and Unger (1998) discuss a prototype for remote data access and show evidence that confidentiality can be protected while enhancing the quantity and quality of the returned output.

Nowadays several statistical agencies allow for remote access of microdata. Statistics Canada, Statistics Denmark and the US National Center of Health Statistics (NCHS), for example, have remote access facilities. They all manually verify the confidentiality of the generated output. The US Census Bureau is developing the data dissemination system called American FactFinder, which allows users to request confidentialised tabular data. See http://factfinder.census.gov , and Hawala (2001).

One of the oldest remote access systems for statistical microdata is the Lissy system of the Luxembourg Income Study (LIS). This project started in 1983 and was extended with the Luxembourg Employment Study (LES) in 1994. The objective of the project is to open up microdata from a large number of countries for comparative research on income and employment. See Rainwater and Smeeding (1988), Cigrang and Rainwater (1990) and http://www.lisproject.org. Queries must be submitted by e-mail and are handled and returned automatically.

Another large-scale remote access system is being developed under the Digital Government Program in the United States by the National Institute of Statistical Science (NISS). See NISS (1998) and http://www.niss.org/dg. This system will allow for interactive queries from a variety of users.

The IPUMS-International project (Integrated Public Use Microdata Series) was set up in 1999 to disclose the census counts of many different countries. Via the web requests for confidentialised tables can be submitted, which can be downloaded afterwards. See http://www.ipums.org/international.

Interesting in this respect is also the Data Extraction System (DES) created by the U.S. Bureau of Census with the objective to make large collections of 'raw' data from large surveys and censuses more readily accessible and easier to use. See http://www.census.gov/DES/www/welcome.html .

Recently, at Statistics Netherlands a pilot was started which ultimately has the goal to construct a remote access system for governmental organisations like ministries and planning offices. The remote access system should give access to various household surveys linked to register data for research purposes. A preliminary study into the advantages and drawbacks of such a facility was done by de Boer, Schouten and Willenborg (2000).

Of course, the use of remote access systems is not restricted to statistical microdata. Examples are given in Kafatos et al. (1998) (earth observations) and in Lees et al. (1999) (medical records). However, the differences between these remote access systems and remote access of statistical microdata are that general systems usually do not involve statistical queries or that the data does not contain sensitive information.

The main objectives of this paper are to discuss the trade-off between disclosure evaluation before and after output is released to a user. We give tentative answers to questions like: What are the possibilities of remote access and what is not possible? What kind of statistical disclosure control can be applied? What kind of restrictions must be imposed? What ways are there to control and combine the output of several queries after they have been returned to the user?

In section 2 we describe various aspects of remote access systems. Next, in section 3 we illustrate these aspects by the Lissy remote access system used at LIS\LES. Finally, we discuss disclosure evaluation for remote access systems in section 4 and draw some general conclusions in section 5.

2. Aspects of remote access systems

There are a great number of aspects that can be considered when implementing a remote access system. The basic features that play a role are confidentiality, user-friendliness and feasibility. In the literature user-friendliness is also referred to as utility. In this section we will elaborate on the various aspects. However, the list of aspects we discuss is by no means exhaustive. We also refer to Keller-McNulty and Unger (1993), Blakemore (2001) and the research proposal in NISS (1998).

Confidentiality

Confidentiality is concerned with the protection of data on individual persons, households, companies and other institutions.

First, one needs to decide what microdata to make accessible in the remote access system. Clearly, this decision depends on the demands of the potential users, the relevance of the microdata for research or policy making, the legislation, the quality of the data and the sensitivity of the information contained in the microdata.

Next, it should be decided whether the statistical database is static or dynamic, i.e. whether it contains longitudinal data and whether it will be updated in time. This has consequences for the disclosure control, see for instance Shieh and Lin (1999).

Legislation now restricts the use of the microdata and the statistical output that may be returned. However, legal conditions are hardly ever unambiguous when it comes to disclosure. In practice statistical agencies need to decide what disclosure is and what is not, i.e. what output is considered to be undesired. For instance in The Netherlands the Personal Data Protection ACT (WBP) states that variables like religion, health, sexual behaviour and membership of a political party or union need to be handled with extra care. See Al and Altena (2000). Nevertheless, it is not directly clear what the consequences are for disclosure control.

Hence, legislation needs to be interpreted so that disclosure can be defined. Especially, when remote access systems are interactive and disclosure evaluation is automatic, clear definitions are essential and inevitable.

When clear definitions of disclosure are available, then the next step is to decide how confidentiality of the data can be assured. This means disclosure must be evaluated and controlled. Disclosure evaluation and disclosure control cannot be viewed separately. The types of output that are not allowed determine the measures that must be taken to change the output. Disclosure control measures by themselves potentially reveal information to the user. Literature describes various methods to deal with disclosure evaluation and control. See Adam and Wortmann (1989), Özsoyoglu and Su (1990) and Willenborg and de Waal (1996 and 2001).

One method of data protection is the conceptual data model, which aims at constructing a statistical database structure that describes the relations between all information contained. The set of possible queries is evaluated and confidentialised

beforehand. However, such a database structure can be very complex and is impractical in case advanced statistical analyses are to be performed.

Another method is the restriction of queries based on their output and the output of previous queries. Effectively this means that queries and output need to be filtered either automatically or manually on a query basis, and that a query history database must be set up and maintained. We will return to the issue of query restriction in section 4.

Evidently, disclosure evaluation and control can in many cases not be constructed so that the remote access system always guarantees the confidentiality of the data. Different users may have different levels of prior knowledge or may combine output with output from previous queries. Therefore, besides the removal of all direct identifiers, the microdata must be locally suppressed, perturbed, masked or recoded. Disclosure is then partially prevented by adapting the microdata or by increasing the uncertainty in the generated output. Many techniques are available for adaptation or perturbation of the data, see for example Duncan and Mukherjee (2000). A promising technique is post-randomisation, see Gouweleeuw, Kooiman, Willenborg and de Wolf (1998) and van den Hout and van der Heijden (2002).

Still, disclosure of individuals cannot be precluded in general. As a last step the statistical agency can demand that users of the remote access system sign a contract in which they agree on a number of rules and in which the consequences of violations are described. For instance, the agency may request that results are always presented to the agency before publication. Setting up a contract does not only serve as a back-up in case of misuse of the system, but will also deter potential misuse because of the consequences.

User-friendliness

Another main feature of a remote access system is its user-friendliness or its utility to users. Important questions are: What users are allowed access? What software can they use? What statistical queries are possible? How detailed is the output that users receive? Of course the answers to such questions are interrelated.

One may distinguish three major groups of users: governmental organisations, universities and other research institutes, and the general public. Employees of governmental organisations often have to comply largely with the same legal conditions as the employees of statistical agencies, so that access to microdata is less perilous. Nevertheless, additional agreements are still necessary. It can be expected that governmental organisations are especially interested in answers to specific, clear-cut questions, and are less interested in long-term research. Researchers from universities or other institutes may not have to comply with legal conditions. However, they will be interested in more advanced statistical analyses and like to use specific software. Finally, the general public may also be allowed access to microdata. Due to the heterogeneity of this group and the lack of clear research

questions, the variety of software, possible queries and output will in most cases be limited.

Other relevant aspects of the remote access system to the user are speed, documentation and the presence of a helpdesk. The service time of the system depends on the extent to which the system is automated, i.e. are queries evaluated and executed manually. Furthermore, the choice for an on-line (webpage) or off-line system (via e-mail), and the number of users also influence the speed with which queries are replied. Finally, and maybe most importantly the speed is determined by the evaluation and control of disclosure. For the user it may be very useful to be able to assess the occupation and busyness of the system.

To both the administrators and the users it is vital that the system is well documented, so that correct queries can be formulated and that the output is clear and interpretable. The documentation may also consist of the availability of test files, examples of queries and pre-defined queries that perform routine actions like weighting. For remaining questions a helpdesk may be created.

Feasibility

Essential to the architecture of the remote access system is the feasibility of its implementation. The most important aspect of the feasibility is the security that can be attained. Although, queries and output are evaluated for disclosure, the statistical database itself and the data transfer between the user and the agency may be the subject of hackers. A fully secure system does not exist of course, but there are many methods to minimize the risk. Passwords, firewalls and encryption may be used, see for instance Denning (1982).

The capacity needed for the maintenance and administration of the system, and to keep up a certain level of service also play a role when it comes to feasibility. Furthermore, the choice for automatic or manual handling and evaluation of queries determines the labour intensiveness of the remote access facility. These aspects are all directly related to budgetary constraints.

A final aspect, which is not really connected to feasibility, is the reliability and accuracy of the output. The statistical agency may consider it to be one of its responsibilities that the generated statistical output is interpreted correctly. Publications and results based on the output must be statistically well founded. Essential in this respect is the quality of the underlying microdata.

Ideally, the three features confidentiality, user-friendliness and feasibility should all be optimised at the same time. However, these features are conflicting. For a given feasibility, as we mentioned earlier, confidentiality and user-friendliness are complementary. The more queries are allowed and the more output is returned, the more risk is present that individual information is disclosed. Obviously, the

implemented system must balance the two concepts given a minimal level of confidentiality protection imposed by legislation.

If one neglects feasibility for a moment, then it may be theoretically possible to find the optimal balance given a minimal level of confidentiality protection. We should, then, be able to fully evaluate the output of every statistical query regarding also all previous output for potential disclosure. Evidently, this asks for a very sophisticated and complex system using extremely fast computers. In this respect we refer to Dobra, Karr and Sanil (2002). They show that in case requests for tabular data can be submitted, scalability soon becomes an almost infeasible problem.

In practice and for general queries we may not be able to build a remote access system that optimally balances confidentiality and user-friendliness. Computers are simply not fast enough and the construction of a system that fully evaluates the risk of disclosure may be too costly and complex and therefore not feasible.

Hence, when setting up a remote access facility one should define a minimal level of confidentiality protection, and balance confidentiality and user-friendliness under feasibility constraints. This means the statistical agency has to decide what disclosure evaluation and control actions to perform before output is released, what actions to perform afterwards, and in what way they rely on the perturbation and adaptation of microdata and the use of contracts. This will be the subject of section 4. First, we demonstrate the various features of a remote access system by the Lissy system used at LIS\LES.

3. Luxembourg Income Study and Luxembourg Employment Study

The Luxembourg Income Study (LIS) is a not-for-profit cooperative research project with a membership that includes 25 countries on four continents: Europe, America, Asia and Oceania. The LIS project began in 1983 and is mainly funded by the national science and social science research foundations of its member countries. The LIS database is a collection of household income surveys. These surveys provide demographic, income and expenditure information on three different levels: household, person and child.

The Luxembourg Employment Study (LES) is a parent project of LIS that was initiated in 1994. This project has been partly funded by the Human capital and Mobility Programme of the European Commission and the Norwegian Research Council. The LES database includes Labour Force Surveys from countries with quite different labour market structures. These surveys provide detailed information on areas like job search, employment characteristics, comparable occupations, investment in education, migration, etc.

The LIS/LES team harmonises and standardises the microdata from the different surveys in order to facilitate comparative research. Users have to agree to a contract in which they describe the objectives of their research and in which they sign a confidentiality pledge.

From the LIS/LES website detailed documentation on the key variables in the microdata can be downloaded. Furthermore, synthetic microdata files are available that can be used to test the syntax of the statistical queries.

The users submit their statistical requests under the form of SAS, SPSS or STATA programs to LIS/LES via the Internet mailing system. The email requests contain the "code" created by the user for the specific statistical package used and a standardized header identifying the user. The processing system automatically accepts these requests, processes them, and then returns the results in the form of email to the registered address of the person making the request. The operating system consists of a series of software components connected through one or more networks. These components work together to receive, process and return statistical requests. The LIS/LES operating systems is a remote job execution system and not a remote access system. All the components of the system are physically separated and at no moment is a user in direct contact with the data.

The diagrammatic presentation in figure 1 shows the main pieces of the LIS operating system.

The heart of the system is the job control component or Post Office. It plays the role of the "traffic cop", which manages the entire access mechanism. At the speed of five seconds interval the Post Office accomplishes the following tasks:

- It retrieves the email requests from the mail server.
- It prepares these requests for processing by checking for all security issues like clearly identifying a user, checking for the use of illegal statistical commands, check for the usage of sequences of commands or variables or any other combinations not allowed.
- It returns any job that breaches security to the sender along with an error message explaining the violation.
- It distributes the requests to the batch processor computers.
- It returns the statistical results to the proper (registered) user email addresses.
- It sends suspicious output to the review queue for manual review instead of returning results to the user.
- And finally it maintains critical databases needed for the overall operation.

Step one in the sequence of the Post Office routine is a query of the mail server to determine if any requests have been received. Only plain text request are accepted. Any encapsulations in HTTP or attachments are immediately rejected. Once received, the request is scanned for the mandatory information located at the beginning of the request. The mandatory information includes the requestor's userid, the requestor's password, the statistical package that is being used and finally the dataset to work on. The user identification and password are checked against the list

of registered users. If this identification is ok, a check is made to determine if this user's access to the database is still active. If the identification is not ok or not active, then the query is filtered.

Following the access check, the syntax of the request is examined to determine any violations regarding the types of statistical procedures used and sequences or combinations of words. The LIS operating system does not provide a set of pre-configured security settings and definitions. The configuration utility of the system allows adding the necessary rules to the security database in order to fully define a project oriented security policy. Any requests that violate the rules established for the project are filtered and sent back to the email address contained in the header of the mail message with a notification of the error. The LIS operating system is therefore flexible when it comes to the types of analyses that are allowed. In principle every type of analysis is allowed except for queries that show individual records of course.

A copy of the request is saved to the archive so that a complete list of all jobs ever submitted is maintained. This archive not only provides a backup in case a request has been lost but also provides evidence of misuse of the database. Finally the file containing the request is moved to an area for further processing by the Batch machines.

Requests are processed by the batch processing computers and the results are packaged and returned to the common area accessible to the Post Office. The security settings of the system define which supplementary types of output can be forwarded together with the standard listing (graphics, intermediate files, spreadsheets, etc.). If the Post Office finds a request file waiting to be sent back to a user, it initiates an examination of the file size and contents. Output that is judged to be acceptable under the project rules is returned to the sender automatically as an email. The results are sent back using the email address stored in the user database and not to the address of the sender of the request, which could be different. If the results have been judged as not corresponding to the security settings, they are put to a special location for manual review. This review consists in the manual editing of the result by a data manager that has to validate whether or not these results can be send back to the user. In this editing process the data manager can add comments to the listing and delete pieces of it.

For all these operations the database of job submissions is updated and can be used to provide numerous statistics concerning system usage.

4. Disclosure evaluation before and after the release of output

In section 2 we mentioned the balancing of data access and data confidentiality given feasibility constraints. In practice, a statistical agency needs to decide to what extent disclosure is evaluated before release of output and to what extent it is

evaluated on a later point in time. Furthermore, then also the perturbation and adaptation of microdata and the use of contracts come into play.

In this section we assume that a desired level of security can be attained. We will focus attention to disclosure evaluation before and after release of output. We only consider remote access facilities that are set up for governmental organisations and research institutes.

4.1 The process of remote access

Figure 2 gives a schematic overview of the process of submitting and handling queries, and returning the corresponding, confidentialised output.

The original statistical database is confidentialised by removing identifiers, and possibly some data masking and recoding is applied. Only the confidentialised database can be approached in the statistical queries.

A user submits a query either interactively or as a batch job. First, the query has to pass a query filter. If it is a correct query, the statistical software executes the query and sends the output to the output filter. If the query is not correct, for example because it contains commands that are never permitted, the query is rejected.

The logfile database contains all queries and output submitted and returned by each user or groups of users. The output filter partially evaluates the disclosure of the particular output using the metadata in the database, and possibly in connection with previous released output in the logfile database. If necessary the output is adapted to control disclosure.

The query filter, the output filter and the execution of the query may be automated or may be manual.

An optional element is a distiller for making summaries of logfiles (queries plus output). At any point in time the logfiles of a particular user can be summarized using software or manually, and the summary can be sent to a monitor. The monitor checks the summaries manually, for instance for the presence of repeated tabular output on the same sensitive variables.

4.2 Query filter and output filter

The query filter, the output filter and the summaries of logfiles are the tools used to evaluate and control disclosure.

Let us suppose that every query consists of a single command or action, e.g. the recoding of a variable into a new variable, the selection of a subpopulation, a frequency table, a factor analysis, etc.

A user of a remote access system is in principle permitted to submit any statistical query that passes the query filter. The integrity of the statistical database must be maintained. Therefore, a minimal requirement is the filtering of all queries that alter the database or that show individual records in the microdata directly. Let us assume that these queries are blocked by the query filter.

Now, the output filter needs to block or alter those queries that lead to disclosure. The output of such queries must be confidentialised before it is returned or the query is simply rejected.

It is clear that we always need to evaluate queries in a historical perspective. If a frequency table is requested, then we need to know what subpopulation is involved and how variables are coded. In fact, recodes and selections alter the metadata, which is needed to evaluate subsequent queries. Hence, an output filter should always take the most current metadata into account.

Given some definition of disclosure, an output filter may minimally evaluate the disclosure of a submitted query using the most actual metadata but irrespective of other previous queries. For specific queries such as quantitative tables and frequency tables, relatively much research has been done to evaluate and control disclosure. See Willenborg and De Waal (2001). Tables have to satisfy a number of rules. If these are violated, then recoding, rounding and cell suppression are often used to confidentialise the output. However, little attention has been paid in the literature to other statistical queries, see Reiter (2003) for disclosure control of regression analysis. In case an output filter must evaluate a large set of general statistical queries, then further research is necessary.

Apart from contextual queries that concern the metadata, the combined output from queries may lead to disclosure while each of the single queries is safe. For example, frequency tables may be linked by partially using the same spanning variables.

Schlörer (1976) was one of the first who investigated query restriction in case several queries are submitted. Queries are rejected in case the output combined with previous output reveals too much information. See also Chin and Özsoyoglu (1982) and Chin, Kossowski and Loh (1984). In the literature query restriction usually concerns relatively simple statistical queries and databases that have a simple structure. Dobra, Karr and Sanil (2002) show that when only tabular data are released dynamically, the evaluation of disclosure is theoretically possible but practically not feasible due to scalability problems.

Another major problem when applying query restriction is the dependence on the order in which queries are submitted. The acceptance of a query depends on the queries that have already been submitted. Therefore, it may occur that a particular query that is scientifically interesting is rejected because previous but less interesting queries are submitted and accepted.

Furthermore, the remote access system must always give output that is consistent with previous output. For example, two different requests for the same table must give identical output. This means that disclosure control methods like data perturbation must be applied consistently. The remote access system thus must keep track of the information that is released to a user or a group of users.

Summarizing we can say that evaluation of single queries is definitely possible, although in general further research is necessary, but that complete evaluation of multiple queries may be too complex, time consuming or restrictive to implement.

4.3 Logfiles and summaries of logfiles

Given the considerations in the previous section we may choose to shift part of the evaluation of disclosure to a moment in time after the output is returned to the user. The motivation is that evaluation does not heavily affect the speed of the remote access system in that case, which makes the system user-friendlier. In addition, the system will be less restrictive in principle. However, we must preserve a minimal level of confidentiality protection.

The freedom of actions by users of the system must, however, be restricted by a contract, and if deemed too risky then supplementary data masking and recoding of the underlying database must be applied or some queries may simply not be permitted.

By shifting disclosure evaluation, we are less restricted by the fact that evaluation is time consuming. Nevertheless, complete evaluation may still not be doable, because it is simply too complex and time consuming. Furthermore, we should also be able to combine queries that are submitted at different time points by different users from the same group of users. Therefore, all queries and output must be logged.

We can distinguish two types of users, users who deliberately seek for sensitive information, and users who do not have such intentions. Say we call these type A and type B users.

An important consequence of returning potentially unsafe output, is that we must shift attention from general disclosure evaluation to evaluation of queries that intendedly aim at disclosure. It means we in principle trust the users of the remote access system. If we conclude at a later point in time that the combined output of several queries reveals individual information, then we cannot retrieve the output from the user. However, if we have some evidence that the user is submitting queries that are not in accordance with the research proposal, then we can address this user and take appropriate actions.

A question that immediately arises, is whether the two types of users can be recognized by the queries they submit. In general, the answer to this question is negative, since in general we do not even know the prior knowledge of the user. Hence, a particular user may only need little information to disclose a particular individual element of the database. We must, therefore, always fix the level of assumed prior knowledge.

Given a certain level of prior knowledge, a detailed and thorough, manual analysis of the queries that are submitted will probably in many cases give substantive evidence whether a user is bona fide or not.

A first requisite for the distinction of users is the objective of the research. A type B user will submit queries that will bring him or her closer to answers to the research questions. A type A user will concentrate on detailed analysis of small subpopulations. It may be necessary, though, that a particular statistical analysis demands detailed information. For instance, when fitting a linear regression model it is essential that one gains insight in the existence of possible outliers.

A second requisite is the availability of the metadata that are part of the statistical database. This implies evaluation is database specific and must be employed by an expert on the statistical issues involved in the database.

Furthermore, in order to classify a user one needs to understand the perspective of users that attempt to compromise the database. See Elliot and Dale (1999).

In the literature, methods have been developed to detect automatically anomalous behaviour of users of some computer account in order to recognize unauthorized users. Profiles of authorized users are analysed and are used to define "normal" behaviour. The actions of a new user are compared to the profiles. See for instance DuMouchel and Schonlau (1998) and Ryan, Lin and Miikkulainen (1998). These methods are not very well suited, however, to the detection of type A users. Research objectives can be expected to be very different for different users and consequently the statistical queries are not comparable. Furthermore, long training periods are required to determine normal research profiles for each kind of research. It seems, therefore, not feasible to implement fully automatic detection of users that search for individual information.

Nevertheless, the evaluation of disclosure may be supported by the creation of some tools. First, all queries and output must be logged. For instance standard statistical software packages like SPSS and SAS offer the option to save all syntax and output in logfiles. One may choose to also add relevant metadata to the logfile, but this is not straightforward in general. Secondly, logfiles may be combined and summarized into a single overview of the queries submitted by a particular user. This may be very helpful, since the information contained in logfiles rapidly grows to large proportions, which makes the evaluation rather burdensome.

Characteristics that may be contained in a summary are: background information to the user, the statistical database, the software, the number of logfiles and the research objective, general characteristics like the number of frequency tables and the number of regressions, relevant metadata like recodes, selections, and the use of sensitive variables. One may, for instance, for every set of microdata construct a file that contains all variables and queries that are considered to be sensitive and need extra care. The summary may then display all queries explicitly that are contained in this file or that operate on a sensitive variable.

The choice of characteristics that is recorded in the summary and the degree of detail of the characteristics are clearly subjective. It may also be that different experts use different characteristics to identify anomalous behaviour.

Furthermore, it seems vital that the summaries contain information about the objectives of the researcher. Ideally, this information should be subtracted from the logfiles themselves. However, this is no easy task. For this purpose it may, therefore, be requested that the users attach clear remarks about the objectives of their statistical analyses to the queries they submit. These remarks can be copied into the summary and can be compared with the output.

Further research into the building of summaries is necessary and definitely interesting in the context of remote access systems.

5. Conclusions

In the previous sections we sketched and illustrated the interplay between user-friendliness, feasibility and confidentiality. The evaluation and control of confidentiality in remote access systems is a mixture of confidentialisation of the underlying statistical database, restriction of queries, detection of mala fide users and the use of contracts.

We draw the following conclusions:

- Query filters that reject queries that are never permitted can be implemented easily.
- Disclosure evaluation of single queries has been analysed for the most common simple statistical queries. Further research is necessary in case general statistical queries are permitted. Since actual metadata is required, the implementation of automatic output filters is not straightforward.
- Filtering of all output that leads to disclosure when combining with previous output is in general too complex and time consuming. One must, therefore, accept that a remote access system is not completely safe in general.
- In order to make the remote access system user-friendlier and less restrictive, part of the disclosure evaluation may take place after the output has been returned to the user. This evaluation should focus on anomalous behaviour.
- Anomalous behaviour of users cannot be formalised easily and knowledge of the research objective and the underlying database are essential.
- Summaries of logfiles may be used to indicate whether queries are not in accordance with the research question. However, further research is necessary, since this is not at all straightforward.
- The evaluation and control of queries is easier to implement in case of remote access via e-mail, ftp or other indirect way.

Contracts and data masking or perturbation of the statistical database can be viewed as measures to reduce the risk that remains after implementing a set of disclosure evaluation techniques. An important question, therefore, is in what respect we may rely on the trustworthy of the users of the remote access system. Especially, when it comes to specific users like governmental organisations, universities and other research institutes, one may argue whether it is necessary to filter all output that only by very sophisticated methods leads to potential disclosure. Such institutions usually have a long-term relationship with the statistical agency and the users are often known to the statistical agency. These are the select few that are authorised to

perform statistical analyses. Hence, such users put a lot at stake when intendedly searching for individual information. One other argument that is often put forward is that historically only very few violations have been reported.

The other side of the picture is that we need at least a reasonable level of disclosure evaluation in order to conclude that violations do not occur and to show that statistical agencies take disclosure very seriously.

In general, we believe that remote access systems are valuable alternatives for on-site facilities. As for most data dissemination some risk for disclosure must be accepted. In order to balance user-friendliness and confidentiality part of the disclosure evaluation must be performed after the output has been released. However, if we want to automate disclosure evaluation by means of an output filter and a summary distiller for logfiles, further research is clearly necessary.

We believe that challenging research objectives for the future are

- the construction of logfiles that are not software-related
- the building of software that makes summaries of all logfiles and that preserve the information needed to evaluate disclosure

the disclosure evaluation and control of sophisticated statistical analyses.

Figure 1: The remote access system used for the Luxemburg Income Study (LIS) and Luxemburg Employment Study (LES).

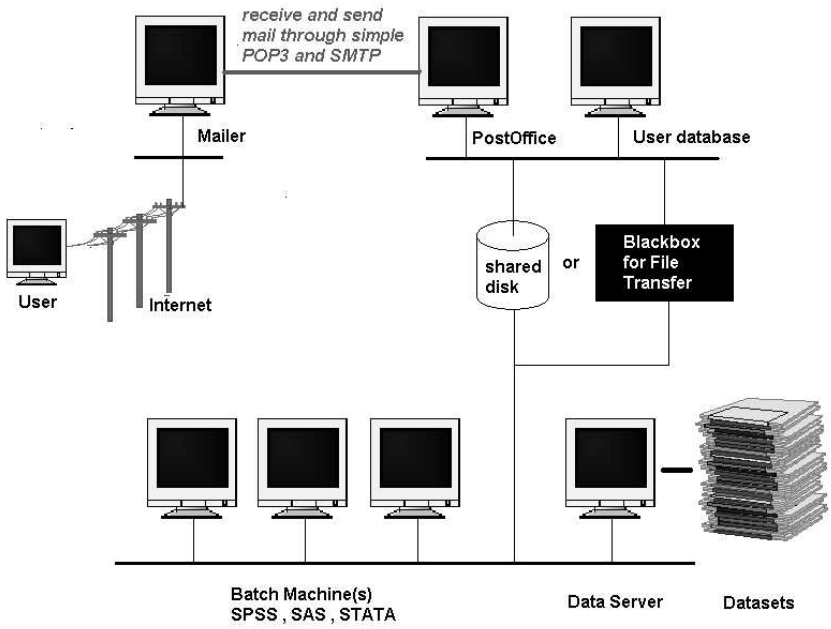
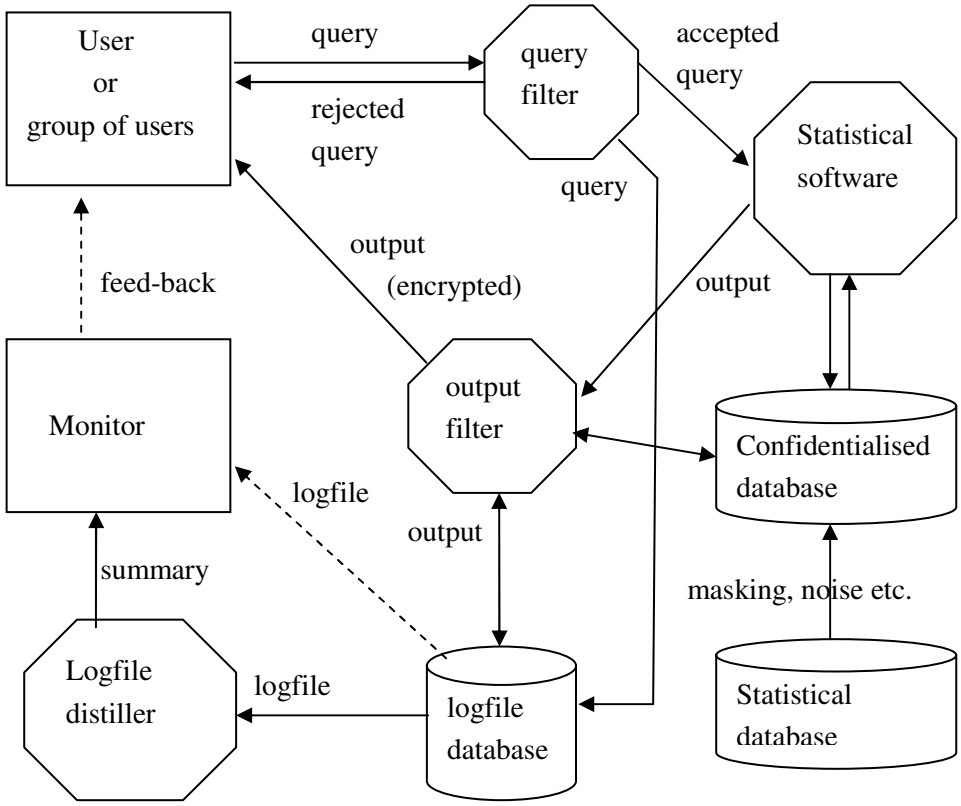


Figure 2: A schematic overview of a general remote access system.



References

Adam, N.R. and Wortmann, J.C. (1989), Security-control methods for statistical databases: A comparative study, ACM Computing Surveys 21, pp. 515-556.

Al, P.G. and Altena, J.W. (2000), Data security, privacy and the SSB, Netherlands Official Statistics 15 (summer), pp. 47-51.

Blakemore, M. (2001), The potentials and perils of remote access. In: *Confidentiality, disclosure, and data access: theory and practical application for statistical agencies* (editors P. Doyle, J.I. Lane, J.J.M. Theeuwes, L.M. Zayatz), Amsterdam, Elsevier.

Boer, P.K. de, Schouten, J.G. and Willenborg, L.C.R.J. (2000), Remote access: een voorstudie, Report, Statistics Netherlands, Voorburg.

Chin, F.Y., Kossowski, P. and Loh, S.C. (1984), Efficient inference control for range sum query's, Theoretical Computer Science 32, pp. 77-86.

Chin, F.Y. and Özsoyoglu, G. (1982), Statistical database design, ACM Transactions on Database Systems 6, pp. 113-139.

Cigrang, M. and Rainwater, L. (1990), Balancing data access and data protection: The Luxembourg Income Study experience, Proceedings of the American Statistical Association, Section on Statistical Computing, pp. 1-4.

Denning, D.E. (1982), *Cryptography and data security*, Reading, MA, Addison-Wesley.

Dobra, A., Karr, A.F., Sanil, A.P. and Fienberg, S.E. (2002. Software systems for tabular data releases, International Journal of Uncertainty, Fuzziness and Knowledge Based Systems 10(5), pp. 529-545.

Doyle, P., Lane, J.I., Theeuwes, J.J.M. and Zayatz, L.V. (2001), *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, Amsterdam, Elsevier.

DuMouchel and W., Schonlau, M. (1998), A fast computer intrusion detection algorithm based on hypothesis testing of command transition probabilities, The 4th International Conference of Knowledge Discovery and Data Mining, August 27-31, New York, pp. 198-193.

Duncan, G.T. and Mukherjee, S. (2000), Optimal disclosure limitation strategy in statistical databases: Deterring tracker attacks through additive noise, Journal of the American Statistical Association 95, pp.720-729.

Duncan, G.T. and Pearson, R.B. (1991), Enhancing access to microdata while protecting confidentiality: Prospects for the future, Statistical Science 6, pp. 219-239.

Elliot, M. and Dale, A. (1999), Scenarios of attack: the data intruder's perspective on statistical disclosure risk. Netherlands Official Statistics 14 (spring), pp. 6-10.

Fienberg, S.E. (1994), Conflicts between the needs for access to statistical information and demands for confidentiality, Journal of Official Statistics 10, pp. 115-132.

Gouweleeuw, J.M., Kooiman, P., Willenborg, L.C.R.J., de Wolf, P.-P. (1998), Post randomisation for statistical disclosure control: theory and implementation, Journal of Official Statistics 14, pp. 463-478.

Hawala, S. (2001), American FactFinder: U.S. Bureau of the Census works towards meeting the needs of users while protecting confidentiality, UNECE Statistical Division, Working Session on Statistical Data Confidentiality, Skopje.

Hout, A. van den and Heijden, P.G.M. van der (2002), Randomised response, statistical disclosure control and misclassification: a review, International Statistical Review 70, pp. 269-288.

Kafatos, M., Wang, X.S., Li, Z., Yang, R., Ziskin, D. (1998), Information technology implementation for a distributed data system serving earth scientists: Seasonal to interannual ESIP, Proceedings of the 10th International Conference on Scientific and Statistical Database Management, IEEE.

Keller-McNulty, S. and Unger, E.A. (1993), Database systems: Inferential security, Journal of Official Statistics 9, pp. 475-499.

Keller-McNulty and S., Unger, E.A. (1998), A database system prototype for remote access to information based on confidential data, Journal of Official Statistics 14, pp. 347-361.

Lees, P.J., Chronaki, C.E., Simatitakis, E.N., Kostomanolakis, S.G., Orphanoudakis, S.C., Vardas, P.E. (1999), Remote access to medical records via the internet: Feasibility, security and multilingual considerations, Computers in Cardiology 26, pp. 89-92.

National Institute of Statistical Sciences (1998), A web-based query system for disclosure-limited statistical analysis of confidential data, Research proposal of Digital Government project available through http://www.niss.org/dg.

Özsoyoglu, G. and Su, T.A. (1990), On inference control in semantic data models for statistical databases, Journal of Computer and System Sciences 40, pp. 405-443.

Rainwater, L. and Smeeding, T.M. (1988), The Luxembourg Income Study: The use of international telecommunications in comparative social research, Annals of the American Academy of Political and Social Science 495, pp. 95-105.

Reiter, J. (2003), Model diagnostics for remote access regression servers, Statistics and Computing, to appear in Statistics and Computing 4.

Ryan, J., Lin, M. and Miikkulainen, R. (1998), Intrusion detection with neural networks. In: *Advances in Neural Information Processing Systems 10*

(NIPS'97, Denver, CO), (edsitors Jordan, M., Kearns, M.J., Solla, S.A.), Cambridge, MA, MIT-press.

Schlörer, J. (1976), Confidentiality of statistical records: A threat-monitoring scheme for on line dialogue, Meth. Inform. Med. 15, pp. 36-42.

Shieh, S.P. and Lin, C.T. (1999), Auditing user query's in dynamical statistical databases, Information Sciences 113, pp.131-146.

Willenborg, L.C.R.J. and De Waal, A.G. (1996), *Statistical disclosure control in practice*, Lecture Notes in Statistics, New York, Springer.

Willenborg, L.C.R.J. and De Waal, A.G. (2001), *Elements of statistical disclosure control*, Lecture Notes in Statistics, New York, Springer.