WP. 20
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Topic (iii): Editing and Imputation in the context of data integration from multiple sources and mixed modes

STUDYING THE OPTIONS OF SUBSTITUTING A REGULAR STATISTICAL SURVEY WITH ADMINISTRATIVE DATA

Prepared by Gergely Horváth, Zoltán Csereháti, Hungarian Central Statistical Office

I. Introduction

1. Using administrative data sources for statistical purposes is wide-spreading in different areas. These kinds of solutions have numerous advantages compared with conventional statistical surveys. From the data providers' perspective this primarily means the reduction of administrative burden because they do not have the obligation to provide the same data for two or more different authorities (e.g. once for the Statistical Office and once for the Tax Authority). Usage of administrative data yields benefits for the Statistical Offices too. They may get data more quickly and do not have the burden to contact many statistical units, instead they may get all the data from one authority. Nowadays, data transfer is done almost exclusively by electronic means.

2. Difficulties: data quality depends mainly on the data collection, validation and editing job done by the administrative authority, As the authority collects the data primarily not for statistical purposes, these processes might not be sufficient for statistical needs. Consequently the resulting data may be biased.

3. Those authorities, that possess administrative databases, often maintain registers on their own. Their internal regulations affecting the update and maintenance practises of these registers may be quite different from the ones being in effect in the Statistical Office. Another problem is that the scope of the collected data does not cover perfectly the population targeted by the Statistical Office. Considering some of the variables, we must face considerable under-coverage, while by other variables some over-coverage can be observed. Regarding the methodology of data editing, there may also be significant discrepancies. Let us consider for example the Tax Authority. Clearly, it focuses on variables closely connected to the tax payments. These variables are validated and edited exhaustively. For another group of the variables, with no significant importance for the Authority, the applied data editing methods may not be so elaborated, or maybe fully omitted. The weighting procedure applied in the selective editing phase may also be different. Considering an administrative data source, the quality assurance policy may be in many aspects not as strict as in the Statistical Office. Some of the quality indicators needed for the Statistical Office are not being calculated by an authority. Furthermore, some of the documentation material describing the applied data editing and processing stages may be not available in an authority.

4. Many governmental bodies and state authorities possess such kind of administrative data that can be readily used for statistical purposes. The problem is that in many cases the Statistical Office does not have any signed agreement for the collaboration with them. This is why we must always take the possibility of being rejected into account, face distrustful behaviour, and reluctant collaboration. There is another feature of these authorities which can make the collaboration difficult. They may be aware of the

fact that some of their processes, databases, documentations may be incomplete, defective, erroneous, or simply not based on sound methodology. Consequently, they do not want to give an inside view of their work. This can be a major obstacle in the collaboration. This can be a problem even in the case if we try to do our best in explaining and clarifying the fact that we do not intend to supervise their work and policies.

5. There are several quality related issues that must be considered when using external administrative data:

(a) Coverage problems: The received dataset does not contain exhaustively all the statistical units, which form the target population of the statistical survey. This issue is not solely about coverage error, but also about the fact that the absence of some elements is not governed by the rules of chance, thus the available dataset can not be considered representative from the aspect of the survey; consequently yielding biased estimations.

(b) Timeliness: Often we must face the fact that not even the coverage problems are the most sensible part of our work. The reference period of administrative data sources is often much different from what is needed. For example, we may get corporation tax data from the Tax Authority for a wide range of enterprises, however the reference period of these data is 1.5-2 years earlier than the desired period, consequently these data can be used only as weak auxiliary information in some kind of regression models, but not as hard data.

(c) Untraceable origin: The source of some administrative data can be sometimes the data of other authorities, i.e. another kind of administrative data. That is why it is often not easy to trace the exact origin of the data. It may also be unclear whether these data have been subject to any data editing procedure in any of the affected authorities.

6. In this paper we will present the plans worked out in the Hungarian Central Statistical Office (HCSO) for the replacement of a survey with administrative data. In this project the needs of a special statistical field, the health statistics, meets the broader interests of the whole statistical office. A short and not too complicated annual survey was selected, as an experimental project, to examine the key elements of such tasks, replacing surveys with secondary data.

II. Survey of general practitioners' services

A. Description of the survey

7. Data collection, concerning the primary health care in the Hungarian statistical system has been carried out since the fifties. In the socialist era the ministry collected data, but following the democratic transition this task was given to the statistical office. Until 2006, there was a separate data collection for pediatricians’ and one for the general practitioners’ (GP’s) services, which were merged into one.

8. The main purpose is to provide statistics on general practitioners’ and family paediatricians’ services, the occupational health service, the activity of general practitioners and the occupational health service, on cared persons and those who are exposed to occupational hazard.

9. The data covers the number of general practitioners’ and family paediatricians’ services, the number of partions registered into each service, the patient flow by gender and age group, and the number of patients sent to hospitals and to specialists.

10. The data collection is complete. It is an annual survey, with a self-administered mail or electronic questionnaire (a downloadable MS Excel file). There is a statistical module in many software solutions used by GPs to record the patient flow, so almost the entire data of the questionnaire could be produced automatically. In most of the publications the data are presented on NUTS2 and NUTS3 levels.

B. The Data Entry and Validation system in the HCSO (ADEL)

11. In the previous section the main goal of this work and the key features of this survey were presented. Now we give a detailed description of the main software which is the processing tool for the survey data. It is one of the most important software tools in the HCSO, and used in almost every business survey in the editing phase of data processing.

12. The Data Entry and Validation system (ADEL, a Hungarian acronym) is an in-house developed framework which was introduced in year 2000. It provides the basic functionality needed to keying data from paper questionnaires and additionally supports input from e-questionnaires plus accepts administrative data from external sources like other Government Agencies.

13. The system is bundled to the databases of HCSO by means of using reference data and storing keyed / loaded / validated data directly. From technical point of view, it was developed within the framework and tools provided by the Oracle system like Pl-sql, Forms and sqlplus.

14. The ADEL system is integrated with different systems like survey control system (GESA), electronic data collection system and the meta system. The latter one is used to control access to the data and functions (i.e. user roles) and handle code validation, error categorization, process control and other meta-driven functions.

15. A role-based access control is implemented in the system. Users with basic rights can only select and report data or errors while the users with extended rights can control and modify data and accept / except specific errors providing explanations.

16. The ADEL framework itself is highly standardized and enforces the use of conventions in the development process on different levels.

(a) It provides a set reusable modules (report format transformers), conventions and standards on object level (compulsory objects in the DB and compulsory folder structure in the file system).

(b) It uses strong naming convention both in the DB and in the file system.

(c) Necessary building blocks (standard first page) and validation modules/codes are generated from meta system.

(d) It provides standard user interface like font type & size, colors (for input and display items), navigation style between tables/rows/items and using of keyboard or mouse actions.

17. The ADEL framework has standard menu structure with functions like data entry, online validation, background validation, data and error reporting (the reports can be plain text or in MS Excel format), management information, quality reports and online help, online user guide. Different validation rules are implemented within ADEL system

(a) inside questionnaire (between tables, rows, columns),

(b) between questionnaires (previous periods, other surveys).

18. The validations are categorized by statisticians and this information is stored in meta system including error levels and error messages. The list below explains the categories and the error handling methods:

(a) warning : information level, can be ignored,

(b) caution: must be corrected or can be accepted by the statistician with explanation ,

(c) error: must be corrected or accepted by super user with an explanation ,

(d) critical: must be corrected immediately, cannot be continue without it.

19. The ADEL framework provides management information about

(a) survey status: data entry or receiving e-questionnaires is allowed or prohibited

(b) processes: audit trail, who & when started the process

(c) questionnaires: who and when entered/loaded, modified, deleted, controlled the questionnaire, duration of process (by user and questionnaire)

(d) errors: levels, status (accepted or not), who and when accepted, explanation text

20. 103 surveys used this framework in HCSO, in 2011.

C. Data processing in the survey of the general practitioners' services

21. The deadline of the onset of the questionnaires is end of January, then data typing and corrections last till the end of June. This work is done by one of the regional directories of HCSO in Miskolc as well as the pressing of the data suppliers. Data process is a fully manual task consisting only of micro editing. There is no imputation; missing entries are corrected by a recall of the data supplier or according to the contexts within the questionnaires.

22. As a first step, data are sight-controlled and compared to the questionnaire from last year. Missing or mistyped data are corrected by the help of the data supplier, if needed. The next step of data process is driven by the electronic system called ADÉL. This system monitors and filters most of the mistakes coming from code validity and from breaking the logical contexts within the questionnaires during data-entering. Mistakes are corrected by the competence centres. Total is revised according to the sum of the internal data.

23. After the editing phase, the next step is the production of some preliminary tables to compare the trends and identify significant changes. There is also outlier detection in that phase. There is also some manual correction, using the knowledge of experts. The finalized data from the ADEL (the previously described data validation system) are transferred into the general production database. The publication tables are made by an Oracle-based application. The data are available via the homepage of the HCSO and the data are also loaded into the public Dissemination database.

24. The data of the survey are published in various forms (yearbooks, thematic publications), with different detail levels, but these data are available via public databases. Below there is a list of some publications:

(a) Statistical Pocketbook of Hungary (print)
(b) Statistical yearbook of Hungary (print)
(c) Yearbook of Health Statistics (print and electronic)
(d) Social Care Systems of Hungary
(e) Yearbook of regional statistics
(f) Pocketbook of the regions of Hungary
(g) Public databases on the website of the HCSO
  - MR-STAR (Database of regional data)
  - T-STAR (Database of settlements)
  - Dissemination Database

D. Review of the data validation / editing rules of the survey

25. Changing the data source of this kind of statistics will have impact on the whole data processing, so we revise the editing and validation rules. As a first step, the survey's 260 validation rules have been examined and categorized. At the end of this work four groups can be formed on these rules. Our goal was to identify the most important attributes of validation / editing rules. The classification of the types of edits will help us to identify the role of edits in the data processing.

(a) PREV: These rules are based on the data of the previous year, in most cases this is a simple comparison of the two data. The general threshold is 25%, anything is lower or higher are marked as erroneous.

(b) VALID: Validity checks for categorical variables to correct typing mismatches as errors.

(c) CONSISTENCY: This is the most frequent type of rule. We can check the internal consistency of the data utilizing these rules. The general form is an IF ( ) THEN ( ) statement.

(d) SUM: In the questionnaire, sums are asked, and these sums are checked.

1. Table Number of validation rules by type and importance

Type of rule	Importance of rule		
	Error (2)	Critical (3)	Total
PREV	15		15
VALID	2	27	29
CONSISTENCY	40	159	199
SUM		17	17
Total	57	203	260

26. In this table one can notice that the most numerous cell, and the most frequent type of edit, is categorized as critical importance and belongs to the CONSISTENCY group. The critical error means it must be corrected before the questionnaire is finalized. These are quite rigid rules, but during the editing phase, it is possible to recontact the doctors, so the correction is not a really big problem.

27. When replacing a survey with secondary data completely we need to review the whole data processing, including the edits. When a statistical office carries out a survey, there is a direct connection with the data supplier. So it was not too complicated to contact the data provider (in most cases, via telephone) if errors were found in the questionnaire. But if we get the data from a government agency (National Health Insurance Fund Administration – NHIFA), we can only use their knowledge. In our opinion – as far as we know now – in the future, many of the edits can be skipped. The main reason to skip the edits is that we will not have the possibility to recontact the final data provider, the general practitioners. Two bigger reasons were found:

(a) The rule defined in the edit never will be broken (this can be the case of the SUM and VALID type of edits)
(b) To find the real cause of an error is not possible.

III. The B300 report

28. Report B300 will be the new data source for the statistical data based on administrative data. NHIFA collects detailed data from GPs on the number of patients in electronic way since September 2006. The aim of the data collection is mainly in connection with financing the GPs. The report must contain each doctor–patient meeting and its details (diagnose or possible hospitalization, etc.). It is easy to fit in the computer aided registers used by the doctors (data framework of the report is publicly available on the webpage of NHIFA).

29. The report contains data of the patient flow’s diary collected in electronic way from the beginning. Almost all the basic data of this statistical report got into one state-owned organization. So it seems to be practical to collect data directly from National Health Insurance Fund Administration, which means the source must be administrative from now on.

A. Data processing and use of the data

30. The report does not contain the variables of gender and age of patients. They are not necessary for the financing, but important for statistics. These data can be reached at National Health Insurance Fund Administration, and can be shared by the HCSO if needed.

(a) There is one row for the identification data of the general practitioner's service

- Code of the service, operator
- Start and finish dates of reporting

(b) Detailed data of the patient flow

- Date and time of the visit
- Location (at home, consultation room)
- Identifiers of the doctor and the patient
- Data on financial issues

(c) Other data of the treatment

- Diagnoses
- If the patient sent to specialist or hospital
- Medicaments, medical supplies

31. From the viewpoint of statistics, there is a quite serious drawback of the B300 report, the absence of the gender and age of the patients. It is obvious that these are key data for any statistics concerning population and social issues. Certainly the NHIFA has these data, and will provide it to HCSO, but the source is not the B300 report, but other register maintained by the NHIFA.

32. The data processing workflow of the B300 report in the NHIFA is different from a standard statistical data process. The main goal of the data collection is to produce solid base for the funding of the general practitioners' services. The completeness and the error-free internal consistency of the data is the most important issue. There is a series of internal checks to achieve this. Error lists are produced and sent back to the doctors. The correction is always made by the doctors, not by the clerks of the NHIFA. The error rate is very low, only 0.1-0.2%.

IV. Issues of replacement of the data source

A. Similarities and differences

33. There are a sort of common attributes in both the statistical, and the administrative data collection. The most important is that the observation unit is the same, the GP's service. The ultimate data source is also the same, the register software of the patient flow. The softwares have the ability to produce both the statistical and the administrative reports. The person responsible for the data is also the same in most cases, the doctor, running the GP service.

34. As the data provided to the NHIFA are more detailed and more frequent, the patient flow register softwares are built to meet the needs of the NHIFA rather than the statistical office. After a series of meetings, the experts of the NHIFA stated that the almost everything covered in the statistical data collection can be provided from their databases. When carrying out the replacement, attention should be given to one important difference. The number of persons registered in a doctor's register and the number of patients insured and registered to a GP service can be different. The main reason that someone can lose the insurance, but still can be in a doctor's register. The NHIFA will provide the number of persons with valid social security insurance and registered to a certain GP service.

35. The statistical office does not want to change the frequency and the contents of these statistics. The change of the data provider only affects the data source, nothing else, except one thing: the data of occurrence of diseases can be an annual data collection, because it can be provided using the data of the B300 report.

36. Preparing the replacement of the data provider, a decision was made with cooperation of the experts of the NHIFA to compare data of the two sources from the previous years, 2009 and 2011. The same set of data will be produced as the statistical data, based on the B300 report, so we can study the data processing workflow and also the data. The aim of this test is to discover the differences (if there is any) and try to explain the changes in the results after the replacement of the data source. This work has not finished yet, we just made the first steps.

37. One of the first tasks will be the comparison of the observation units. We thought it is a very simple task, because both dataset contains the code of the GP service. Therefore it seemed to be a very easy task, using a key. It turned out, however, that in some cases the identifier is not identical. We had to use other methods to connect the datasets.

(a) The possible connection methods (elements of generated key)
- HSZKOD (official code of the GP service)
- Name of the doctor
- Address of the general practitioner's service

38. When matching identifiers are found, it is simple to make the connection, but in other cases, we have to use more complex data integration techniques. For the first experiments we used the most simple manual techniques to find the matching records. Here are some of the possible inconveniences

(a) Form and writing of the name
- “Dr” prefix (is it in the database or not)
- Variants of married name in the case of women
- Spelling issues: “t” or “th” in the end of the names (as in Horváth) or “i” versus “y”

(b) Accuracy and spelling of the address

39. Before the final decision of the replacement is made, we have to produce the most frequently used estimations and tables from the two different sources in a comparable form, to find out in which domains are he biggest differences. We have to draw attention of the users to the changes of the source and the possible consequences.

B. Suggestion for data processing

40. While planning the data processing workflow of the changed data source, we have to take the attributions of the data collection into account. In a traditional statistical data collection, there is a direct connection between the data provider and the statistical office, but in the case of using secondary data, we do not have this connection. Speaking a little technically, this form of data collection is just a series of queries running on the databases of another government agency.

41. Previously it was mentioned that the edits and validation rules applied have to be reviewed. These rules will be applied on the 2009 and 2011 experimental datasets of the NHIFA. Based on the results, we can asses the applicability and usefulness of these 260 rules. The most important difference in the workflow is that, using secondary data, a great amount of edits will be applied to the basic data before the statistical office receives them. We have to take the complete data processing flow also into account, from the doctor's register through the NHIFA to the last step, the queries, producing the statistical data.

42. In the preparatory phase, all of the edits will be applied to the NHIFA data. We have some prior hypotheses on the applicability of the edits, but we have to check and analyze the results carefully. Some remarks on the edits:

(a) SUM type edits: The variables checked in these edits can be defined as functions of other variables. If parts of the functions are valid, the result will always be acceptable.

(b) PREV type edits: What if there is a change in the threshold? We will not be able to contact the general practitioner directly, so we can ask the experts of the NHIFA for special support (a deeper query in the database), or we can simply accept the data.

(c) CONSISTENCY and VALID type: Examining these edits, many of them just checks the fill in, but if the production of the "questionnaire" (a record in a dataset) is made by queries, we also have to check the definition of the query, and then the data (but an erroneous query introduces great amount of systematic error into our data).

V. Conclusions

43. In the HCSO, as in every statistical office there is an extensive use of secondary data. But until now, there has been no real need to completely replace a statistical survey with administrative data. The Survey of General Practitioners' Services will serve as the first example on this field. This experience might also serve as a good example for future work with data from secondary sources and also to gain invaluable experimental and practical knowledge.

44. One of the biggest lessons so far can be that no matter how a "replacement" task looks (transparent statistical process and a good potential source of administrative information), such an undertaking is never easy and straightforward. In this experimental case, even though the ultimate source of the data is the same (the patient-flow register of the GPs), using NHIFA data might reduce administrative burden on a special group of our data providers, altogether with data quality improved (we expect more timely and more accurate data), the risk of introduction of bias and distortion is always there. It is also clear that even during the preparatory work, in order to maintain confidence in our users, we have to discover the possible changes in the data, and find the right explanations.

References

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE: Using Administrative and Secondary Sources for Official Statistics, New York and Geneva, 2011 (http://www.unece.org/fileadmin/DAM/stats/publications/Using_Administrative_Sources_Final_for_web.pdf)

Erzsébet Kómár: User Guide of ADEL data processing system (only in Hungarian), 2009

Anders Wallgren, Britt Wallgren: To understand the Possibilities of Administrative Data you must change your Statistical Paradigm!, Joint Statistical Meetings Section on Survey Research Methods – JSM 2011 (http://www.amstat.org/sections/srms/proceedings/y2011/files/300347_64422.pdf)