Working Paper No. 1
ENGLISH ONLY

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint UNECE-EUROSTAT Work Session on Registers and Administrative Records for Social and Demographic Statistics
(Geneva, 9-11 December 2002)

Supporting paper – SESSION 1

TOWARDS THE USE OF ADMINISTRATIVE RECORDS AND REGISTERS FOR STATISTICAL PURPOSES AT THE HUNGARIAN CENTRAL STATISTICAL OFFICE

Submitted by József Sánta, Hungarian Central Statistical Office

I. INTRODUCTION

1. In Hungary administrative data sources are not used in demographic and social statistics as much as they could be. This year a Committee and Working Groups were set up at the Hungarian Central Statistical Office (HCSO) to deal with the problems of data collection (surveys) in social statistics. The commission of one of the Working Groups is to study the situation of the use of administrative data sources for statistical purposes.

2. The main objectives of the Working Group in question are as follows:

- exploring those administrative data sources which could be used for statistical purposes;
- developing a meta-data base to store the most important documentations, instructions and description of methodological procedures of the data sources;
- continuous analysis of the information received;
- submission of proposals to the management of the HCSO for using the administrative data for statistical purposes;
- elaboration of methodology, testing, proposals for matching of statistical and administrative data;
- clarification of data protection issues from legal and technical points of view;
- consultations with the management of the administrative data sources to consider statistical needs in maintaining their data.

3. There is a marked backlog in the field for two reasons: one is the legal regulation of data transfers in the public administration, the other one is internal as, in the past, the HCSO was not compelled to exploit administrative data sources to a greater extent. In parallel with this process the matching of data sources was also a neglected area. Even in the case of statistical files as own, matching has almost never been applied.

4. The literature distinguishes exact and statistical matching (Federal Committee on Statistical Methodology, 1980 and Winkler: Matching and Record Linkage, Business Survey Methods, 1995, John Wiley and & Sons). An exact match is a linkage of data for the same unit from different files; linkages for units that are not the same occur because of error (e.g. a mistyped address). Statistical matching attempts to link files that have few or no units in common and the linkage is based on similar characteristics.

5. This paper provides an overview of the state of affairs, laying emphasis on matching. The HCSO has only now begun the process of using administrative records and registers for statistical purposes and much effort will be required to improve this situation. It should be mentioned, however, that in the field of economic statistics there is well-established cooperation between the major agencies of the public administration in this respect.

II. LEGAL BACKGROUND

6. The operation of the HCSO is basically governed by Act XLVI of 1993 (modified by Act CVIII of 1999) on Statistics (briefly Statistics Act). The Statistics Act refers to and is "in harmony with Act LXIII of 1992 on Protection of Personal Data and Disclosure of Data of Public Interest" (briefly Data Protection Act). Both Acts are the fruits of the transition period and the Data Protection Act was born first.

7. In order to understand the constraints under which the HCSO can use administrative data sources, it is worth mentioning certain articles and paragraphs from the Data Protection Act:

- Article 7 para 2: Unlimited, general and uniform personal identification code shall not be used.
- Article 8 para 1 : Data shall not be transferred and files shall not be connected unless consented to by data subject or provided for by law. The conditions for data processing shall meet in each case with regard to each personal data.

- Article 8 para 2: Connection of files processed by the same controller, as well as those of state organization and self-government shall likewise be governed as in para 1.

8. As a state organization, the HCSO has to comply with para 1, which could make its functioning impossible. In order for the HCSO to surmount this problem, some "concessions" are coded into the Statistics Act:

- Section 19 subsection 2: In case of surveys … covering a period of over one year, the data- stock shall be given an inner identification code on basis of which the identity of the person concerned cannot be established. The personal identification data of the person shall be managed separately from the data stock. …
- Section 19 subsection 3: For the time of adding new data to the data-stock and of carrying out a sampling process in order to collect statistical data for the same purpose the personal identification data may be temporarily linked with the data-stock. The rules of data linkage shall be established taking into consideration the standpoint of the Commissioner (Ombudsman) of Data Protection, and be submitted to the National Statistical Council to request the latter's opinion about it.

III. POSSIBILITIES AND LIMITATIONS IN THE USE OF ADMINISTRATIVE DATA

III.1 A strange Population Census

9. In general, the population censuses provide a good basis for framing, sampling or simply comparing data from other sources due to the good coverage, independently of the methods used (traditional, register-based or combined) to take the census. Obviously one of the essentials is the possibility of identification. The 2001 Hungarian Population Census had to be conducted under unprecedented circumstances. Firstly, at the request of the representatives of the registered national minorities, no names could be used on the census forms and later on, at the request of the Data Protection Ombudsman, the same applied to the addresses. Now the only linkage between the Census and the "external world" is the address file, which was used for conducting the Census. There is no possible way to exactly match persons between the Census and another data source. An exact matching is limited to addresses, which is known to be a potential source of error.

III.2 Social Sampling Surveys

10. In the case of social surveys the situation is somewhat better. Here the target sampling unit is a household within an address. In Hungary multi-household addresses are not unusual. For repeated or continuous surveys (e.g. labour force, household budget) the persons interviewed are identified with an inner code according to the method laid down in Section 19 above.

III.3 Businesses

11. In the business world of Hungary, businesses are identified by a unique identifier, which is used in several registers and administrative data sources. A single-window registration system was introduced in 1999 in the public administration, in which data-flow between different registration agencies is permitted under the control of laws and high level decrees. A part of the data registered is public.

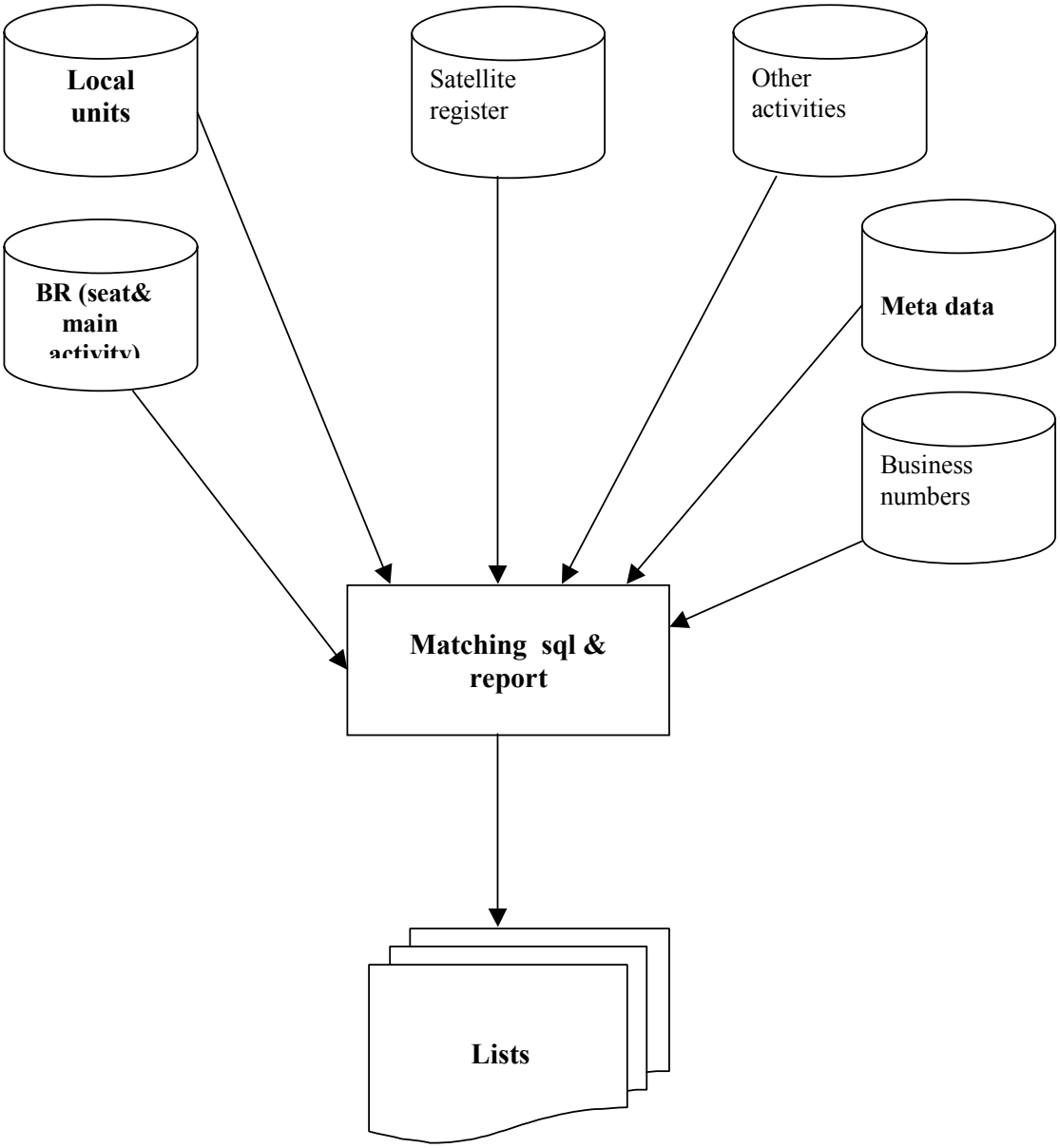
IV. EXPERIMENTS WITH MATCHING

12. According to the program of the Working Group, matching experiments began this year, focusing on census data. For the time being quality analyses and findings are far from complete. The experiments are intended to determine possibilities, to justify their feasibility and expected benefit. The main obstacle so far is that some matching variables are derived variables (e.g. highest degree of education, economic activity created from livelihood sources) and require the complete processing of other variables.

IV.1 Subtask 1: Linkage of business data

13. As this subtask deals with economic data and registers, the only aim here is to present a solution under ideal circumstances. Exact matching could be achieved between different data sources by using a common business identity number. The data sources linked were as follows: the HCSO Business Register (BR), its satellite register consisting of businesses operating in retail, restaurant, and hotel/accommodation activities, as well as the Tax Register maintained by the Tax and Auditing Office. The BR, as a part of the single-window system, uses a lot of elements from administrative sources (see Figure 1).

Figure 1 - Links for the Business Register in Oracle



IV.2 Subtask 2: Linkage between Population Census and Business Register data

14. It is well-known that many statistical offices play with the idea of using administrative data to a greater extent than before to combine with or substitute for census data. In 2001 the HCSO conducted a traditional population census using enumerators and coders, but took the opportunity to take the first faltering step to use data from the BR for post evaluation purposes and towards a potential future combined census. In past censuses, the coders used printed manuals for coding the variables of mostly state-owned enterprises. In 2001 it was impossible to apply the old solution because of the large number of businesses established in the transition period. It could be done just for enterprises with more than 20 employees (some 30 thousand). For these enterprises the business identity number was mapped to an inner identifier, which could be used by the coder from the shortened and manageable manual. The optical reader worked with this inner identifier.

15. The variables involved in the post evaluation are as follows. On the census questionnaire there were questions about the employer : settlement, where the employer is located, main activity, rough scale of number of employed, settlement of the work place (local unit in BR terminology). BR contains the main activity (and secondary activities, if they exist), registered addresses for the seat and the local unit(s), number of employed.

16. The main goal of the post evaluation is to explore the differences between the data registered and the data supplied by people enumerated and coded by coders.

17. By now two companies (with about 500 employees) were selected for comparison, one from a small town and another one from the capital city. The following differences were found:

- 0.5 % difference in number of employed in the small town, while in the capital city the difference was much greater;
- identical main activity;
- dispersion in the data of workplaces reported;
- suspicious commuting data.

18. As the link has been established through the identifiers the comparison can be easily extended for all the companies concerned in order for us to see the main reasons for the considerable differences in the number of employed.

IV.3 Subtask 3: Matching Population and Agriculture Census data by Addresses

19. Prior to the 2001 Population Census (PC), a General Agricultural Census (AC) was conducted in 2000. The farms (agricultural households) are identified by addresses. The AC form contains a few demographic and employment variables about the persons living in the household dealing with agriculture. These variables are, among others, the following: sex, age, highest degree of education, economic activity, worked days in agriculture. It is reasonable to conclude that, for deeper analyses, the AC data should be combined with more variables from the PC. This requires a match based on addresses.

20. In the HCSO there is neither commercial (e.g. Search Software America) nor in-house developed software in use for address matching. In spite of some efforts to standardize the form of addresses, there still exist various differences in the same addresses used in different data sources, not to mention the diversity of databases. Nowadays the change of street names is quite usual.

21. For experimental purpose a few rural settlements were selected for matching. In the experiment an address consisted of the name of a public domain, a number and an additional special character, where necessary. It should be noted that in urban areas an address is more complicated, e.g. it may contain floor-number, door-number.

22. The subtask was carried out in the following steps:

- Creation of clean and common structure of addresses – this involved manual corrections and adjustments;
- Matching PC and AC records by address;
- Matching persons within an address by demographic/social variables - the matching variables were as follows: sex and age for exact matching; highest degree of education and economic activity for "loose" matching. The looseness of the matching stems from the fact that the structure and the length of the code-list of highest degree of education and economic activity are not the same for the two censuses.

23. Results of the address matching:

- PC addresses: 3552
- AC addresses: 1425
- Rate of matched addresses: 99.8 %
- Rate of addresses not found in the PC address file: 0.2 %

24. It is not to say that the address matching always works with this high efficiency but rather that in the rural environment the address structure is quite simple. The step iii will be done later once the personal census data are available.

IV.4 Subtask 4: Matching labour force survey and population census data

25. Typically this is a task that could have been done a long time ago. The social surveys are not mandatory and therefore the non-response rate may exceed an undesirable threshold. As a population census provides lot of information about people, there is a plan in the HCSO according to which the social environment of the non-responsive households is to be depicted with the help of census information. The linkage of the two data sources can offer a solution to the problem. The data sources can be linked by addresses. In the case of multi-household addresses, only the smallest household is selected. For the execution of the task a non-response "flag" must be present in the survey data set, while the descriptive variables in the census data set are the age, highest degree of education and labour force status of the head of household, as well as the environment of the dwelling.

26. In 2002 a deep analysis dealt with the problems of non-response.

IV.5 Subtask 5: An Attempt at Statistical Matching between Social Survey and Population Census Households

27. The sample frame and consequently the samples of social surveys are not too large. This is why the census data sets had to be sufficiently large in order to obtain as many matched pairs as possible. Four county centres (towns) were selected for the experiment. To avoid linkage of data of same households, data from separate towns were intentionally linked on the basis that two households are matched if the household's characteristics are similar (a kind of nearest neighbour method). From multi-household addresses only the smallest household is selected, as in Subtask 4. The matching characteristics were: number of persons in the household, household composition by family status and a scale of average age. These indicators are mapped into a single number that is really the sorting key.

28. The method can be used – with different criteria – in donor imputation also. It was used in the case of the Population Census for imputing missing personal data.

V. OTHER ACTIVITIES IN SCOPE

29. An important task of the Working Group is to explore potential administrative data sources. This work is facilitated by the help of the Data Protection Register, which operates at the Parliamentary Commissioners's Office of Hungary for the Data Protection Ombudsman, and which contains information about all records and registers maintaining data on persons. Many offices holding register as members of the national statistical service are in regular contact with the HCSO. The work necessary in order to collect information from these agencies is considerable and requires human resources and time.

30. In the meantime and in parallel, statisticians should prepare legal proposals and try to convince the legislators to support the HCSO in prevailing its interest.

- - - - -