Working Paper No. 10
2 May 2003

ENGLISH ONLY

UN STATISTICAL COMMISSION and
UN ECONOMIC COMMISSION FOR EUROPE

STATISTICAL OFFICE OF THE
EUROPEAN COMMUNITIES (EUROSTAT)

CONFERENCE OF EUROPEAN STATISTICIANS

Joint ECE-EUROSTAT Work Session on Population and Housing Censuses
(Ohrid, The former Yugoslav Republic of Macedonia, 21-23 May 2003)

Session I – Supporting paper

CENSUS 2002: AUTOMATIC AND SEMI-AUTOMATIC CODING

Submitted by Serbia and Montenegro[1]

1. The seventh Census of population, households and dwellings since World War II was conducted in the Republic of Serbia on 1-15 April 2002. After completed terrain work in which 250 republic instructors, 4700 municipality instructors and over 38000 enumerators participated, the ***First results of the Census*** were released, giving preliminary data about number of population and dwellings by municipalities.

2. In the next phase, we worked on data entry into the united base. Data entry itself took about five months and, during that time, about 7,5 million personal questionnaires (form P-1 that refers to person) and 2,5 million questionnaires for household and dwelling (form P-2 that refers to household and dwelling) were entered.

3. Data were delivered in authentic form for the data base from forms P-1 and P-2, but during the process of entering, a few fields were coded where the name of settlement or municipality should have been entered.

4. After data entry into the united base, the coding of three characteristics from the personal questionnaire was completed:

- educational attainment
- occupation and
- economic activity.

5. Data entry operators retyped data from personal questionnaires in original form. Passing the complete entered material through software designed for automatic coding, i.e. recognizing the text, resulted in a much accelerated process of coding. **This is the first time that such a method of coding of Census material has been used in the Republic Statistical Office**.

[1] Paper prepared by Dragana Djokovic-Papic and Mirjana Popovic of the Statistical Office of the Republic of Serbia.

6. Experts of the Republic Statistical Office realised the program for automatic and semi-automatic codifying in the program Visual basic for SQL server base. The coding of material has been undertaken on the municipality level. In this way, the Republic Statistical Office has coded 100 municipalities out of a total of 161 municipalities in the territory of Central Serbia and Vojvodina, while other municipalities have been codified in another two destinations. We shall present the work on coding these 100 municipalities, which represent a little more than half of the total population (52,9%).

7. As far as scope of work is concerned, it should be noted that, during two phases of work in case of these 100 municipalities, over 2.600.000 syllables have been coded in total, using automatic and semi-automatic coding (see Table 1 below).

	AUTOMATIC CODIFYING	SEMI-AUTOMATIC CODIFYING	TOTAL
Educational attainment	177779	565491	743270
Occupation	604100	361555	965655
Economic activity	719322	189042	908364
All	1501201	1116088	2617289

Table 1

8. Software for coding has made ***automatic coding*** possible as a first step. It is important that text in certain fields of the personal questionnaire correspond with text in the respective software code base.

9. We give below an example of information that, for each of the three fields on the territory of Central Serbia without Belgrade, automatic coding was carried out:

- educational attainment 23,9%;
- occupation 62,6%;
- economic activity 79,2%.

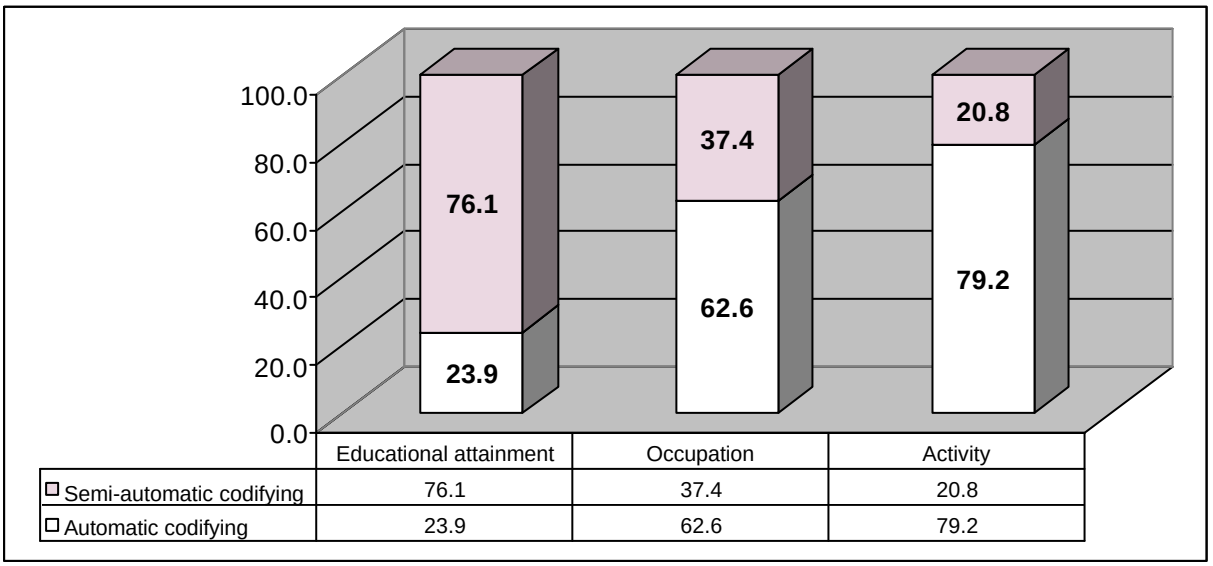

Chart 1

10. ***Semi-automatic coding*** is the second step. This enables the coding clerk to code those syllables which did not correspond completely with those offered in classifications. It should be mentioned that if more syllables contained identical text, this software, by coding this one concept, could accomplish coding of all syllables.

11. On the basis of a selected part of the text (part of the word, word, more words), the software, after a complete search in the software code base, offered a list of possible denominations. Except for the textual part of the code, the code clerk had an opportunity to see the numerical part of the code, too, and this was of great help in making final decisions for those who were well acquainted with specific classifications.

12. If there were syllables for which the text for a specific field was not of sufficiently high quality for coding, there is an option to look at each syllable. Thus, more data is visible for each person that may help in coding a certain field:

- Identification of a syllable (municipality, enumeration area, dwelling, household, ordinal number of a person):
- Gender:
- Date of birth:
- The highest completed schooling:
- Name of a secondary, high or higher school:
- School that person is attending at the moment:
- Economic activity:
- Occupation:
- Sector where the person works.

13. The coding has been undertaken by statisticians specialised in individual areas and by the best operators who entered Census material. After the coding was completed, the software made it possible to survey the coded material, as well as to make any changes deemed necessary by the supervisor.

14. The control of coded material was conducted in two ways:

- by visual control; and
- by logical control.

15. Supervisors were specialised experts in each individual field who also participated in preparations and adaptations of the three classifications for the Census needs. Checking of the material was adapted to each individually coded field and what is most important is that some basic rules of logical control for those fields were included, so that the material was checked and updated already on that level depending on specific needs. Amendment of checked material was made by mask choosing codes on the level of individual syllables.

I. EDUCATIONAL ATTAINMENT

16. As far as educational attainment is concerned, the question is *"What is the highest completed schooling of person that is being listed"*, and if it is secondary, high, or higher, it is important to specify their name or course. Each of these schools is coded by classification ***List of schools***, which is adjusted to international classification ***The International Standard Classification of Education (ISCED 1997).***

17. If the two fields are conflictual (numeric part of the answer of the highest completed school and its nomination), or if the answer was obtained on the basis of a statement about highest

completed school during work on software for coding, then there was a need for a cross-breed of those two attributes.

18. A conflict of these two fields is possible for other reasons:

- Subjective nature of statement;
- Mistake of enumerators during encirclement;
- Operator's mistake at input.

19. The low level of automatic coding efficiency in the field of educational attainment (23,9%) is the result of incompatible and unequal denominations of completed secondary, high and higher schools, or different textual forms of denominations, all of which made automatic coding difficult. For example, there were also answers with full denomination and place of educational center, so it was not possible to conduct automatic coding.

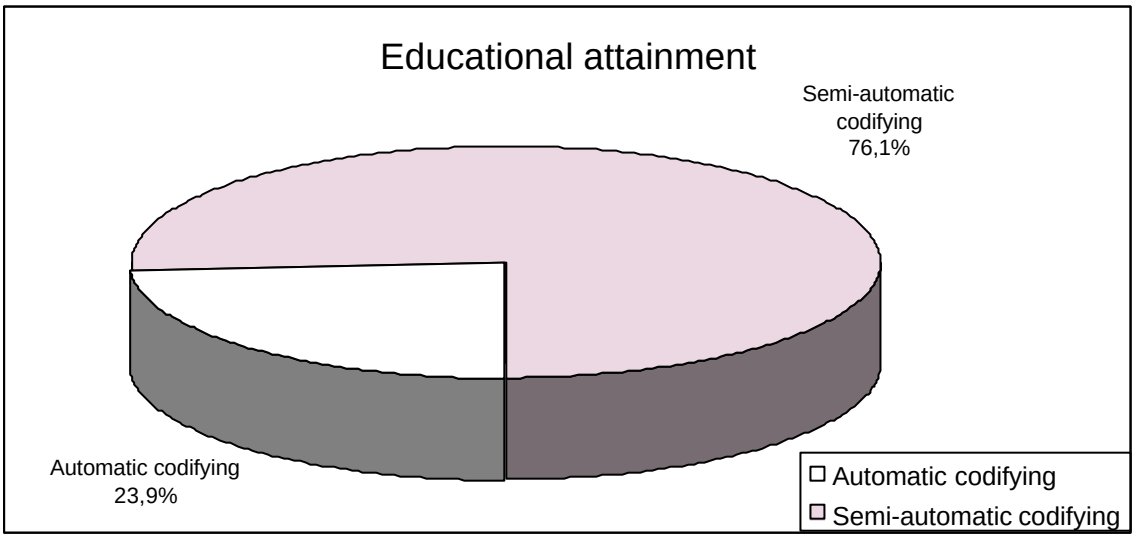

Chart 2

Automatic codifying is successful if complete entered text matches with specific modality from software code base.

II. OCCUPATION

20. As far as denomination of occupation is concerned, coding was conducted according to new ***Classification of occupations*** that is adjusted to international classification ***International Standard Classification of Occupations*** (ISCO-88). Experts of the Republic Statistical Office (RSO) and the Federal Statistical Office have translated and adjusted the ***ISCO*** classification. The ***Methodical list of occupations***, which was prepared for the Census 1991 and it is still used in many statistical surveys conducted by RSO, was also used for completion. In 1999, the classification of occupations was tested in a Trial Census accomplished on a chosen sample.

21. To make coding of material as successful as possible, keys with valid ***Unique Classification of Occupations*** are made. ***Unique Classification of Occupations***, as well as the official classification, is used in registrations of employees. In that way, a rich fund of denominations of occupations is available (almost six thousand).

22. However, what makes automatic coding of classification of occupations difficult is:

- Different combinations of complex denominations of occupations;

- Imprecise denominations of occupations;
- Specifying of title ranks;
- Specifying of denominations of activities instead of occupations;
- Mistakes in typing.

23. As we already mentioned, it is possible to survey and amend coded material. Logical control is conducted according to rules of new classification of occupations, which demands, along with specific occupations, appropriate educational attainment, too. Noticeable mistakes in coding can be corrected immediately.

Efficacy of automatic codifying, that amounted 62,6%, can be regarded as very satisfactory.

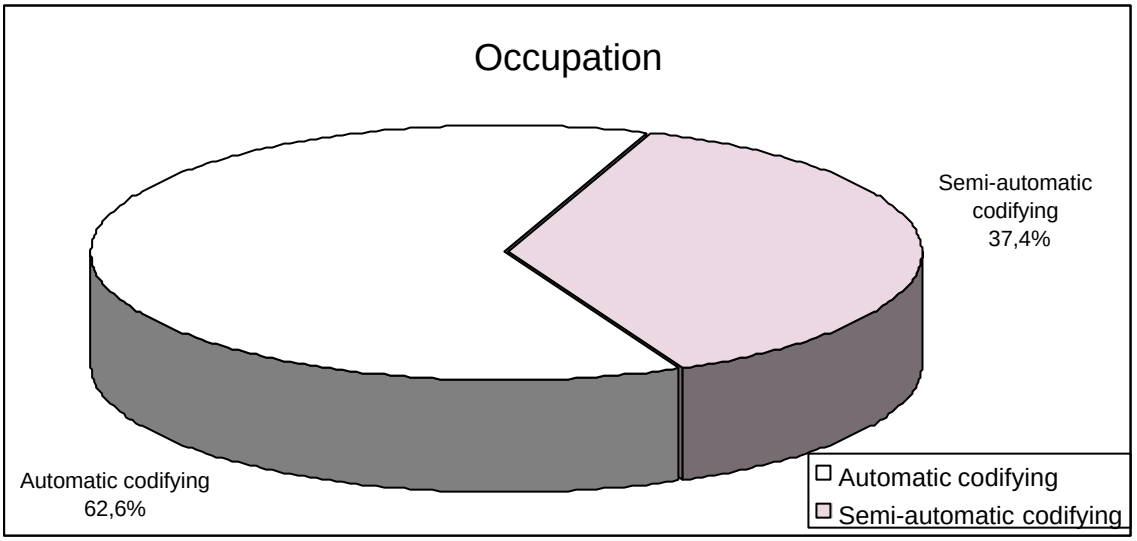

Chart 3

III. ECONOMIC ACTIVITY

24. Ultimately, as data on efficacy of automatic coding also demonstrate (79,2%), coding of activities was most efficiently accomplished, in great part thanks to the PL form that was an auxiliary application in Census 2002. Companies and enterprises were obliged to complete this application for each employee and to give official code and denomination of economic activity that the enterprise practiced. This had to be done before conducting the Census.

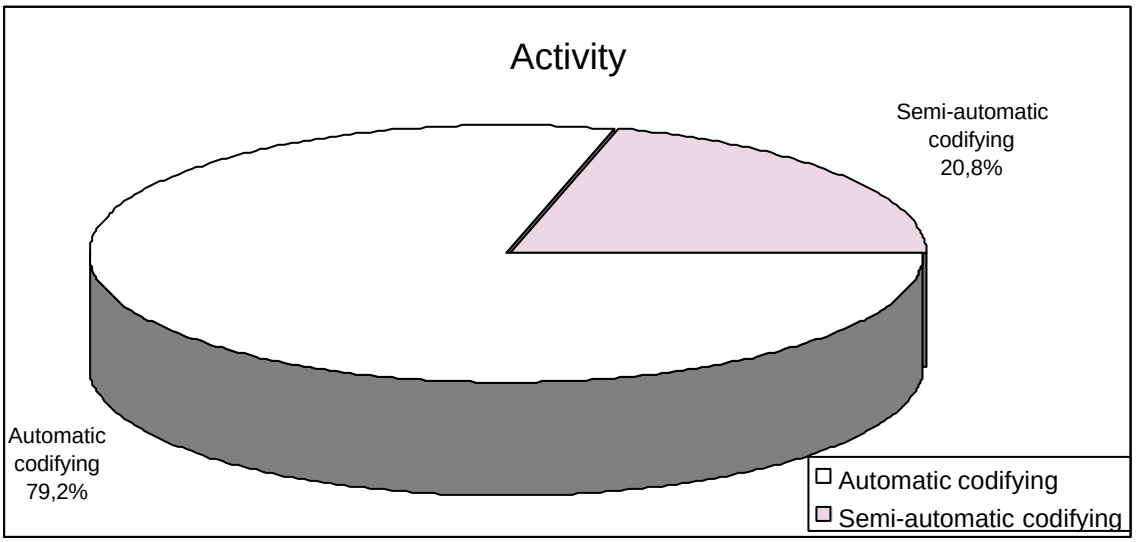

Chart 4

25. The only problem that we confronted in this field was specifying of old codes of economic activities from ***Uniform Classification of Activities*** that consisted of six numerical places, not new ones with five, from ***Classification of Activities*** that is completely comparable with ***Statistical Classification of Economic Activities in European Community NACE*** Rev.1.

26. Disregarding problems that occurred in this field, the efficacy of automatic recognition is very high, above all due to the numerical part of characteristics of economic activities. Even in the phase of automatic, as well as in the phase of semi-automatic, coding, the software enabled work with old as well as new classification of activities.

27. Experts from the Republic Statistical Office prepared keys for the transferal of old to new codes of economic activities to be used for Census purposes after coding is completed, as well as for the requirements of other statistical surveys and registers.

28. The important thing during the process of coding of this characteristic is that the software approach to ***Register of organizational units*** was made possible, so the code clerk could also have access to data on the basis of enterprise's and shop's name and they took over the appropriate denomination or the code of economic activity.

29. Taking everythin into consideration, the Republic Statistical Office of Serbia has a very positive experience with such methods of work on the coding of Census material.

30. Automatic coding has improved and accelerated this phase of data processing through appropriate and comfortable work, because, in the real sense of the word, we are talking here about minutes for its realization on the municipality level.

31. On the other hand, semi-automatic coding, access of individual syllables, control and making of amendments represent complex and responsible phases of work that have to be improved further.

32. Such methods should also be used for coding denominations of educational attainment, occupations, and activities in the other regular statistical surveys. Gained experience of code clerks and supervisors during 4 months of work will be of great help for further development of software and data processing.