WP.36
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Vienna, Austria, 21-23 April 2008)

Topic (vi) Censuses

DATA IMPUTATION AND ESTIMATION FOR THE AUSTRIAN REGISTER-BASED CENSUS

Invited Paper

Prepared by R. Fiedler, P. Schodl, Statistics Austria

Abstract

Because of the transition from a conservative census in 2001 to a register based census in 2010, Statistics Austria has been facing with new challenges concerning data collection, data editing and imputation. As preparation for census 2010 we are currently in the end phase to accomplish a Test Register Based Census (TRBC) and therefore able to present first experiences. One of the main tasks is to estimate incomplete data. We want to give a brief overview of problems we were facing and then, as an example, focus on the estimation of the attributes "occupation" and "education". "Occupation" is a core topic of the census recommendations, but unfortunately not included in any administrative register. We used a hot-deck estimation technique based on the labour force survey data (LFS). Whereas the goal for "occupation" is just to produce tables on aggregated level, the imputation for "education" is made on micro-level for all missing data. We use a similar idea as for the estimation of "occupation", but can set up on a much better situation since data is only missing for some thousand people.

In comparison with the traditional census of 2001, which contains occupation, the TRBC shows satisfying results. Moreover, we give an estimation of the expected deviation and give conditions for the worst-case of the expected deviation. In particular, we show that the estimation is getting better the higher the aggregation level is. Furthermore, the estimation is not getting worse if the distribution of occupation contains very small groups.

I. INTRODUCTION

1. In preparation for the first register based census in 2010, Statistics Austria is in the final phase to accomplish the Test Register Based Census 2006 (TRBC) with reference date 31.10.2006. The final report will be presented at the end of April 2008. To give a better understanding of our strategies to estimate incomplete data we want to give a brief overview of our underlying registers.

II. DATA COLLECTION

2. For the TRBC eight basis registers and additionally seven comparison registers are merged. The basis registers are used to determine the masses like the number of buildings and dwellings, the number of local units or the number of people with main residence in Austria. They also provide information

about the core topics needed for the census. The comparison registers are mainly used for cross checks and quality issues, for example to complete information not or only partly included in base registers. The basis registers function as comparison registers for other attributes, too. The “backbones” of the census are the Central Population Register (CPR) and the Central Social Security Register (CSSR). Other basis registers are the Tax Register (TR), the Unemployment Register (UR), the Register of Educational Attainment (EAR), Register of enrolled Pupils & Students (PSR), Business Register of enterprises and their local units (BR) and Housing Register of buildings and dwellings (HR). All these registers can be linked with unique keys.

3. It would go beyond the scope of this paper to describe the estimation methods for all attributes. Therefore we want to concentrate on some demographic attributes and then in special focus on estimation of “occupation” and “education”, which illustrates our main problems very well.

III. RECORD LINKAGE

4. Since we face the problem of imperfect linkage of registers for most attributes we use record linkage methods first. For many registers there exists a residual mass of persons who cannot be linked because of false or missing linkage keys. The main attributes for our linkage procedures are sex, date of birth, and address. The most important field is the address, because it is the most reliable attribute to identify a person precisely. On the other hand it is also problematical because of different notations in different registers; hence it is necessary to standardize notations. Some of the steps of standardization are:

- Standardization of the notation of “strasse”, “platz”, “gasse”
  - E.g. "straße", "str.", "Str." or "STR." become „strasse“
- Converting of all characters to lower-case characters
- Removal of spezial characters like “.” and “-“
- Removal of all blanks
- Converting of umlauts
  - E.g. “ö“ becomes “oe“
- Removal of all characters after the last blank or the last point
  - E.g. "Mozart Platz 12/2" becomes "Mozart Platz".

5. This is only an excerpt of all standardization rules. At the end the standardized street names look like:

Street name	Standardized street name
Siezenheimer Straße 64	siezenheimerstrasse
Hausnummer 157	hausnummer
Poellach 92	poellach
Dauphinestrasse 94	dauphinestrasse
Schoenbrunnerstr.85/31	schoenbrunnerstrasse
Kapellenweg 10	kapellenweg
Montecuccolistr. 24/2/4	montecuccolistrasse
Vierzehnerstr. 14	vierzehnerstrasse

6. At a first glance the removals of all characters after the last blank or the last point seem to be too crude to produce one-to-one matches, because within this step all house numbers are removed. However, we work on restricted number of people, namely those who have corrupted keys, so tests have shown that for the great masses matches are correct. The reason for implementing this step in our linkage procedures was that we got too few matches. By automatic linkage not all persons can be linked and in some cases there exist several possibilities to link them. Therefore all unclear matches are also controlled manually to minimize error rate.

IV. ESTIMATION PROCEDURES

7. After the record linkage other estimation procedures are used to impute data. The next section focuses on the estimation of "occupation". This is a special case, because it is not included in any register. A similar idea is also used to impute the attribute "education", although it is an easier case because we have only to impute missing data.

A. Occupation

8. The goal is to impute the attribute occupation for the TRBC 2006 by using information of the labour force survey (LFS) and a form of hot-deck procedure. The labour force survey is a quarterly sample survey containing occupation along with other demographic attributes and covering about 50.000 persons. This represents about 0.6% of the whole Austrian population. A classical hot-deck procedure is normally used to impute a small number of missing data. Therefore no micro-level predictions can be made, e.g. no tables on the level of municipalities will be published.

9. The imputation of the occupation was implemented as follows: In the first step, the occupation for some special groups, like soldiers, public officials etc. is derived by attributes from other registers.

10. The second step is to apply a hot-deck technique: the people in the LFS are aggregated to groups by attributes which are in the LFS as well as in the TRBC, and which are presumed to be correlated with occupation. Now the decks were formed: For people of a certain deck the frequency of the attribute occupation from the labour force survey is used to estimate the distribution of the occupation of the people in the TRBC of the same attributes. For example, if we take all people from 30 to 34 years in Tyrol who are included in the labour force survey we got a weighting scheme, e.g. 200 with occupation A, 300 with occupation B and 500 with occupation C. Now every person in the TRBC in Tyrol randomly gets an occupation according to that weighting scheme, hence for 20% the occupation *A* is imputed, for 30% the occupation *B* and for 50% the occupation *C*.

11. The selection of the attributes to form the decks is non-trivial. If we take many fine-levelled attributes the distribution of occupation would be very consistent between the labour force survey and the TRBC, but the decks are very specific and in many decks there would be no donor from the labour force survey. If we select too big levels no people without occupation would be left over, but the distribution of labour force survey and TRBC are probably not the same any more. To bypass this problem we used only those attributes that show the highest correlation to occupation and abandon others to minimize the number of non-imputed people.

Estimation of the variance

12. First we estimate the variance at finest level of our decks. Let N be the number of persons in the TRBC. To assign an occupation to a person a uniformly distributed random variable between 0 and 1 is produced. Per deck the interval [0,1] is divided into parts so that the length of the part corresponds to the probability of the assignment. According to our example: in the deck {Tyrol, between 30 and 34 years} the interval [0,0.2) is assigned to occupation *A*, the interval [0.2,0.5) is assigned to occupation *B* and the interval [0.5,1) is assigned to occupation *C*. Let X be the number of persons that a certain occupation i is assigned, N_j the number of persons of the TRBC in the deck j. For the occupation i we have the interval $[r_i, R_i)$ of the length $l_i=R_i-r_i$ and per person a value of the uniformly distributed random variable Y. The probability, that a person in the deck j gets an occupation i is $P(Y<l_i)=l_i$. Therefore the expected value is

$$E(X)=N_j l_i. \qquad (1)$$

Every person has the same probability to be assigned to a certain occupation (contributes to X), therefore X is binomial distributed and has the variance

$Var(X)=E(E(X)-X))2=N_j l_i(1-l_i)$. (2)

To maximize the expected deviation we differentiate the radix of equation (2) and equate it to zero, which gives:

$0=1/2*(1-2l_i)(l_i-l_i^2)^{-1/2}=1-2l_i$. (3)

Therefore the expected deviation of X is highest when l_i=½. In decks with many different occupations the assigned intervals should become very small and therefore the expected deviation should be quite small, too. However, the variance in the worst case shouldn't be ignored:

σ_{max}^2=VAR$(X)=N_j$½½

13. The variation coefficient σ/μ becomes quickly small for growing N_j.

N_j	10	100	1000	10000
$\sigma_{max/}\ \mu$	0.3	0.1	0.03	0.01

With a population of more than 8 Million people and a few hundred decks an average deck with 20000 people can be expected which would be equal to a variation coefficient of 0.7%.

B. Education

14. Another core topic, which requires estimation, is education. Education has two aspects: current enrolment and graduates. Here, we only discuss enrolment.

Record linkage

15. In contrary to occupation, enrolment is widely covered by the Register of enrolled Pupils & Students (PSR). But not all enrolments can be matched to persons in the TRBC. Hence we have enrolments without a person, and on the other hand we have persons without enrolment.

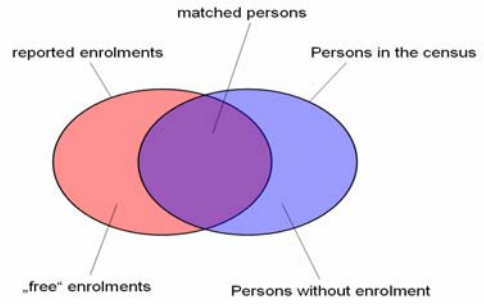


Figure 1: Situation after merging Register of enrolled Pupils & Students (PSR) and TRBC.

16. Figure 1 shows the situation after merging the PSR and the TRBC. On the left side we have about 10.500 'free enrolments' that could not be linked to any person with main residence in the TRBC. Some of these enrolments belong to people that have no residence in Austria but attend an Austrian school or university. For these people it is justified that they could not be linked because they do not belong to the Austrian population according to the census regulations. Unfortunately it is quite difficult to distinguish between these 'justified' non-matches and those that could not be linked just because of corrupted keys. For about 500 enrolments we know that they belong to foreign people because of address information. About 1500 enrolments can be linked to the TRBC with record linkage methods by using demographic attributes, and 7000 enrolments can be matched to persons by statistical matching. For the 1500 enrolments left over it just can be assumed that they belong to the mass of in-commuters.

17. On the right side we have about 23500 people in school-age with main residence in Austria but no enrolment. As described above, 8500 people can be linked to an enrolment, and for the rest we have to estimate an enrolment, taking into account, that some of them don't visit an Austrian school or are home-educated.

Imputation / Estimation

18. The idea of the imputation is the same as for the occupation with the following differences:
- We do not use the labour force survey data but the TRBC itself; hence those persons, who already are matched with some enrolment.
- As we use here a very large number of correlated variables, we first make a clustering of the data base

19. Since the goal is to bring forward the distribution of a target variable to the complete mass of the census, we want to form groups which share the same distribution of the target variable. For example, we assume that the variables age, sex and nationality, which are included in the TRBC, correlate strongly to the type of enrolment. The method is to group the persons of the TRBC in clusters which share similar enrolment, and then apply this enrolment to the persons in the same group of the census.

20. Here the question arises, how many groups should be built from the data? If too few groups are formed, one group contains very different persons, and the correlation is lost. If too many groups are formed, we might carry forward only single persons from the data base, not distributions. We do not want every person of the same demographic attributes to obtain the same education or occupation. Hence we have to pick groups large enough to contain a representative distribution of the target-variable. To gain this, we use a multiple cluster-analysis.

21. The multiple clustering works as follows: First, each variable of importance is clustered, by the measure of similar distributions of educations. Second, the possible combinations of groups are clustered (figure 2).

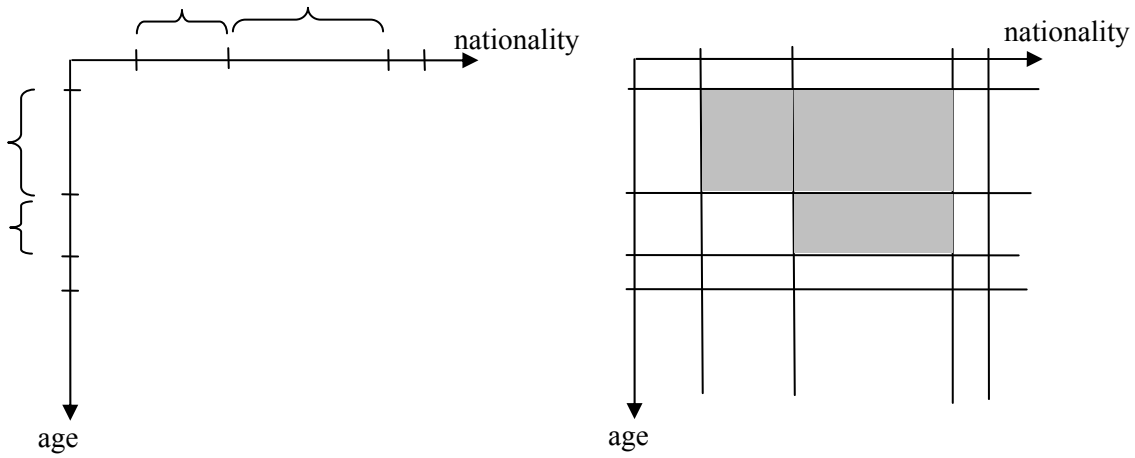


Figure 2: Multiple clustering

22. Why not only apply the second stage? Because applying the second stage without the first one would only gain groups for the observations with education, not groups for the whole TRBC-dataset, illustrated by figure 3:

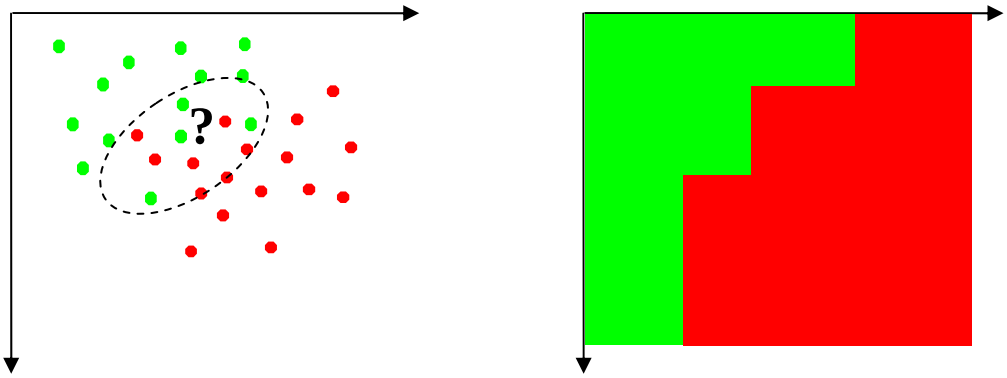


Figure 3: One-staged Two-staged clustering

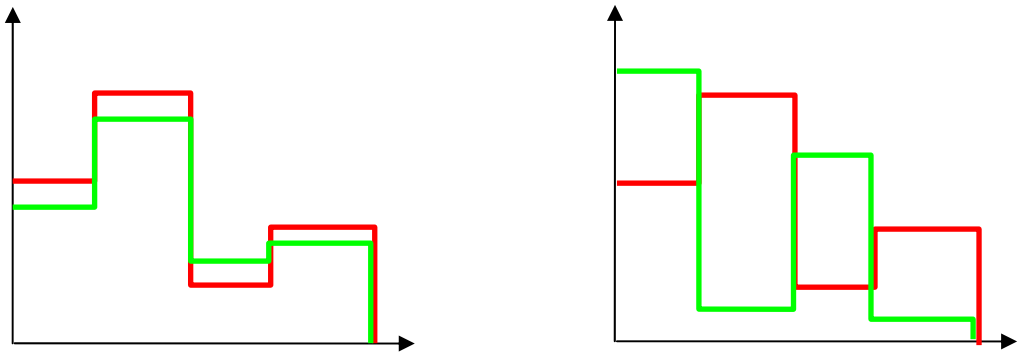
23. With the red dots belonging to one cluster, the green dots to another. Here, one cannot determine to which cluster one point belongs that is located in the region where both red and green dots appear. In other words, this method does not provide a partition of all possible persons, while the two-staged method does.

24. Moreover, the two-staged method somewhat 'smoothens' the borders of the groups, since if we apply cluster analysis to the data and then determine for every TRBC-observation a group (e.g. by the 'neighbour'-observations), the groups would get fuzzy and very hard to illustrate.

25. The measure for the 'nearness' in the cluster analysis is the distance of the distribution-mappings of one group. If two groups have a similar distribution of the target variable, their distance is small, if the distributions are different, they should have a large distance. We use the common 1-norm for function spaces, i.e.

$d(f,g) = \int |f(x)-g(x)| \, dx.$

It can be interpreted as the area in between the graphs of the two functions (figure 4).
Note that all distribution functions are scaled, i.e. $\int f = 1$, and obviously $f(x) \geq 0$ for all x.



Similar distributions – small distance

Different distributions - large distance

Figure 4: Distance Function

26. Since we do not want to start with single observation, for the first step of the cluster analysis we use a fixed number of clusters to be built. For the second stage we use Ward's minimum variance method. Outliers are eliminated the following way: Clusters which do not contain a certain number of observations are not kept. These are put together to an 'outlier-group' and get applied aggregated to the outliers of the TRBC.

Hot-deck technique:

27. As soon as we obtained suitable groups, we apply a hot-deck technique analogous to the one used for occupation (see above).

V. FURTHER WORK

28. With the experiences gained in the field of the persons in school-age, we have to perform record linkage, statistical matching and estimation for persons out of school-age. Obviously the problem here is that we do not know whether a person of age e.g. 17 years is still in school or not. We will have to be more rigid on the record linkage, and for the estimation we will have to use the activity status.
