WP.41
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 (vii): Editing and imputation of census data

# THE DATA IMPUTATION PROCESS OF THE AUSTRIAN REGISTER-BASED CENSUS

Prepared by Kausl Alexander, Statistics Austria

## I. Introduction

Due to the transition from a traditional census in 2001 to a register–based census in 2011, Statistics Austria is facing new challenges concerning data collection, data editing, imputation and quality management. Unlike in some Nordic countries the transition period from a traditional to an administrative census is very short. As a preparation for the census 2011 a Test Register Based Census (TRBC) was conducted in 2006, where new strategies and editing & imputation techniques were developed.

In this paper we want to give an overview about the general data work flow of the Austrian census with special focus on the imputation process. One of the main tasks was to develop a hierarchical estimation order of the variables, to guarantee a structured imputation procedure. This was necessary because of different data delivery times as well as the possibility to assess the quality of the imputation step.

Existing missing values of different variables in the registers will be estimated by different types of imputation methods. We are using deterministic editing as well as statistical methods (hot-deck technique, logistic regression) to impute missing values on the micro-data-level. Some examples, given in this paper, will show the practical realization.

Finally some aspects of the quality framework will be described. In particular the quality assessment of the imputation step will be discussed.

## II. Data process

In this section we want to give an overview about the general data work flow of the Austrian census, which can be classified into three levels (see Figure 1).

In a first step Statistics Austria receives raw data from different administrative sources. There exist eight basis registers and additionally seven comparison registers which are mainly used for cross checks and quality issues, for example to complete information that is not or only partly available in the base registers.

The Central Population Register (CPR), developed in 2001, forms the backbone of the census, since the units of analysis are individuals with their main residence in Austria. Other base registers are the Housing Register of Buildings and Dwellings (HR), the Register of Educational Attainment (EAR), the Register of enrolled Pupils & Students (PSR), the Unemployment Register (UR), the Central Social Security Register (CSSR), the Tax Register (TR) and the Business Register of enterprises and their local units (BR).

In a second step these different sources are combined to the Census Database (*CDB*), by using unique IDs only. The *CDB* only includes information available from the registers (raw data).

Finally we enrich the *CDB* with imputations of item non-response. These steps result in the Final Data Pool (*FDP*), which consists of both real and estimated values. In each of these three steps (Register, *CDB* and *FDP*) the dataflow is linked to the quality assessment (see section IV), so that changes can be monitored from a quality perspective.

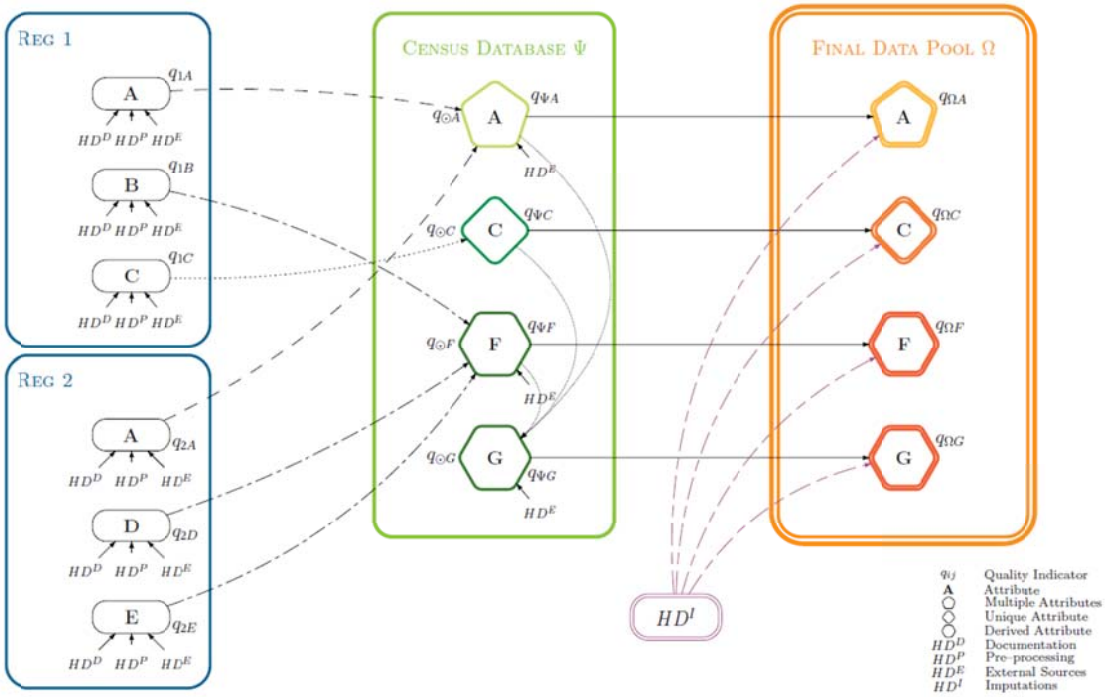


**Figure 1:** *Data process and quality assessment for register-based statistics*

# III. Imputation process

## A. Estimation order

The imputation process was developed at the TRBC in 2006. There was rather low experience at this time. Each variable of the test census was imputed separately. Some variables were imputed by correlated predictors, which had to be adapted and changed in a later process. However the imputation step didn't run again after the modification. The TRBC was a helpful learning process for handling all the arising problems.

To guarantee a structured imputation procedure, one of the main tasks was to develop a hierarchical estimation order, so that all imputation steps are connected. We had to regard several aspects:

- A variety of registers are used in the census process to ensure sufficient quality for all variables. Because of different data delivery times, it was necessary to check at which time each item can be edited.

- The choice of predictors used for imputation should mainly be based on the strength of their association with the variables to be imputed. Therefore it was important, to analyse the highest correlations between the variables to develop optimal estimation models for each imputation step. Variables which will be imputed and will be used as predictors to estimate other items received special interest.

- The structured imputation procedure also provides the possibility to assess the quality of the imputation step, as a part of the quality framework.

In Figure 2, a general description of the dependencies of all variables, which will be imputed or used as predictors, is shown. The hierarchical work flow is marked by arrows. There are correlations of variables not only within each subject (i.e. LMS ← Age Sex POP), but also between the subjects of the census (i.e. EDU ← Age Sex COC PFE).

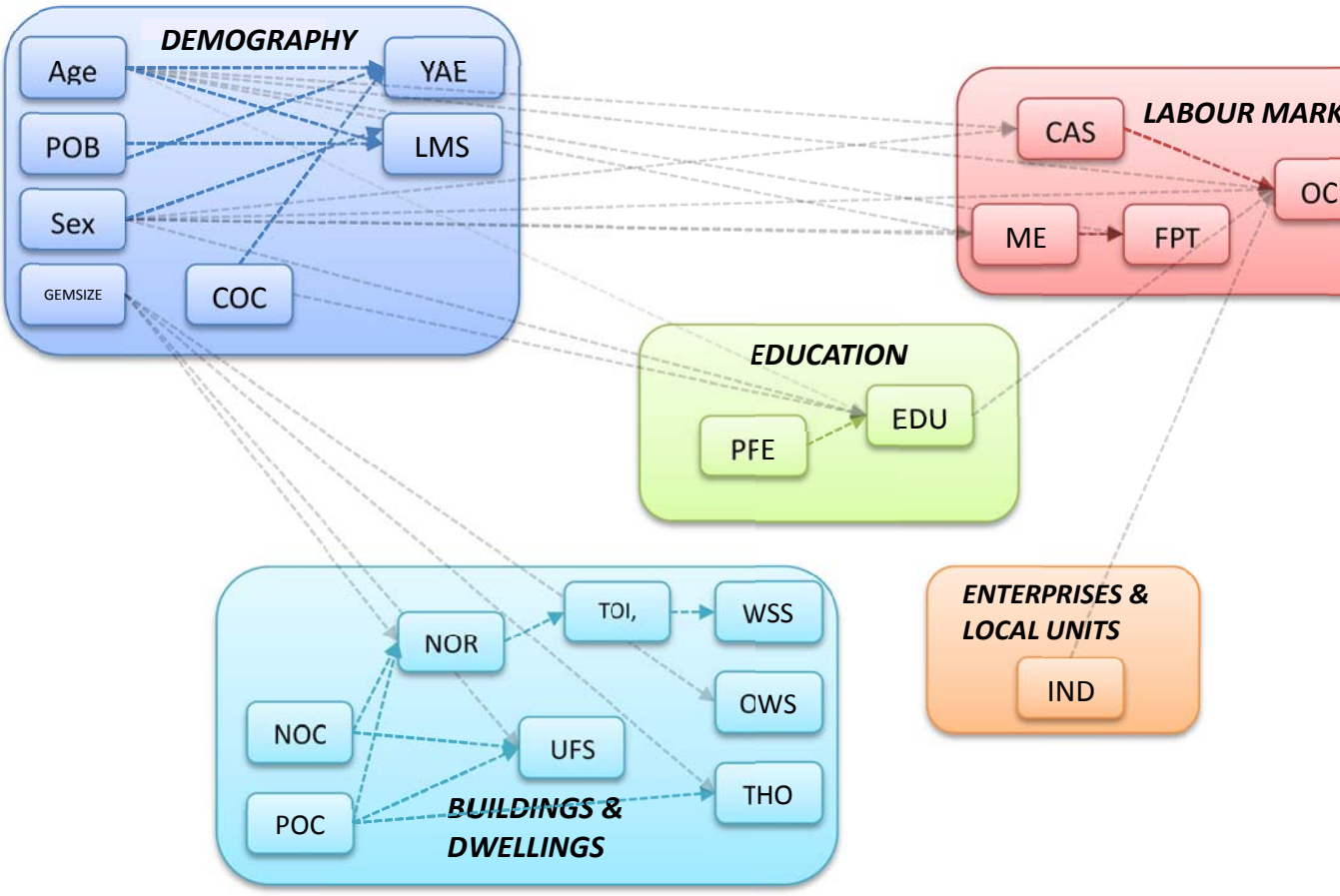


**Figure 2:** *Structured work flow of the imputation process*

*(POB … place of birth, GEMSIZE … size of locality, COC … country of citizenship, YAE …year of arrival in the country since 1980, LMS … legal marital status, CAS … current activity status, ME … marginal employment, FPT … full-part-time employment, OCC … occupation, PFE … participation in formal education, EDU … educational attainment (highest completed level), NOC … number of occupants, POC ... period of construction, NOR … number of rooms, TOI/BAT … toilet/bath facilities, WSS ... water supply system, UFS ... useful floor space, OWS ... type of ownership, THO ... type of heating, IND … industry (branch of economic activity))*

## B. Deterministic editing

Existing missing values of different variables in the registers have to be imputed to gain a complete data set. Before statistical methods are used, we primarily look for alternative deterministic rules. Because of the possibility of using a variety of register data we are often able to derive missing values from auxiliary data:

i. We have information about persons receiving a widow's/widower's pension in the Central Social Security Register. Furthermore we have data about relationships between family members. From this information missing values of the variable "Legal Marital Status" (LMS) can be deduced to *"widowed"*.

ii. People who have a lower age than 15 are easy to handle. With regard to "Educational Attainment" (EDU) they are classified under *"not applicable (persons under 15 years of age)"*. Their "Current Activity Status" (CAS) is "*persons below the age of 15"* and they get *"never married"* as marital status.

iii. The Central Population Register has information about the attribute "Place of Birth" (POB). Missing values are filled up with the value of "Country of citizenship" (COC), if the person has a foreign citizenship. It is a reasonable assumption that people with foreign citizenships were born in that country.

iv. For the variable "Water Supply System" (WSS) it's clear to deduce that there is piped water in the housing unit, if a flush toilet or a fixed bath or shower exists.

## C. Correlation and Clustering

To obtain optimal information for imputation models it is necessary to check correlations between variables to select the attributes, which have the greatest strength of association. Different types of measures have to be used for different types of variables. In most of our cases we are dealing with qualitative variables. Therefore we are using Pearson's $X^2$-test of independence to assess whether paired observations on two categorical variables, expressed in a contingency table, are independent of each other. Since the $X^2$-statistics automatically leads to high values in case of a great number of observations, the Cramer's V has to be used to assess the strength of dependency. It is a value between [-1,1] for 2x2 tables and a value between [0,1] for greater tables. A value around zero implies that there is no significant dependency between two variables.

For ordinal variables the Spearman's rank correlation coefficient can be used, for continuous variables Pearson's correlation coefficient is the most common measure of linear correlation. In some cases it is better to use Spearman's rank correlation coefficient also for continuous variables, because it is more robust against outliers.

As a preparation for imputation methods, which will be described below, it is sometimes necessary to cluster variables. This can be done by logical considerations or by statistical methods, like the Average Linkage method or Ward's minimum variance method. In Average Linkage the distance between two clusters is defined as the average distance between pairs of observations, one in each cluster. Ward's minimum variance criterion minimizes the total within-cluster variance. At each step one has to find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. Ward's method tends to produce clusters with roughly the same number of observations, which can be a benefit for imputation methods. On the other hand, Average Linkage is more robust against outliers.

In the next table you can find correlation measures for the variable "Educational Attainment". The predictors "Sex", "Age (clustered with Ward's method)", "Participation in formal education" and "LOC (Size of locality clustered with Ward's method)" are selected to build up an imputation model. For all variables the p-value illustrates a significant dependency. Cramer's V shows the strength of association.

| | Sex | Age clustered | PFE | GEMSIZE clustered |
|---|---|---|---|---|
| **Chi-square** | 12.151 | 610.300 | 831.925 | 182.488 |
| **P-value** | < .0001 | < .0001 | < .0001 | < .0001 |
| **Cramer's V** | 0.1467 | 0.46493 | 0.35034 | 0.15923 |
| **Conclusion** | weak association | medium association | medium association | weak association |

**Table 1:** *Correlation measures for education attainment*

## D. Hot-Deck technique

Based on the variable "Legal Marital Status (LMS)" the general functionality of the Hot-Deck technique will be described (see also [1]). People are aggregated to groups by attributes, which are strongly correlated to LMS. These groups are also called "decks" and the method is called "Hot-Deck" because it is using information from observations in the same data set. For people of a certain deck the frequency of LMS is used to estimate the distribution of this variable and to impute missing values for people, which are characterised through this deck.

For example, if we take the female population from 30 to 35 years in Tyrol we get a weighting scheme, e.g. 200 with "LMS = never married", 700 with "LMS = married" and 100 with "LMS = divorced". Now every woman from 30 to 35 years in Tyrol with a missing LMS randomly gets an LMS according to that weighting scheme, hence for 20% the "LMS = single" is imputed, for 70% the "LMS = married" and for 10% the "LMS = divorced".

The selection of the predictors to form the decks is non-trivial. If we take to many fine-levelled attributes the distribution of LMS can be preserved on a very deep level, but the decks are very specific and in many decks there would be no donor at all. If we select too high levels it's possible to estimate the LMS for every none-response, but the distribution of the complete data can be distorted in deeper levels. To bypass this problem we use only those predictors that show the highest correlation to LMS and cluster fine-levelled variables in an optimal way.

***Estimation of the variance***

Let $N$ be the number of persons in the register. To assign a LMS 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, female, between 30 and 35 years} the interval [0,0.2) is assigned to "LMS =never married", the interval [0.2,0.9) is assigned to "LMS = married" and the interval [0.9,1) is assigned "LMS = divorced".

Let $X$ be the number of persons that a certain LMS $i$ is assigned, and $N_j$ the number of persons in the deck $j$. For LMS $i$ 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 LMS $i$ is $P(Y<l_i)=l_i$. Therefore the expected value is

$$E(X)=N_j l_i.$$

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

$$Var(X)= N_j l_i(1-l_i)$$

You can see that the variance of $X$ is highest when $l_i=½$ and that the variation coefficient $VarK(X) = \sqrt{Var(X)}/E(X)$ becomes quickly small for growing $N_j$.

| $N_j$ | 10 | 100 | 1000 | 10000 |
|---|---|---|---|---|
| $VarK(X)$ | 0.3 | 0.1 | 0.03 | 0.01 |

## E. Logistic Regression

Logistic regression is a type of regression analysis used for predicting the outcome of a categorical variable based on one or more predictor variables. The probabilities describing the possible outcome of a single trial are modelled, as a function of explanatory variables, using a logistic function. Unlike ordinary linear regression, logistic regression is used for predicting binary outcomes rather than continuous outcomes. Given this difference, it is necessary that logistic regression take the natural logarithm of the odds (referred to as the logit or log-odds) to create a continuous criterion. The logit of success is then fit to the predictors using regression analysis. For binary response variables the logistic model has the form

$$logit(p) := ln\left(\frac{p(y_k=1)}{1-p(y_k=1)}\right) = \beta X.$$

Receiving probabilities from the logistic regression, a uniformly distributed random variable between 0 and 1 is produced, to assign a value for the imputed variable.

Logistic regression can be used for binomial, ordinal or multinomial response variables. A set of different measures can be analysed to assess the goodness of fit of the logistic method. Deviance or Pseudo $R^2$ is used to assess the global goodness of fit. To assess the contribution of individual predictors, one may examine the significance of the Wald statistic. You can compute the classification rate to get further details of the accuracy of your imputation model.
In the binomial case it's possible to do a residual diagnostic to identify leverage points and outliers, which should be excluded from the estimation model. You can analyse the ROC-curve and the Rank Correlation of Observed Responses and Predicted Probabilities to get further impressions of your model.

We are using the logistic regression method for imputing the variable "Educational Attainment" and for some variables of the Buildings and Dwellings statistics.

The use of specific imputations methods depends on the characteristics of the data. For our hot-deck technique approach, every deck should contain a relative high amount of observations to ensure that the deviation becomes small (see section D). In the worst case, if there exist only one observation in a deck, all non-response values of this deck, will be deterministically imputed by the same response value of the one observation.

Generally logistic regression can handle small subpopulations. On the other hand some quality measures are not valid if the amount of covariance patterns is too high, relatively to the amount of observations. Also sparseness in the data refers to having a large proportion of empty cells (cells with zero counts). Zero cell counts are particularly problematic with categorical predictors. As a consequence the logistic model doesn't converge and so the model is invalid. A next problem of the logistic regression model is that it produces probability values for response variables, which are not in the donor set. So one has to check the output and if it's necessary one has to modify the data. This problem cannot occur if you use the hot-deck model.

## F. Further Work

For the majority of our variables the amount of missing values is rather low, because most of them are covered in multiple registers. "Age" or "Sex" contains only a couple of hundred

missing values. For other variables the amount of missing values is often lower than 4%, only for a few numbers of attributes in the Buildings and Dwellings statistics the amount is about 10%. In all these cases it is actually easy to handle the existing lacks.

For the variable "Year of arrival in the country since 1980" (YAE) only a minor part of the data is covered by the registers. Because of the fact that our Population Register was set up in the year 2001, we don't have information about earlier years.

For the variable "Occupation" (OCC) the quality of the register data has considerably improved since the last years, especially since the TRBC. We are receiving data from the Unemployment Register, the Register of the Austrian Federal Economic Chamber, the Tax Register and several Social Security Registers. However the amount of missing values is expected to be more than 10%.

For both variables (YAE, OCC) we are currently focused to handle the existing problems. We are investigating the possibility to estimate the variable OCC from the existing register information. The results will be compared with the Labour Force Survey to check if the quality of the variable is efficient enough. For the variable YAE it has not yet been decided how this variable will be derived.

The Register of Buildings and Dwellings (HR) has to be checked for incorrect data. Furthermore it contains objects with default-values, which are not valid. These incorrect and invalid data will be set to missing values and therefore they have to be imputed as well.

## IV. Assessment of quality

Statistics Austria has developed a quality framework for statistics based on administrative sources. It provides quality measures, which combine information from three so-called hyperdimensions (*Documentation $HD^D$, Pre-processing $HD^P$ and External Source $HD^E$).* An additional hyperdimension $HD^I$ will assess the quality of the imputation step. The framework results in exactly one quality indicator for accuracy for each attribute in each register or data pool (see Figure 1).

By using the framework, an external view on the data processing is possible and an increase or decrease in quality during the data generation process can be monitored. In a three-stage process (raw data, combined data, imputed data) we derive quality indicators that aim to cover all available quality information. The quality assessment is closely tied to the data flow, but independent from data processing. Because of this separation, the quality assessment can evaluate the processing procedure without influencing it.

A detailed description of the realization of $HD^D$, $HD^P$ *and* $HD^E$ you can find in [2], [3] and [4].

Current research is focused on the calculation of quality for the hyperdimension $HD^I$. The imputation process itself has to be monitored to get a quality assessment in the Final Data Pool. The classification rate probably will be used to get a quality measure for the imputation model. The benefit would be, that the classification rate is a general measure for the goodness of fit and it can be determined for the Hot-Deck and logistic regression method.

The basic idea, to get a quality indicator for $HD^I$, is to set the quality for item non-response to 0 on the *CDB*-level. Finally (on *FDP*-level) the quality measure will be updated with the quality of the imputed value.

## V. Conclusion

This paper presents the general work flow of the data imputation process for the Austrian register-based census. The structured hierarchical procedure is described, as well as the imputation methods are presented. Some current challenges are shown and finally a short outlook about the assessment of the quality is given.

## VI. References

[1] Fiedler R., Schodl P. (2008), *Data Imputation and estimation for the austrian register-based census*. Unece Conference of European Statisticians, Vienna.

[2] Ćetković P., Humer S., Lenk M., Moser M., Schnetzer M., Schwerer E. (2012), *A quality monitoring system for statistics based on administrative data*. Conference contribution on Quality in Official Statistics, Athen.

[3] Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H. & Schwerer, E. (2012), Combination of evidence from multiple administrative data sources: quality assessment of the austrian register-based census 2011. *Statistica Neerlandica*, Vol. 66, No. 1, 18--33.

[4] Ćetković P., Humer S., Lenk M., Moser M., Rechta H., Schnetzer M. & Schwerer E. (2011), *Quality Assessment of Register-Based Statistics - Preliminary Results for the Austrian Census 2011*. Conference contribution at ESSnet on Data Integration, Madrid.