Distr.: General
16 May 2012

Original: English

Economic Commission for Europe

Conference of European Statisticians

UNECE-Eurostat Expert Group Meeting on Censuses Using Registers

Geneva, 22-23 May 2012
session 3: Availability, completeness and quality of data from registers and other sources

Quality assessment of register-based census data in Austria

Note by Statistics Austria [1]

Summary

This article provides a brief history of the introduction of a register-based census in Austria. Furthermore, a structural approach to quality assessment and quality-related aspects of register-based statistics are presented. In a three-stage process (raw data, combined data, imputed data) we derive quality indicators that aim to cover all available quality information. The advantage of a single quality measure for each attribute in each register is its simplicity which offers the possibility for uncomplicated comparison of attributes. Finally, the derivation of quality indicators for three types of attributes (unique, multiple and derived) are discussed in detail. The quality process itself remains independent from data processing which guarantees its applicability for other register-based statistics. The experience gained with the new census type and the quality assessment methods can be of use for other population and household surveys.

[1] Prepared by Manuela Lenk.

I. The evolution from traditional to register–based census in Austria

1. In May 2001, the last traditional population census was carried out in Austria, accompanied by a building and housing census as well as a census of local units of employment. This combined census covered approximately 2.0 million buildings, 3.8 million dwellings and 3.3 million households containing 8.1 million persons by using paper questionnaires. Thus traditional census was a sophisticated and costly effort.

2. Given the challenging requirements for traditional censuses, the importance of administrative data sources for statistical purposes has recently been rising. The processing of data which has already been recorded by administrative authorities offers numerous advantages compared to survey data, such as diminishing costs, removed burden for respondents and the prompt availability of the data. Using data from existing registers should ensure an optimal reflection of reality at reasonable expenditures by combining registers via unique linking variables, enhancing data quality and harmonizing definitions. An obvious advantage is exhibited by the regular updates of register information in order to keep track of any changes in the data describing the units and their attributes (Statistics Finland, 2004, p. 10).

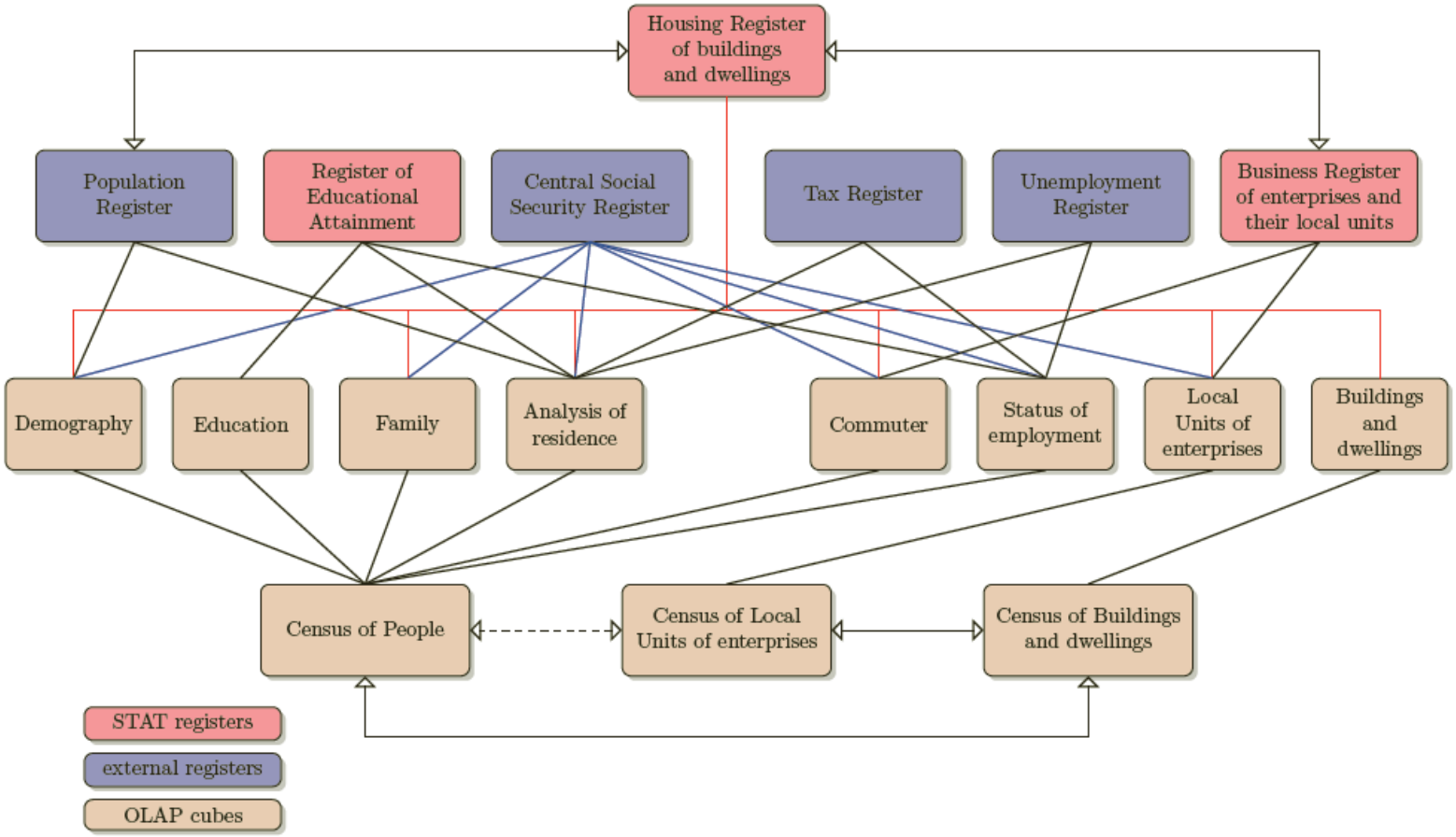
3. Thus, an increasing number of National Statistics Offices (NSOs) promote register-based censuses as a replacement for costly conventional censuses. In 2000, the Austrian council of ministers decided to establish the new method for the census 2011 and in 2006, the regulatory framework came in force by the juridical authorities. Austria is among six European countries (besides Sweden, Finland, Norway, Denmark and Slovenia) that carried out a register-based census in 2011.

4. Based on the number of population from this census, the monetary amounts of fiscal equalization between municipalities and the federal financial authorities as well as the number of eligible voters in the prospective elections are determined. Furthermore, information on commuters, education and employment offer important insights for economic and social policies. However, administrative data may apply definitions that differ from the needs of the NSO even though the data are of good quality (see United Nations, 2007, p. 3). Therefore, the NSO has to decide whether the data is adequate for the issue of interest.

A. The principle of redundancy

5. A key problem arises with the selection of appropriate data sources for supplying the required information. The register-based census aims at covering all relevant variables that were formerly provided by a traditional census. In this respect, the census in 2001 was the initial spark for the creation of some data sources, e.g. the Central Population Register (CPR), the Housing Register of Buildings and Dwellings (HR) or the Register of Educational Attainment (EAR). Prior to 2001, an interconnected network of population records did not exist in Austria. Each municipality administrated its own records and usually the data was not even entered into electronic systems. Introducing the housing register in 2001, a centralized population register evolved and the municipalities had to provide their records. The last traditional Austrian census was already based on information from this newly created population register. In 2004, the Housing Register of Buildings and Dwellings was synchronized with the population register for the first time. Moreover, the Register of Educational Attainment was founded during the census process of 2001.

Figure 1. Registers and Topics in the Austrian register-based census



6. Figure 1 shows all base registers of the census and their links to the respective topics. The red-shaded data sources are maintained by Statistics Austria, the remaining information is provided by external data holders, like the Unemployment Register (UR) or the Central Social Security Register (CSSR). The Central Population Register (CPR) forms the backbone of the census, since the units of analysis are individuals with their main residence in Austria. To assure the quality of the census results, the base registers are backed up by seven comparison registers. These seven fields of administrative units are provided by 35 data holders and are mainly used for cross checks as well as to supply information that is not or only partly available in the base registers (Berka et al., 2010, p. 300).

7. Given the independence of the various registers as well as the autonomous process of data collection, the sources sometimes contain contradictory values for the same attribute. Therefore, the principle of redundancy is used to ensure sufficient quality by acquiring information on sex, nationality or date of birth from as many registers as possible. A particular method developed by Statistics Austria aims at identifying one particular base register to provide the information for a certain variable, whereas the comparison registers are used to confirm the values in the base registers (Lenk, 2008, p. 3). While the registers offer sufficient information for most of the characteristics in the conventional census, some variables could not be included in the register-based census. For instance, the duration of the daily commute, colloquial language or religion are not captured by any of the registers at hand.

B. Residence analysis

8. Regarding the quality of the register-based census, it is important to detect inactive records in the Central Population Register (CPR) and eliminate them for counting purposes to avoid overcoverage. This procedure is known as residence analysis and assures that only individuals with a pre-defined number of signs of life are counted in the census.

Table 1. Signs of life in residence analysis

bPIN OS	CPR	CSSR	TR	UR	SWR	CAR	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
ID3457	✓	✓	✓	✗	✗	✗	
ID3458	✓	✓	✗	✓	✗	✗	
ID3459	✓	✗	✗	✗	✗	✗	
ID3460	✓	✗	✗	✗	✓	✗	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

CPR=Central Population Register, CSSR=Central Social Security Register, TR=Tax Register, UR=Unemployment Register, SWR=Register of Social Welfare Recipients, CAR=Child Allowance Register

9. Individuals who are only covered by the CPR but in no other administrative source implicitly need clarification and are questioned in a written form, which would be the case for individual ID3459 in **Table 1**. In a test census procedure in 2006, some 45,000 letters were sent to cases of clarification, whereof 9,000 individuals affirmed their main residence in Austria. Finally, approximately 0.5 percent of the initial population was not counted. Austrian municipalities have to be informed about non-counted individuals whereby

registration authorities have the chance to prove the residence of these cases and possibly remove the individuals from the residence register. Due to the results of the test census in 2006, about 80 percent of the non-counted cases were removed from the residence registers by local municipalities.

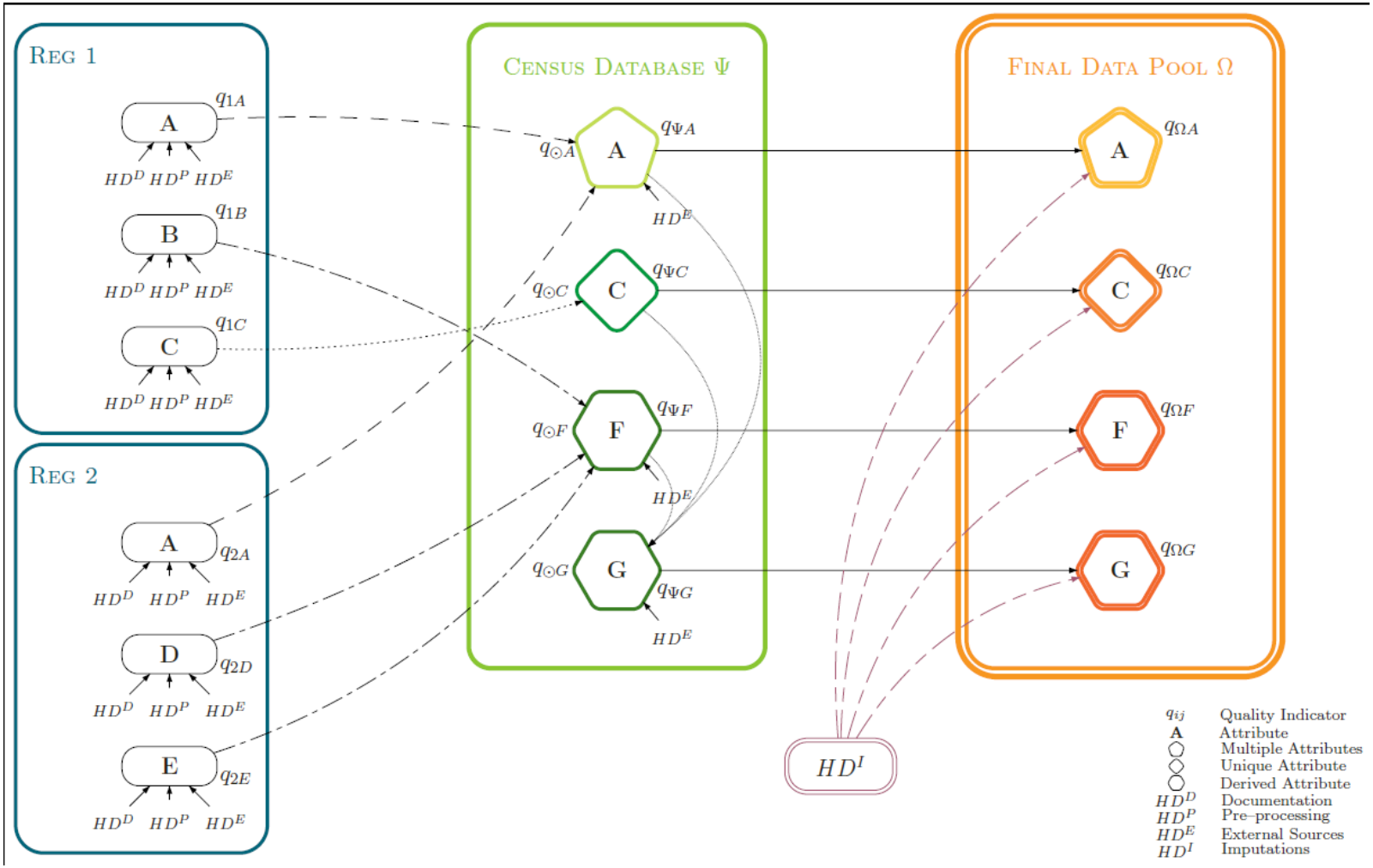
10. For the register–based census of 2011, Statistics Austria applies the same procedure like in the test census. The residence analysis has been started in January 2012. In a first round, about 54,000 letters to clarify main residence were conveyed. The second round is planned in September 2012, because the data transmission from the external data holders will be completed by that date.

II. Considerations of a framework for quality assessment

11. The short transition time from a traditional to the register-based census has been a challenging task in Austria. The interim period for gradually substituting survey data with administrative data lasted about 20 years in some European countries (see Ruotsalainen 2008, for the example of Finland). This allowed an intensive discussion on quality assessment by the NSOs and the data owners in some countries, whereas the transition schedule was very ambitious in Austria. Since the NSO is not responsible for the maintenance of the external data, the necessity of quality assessment in the process of register-based censuses has to be singled out. The quality analysis of register data has to satisfy several requirements such as transparency, accuracy and feasibility. Our approach contributes a quality framework for the analysis of administrative data using three different hyperdimensions for the derivation of quality indices. The framework is closely tied to the data flow yet independent from data processing, which ensures that the processing is not influenced but evaluated.

12. The data flow of the register-based census in Austria consists of three levels: raw data (i.e. the registers), combined dataset (Census Database, henceforth CDB) and imputed dataset (Final Data Pool, FDP). Figure 2 illustrates the data processing, beginning with the receipt of raw data from the various administrative data holders. The information is connected via unique keys (branch-specific personal identification numbers for official statistics, bPIN OS) and merged to the CDB. Further, the CDB data is enriched with imputations of item non-response which complete the FDP. The FDP thus contains real and estimated information.

Figure 2. Quality framework for register-based census



A. Quality assessment on the register level

13. Information on quality at the raw level is obtained by three hyperdimensions: Documentation (HD^D), Pre-processing (HD^P) and External Source (HD^E). Prior to seeing the data, HD^D describes quality-related processes in the register authority as well as the documentation of the data (i.e. metadata). The degrees of confidence and reliability of the data holders are monitored by the use of a questionnaire containing 16 open-ended and nine scored questions (see

Table **2**). The NSO is therefore able to check for data collection methods or legal enforcements of data recording which may significantly influence the quality of the data. The questionnaires are answered by experts from the respective data holders and should thus deliver convincing results. The quality indicator hd_{ij}^{D} is a simple ratio between the *obtained score* and the *maximum score* of the questionnaire for each attribute *j* in register *i*.

Table 2. Scored Questions – HD Documentation

DATA HISTORIOGRAPHY
Can we detect data changes over time?
Is the information available for the reference date?
DEFINITIONS
Are the data definitions for the attribute compatible to those of Statistics Austria?
ADMINISTRATIVE PURPOSE
Is the attribute relevant for the data source keeper?
Does a legal basis for the attribute exist?
DATA TREATMENT
How fast are changes edited in the register?
Are the data verified on entry?
Are technical input checks applied?
How good is the data management, i.e. ex post consistency checks?

14. The second hyperdimension HD^P is concerned with formal errors in the raw data. Range errors, item non-response and missing primary keys are detected in this step of the quality framework. Subtracting all these erroneous records from the total number of entries leads to the number of usable records. The final indicator hd_{ij}^{P} of this hyperdimension is given by the ratio of *usable records* to the *total number of records*. Again, this procedure is carried out for each attribute in each register. If the proportion of usable records for an attribute in a certain register is smaller than that of the same attribute within another register, the quality measure will accordingly be lower.

15. Finally, the third hyperdimension HD^E provides a comparison between the register–based data and an external source. In Austria, the microcensus is a common benchmark for representative surveys and is assumed to be the best comparative dataset available.

Checking for consistency with the external source gives the third quality measure hd_{ij}^{E} which is the ratio between the number of *congruent values* and the *total number of linked records*. If the attribute of interest is not covered by the benchmark, we rely on local expert opinions.

16. Given these three quality measures, an overall quality indicator for each attribute and register can be derived as a weighted average,

$$q_{ij} = v^{D} \cdot hd_{ij}^{D} + v^{P} \cdot hd_{ij}^{P} + v^{E} \cdot hd_{ij}^{E} = \sum_{k \in D,P,E} v^{k} \cdot hd_{ij}^{k}$$

where hd_{ij}^{k} are the hyperdimensions scores and v^{k} are the weights. The advantage of this measurement is its simplicity which offers the possibility for an uncomplicated comparison of attributes. This indicator is able to capture quality-related effects ranging from the data generation to the raw data in the registers. **Table 3** shows some results for the attributes sex, full- or part-time employment and the highest level of education in five registers for data of 2008. Here, we suggest an equal weighting of the three hyperdimensions ($v^{k} = 1/3$). It can be seen that the quality of the attribute sex is very good in all of these registers. However, there are some notable differences between the hyperdimensions. For example, Register 3 has a very low value for the hyperdimension Documentation (HD^{D}), due to the fact that the attribute sex is not relevant for this specific register authority. The indicators for the hyperdimension Pre-processing (HD^{P}) are mostly influenced by missing primary keys, while in this case range and definition errors only play a minor role. For the last hyperdimension External Source (HD^{E}) the framework returns very high quality measures, which means that there is a high degree of agreement between the register data and the microcensus.

Table 3. Quality measures for data of 2008

Register	Attribute	HD^{D}	HD^{P}	HD^{E}	$q^{(33,33,33)}$
REG 1	SEX	1.000	1.000	0.998	0.999
REG 2	SEX	0.792	0.942	0.999	0.911
REG 3	SEX	0.444	0.746	0.997	0.729
REG 4	SEX	0.792	0.993	1.000	0.928
REG 3	FT/PT	0.381	0.698	0.847	0.642
REG 5	EDU	0.928	0.950	0.800	0.891

B. Quality assessment in the census database

17. The entire information from the registers is combined in the CDB which covers all attributes of interest for the census. Since there may be more than one data source providing a certain attribute, a ruleset predefined by the NSO picks the most appropriate information from the underlying registers. Concerning the evaluation of quality for the CDB we distinguish three types of attributes in this linking process:

(a) Unique attributes exist in exactly one register, e.g. highest level of education (cf. attribute C in **Figure 2**);

(b) Multiple attributes show up in several registers, e.g. sex (cf. attribute A in **Figure 2**). The information from multiple sources is combined by a ruleset to derive the most appropriate value in the CDB attribute;

(c) Derived attributes are created based on different attributes, e.g. family and household status (cf. attributes F and G in **Figure 2**). The registers do not contain any information for these attributes in the required specification.

18. It is trivial to assess a quality measure for *unique* attributes, since it is equal to the quality indicator q_{ij} from the raw data. Hence quality indicators for unique attributes are directly transferred to the CDB, e.g. $q_{1C} = q_{\odot C} = q_{\Psi C}$ in Figure 2. For the case of *multiple* attributes, conflicts among registers with reference to a particular value are associated with epistemic uncertainty. If registers provide contradictory information, it does not necessarily show which register is wrong. However, it may express a degree of uncertainty of the value in the Census Database. Applying simple weighted averages and neglecting uncertainty associated with coinciding and opposing evidence could lead to delusive conclusions, whereas more sophisticated methods like the Dempster-Shafer theory are able to deal with this special type of uncertainty (i.e. probability, belief and plausibility).

19. Table 4 shows an example with the multiple attribute sex to illustrate the different outcomes of the methods. Suppose that the quality indicator for the attribute sex in register one is 0.9, the quality indicator for the same attribute in register two is 0.7. To calculate the average of the quality indicators would not make any sense, since consistent and conflicting values would be treated equally. The mean is 0.8, regardless of consistency or conflict between the registers. Dempster-Shafer theory, in contrast, allows combining information from different registers while the degree of belief in the data source is taken into account. Thus, quality indicators increase if there is consistency between the sources, however the indicator decreases when conflicts occur.

Table 4: Quality indicators for the multiple attribute sex

PIN	REG1	REG2	CDB	Average q_{CDB}	Dempster Shafer q_{CDB}
9845	Male	Male	Male	0.80	0.99
4866	Male	Female	Male	0.80	0.77
2047	Female	Male	Female	0.80	0.77

20. A detailed application of Dempster-Shafer theory on register-based census data is given by Berka et al. (2012).

21. The quality assessment for *derived* attributes needs special attention. If the attribute is derived from more than one attribute, the quality of each attribute used in the process has to be assessed. Apart from the extended number of quality indicators, no further problems are assumed to arise for the hyperdimensions HD^D and HD^P. In contrast, complications could arise with the hyperdimension HD^E on register level, since the derived attribute is not included within the administrative data. Therefore, the standard framework cannot be applied because there are no possibilities to calculate hd_{ij}^E on the register level. Alternatively, we suggest deriving hd_{ij}^E on a CDB level and hd_{ij}^D as well as hd_{ij}^P on a raw data level (see Ćetković et al. 2011).

22. A detailed description of the quality assessment for the three types of attributes in the CDB is given by Berka et al. (2010) and Berka et al. (2012).

23. Current research is focused on the calculation of quality indicators for derived attributes and on the Final Data Pool (FDP), which corresponds to the Census Database after the imputations are applied. The amount of item non–response is effectively reduced by imputations; however, the imputation process itself has to be monitored. This is done by using information from the hyperdimension Imputation (HD^I), which is an ongoing task.

III. Conclusion

24. This paper presents a structural approach for the quality assessment of administrative data. A process based on three stages (raw data, combined data and imputed data) derives quality indicators for three hyperdimensions. These measures aim to cover all available quality information for each attribute. To guarantee the applicability of the quality framework for the register–based census 2011, the procedure was tested with register-based labor market data from 2008.

25. A decisive advantage of the quality framework at hand is the independence of quality assessment and data processing. The separation from the processing procedure is required to evaluate the process without exerting influence on it. This offers the possibility to apply the methods on other register–based data sets. Moreover, the cooperation between the NSO and the register authorities is intensified because the data holder is integrated in the quality assessment process.

IV. References

Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H., & Schwerer, E. (2010). A quality framework for statistics based on administrative data sources using the example of the Austrian census 2011. Austrian Journal of Statistics, Volume 39, Number 4, 299-308.

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, Volume 66, Issue 1, 18-33.

Ćetković, P., Humer, S., Lenk, M., Moser, M., 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

Lenk, M. (2008). Methods of Register-based Census in Austria (Tech. Rep.). Statistik Austria, Wien.

Ruotsalainen, K. (2008). Finnish Register-based Census System (Tech. Rep.). Statistics Finland.

Statistics Finland. (2004). Use of registers and administrative data sources for statistical purposes (Tech. Rep.). Statistics Finland.

United Nations. (2007). Register-based statistics in the Nordic countries. United Nations Economic Commission for Europe, New York and Geneva.