Quality assessment for register-based statistics - Results for the Austrian census 2011

Q2014 conference Vienna

Session: II.2. - New quality measurement methods; quality assessment and quality indicators; methods for assessing quality of administrative data

Eva-Maria Asamer *, Franz Astleithner **, Predrag Ćetković**, Stefan Humer**, Manuela Lenk*, Mathias Moser**, Henrik Rechta *

In 2011, Statistics Austria carried out the first register-based census. The use of administrative data for statistical purposes is accompanied by various advantages like a reduced burden for the respondents and less costs for the NSI. However, new challenges, like the quality assessment of this kind of data, arise. Therefore, Statistics Austria developed a comprehensive standardized framework for the evaluation of the data quality for register-based statistics.

In this paper, we present the principle of the quality framework and detailed results from the quality evaluation of the 2011 Austrian census. For each attribute in the census a quality measure is derived from four hyperdimensions. The first three hyperdimensions focus on the documentation of data, the usability of the records and the comparison of data to an external source. The fourth hyperdimension assesses the quality of the imputations. In the framework all the available information on each attribute can be combined to form one final quality indicator. This procedure allows to track changes in quality during data processing and to compare the quality of different census generations.

Keywords: Administrative data, Register–based census, Quality assessment, Austrian census 2011.

*Statistics Austria, Unit Register–based Census, Guglgasse 13, 1110 Vienna

**Vienna University of Economics and Business, Department of Economics, Welthandelsplatz 1, 1020 Vienna

1 Introduction

The importance of administrative data as input for statistical purposes has increased steadily in the last decades. Following the Scandinavian countries, about one third of the United Nations Economic Commission for Europe (UNECE) members now base their census at least partially on administrative data (UNECE, 2014). In Austria, the last survey-based census in 2001 was replaced by the first register-based census in 2011. The advantages of the new approach comprise inter alia reduced burden for the respondents and lower costs. However, new challenges like the assessment of the data quality arise. For this reason, Statistics Austria developed a standardized quality framework for the assessment of administrative data. In section 2, we will introduce the sources of the register based census. In section 3, the quality framework is explained using the example from the quality assessment for the *Legal Marital Status LMS*. Section 4 provides results for the quality assessment of the Austrian census of 2011.

2 Sources for the register-based census

A decisive quality–related topic for register–based statistics is the selection of appropriate data sources for the supply with required information.

Figure 1 illustrates the connections between the data sources and topics of the census. Statistics Austria distinguishes between 7 base registers and 8 comparison registers. The base registers contain, in principle, the attributes of interest for the register-based census. The red shaded registers form the backbones of the census. They determine the population number, the number of buildings and dwellings and the number of enterprises and their local units. To improve the quality of the results, the base registers are backed up by eight comparison registers which gather information from more than 50 data holders. They are mainly used for cross-checks and validation.[1] If there is more than one source for an attribute, the registers serve as instruments for cross-checks and validation because of the autonomous data delivery. This *principle of redundancy* helps to improve quality of data (Lenk, 2008, p. 3).

[1]If data is not or only partly available in the base registers, information is derived from the comparison registers as well (Berka et al., 2010, p. 300).

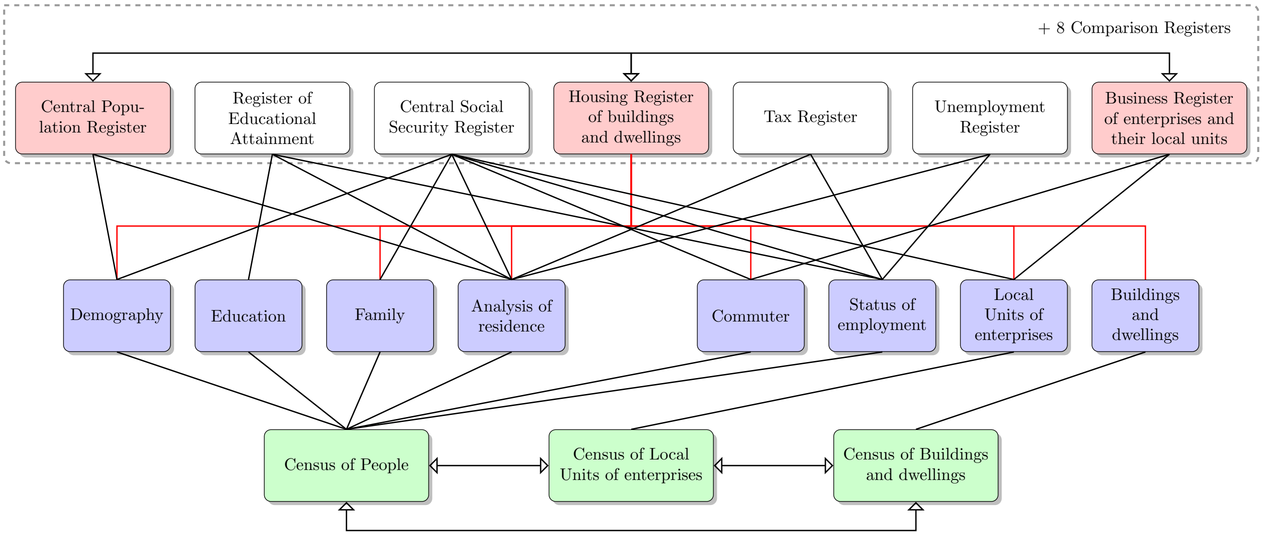

Figure 1: Data sources for the register–based census

3 The quality assessment of administrative data

Statistik Austria is not responsible for the data maintenance of the external data sources which contribute the majority of the required information. Hence, the relevance of quality assessment in the process of register–based statistics has to be emphasized. Our approach for the assessment of administrative data was inspired by work from other National Statistical Institutes NSI (Daas, Ossen, Vis-Visschers, and Arends-Tóth (2009); Daas and Fonville (2007)) and relies on four quality-related hyperdimensions (Berka et al., 2010, 2012).

The data processing for the Austrian census is divided in three levels that have to be considered in the quality assessment: the raw data (i.e. *the registers* i), the combined dataset (*Central Database CDB*) and the imputed dataset (*Final Data Pool FDP*). Four hyperdimensions (HD^D, HD^P, HD^E, HD^I) aim to assess the quality for different types of attributes at all stages of the data processing. Figure 2 illustrates the data processing, beginning with the delivery of raw data from the various administrative data holders. The data is connected via a unique personal key (*branch-specific personal identification number bPIN*) and merged to data cubes in the CDB. Finally, missing values in the CDB are imputed in the FDP where every attribute j for every statistical unit n in the statistics

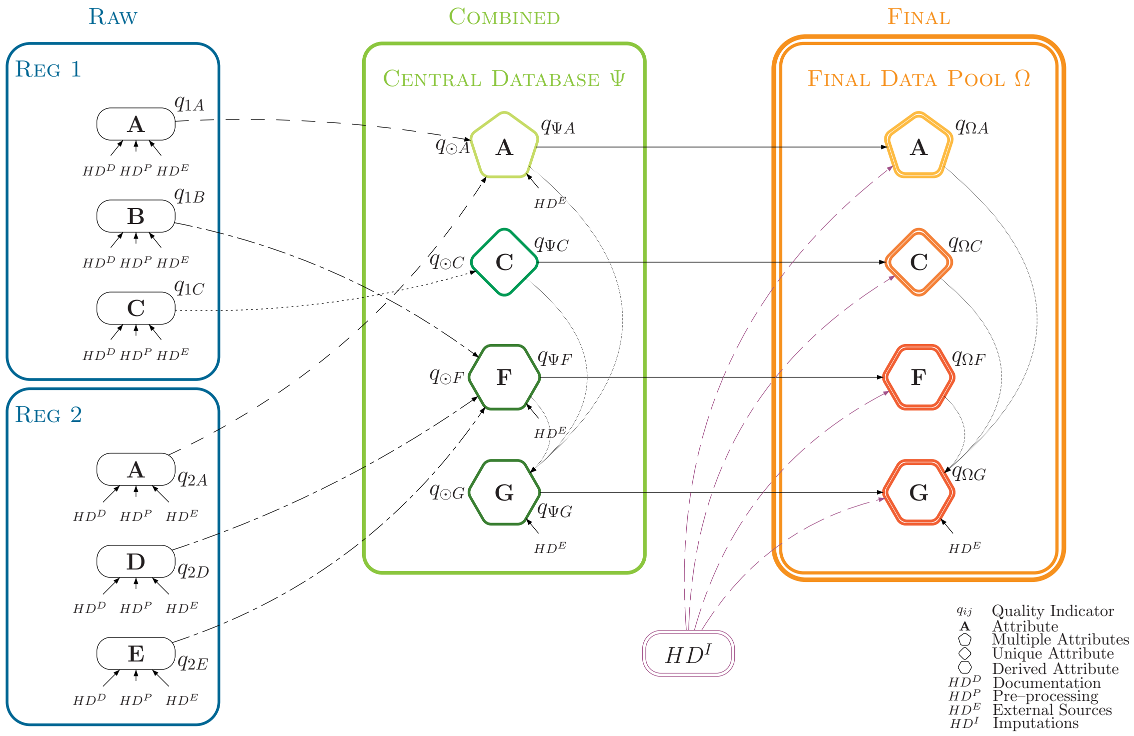

Figure 2: Quality framework for register–based censuses

of administrative data obtains a quality indicator $q^n_{\Omega j}$. In the following, we will explain the quality framework using the example of the calculation of the quality measure for the *Legal marital Status LMS*.

3.1 The Raw Data Level

We start our considerations on the quality assessment at the first level of the framework. Information on quality at the raw data level (see blue boxes in Figure 2) is obtained via three hyperdimensions: Documentation (HD^D), Pre-processing (HD^P) and External Source (HD^E). The derivation of the quality measures for the LMS on the raw-data level can be retraced in the following tables.

Hyperdimension HD^D

HD^D describes quality-related processes as well as the documentation of the data (metadata) for the administrative authorities. The degrees of confidence and reliability of the data holders are monitored by the use of a questionnaire containing several open and scored questions. Table 1 shows the scored questions as they were used for the quality-assessment of the Austrian census of 2011.

Table 1: Scored Questions — HD Documentation

DATA HISTORIOGRAPHY	
	Can we detect data changes over time?
	Is the information available for the cut–off 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?

Data for the LMS are obtained in eleven source registers which have to be assessed.[2] The calculation of the hyperdimension documentation HD^D for each source register is illustrated in table 2. The data holders answer quality related questions on a dichotomous (Yes or No) or ordinal scale. The higher the value for each question the better should be the quality-related performance of the register. According to theoretical considerations each question is weighted differently. The metadata for each register is summarized as the weighted average of these scored questions. For example, the value of 1 for the question definitions in the central population register (CPR) means that the definition of the legal marital status is the same in the CPR than in the register-based census. In practice, data for comparison registers are delivered from up to 20 data holders (regional offices). On raw data level, documentation is done for each delivery. According to our data processing, these sources are aggregated to one comparison register. For HD^D the relative contribution to each comparison register is used to compute a weighted average out of the single answers for each comparison register. Table 2 illustrates, that in the Social Welfare Register (SWR) a data copy cannot be produced for the exact cut-off-date for all delivered records, yielding to a value of 0.47 for the sub-dimension cut-off-date for this comparison register. Now we summarize the available metadata as the weighted

[2] Source registers: *ASR*: Asylum Seekers Register, *UR*: Unemployment Register, *RPS*: Register of Public Servants of the Federal State and the Länder, *CAR*: Child Allowance Register, *CFR*: Central Foreigner Register, *CSSR*: Central Social Security Register, *CHR*: Chambers Register, *HPSR*: Hospital for Public Servants Register, *SWR*: Register of Social Welfare Recipients, *CPR*: Central Population Register, *TR*: Tax Register.

average of the sub-dimension. This yields to exactly one quality measure for HD^D for the LMS for each register.

$$HD^D_{ij} = \frac{\text{obtained score}}{\text{achievable score}} \qquad \begin{matrix}\text{i ... Register}\\ \text{j ... Attribute}\end{matrix} \tag{1}$$

Table 2: Calculation of the hyperdimension documentation HD^D for the legal marital status (*LMS*)

HD	Weight	ASR	UR	RPS	CAR	CFR	CSSR	CHR	HPSR	SWR	CPR	TR
Detect Changes	1	0	1	0.87	1	0	1	0.67	0.35	0.51	1	1
Cut-off date	2	0	1	0.87	1	0	1	0.67	0.35	0.47	1	1
Definitions	2	1	1	1	1	1	1	1	1	1	1	1
Relevance	4	0	0	0.62	1	0	1	0.67	0.7	0.83	0	1
Legal basis	1	0	1	1	1	0	1	0.67	0.35	1	1	1
Timeliness	3	1	1	0.80	1	1	1	1	0.81	0.85	1	1
Administrative Contr	2	0.33	0.67	0.73	1	0.33	1	0.67	0.73	0.81	1	1
Technical Contr	2	0.67	0	0.70	1	0.67	1	0.78	0.49	0.77	1	1
Data management	4	0.33	1	0.63	0.67	0.33	0.67	0.78	0.59	0.64	1	1
HD^D		0.397	0.683	0.864	0.936	0.397	0.936	0.778	0.706	0.746	0.810	1

Hyperdimension Pre-processing HD^P

The hyperdimension pre-processing HD^P is based on the share of useless records (missing identification keys, missing values, values out of range, see table 3). The final result of

Table 3: HD Pre–processing

	Number of observations
—	Records without unique ID
—	Records with item non–response (but including unique IDs)
—	Records with wrong values or values out of range
=	Usable records

this hyperdimension is given by the ratio of *usable records* to the *total number of records*.

$$HD^P_{ij} = \frac{\text{usable records}}{\text{total number of records}} \qquad \begin{matrix}\text{i ... Register}\\ \text{j ... Attribute}\end{matrix} \tag{2}$$

HD^P for the LMS in the source registers are shown in table 4. Most data sources provided formally correct information on the LMS. However, for data from the Asylm-seekers Register(ASR) and Social Welfare Register (SWR) there was a significant amount of missing unique personal identifiers (56.1% and 14.4%, resp.) yielding a lower quality indicator.

Table 4: Calculation of the hyperdimension HD^P for the legal marital status (*LMS*)

Register	Observations	Missing bPIN %	Non resp. & Out of range %	HD^P
ASR	66411	56.12	3.73	0.402
UR	327702	1.30	7.74	0.910
RPS	640155	1.66	2.85	0.955
CAR	3658263	2.72	0.01	0.973
CFR	747688	7.67	2.58	0.898
CSSR	8811838	6.30	48.30	0.454
CHR	23904	3.40	41.51	0.551
HPSR	87954	6.23	38.60	0.552
SWR	263134	14.44	7.24	0.783
CPR	9605679	0.0	33.04	0.670
TR	9359027	6.28	9.31	0.844

Hyperdimension External Source HD^E

The last hyperdimension (HD^E) on raw-data level assesses the data-quality of the source registers in comparison to an external source, in our case, the Austrian microcensus. It is calculated as the number of consistent values divided by the number of all records that could be linked to the microcensus.

$$HD^E_{ij} = \frac{\text{number of consistent values}}{\text{total number of linked records}} \qquad \begin{matrix} \text{i ... Register} \\ \text{j ... Attribute} \end{matrix} \tag{3}$$

Table 5: Calculation of the hyperdimension HD^E for the legal marital status (*LMS*)

Register	Linked observations	Conflicting observations %	HD^E
ASR	10	50.0	0.500
UR	1239	1.9	0.981
RPS	2993	4.1	0.959
CAR	13905	3.0	0.970
CFR	2235	11.5	0.885
CSSR	20346	5.8	0.942
CHR	71	11.3	0.887
HPSR	194	2.6	0.974
SWR	576	5.2	0.948
CPR	27959	2.9	0.971
TR	24332	8.9	0.910

In table 5, we see the results of the comparison to an external source for the LMS. For example, 1.239 individuals from the Unemployment Register could be linked to the microcensus. Out of these observations, 1.9 per cent were classified wrong. This yields to a HD^E value of 0.981 for the LMS in the UR.

Final quality on the raw-data level

Given these three quality measures, an overall quality indicator for each attribute on register-level can be derived as a weighted average. In our framework, each hyperdimension has the same weight ($v^D = v^P = v^E$), and therefore an equal impact on the quality measure. The resulting value summarizes the existing quality-related information for each attribute j in each register i. Hence, this indicator is able to capture quality-related effects from the data generation through to the raw data in the registers.

$$q_{ij} = v^D \cdot HD_{ij}^D + v^P \cdot HD_{ij}^P + v^E \cdot HD_{ij}^E = \sum_{k \in D,P,E} v^k \cdot hd_{ij}^k \tag{4}$$

i ... Register, j ...Attribute

Table 6 summarizes the information for the attribute LMS for each register. Hence, we obtained eleven quality indicators. ASR has the lowest quality-measure, while CAR delivers the best quality for the LMS. The quality differs partly because of the different subgroups covered by the registers (families with young children vs. foreign people), but also because the LMS is not relevant for the ASR but it is relevant for the CAR. In the next step this information on quality of the data in the registers is used to evaluate the quality of the value chosen for the *CDB*.

Table 6: Calculation of the quality indicator for the (*LMS*) for the registers

Register	HD^D	HD^P	HD^E	q
ASR	0.397	0.402	0.500	0.433
UR	0.683	0.910	0.981	0.858
RPS	0.864	0.955	0.959	0.926
CAR	0.936	0.973	0.970	0.960
CFR	0.397	0.898	0.885	0.726
CSSR	0.936	0.454	0.942	0.777
CHR	0.778	0.551	0.887	0.739
HPSR	0.706	0.552	0.974	0.744
SWR	0.746	0.783	0.948	0.826
CPR	0.810	0.670	0.971	0.817
TR	1.000	0.844	0.910	0.918

3.2 The Central Data Base CDB

The entire information from the registers is combined in the Central Database (CDB, green box in Figure 2) which covers all attributes of interest for the register–based census. At this level, a quality indicatorq_j^n for each attribute j for each statistical unit n is computed for the first time. Concerning the evaluation of quality for the CDB we distinguish three types of attributes by their origin.[3] *Unique attributes* exist in exactly one register, e.g. educational attainment (cf. attribute C in figure 2). For this reason, the measure of quality in the CDB is the same as in the raw data. *Derived attributes* are based on different attributes, e.g. current activity status (cf. attributes F and G in figure 2). The registers do not contain any information for these attributes in the required specification, but related information. *Multiple attributes* show up in several registers, e.g. LMS (cf. attribute A in figure 2). Since there are multiple data sources providing a certain attribute, a predefined ruleset, based on experience of *Statistik Austria*, picks the most appropriate value from the underlying registers according to the constellation in the source registers. To assess the validity of this chosen value, all the available information is taken into account. The Dempster-Shafer Theory (DST) for the combination of evidence is applied to derive a quality measure for this attributes for each statistical unit.

The quality measures on the raw data level are considered as beliefs in the correctness of the value. DST for the combination of evidence takes into account all availabile evidence from the registers to form one quality-indicator on the CDB-level $q^n_\odot$ for each statistical unit n. In the next step, the values in the CDB are compared to an external source HD^E. This yields to the last quality indicator in the CDB q^n_Ψ. Table 7 shows the last quality measures on CDB-level $\overline{q}^n_\Psi$, which is the weighted average of $\overline{q}^n_\odot$ and HD^E. In our example $\overline{q}^n_\Psi$ is 0.728. Hence, HD^E slightly increases the quality indicator.

Table 7: The quality for the *LMS* on CDB level

	$\overline{q}^n_\odot$	HD^E	$\overline{q}^n_\Psi$
q	0.721	0.973	0.728

3.3 The Final Data Pool FDP

In the last step of the data generation missing values in the CDB are imputed in the FDP. For the assement of the data quality in the FDP the fourth Hyperdimension HD^I is computed. For that, the distinction of methods is crucial (see Kausl, 2012). In the Austrian census deterministic editing, Hot-Deck techniques and logistic regressions are

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

applied. However, the principle for the evaluation of the imputations is the same for all methods. It is based on the quality of the inputs and the quality of the imputation model. The quality of the input is assessed as a weighted average of the quality of the input variables, that are used for each statistical unit n.

$$HD^{In} = \underbrace{\frac{1}{N}\sum_{j=1}^{N} q_{\Omega j}}_{\overline{q}_{Input}} \cdot \Phi^{m} \tag{5}$$

I ... Imputation, n ... Statistical unit, N ... Number of Inputs for m,

m ... Imputation method, Φ^m ... Classification rate for m

The accuracy of the imputation models m is assessed using classification rates Φ. The classification rate is the number of correct imputed values, if the model is applied to existing data.[4] Finally, the quality of the imputations is the product of the quality of the input and the accuracy of the output of the model. For a detailed explanation of the quality assessment for the different imputation techniques see Astleithner et al. (forthcoming).

Table 8 shows the improvement of the average quality from CDB to FDP level. The average quality in the CDB, where missing values obtain the quality of zero, for the attribute is $\overline{q}^n_\Psi$. Now these missing records are imputed and obtain a quality measure according to their method of imputation. The average of the imputation quality HD^I for the *LMS* is 0.956[5] Formerly missing values now have a quality indicator higher than zero. For this reason, the average quality of the *LMS* is higher in the FDP ($\overline{q}^n_\Omega$) than in the CDB ($\overline{q}^n_\Psi$).

Table 8: The quality for the *LMS* on FDP level

	$\overline{q}^n_\Psi$	$\overline{HD^I}$	$\overline{q}^n_\Omega$
q	0.728	0.956	0.949

[4]For ordinal variables the distance between the true value and the estimated value is taken into account. For numerical variables, the accuracy of the model is simply the correlation coefficient between the true and the imputed values.

[5]Due to space constraints the comprehensive calculation of the quality framework can not be illustrated at this place. For further information see the documentation of data quality for the Austrian census.

4 Results of the quality assessment for the Austrian Census 2011

The introduced quality framework is used for the quality assessment of the Austrian register-based census of 2011.The resulting quality measures for selected attributes on CDB- and FDP-level are presented in table 9.[6] We will explain the different possible combinations of types of attributes and imputations using examples. The quality of the multiple attribute Age, is compared to an external source HD^E on CDB level $\overline{q}^n_\Psi$ and 415 records are imputed ($Impутions$) on FDP-level $\overline{q}^n_\Omega$. Sex, a multiple attribute has the best data quality according to our framework. As there are no imputations the data quality on FDP-level is the same as on CDB-level. The educational attainment is a single attribute. This means, that the quality is the same in the source register and in the CDB. However, on FDP-level the average quality increases due to 293698 imputations. This distinguishes it from the attribute field of educational attainment, which is not imputed. This means, that the quality-indicator is the same from the register to the FDP. The family status is the last type of attribute, as there are derivations on FDP-level. In that case, the comparison to an external source is carried out in the FDP.

Table 9: Results for the Austrian census on FDP-level

Attribute	$\overline{q}^n_\odot$	HD^E	$\overline{q}^n_\Psi$	$\overline{HD^I n}$	HD^E	$Imputations$	$\overline{q}^n_\Omega$
Age	0.999	0.997	0.998	0.731	.	415	0.998
Sex	1.000	0.999	0.999	.	.	.	0.999
Country of birth	0.987	0.982	0.986	.	.	.	0.986
Country of citizenship	0.985	0.995	0.988	.	.	.	0.988
Legal marital status	0.721	0.973	0.728	0.956	.	1929346	0.949
Educational attainment	0.791	.	0.791	0.595	.	293698	0.815
Field of educational attainment	0.819	.	0.819	.	.	.	0.819
Family status	0.820	.	0.820	0.799	0.999	923910	0.948
Current activity status	0.909	0.923	0.913	.	.	.	0.913
Status in employment	0.930	0.955	0.930	.	.	.	0.930
Occupation	0.535	0.492	0.535	0.645	.	1036480	0.699
Full/part time employment	0.681	.	0.681	0.788	.	133421	0.707

5 Conclusion and Outlook

The comprehensive quality-framework enables to assess the quality of data in every step of the data-generation. Even though it was developed around the first register-based cen-

[6]Because of space constraints the underlying assessment of raw-data is not illustrated.

sus in Austria, the aim was to realize a generalizable procedure for the evaluation of all kind of administrative data. According to theoretical considerations, the weights can be chosen and due to the modular design, each step can be carried out individually. The application of the quality framework for the register-based census comprises various possibilities. From one final quality indicator the user can decompose the value and find the underlaying quality related information. As the quality indicator is calculated on the level of statistical units data quality can be analyzed for sub-groups of the census. Furthermore, it can be used as an additional factor of uncertainty in statistical analysis. The possibility to use the quality indicator for statistical purposes is, however, still an ongoing research task. A very simple, but nevertheless important application is the comparison and monitoring of data-quality. Both, between different data sources and between different census-years.

References

Astleithner, F., Ćetković, P., Humer, S., Lenk, M., Moser, M., Schnetzer, M., et al. (forthcoming). Quality measurement in administrative statistics and the assessment of imputations. *Journal of Official Statistics*.

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.

Daas, P., & Fonville, T. (2007). *Quality control of dutch administrative registers: An inventory of quality aspects* (Tech. Rep.). Statistics Netherlands.

Daas, P., Ossen, S., Vis-Visschers, R., & Arends-Tóth, J. (2009). Checklist for the quality evaluation of administrative data sources. *Statistics Netherlands Discussion Paper*(09042).

Kausl, A. (2012). The data imputation process of the Austrian register-based census.

Conference contribution at UNECE Work Session on Statistical Data Editing, Oslo, 24-26 September 2012.

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

UNECE. (2014). *Measuring population and housing – Practices of UNECE countries in the 2010 round of censuses* (Tech. Rep.). United Nations Economic Commission for Europe.