Evaluation of the register-based census

Quality assessment of administrative data

Documentation of Methods

STATISTICS AUSTRIA

Registers, Classifications and Methods Division

30th June 2014

Preface

The quality-framework for the assessment of administrative data was developed in cooperation between Statistics Austria, division of register-based statistics, and Vienna University of Economics and Business, department for economics.

This report describes the principle of the quality-assessment for administrative data as it was used for the evaluation of the Austrian census of 2011. The detailed results for 2011 can be downloaded from the Homepage of Statistics Austria.[1] We would like to thank Prof. Wilfried Grossman (University of Vienna) and Thomas Burg (Statistics Austria) for their important contribution to the framework. Furthermore, we would like to thank Christopher Berka, Reinhard Fiedler, Matthias Schnetzer and Christoph Waldner for their work in the development of the framework. Any errors that remain are in the responsibility of the authors.

Manuela Lenk — Franz Astleithner
Eva-Maria Asamer — Predrag Ćetković
Henrik Rechta — Stefan Humer
Eliane Schwerer — Mathias Moser
STATISTIK AUSTRIA — WU WIEN

[1] http://www.statistik.at/web_de/Redirect/index.htm?dDocName=076880

Contents

Chapter 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 this 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, various books and articles were published in the last decade. Departing from Pipino, Lee, and Wang (2002); Batini and Scannapieco (2006); Karr, Sanil, and Banks (2006) who have a broad understanding of data-quality, Wallgren and Wallgren (2007) developed a guideline for the assessment of the different dimensions of data-quality. The Scandinavian countries have a long tradition in the use of administrative data (UNECE, 2007; Zhang, 2011; P. J. Daas, Ossen, Tennekes, & Nordholt, 2012; Hendriks, 2012; Axelson, Holmberg, Jansson, Werner, & Westling, 2012; Zhang, 2012). Based on this experience, Statistics Austria developed a standardized quality framework for the assessment of administrative data.

This report describes the conception of the quality-framework of administrative data, as it was developed and applied for the register based census of 2011. In every stage of the data processing a quality-indicator is derived for each attribute. Even though the framework was developed around the register-based census, it was designed for general applicability. Due to the modular design, every step of the quality-framework can be applied individually. In the chapter 2, we will introduce the sources of the register based census. In chapter 3, the quality framework is explained using the example from the quality assessment for the *Legal Marital Status LMS*. If an attribute is obtained from multiple register, the information from the data sources have to combined. Chapter 4 focuses on the application of Dempster-Shafer-Theory for this purpose. In chapter 5 the quality assessment of imputation is explained in detail.

Chapter 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.

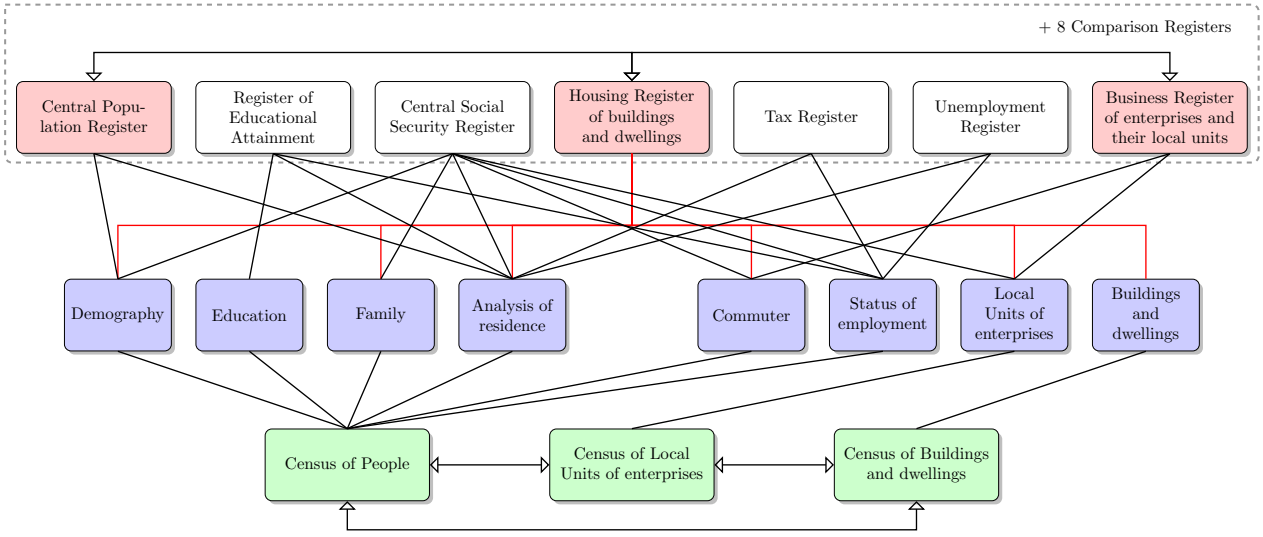

Figure 2.1: Data sources for the register–based census

Figure 2.1 illustrates the connections between the data sources and topics of the census. Statistics Austria distinguishes between seven base registers and eight 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).

Chapter 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 (P. Daas, Ossen, Vis-Visschers, & Arends-Tóth, 2009; P. Daas & 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 3.1 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 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 (registers i; see blue boxes in Figure 3.1) is obtained via three hyperdimensions: Documentation (HD^D), Pre-processing (HD^P) and

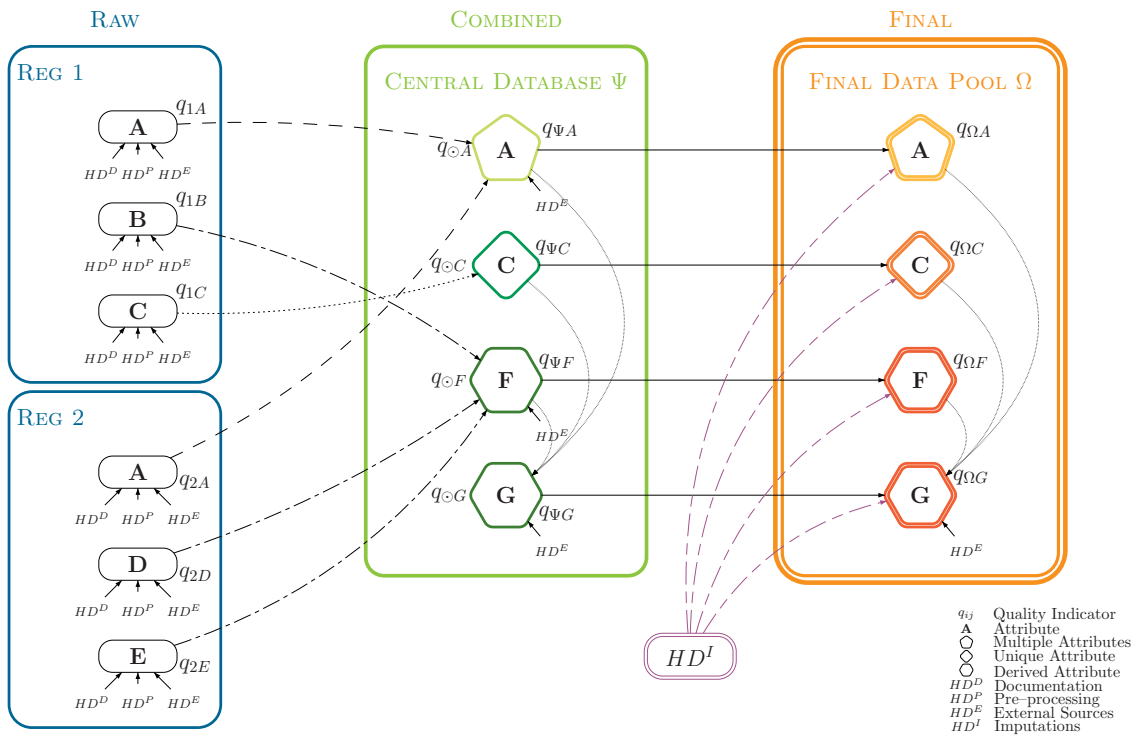

Figure 3.1: Quality framework for register–based censuses

External Source (HD^E). The derivation of the quality measures for the *Legal Marital Status 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 (meta-data) at 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. The open questions gather information of general interest, like the timeliness of data delivery or on the definition of the sample. This information is important for the documentation of the delivery but is not used for the quality assessment of the census. Table 3.1 shows the scored questions and the corresponding weights as they were used for the quality-assessment of the Austrian census of 2011.

Data for the LMS are obtained in eleven source registers i which have to be assessed.[1] The calculation of the hyperdimension documentation HD^D for each source register is illustrated in table 3.2. 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

[1] 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.

Table 3.1: Scored Questions — HD Documentation

Hyperdimension Documentation	Weight
DATA HISTORIOGRAPHY	
Can we detect data changes over time?	1
Is the information available for the cut–off date?	2
DEFINITIONS	
Are the data definitions for the attribute compatible to those of STATISTICS AUSTRIA?	2
ADMINISTRATIVE PURPOSE	
Is the attribute relevant for the data source keeper?	4
Does a legal basis for the attribute exist?1	
DATA TREATMENT	
How fast are changes edited in the register?	3
Are the data verified on entry?	2
Are technical input checks applied?	2
How good is the data management, i.e. ex post consistency checks?	4

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 average out of the single answers for each comparison register is computed. Table 3.2 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 average of the sub-dimension. This yields to exactly one quality measure for HD^D for the LMS for each register (Table 3.2).

$$HD^D_{ij} = \frac{\text{obtained score}}{\text{achievable score}} \qquad \begin{matrix}\text{i ... Register}\\ \text{j ... Attribute}\end{matrix} \tag{3.1}$$

Hyperdimension Pre-processing HD^P

The hyperdimension pre-processing HD^P is based on the share of useless records (missing unique personal identifiers, missing values, values out of range; see table 3.3). The final result of this hyperdimension is given by the ratio of *usable records* to the *total number of records*.

$$HD^P_{ij} = \frac{\text{usable records}}{\text{total number of records}} \qquad \begin{matrix}\text{i ... Register}\\ \text{j ... Attribute}\end{matrix} \tag{3.2}$$

Table 3.2: 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

Table 3.3: HD Pre–processing

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

The results for this hyperdimension for the LMS in the source registers are shown in table 3.4. Most data sources provided formally correct information on the LMS. However, data from the Asylum Seekers Register (ASR) and Social Welfare Register (SWR) have a significant amount of missing unique personal identifiers (56.1% and 14.4%, resp.) lowering the quality indicator.

Table 3.4: Calculation of the hyperdimension HD^P for the legal marital status (*LMS*)

Register	Observations	Missing bPIN %	Non resp. & Out of range %	HD^P
ASR	66,411	56.12	3.73	0.402
UR	327,702	1.30	7.74	0.910
RPS	640,155	1.66	2.85	0.955
CAR	3,658,263	2.72	0.01	0.973
CFR	747,688	7.67	2.58	0.898
CSSR	8,811,838	6.30	48.30	0.454
CHR	23,904	3.40	41.51	0.551
HPSR	87,954	6.23	38.60	0.552
SWR	263,134	14.44	7.24	0.783
CPR	9,605,679	0.0	33.04	0.670
TR	9,359,027	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. If the attribute is not covered by the microcensus or another suitable survey, an expert on the specific dataset is asked to assess the validity of the data on a scale between zero and one.

$$HD_{ij}^{E} = \frac{\text{number of consistent values}}{\text{total number of linked records}} \qquad \begin{matrix} \text{i ... Register} \\ \text{j ... Attribute} \end{matrix} \tag{3.3}$$

Table 3.5: 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	1,239	1.9	0.981
RPS	2,993	4.1	0.959
CAR	13,905	3.0	0.970
CFR	2,235	11.5	0.885
CSSR	20,346	5.8	0.942
CHR	71	11.3	0.887
HPSR	194	2.6	0.974
SWR	576	5.2	0.948
CPR	27,959	2.9	0.971
TR	24,332	8.9	0.910

In table 3.5, 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.

$$q_{ij} = v^D \cdot HD_{ij}^D + v^P \cdot HD_{ij}^P + v^E \cdot HD_{ij}^E = \textstyle\sum_{k \in D,P,E} v^k \cdot hd_{ij}^k \qquad \text{i ... Register, j ...Attribute} \tag{3.4}$$

Table 3.6 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 relevant for the CAR but it is not for the ASR. In the next step

this information on data quality in the registers is used to evaluate the quality of the value chosen for the *CDB*.

Table 3.6: 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 3.1) which covers all attributes of interest for the register–based census. At this level, a quality indicator q_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.[2]

Unique attributes exist in exactly one register, e.g. educational attainment (cf. attribute C in figure 3.1). 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 3.1). 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 3.1). 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 (Chapter 4) is applied to derive a quality measure for these attributes for each statistical unit.

Depending on the type of the assessed attribute an additional comparison to an external source is carried out in this step.Multiple attributes, attributes that couldn't be compared

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

to an external source on the raw-data level and attributes that are derived on CDB-level are compared to an external source at this stage.

If we focus on our example, the LMS, 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 available 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.[3] This yields to the last quality indicator in the CDB q^n_{Ψ}. Table 3.7 shows the last quality measures on CDB-level $\overline{q}^n_{\Psi}$, which is the weighted average of $\overline{q}^n_{\odot}$ (Weight=0.75) and HD^E (Weight=0.25). In our example $\overline{q}^n_{\Psi}$ is 0.728. Hence, HD^E slightly increases the quality indicator.

Table 3.7: 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 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.

$$HD^{In} = \Phi^m \cdot \underbrace{\frac{1}{N}\sum_{j=1}^{N} q_{\Omega j}}_{\overline{q}_{Input}} \tag{3.5}$$

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

[3] This additional comparison to an external source is only carried out for multiple and derived attributes. If an attribute is derived in the FDP, the additional external source is carried out in the FDP

[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.

$\overline{q}_{Input}$ and the accuracy of the output of the model Φ^m. For a detailed explanation of the quality assessment for the different imputation techniques see Chapter 5 and Astleithner et al. (forthcoming).

Table 3.8 shows the improvement of the average quality from CDB to FDP level. The average quality in the CDB, where missing values have 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. 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 3.8: 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

Chapter 4

Dempster-Shafer theory for the combination of evidence

If attributes are obtained in multiple sources, the information on the values in the source registers and their quality can be used for the assessment of the value in the CDB. For the combination of evidence the Dempster-Shafer theory (DST) is applied. First, we will give an introduction to the DST. Second, we show its application for administrative data. Finally we give an artificial example for the calculation of the quality-indicator for a multiple attribute.

4.1 Dempster-Shafer Theory

Uncertainty plays an essential role in the analysis of complex systems. Nonetheless the definition of uncertainty in such research tasks often remains ambiguous or unclear. Usually the researcher encounters uncertainty as a dual phenomenon (Helton, 1997):

Stochastic Uncertainty results from the fact that systems can behave in different ways, i.e. it is a property of the system itself.

Epistemic Uncertainty occurs due to a lack of knowledge about the system and is thus a methodological problem when performing an analysis. Accordingly epistemic uncertainty deals with the lack of knowledge about the distribution of a certain variable itself, e.g. whether the register represents the 'true' values.

(Hacking, 1975) traces this very important distinction between the two types of uncertainty back to the beginnings of probability theory. (Helton, 1997) states that as long as the separation between stochastic and epistemic uncertainty is not maintained carefully, an evaluation of the systems behavior and characteristics on rational basis becomes difficult or even impossible.

It is common sense that the stochastic part of uncertainty is best dealt within the so–called frequentist approach, the most important discipline of the traditional probability theory. In contrast, the epistemic uncertainty is not considered carefully enough by such a theory.

In order to deal with these shortcomings of traditional probability theory, we apply a 'fuzzy approach'. Statistical fuzzy logic aims to explain epistemic uncertainty and tries to implement models for it. It can be regarded as an extension to the classical probability theory and will therefore yield the same results when no uncertainty is present. Platon already mentioned that besides the dual approach of either TRUE or FALSE there has to be a way to express uncertainty. Current applications of the fuzzy logic are mainly based on the ideas of (Zadeh, 1965). He introduces fuzzy sets, in which an element can be included or excluded, but he also allows for partial inclusion in the set. The degree of inclusion is given by a so-called membership function as a value in the interval [0,1]. These functions exist for each element and combined they yield the so-called fuzzy functions. These functions are generated either through statistics or opinions of experts.

To derive these 'expert opinions', a special form of fuzzy logic can be applied. More specifically we use an evidence theory that was proposed by (Dempster, 1968) and extended by (Shafer, 1992). This so–called Dempster–Shafer Theory focuses on a field of probability theory that is closely related to fuzzy logic. It allows to combine different beliefs about the reality, i.e. expert opinions. Eventually this results in a measure of evidence, which can be interpreted as a probability. This approach is specifically useful when an expert cannot make a definitive statement about the probability that a specific event will occur. What he or she has is a fuzzy belief about the probability that a certain event will arise. In this case the belief of an expert may differ from that of other experts. The Dempster-Shafer Theory aims to combine these different beliefs to come to an overall idea of the probability, taking the uncertainty among different beliefs into account. Consider the treatment of an ill patient in a hospital. Some doctors may have different beliefs about the true reason for the sickness. One doctor might consider a malfunction of the liver as the reason and has a degree of belief of 90%. It could be that another doctor thinks of some other reasons and therefore beliefs that the malfunction of the liver is not the main cause. An easy way to evaluate the overall belief would be a simple averaging of the different beliefs. However this approach does not consider the uncertainty nor possible conflicts between expert opinions that are closely connected to the different beliefs. The Dempster–Shafer Theory tries to overcome these shortcomings by considering the role of uncertainty and conflicts within its framework.

The theory consists of three fundamental functions: the *basic probability assignment function* (bpa), the *Belief function* (Bel) and the *Plausibility function* (Pl) (Sentz & Ferson, 2002).

More formally, the power set 2^X is the set of all subsets of X, including the empty set $\emptyset$ and X itself. The elements of the power set 2^X can be considered as hypothesis of the condition of a complex system. For the Census 2^X can be seen as all possible constellations of certainty and uncertainty between registers or merely a subset of them (see table 4.3 and the explanation in the following chapter). The Dempster-Shafer Theory of evidence assigns a specific degree of belief to each element of 2^X. Formally, this is represented by the *basic probability assignment function (bpa)* (?, ?, see)]klir:1998. The *bpa* defines a mapping of the power set to the interval between 0 and 1.

$$bpa : 2^X \rightarrow [0, 1]$$

The *bpa* of the empty set is 0 and the summation of the *bpa*s of all elements of the power set 2^X equals 1.

$$bpa(\emptyset) = 0 \qquad \sum_{A \in 2^X} bpa(A) = 1$$

The value of the *bpa* describes the proportion of evidence that encourages the hypothesis that a particular element of X belongs to a specific element $A \in 2^X$ of the power set, but not a particular subset of A. It is important to note that the $bpa(A)$ of the set A is not imply any statement about the value of the bpa for a subset of A.

Based on the *basic probability assignment function* we derive an interval that contains the precise probability of the condition A of a system.

$$Bel(A) \leq P(A) \leq Pl(A)$$

The lower bound *Belief* for a certain set A is the sum of all *bpa*s of subsets B of the set of interest A.

$$Bel(A) = \sum_{B|B \subseteq A} bpa(B)$$

The upper bound of the interval is defined as the measure of *Plausibility* and is evaluated by summing up all *bpa*s of the set B that intersect the set of interest A.

$$Pl(A) = \sum_{B|B \cap A \neq \emptyset} bpa(B)$$

The *Beliefs* (Bel) as well as the *Plausibilities* (Pl) do not have to sum up to 1. They are non-additive since they are sums over an arbitrary number of subsets B out of A. Furthermore, due to the fact that the *basic probability assignments* sum up to 1, the *Plausibility* can be derived from the measure of *Belief* and vice versa.

$$Pl(A) = 1 - Bel(\neg A)$$

Furthermore we can compute also *bpa*s from e.g. give Beliefs. If $Bel(A)$ equals $Pl(A)$ the probability of the condition A of a process is explicitly determined. Accordingly the Dempster-Shafer Theory of Evidence yields the same results as traditional probability theory. In the presence of epistemic uncertainty the values of the two measures differ and form an interval of lower and upper bounds of probabilities. The actual value of probability is included in the interval composed of the *Belief* and *Plausibility*. Some examples for *basic probability assignment function* (bpa), the *Belief function* (Bel) and the *Plausibility function* (Pl) can be found in the Appendix.

An advantage of Dempster–Shafer's Theory of Evidence is its capability of combining information from independent sources when epistemic uncertainty is present. Generally, the intention of data aggregation is to summarize and simplify information. Widely used aggregation methods are the evaluation of averages (either in arithmetic, geometric or harmonic form) or the selection of particular properties of the data (e.g. minimum, maximum or median of an empirical distribution). Combination rules can be seen as a derivation of such rather simple aggregation techniques. Their purpose is to aggregate evidence about the condition of a system obtained from multiple data origins. Examples for different sources of information depend strongly on the field of application. Their role can be taken by a group of experts (e.g. doctors), a number of sensors (airborne radar stations) or various administrative registers, which deliver information on certain attributes of statistical units of the population.

The initial rule for the combination of evidence within the Dempster–Shafer Theory is the so–called Dempster Rule (Dempster, 1967). It can be regarded as a generalization of Bayes' rule and conflates multiple *Belief functions* by aggregating their *basic probability assignment* functions (see equation 4.1). Dempsters Rule is a strictly conjunctive procedure and accents agreement of multiple sources of information.

$$bpa_{1,2}(A) = (bpa_1 \oplus bpa_2)(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \emptyset} bpa_1(B) \cdot bpa_2(C) \tag{4.1}$$

Conflicting evidence is considered through the normalisation factor $1 - K$, whereas K stands for the sum of *bpa*s assorted with conflict, as can be seen in equation 4.2. Set-theoretic, these are all products of *bpa*s where the intersection equals $\emptyset$.

$$K = \sum_{B \cap C = \emptyset} bpa_1(B) \cdot bpa_2(C) \tag{4.2}$$

This property of Dempsters Rule induced heavy criticism by (Zadeh, 1986) and (Yager, 1987). As a consequence, various combination rules were proposed in the literature, like for example Yager's rule or Dubois and Prade's disjunctive pooling rule. In the application on the quality measurement of combined administrative data sources, we disregard their conceivable arguments because Dempsters Rule is associative (Joshi, Sahasarabudhe, &

Shankar, 1995). Consequently the succession of the multiple sources has no impact on the results of our analysis. Examples for coinciding and conflicting evidence is provided in the Appendix.

4.2 Application

For the register-based census we use the Dempster-Shafer Theory to combine quality indicators from different data sources. The quality framework aims to deliver a quality indicator for each attribute in the Census Database (CDB), which contains information on the population in Austria (e.g. residency, sex, status of employment). It is filled from different administrative data sources (registers) based on a predefined ruleset. The quality indicators of the attributes in this CDB are derived based on the quality measures from the original registers. If there is only one base register available to compare with, the indicator also resembles the quality of the CDB. In the case of multiple attributes several registers have information over the same attribute, e.g. sex may be included in four registers. Since these sources can be regarded as different opinions (or beliefs) on a common subject (the attribute) it allows for the implementation of the Dempster-Shafer Theory.

In a first step we assign a certain mass of certainty (C) and uncertainty (U) to each attribute in each register, which is based on the quality measures of these attributes q_{ij}. This yields 2^n (n being the number of registers with the same attribute) possible combinations of certainty and uncertainty. For the case of $n = 2$ that would be: CC (both certain), CU or UC (one register uncertain), UU (both registers uncertain). These different cases can be grouped into agreement, uncertainty, logical impossibility and conflicting evidence. This process depends on the values of the attribute in the different registers, e.g. CC can be agreement (if both registers are sure the person is 'female') but in another case it can be a logical impossibility (if one register is sure that the person is 'male' and the other is sure that the person is 'female').

In some cases there are a lot of registers which contain information about the same attribute. If the number of registers becomes large, computational difficulties will arise because of the assignment of the constellations of certainty and uncertainty to the corresponding groups mentioned above. We solve them by creating (i.e. from REG_1 to REG_n) a look-up table for each case, which contains possible combinations of certainty (C) and uncertainty (U). In the second step we simply take the combination (e.g CCCC) from the look-up table, that corresponds to our actual case.

Note that different registers may show differing values for the same observation. It is possible that register 1 is absolutely certain that an individual is male, while register 2 is sure that this person is female (case CC), which would be a logical impossibility. Therefore it depends on the values within the different registers if CC can be regarded as agreement or logical impossibility. Accordingly it is possible to calculate different

Table 4.1: Table of used symbols

Symbol	Name
Bel	Degree of Belief
ξ	Normalisation of uncertainty
$\lightning$	Logical impossibility
ω	Conflict
U^n	Uncertainty

beliefs, e.g. the belief that register 1 shows the true value or that register 2 is correct.

The combination rules are calculated based on the degree of agreement, uncertainty, logical impossibility and conflicting evidence.

$$Bel = \frac{1 - \lightning - \omega - U^n}{1 - \lightning} \tag{4.3}$$

The general equations 4.3 and 4.4 are applied to each observation for each attribute. The CDB marks the actual belief for a certain observation and therefore defines which belief is calculated.

$$\xi = \frac{U^n}{1 - \lightning} \tag{4.4}$$

Accordingly if an individual is male according to the CDB we check the reliability of this information using the comparison registers. For each observation measures of *Belief* and *Plausibility* are constructed. These figures can be interpreted as a confidence interval for the accuracy of each observation k. The quality indicator for each observation is now computed as the mean of belief and plausibility.

$$q_{\Psi,A_k} = \frac{Bel(A_k) + Pl(A_k)}{2}$$

The overall quality indicator for attribute A is computed as the average over the whole population.

$$q_{\Psi,A} = \frac{1}{2n} \sum_{k=1}^{n} (Bel(A_k) + Pl(A_k))$$

We will present not only the mean for each attribute (although it is the most important figure for our application as it represents the quality indicator for a multiple attribute within the CDB) but also other moments and distribution measures such as the standard deviation or quantiles. These indicators give information on the accuracy and deviations of our results and therefore deliver a more sophisticated picture of the quality assessment for an attribute in the CDB.

4.3 Artificial example for the combination of evidence

Table 4.2: Quality Indicators for sex in four selected registers

Register	HD^D	HD^P	HD^E	$q_{i,sex}$
REG_1	0.7916	0.9424	0.9985	0.9108
REG_2	0.4444	0.7459	0.9966	0.7290
REG_3	1.0000	1.0000	0.9982	0.9994
REG_4	0.7916	0.9927	1.0000	0.9281

Suppose we derived the following quality indicators for the attribute sex in four registers (REG_1 - REG_4) in the first step of the quality framework (see Table 4.2). Where the columns represent different quality aspects. These quality measures are combined using weighted averages. In this case we weighted each quality aspect equally. $q_{i,sex}$ is thus given by

$$q_{i,sex} = \frac{1}{3}HD^D + \frac{1}{3}HD^P + \frac{1}{3}HD^E$$

There is no specific rational behind this weighting. One could apply sensitivity analyses to get an idea of the impact of the weights. Since in this case we can use information on sex from four registers there exist 2^4 possible constellations of certainty (C) and uncertainty (U), as can be seen in table 4.3. The *bpas* can then be derived by multiplying the q_{ij} of the registers according to the certainty–uncertainty setting, e.g. for UUCC using the values from table 4.2:

$$bpa_{UUCC} = (1 - 0.9108) \cdot (1 - 0.7290) \cdot 0.9994 \cdot 0.9281 = 0.0224$$

The decision whether we need to use q_{ij} or its complementary probability is made by the CDB. Suppose the CDB regards a specific person as 'male'. If this person is also 'male' in REG_3 then the register has a certainty (C) of q_{ij} that this is correct. Another register REG_1 may believe the person is 'female', hence it is uncertain (U) with the value $(1-q_{ij})$ about the person being 'male'. Accordingly certainty and uncertainty are defined by the values of the register's attribute compared to the CDB.

Following this short example we will present first results of the application of the Dempster-Shafer Theory on the quality framework for the Austrian census. We will again focus on the attribute sex for reasons of simplifications. Table 4.4 shows some distribution figures for the average $q_{\Psi,sex}$ of the upper (*Plausibility*) and lower bound (*Belief*), which can be interpreted as an aggregated quality indicator.

The CDB contains 8.363.820 observations on the attribute sex. The most important moment is the mean (μ) which gives an idea of the overall quality of the attribute sex within the CDB. However, the other measures show that even on the unit level the quality

Table 4.3: Possible Combinations of Certainty and Uncertainty for four Registers

Constellation	*bpa*	Constellation	*bpa*
UUUU	0.00001	CUUU	0.00001
UUUC	0.00001	CUUC	0.00014
UUCU	0.00174	CUCU	0.01773
UUCC	0.02240	CUCC	0.22889
UCUU	0.00001	CCUU	0.00003
UCUC	0.00004	CCUC	0.00038
UCCU	0.00467	CCCU	0.04770
UCCC	0.06029	CCCC	0.61597

Table 4.4: Results of the Dempster-Shafer Application for the attribute sex

Measure of $q_{\Psi,sex}$	Value	Measure of $q_{\Psi,sex}$	Value
Observations	8363820	Percentile_{05}	0.99997
μ	0.99873	Percentile_{25}	0.99997
σ	0.03485	Median	0.99999
Min	0.00002	Percentile_{75}	0.99999
Max	1	Percentile_{95}	0.99999

indicators are very high. For the 5% percentile the quality indicator is already very close to 1. This concentration is also supported by a very low standard deviation (σ). We get a few observations with extremely low quality measures while the majority has rather high quality measures. Accordingly the mean is shifted to the left and falls below the 5% percentile. On the whole our quality measures are extremely left-skewed. Consequently we reach a high degree of confidence that the attribute sex has a very high quality within the CDB.

Both measures, the mean as well as the deviation, provide important information on the quality. For an other attribute we may find a high value for the mean but rather high deviations, which could indicate that one should take a further look at a certain subsample of the population.

Chapter 5

Quality assessment of imputations

After the calculation of the quality indicator for the real values in the CDB, the quality of the imputations has to be assessed. First, we discuss some theoretical considerations on the quality assessment of imputations. Second, we give an overview over the applied imputations methods of the register-based census. In the last section, we show the calculation of the quality indicator for the different types of attributes.

5.1 Imputation process and estimating order

Due to the principle of redundancy, the amount of missing values in register–based statistics is generally considered to be rather low, since a large part of variables is covered in multiple registers. For instance, in the Austrian register–based census of 2011 the level of item non-response for most attributes does not exceed 10% by far. Especially for demographic variables, like sex or age, the number of missing values is considerably lower. Nevertheless, some values need to be imputed due to different reasons.

The EU Commission Regulation 1151/2010 distinguishes between *item imputation* and *record editing* (see European Commission, 2010). Item imputation refers to the insertion of artificial but plausible information into a data record with a missing value in this specific attribute. More specifically, imputations try to set a value in accordance with information already available either in the same record or in the rest of the database. Record editing is the process of checking and modifying data records to make them plausible while preserving major parts of these records. However, record editing is often accomplished by deleting implausible (or out–of–range) values and subsequently re–imputing the missing entries. On the contrary, Chambers (2001, p. 11) does not distinguish "between imputation due to missingness or imputation as a method for correcting for edit failure". He argues that in both cases the true values are missing. For the quality assessment in Austria, both types are treated the same way irrespective of the reason for the imputation.

For example Chambers (2001, p. 11f) distinguishes five quality-related properties that imputations should fulfill:

(1) Predictive Accuracy: The imputed values should be as "close" as possible to the true values.

(2) Ranking Accuracy: The imputation process should preserve the order of imputed values (for attributes which are at least ordinal).

(3) Distributional Accuracy: The imputation procedure should preserve the distribution of the true data values.

(4) Estimation Accuracy: The lower order moments of the distribution of the true values should be reproduced by the imputation process (for scalar attributes).

(5) Imputation Plausibility: The imputation procedure should result in imputed values that are plausible.

These conditions may serve as a reference point for the quality assessment of imputations. Furthermore, the imputation procedure requires a hierarchical estimation order to connect all necessary steps in a chronological way. In this respect, two aspects have to be considered on a theoretical basis (see Kausl, 2012):

- In most statistics based on administrative data, a variety of registers is used in order to ensure sufficient quality for all required attributes. Due to possible differences in the data delivery (delays) it is necessary to check at which time each item can be edited.
- The choice of predictors used for imputations should be based on their association with the variables to be imputed. Therefore, it is imperative to analyze the highest correlations between the variables to develop optimal estimation models for each imputation step. Already imputed variables can be used as predictors to estimate other items.

As an example, Figure 5.1 illustrates the imputation interdependencies between the variables of the Austrian census topics. The hierarchical work flow is indicated by the arrows from one to another attribute. The relationships between variables are not confined within the topics (e.g. *LMS* $\leftarrow$ *AGE, SEX* and *POB*), but also connect variables between the topics (e.g. *EDU* $\leftarrow$ *AGE, SEX, COC* and *PFE*). Demographic attributes, like age and sex, are the first ones in the estimation order, variables concerning the labour market are the last. Followingly, many other variables are required to impute missing values in labour market variables, such as occupation (*OCC*). In the next step, the quality of the imputations has to be evaluated.

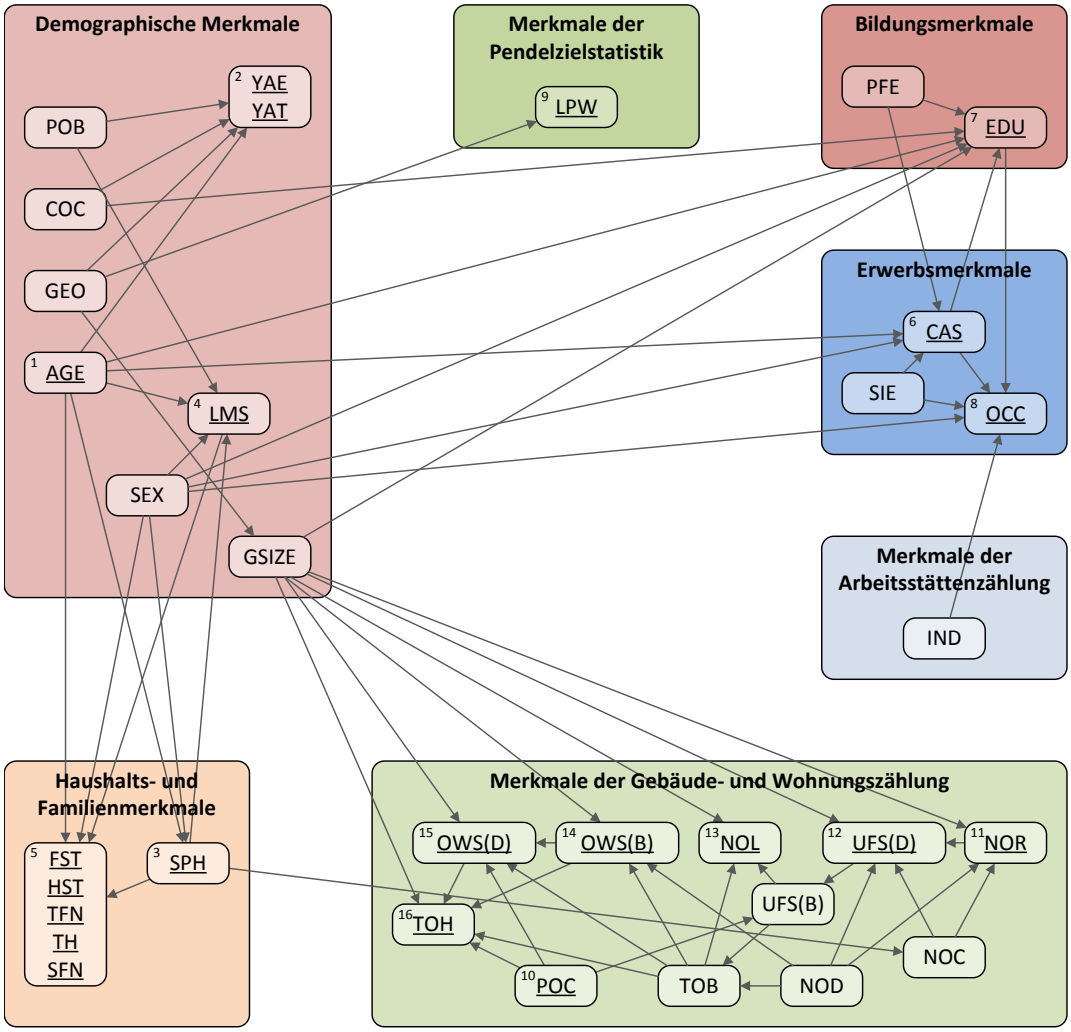

Figure 5.1: Estimation Order

AGE Age, *COC* Country of Citizenship, *CAS* Country of Citizenship, *EDU* Educational attainment (highest completed level), *FST* Family status, *GEO* Geographical area, *GSIZE* Size of the locality, *HST* Household status, *IND* Industry (branch of economic activity), *LMS* Legal marital status, *LPW* Location of place of work, *NOC* Number of occupants, *NOD* Number of dwellings in the building, *NOL* Number of floors of the building, *NOR* Number of rooms, *OCC* Occupation, *OWS(B)* Type of ownership for the building, *OWS(D)* Type of ownership for the dwelling, *PFE* Current education, *POB* Place of birth, *POC* Period of construction, *SEX* Sex, *SFN* Size of family nucleus, *SIE* Status in employment, *SPH* Size of private household, *TFN* Type of family nucleus, *TOB* Dwellings by type of building, *TOH* Type of heating, *TH* Type of household, *UFS(B)* Useful floor space of the building, *UFS(D)* Useful floor space of the dwellings, *YAE* Year of arrival in the country since 2000, *YAT* Year of arrival in the country since 1980

5.2 Applied imputation methods

For the quality assessment of imputations in register–based statistics (HD^I in our framework), the distinction of methods is crucial (see Kausl, 2012). We distinguish between *deterministic editing*, *statistical estimation* (primarily Hot-Deck technique, but also logis-

tic regressions) and *statistical matching*. To begin with the first, missing values in the data can be imputed by deterministic rules, even before applying statistical methods because we are able to derive missing values from auxiliary data. Two examples from the Austrian register–based census illustrate such cases:

- Missing values in the legal marital status (*LMS*) are classified according to the Central Social Security Register. Information on individuals receiving a widow's or widower's pension is provided by this register. The relevant information is gathered to change missing values of the attribute *LMS* to *"widowed"*, if a person receives a widow's or widower's pension.
- People younger than 15 years are classified *"not applicable (persons under 15 years of age)"* with regard to the educational attainment (*EDU*). Their current activity status (*CAS*) is *"persons below the age of 15"* and their marital status (*LMS*) is *"never married"*.

We do not consider such derivations with the utmost matching probability as an estimation in the narrower sense but rather as plausibility steps. However, there are also derivations with substantial uncertainty due to a lack of information. Still, in the following cases taken from the Austrian register–based census no statistical imputation method is necessary:

- The Central Population Register has information on the place of birth (*POB*). Missing values are filled up with information on the country of citizenship (*COC*), if the person has a foreign citizenship. The available data justify this assumption: 77% of individuals with a foreign *COC* were also born in this (foreign) country. Hence, even though there is uncertainty, this imputation method classifies 77% of the attribute *POB* as correct when it is applied to observed data for 2011.
- Suppose the marital status (*LMS*) is missing and there is another individual living in the same household. If the other person is *"married"*, the age difference between the two individuals is less than 18 years, and their sex differs, then the missing marital status is set to *"married"*.

Another important imputation method for the Austrian census is Hot-deck imputation. This method choses the imputed value from an assumed or estimated distribution, that is taken from existing data (Little & Rubin, 2002). It is suitable for all scenarios of missing data, except for missing not at random higher than 10% (Roth, 1994). A detailed review on hot-deck methods is given by Andridge and Little (2010). For the Austrian case individuals are aggregated to groups (*"decks"*) by attributes which are strongly correlated to the response variable. The distribution in the decks of the source data, derived from the FDP, is transferred to the corresponding group of the target data. Table 5.1 gives an

example of artificial data for the *LMS*. The distribution of the existing values in the census of the same year is applied on the missing values for the same attribute. As an example, 55.6% of all females aged 30 to 40 years with their main residence in the federal state Tyrol and a missing value for *LMS* will be considered as married women. Since we cannot be sure which women with a missing LMS are actually married, a uniformly distributed random variable with the interval [0,1] determines the assignment of the LMS. According to our example in Table 5.1, the interval [0,0.37) is assigned to "LMS never married", the interval [0.37,0.926) is assigned to "married", the interval [0.926,0.996) is assigned "divorced" and finally the interval [0.996,1) is assigned to "widowed".

Table 5.1: Artificial example of the Deck for legal marital status (*LMS*)

Sex	Age	Federal State	Size of deck	$P_{never\,married}$	$P_{married}$	$P_{divorced}$	$P_{widowed}$
female	30-40	Tyrol	50.000	37%	55.6%	7%	0.4%
male	50-60	Vienna	100.000	12%	66%	20%	2%
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Finally, statistical matching is the last applied imputation method in the Austrian census. It is based on the combination of two incomplete records. We will explain the procedure using an example of a missing observation for the educational attainment (*EDU*). Register–based statistics rely on unique identification keys for every individual in order to combine the information from multiple data sources. Consider a data record with a missing value for the educational attainment (*EDU*). Consider another data record with a missing unique identification key but information on several other attributes, among which is the (*EDU*). Statistical matching searches these loose observations and connects them with individuals who have a missing value for *EDU* but else the same characteristics. Two incomplete records, one of them useless because of the missing identification key, can be merged to one complete record.

5.3 Quality assessment of imputation models

In general an overall quality measure for imputations requires the evaluation of two parts, the input of the estimation model as well as the output (i.e. the accuracy of the model). The inputs of the estimation models are assessed with the three hyperdimensions HD^D, HD^P, and HD^E that are combined in the CDB.

For the evaluation of the model itself, the so–called *classification rate* Φ is used to obtain a quality measure for the imputations.[1] It is a general measure for the goodness of fit and can also be calculated for a variety of imputation techniques. Its principle is

[1]The statistical measures to evaluate the imputation performance were adopted from Hui and AlDarmaki (2012) as well as Chambers (2001).

to apply the imputation model to already existing data and compare the results of the imputation process with the true values of these observations. The classification rate equals the ratio between the matching values and the number of all compared entries.

This measure can be applied specifically to categorical variables and is shown in equation (5.1), where $\widehat{Y}_i$ is the estimated value for the observed value Y_i^* of person i. n is the sample size and I is an indicator function. Take the legal marital status (*LMS*) as an example for a categorical variable. In this case the quality assessment should measure the hit ratio, i.e. the probability that the estimation model picks exactly the right category of the true value.

$$\Phi^m = 1 - n^{-1} \sum_{j=1}^{N} I(\widehat{Y}_j \neq Y_j^*) \tag{5.1}$$

For ordinal variables, the distance of the imputed value to the true value is relevant, hence equation (5.2) is a modification of the classification rate Φ that measures and standardizes this gap. A satisfactory quality indicator has to consider the accuracy of the model which means measuring the contiguity of the estimated value to the true value. Assume several categories of the attribute educational attainment (*EDU*), ranging from primary to higher tertiary education. If the true value was *higher tertiary education*, an estimated value of *lower tertiary education* would be more accurate than an estimated value of *lower secondary education*.

$$\Phi^m = 1 - n^{-1} \sum_{j=1}^{N} \left(\frac{1}{2} \left[\frac{|\widehat{Y}_j - Y_j^*|}{max(Y) - min(Y)} + I(\widehat{Y}_j \neq Y_j^*) \right] \right) \tag{5.2}$$

For the case of numerical variables both concepts (5.1) and (5.2) can be applied,[2] however a simple correlation coefficient between estimated and true values is considered to be a rather intuitive approach. One example for a metric attribute is the variable "useful floor space" (*UFS*) of a household. The correlation coefficient between the estimated and the true *UFS* can be applied analogously to the classification rate for the evaluation of the imputation model.

Finally, we explain the application of the assessment of imputation methods described above: deterministic editing with and without uncertainty, statistical estimation as well as statistical matching. As already mentioned, the source variables for the imputation process are the attributes in the FDP rather than attributes in the raw data. Therefore, the quality indicator from the FDP delivers the quality information for the source variables whereby we use the values for the single statistical units. According to the type of imputation we distinguish the following quality assessment rules:

[2]Chambers (2001, p. 15) suggests that the methods which are developed for categorical variables could also be applied on scalar attributes by first categorizing them. If the arbitrariness of categorizing variables should be avoided, an applicable imputation performance measure has to be constructed.

- **Deterministic editing without uncertainty:** The input quality equals the quality of the source variables $q_{\Omega,i}$ where i denotes the attribute. The output quality equals 1, as there is no uncertainty about the correctness of the model. The overall quality of the imputation yields

$$HD^{In} = \Phi \cdot \underbrace{\frac{1}{n} \sum_{j=1}^{N} q_{\Omega,i}^{n}}_{\overline{q^n}_{Input}} \tag{5.3}$$

  where $\Phi = 1$.

- **Deterministic editing with uncertainty:** The input quality equals again the average quality of the source variables $q_{\Omega,i}$, while the output quality equals the classification rate Φ, as shown in equation (5.3).

- **Statistical estimation:** We define imputation quality as the average quality of the predictors $q_{\Omega,i}$ (input quality) times the classification rate (output quality) for the imputations (see again equation 5.3). This measure is independent of the number of predictors and includes both the quality of the data used for the imputations as well as their ex–post fit.

- **Statistical matching:** Two incomplete records — one without unique identification key, another one with the missing value — are merged. Therefore, no imputation in the narrower sense is carried out. The formerly missing value in the merged records is from now on treated as any other non–missing value. The quality measure is obtained via the quality of the used data source.

Chapter 6

Conclusion

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 census 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.

The detailed results for the Austrian census of 2011 can be downloaded from the Homepage of Statistics Austria.[1]

[1] http://www.statistik.at/web_de/Redirect/index.htm?dDocName=076880

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *Volume 78, Number 1*, 40-64.

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*.

Axelson, M., Holmberg, A., Jansson, I., Werner, P., & Westling, S. (2012). Doing a register-based census for the first time: The swedish experiences. In A. S. Association (Ed.), *Jsm proceedings, survey statistics section* (p. 1473-1480).

Batini, C., & Scannapieco, M. (2006). *Data quality: concepts, methodologies and techniques*. Springer.

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.

Chambers, R. (2001). Evaluation criteria for statistical editing and imputation. *National Statistics Methodological Series No. 28*.

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).

Daas, P. J., Ossen, S. J., Tennekes, M., & Nordholt, E. S. (2012). Evaluation of the quality of administrative data used in the dutch virtual census. In A. S. Association (Ed.), *Jsm proceedings, survey statistics section* (p. 1462-1472).

Dempster, A. (1967). Upper and lower probabilities induced by a multivariate mapping. *Annals of Mathematical Statistics*, *38*, 325–339.

Dempster, A. (1968). A generalization of bayesian inference. *Journal of the Royal*

Statistical Society. Series B (Methodological), *30*(2), 205–247.

European Commission. (2010). Commission Regulation (EU) No 1151/2010. *Official Journal of the European Union*, *Volume 53, L 324*.

Hacking, I. (1975). The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference.

Helton, J. (1997). Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. *Journal of Statistical Compution and Simulation*, *57*, 3–76.

Hendriks, C. (2012). Input data quality in register based statistics – the norwegian experience. In A. S. Association (Ed.), *Jsm proceedings, survey statistics section* (p. 1473-1480).

Hui, G., & AlDarmaki, H. I. (2012). Editing and Imputation of the 2011 Abu Dhabi Census. Conference contribution at UNECE Work Session on Statistical Data Editing, Oslo, 24-26 September 2012.

Joshi, A., Sahasarabudhe, S., & Shankar, K. (1995). Sensitivity of combination schemes under conflicting conditions and a new method. In J. Wainer & A. Carvalho (Eds.), *Advances in artificial intelligence: 12th brazilian symposium on artificial intelligence.* Springer.

Karr, A., Sanil, A., & Banks, D. (2006). Data quality: A statistical perspective. *Statistical Methodology*, *3*(2), 137–173.

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.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4).

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, *47*(3), 537-560.

Sentz, K., & Ferson, S. (2002). *Combination of evidence in dempster–shafer theory*. Sandia National Laboratories.

Shafer, G. (1992). Dempster-Shafer Theory. In S. C. Shapiro (Ed.), *Encyclopedia of artificial intelligence* (p. 330-331). Wiley.

UNECE. (2007). *Register based statistics in the Nordic countries. Review on best practices with focus on population and social statistics* (Tech. Rep.). United Nations Economic Commission for Europe.

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

Wallgren, A., & Wallgren, B. (2007). *Register–based statistics*. John Wiley & Sons, Ltd.

Yager, R. (1987). On the dempster-shafer framework and new combination rules. *Information sciences*, *41*(2), 93–137.

Zadeh, L. (1965). Fuzzy sets. *Information and control*, *8*(3), 338–353.

Zadeh, L. (1986). A simple view of the dempster-shafer theory of evidence and its implication for the rule of combination. *AI magazine*, *7*(2), 85.

Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, *27*(3), 415–432.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.

Appendix A

Translations

Table A.1: Translation of the German names for Raw-data

German Abbreviation	Full German Name	English Abbreviation	Full English Name
Demografie		**Demography**	
DEM_ALTER	Alter in Jahren	AGE	Age
DEM_FAMST	Familienstand	LMS	Legal Marital Status
DEM_GEBSTAAT	Geburtsland	POB	Country/place of birth
DEM_STAATB	Staatsangehörigkeit	COC	Country of citizenship
DEM_ZUWAN_1980	Jahr der Ankunft im Meldeland seit 1980	YAE	Year of arrival in the country since 1980
DEM_ZUWAN_2000	Jahr der Ankunft im Meldeland seit 2000	YAT	Year of arrival in the country since 2000
BEZ	Beziehungen zwischen den Haushaltsmitgliedern	RHM	Relation of household members
GKZ_HWS	Üblicher Aufenthaltsort	GEO	Place of usual residence
Ausbildung		**Education**	
EDU_HAB_ABFELD_NAT	Feld des höchsten abgeschlossenen Bildungsniveaus		Field of highest educational attainment
EDU_HAB_NAT	Höchstes abgeschlossenes Bildungsniveau	EDU	Educational attainment (highest completed level)
EDU_WLAU_ABFELD	Wichtigste laufende Ausbildung		Most important current education
Erwerbstätitigkeit		**Employment**	
ERW_GERINGF	Geringfügig		Minor employment
ERW_STATUS	Derzeitiger Erwerbsstatus	CAS	Current activity status
ERW_VTBESCH	Voll/Teilzeit		Full-Time or Part-Time
Arbeitsstätten und Unternehmungen		**Enterprises and their local units**	
KZA_GKZ	Gemeindekennziffer der Arbeitsstätte	LPW	Location place of work (local unit)
KZA_OENACE	ÖNACE des Unternehmens	IND	Industry (branch of economic activity)
KZA_SELBST	Anzahl der Selbstständigen an der Arbeitstätte		Self-employed at the local unit
KZA_UNSELBST	Anzahl Unselbstständige im Unternehmen		Employees at the the local unit
KZU_GKZ	ÖNACE des Unternehmens	IND	Industry (branch of economic activity)
KZU_RF	Rechtsform des Unternehmens		Legal form of the enterprise
KZU_SELBST	Anzahl der Selbstständigen im Unternehmen		Self-employed at the enterprise
KZU_UNSELBST	Anzahl Unselbstständige im Unternehmen		Employees at the enterprise
Wohnungen und Nutzungseinheiten		**Dwellings and other units of buildings**	
NTZ_ANZRAUM	Anzahl der Räume in der Wohnung	NOR	Number of rooms
NTZ_BAD	Vorhandensein eines Badezimmers oder einer Duschecke	BAT	Bathing facilities
NTZ_GKZ	Gemeindekennziffer	GEO	Geographical area
NTZ_NTZART	Typ der Nutzungseinheit		Type of Dwelling
NTZ-NTZFL	Nutzfläche	UFS	Useful floor space
NTZ-RECHTSVERH	Eigentumstyp der Nutzungseinheit	OWS	Legal form of the dwelling
NTZ_TOI	Vorhandensein eines WCs	TOI	Toilet facilities
Gebäude		**Buildings**	
OBJ_ANSGESCHO	Anzahl der oberirdischen Geschoße		Number of floors
OBJ_BAUP	Bauperiode	POC	Period of construction
OBJ_EIG	Gebäudeeigentümertyp	OWS	Type of ownership
OBJ_FL	Nutzfläche des Gebäudes	UFS	Useful floor space
OBJ_GKZ	Gemeindekennziffer	GEO	Geographical area
OBJ_STATUS	Gebäudestatus		Status of the building

Table A.2: Translation of the German names for CDB & FDP

Full German Name	Full English Name
Arbeitsstätten	**Local units of enterprises**
Gemeindekennziffer der Arbeitsstätte	Location place of work (local unit)
ÖNACE der Arbeitsstätte	Industry (branch of economic activity)
Anzahl Selbstständiger an der Arbeitsstätte	Self-employed at the local unit
Anzahl Unselbstständiger an der Arbeitstätte	Employees at the the local unit
Demografie	**Demography**
Alter in Jahren	Age
Gesetzlicher Familienstand	Legal marital status
Geburtsland	Country/place of birth
Geschlecht	Sex
Staatsangehörigkeit	Country of citizenship
Vorheriger üblicher Aufenthaltsort	Place of usual residence one year prior to the census
Jahr der Ankunft im Meldeland (seit 1980)	Year of arrival in the country since 1980
Jahr der Ankunft im Meldeland (seit 2000)	Year of arrival in the country since 2000
Feld des höchsten abgeschlossenen Bildungsniveaus	Field of highest educational attainment
Höchstes abgeschlossenes Bildungsniveau	Educational attainment (highest completed level)
Wichtigste laufende Ausbildung	Most important current education
Feld der wichtigsten laufenden Ausbildung	Field of the most important current education
Beschäftigung (Beruf)	Occupation
Geringfügig	Minor employment
Derzeitiger Erwerbsstatus	Current activity status
Stellung im Beruf	Status in employment
Voll/Teilzeit	Full-Time or Part-Time
Stellung in der Familie	Family status
Üblicher Aufenthaltsort	Place of usual residence
Größe des Ortes	Size of the locality
Stellung im Haushalt	Household status
Pendelentfernung	Distance of commute
Pendeltyp	Type of Commuting
Adresse der Bildungseinrichtung	Address of educational estabilishment
Familien	**Families**
Beziehungen zwischen den Haushaltsmitgliedern	Relation of household members
Größe der Kernfamilie	Size of family nucleus
Typ der Kernfamilie	Type of family nucleus
Haushalte	**Households**
Wohnbesitzverhältnisse der Haushalte	Tenure status of households
Größe des privaten Haushalts	Size of private household
Typ des privaten Haushalts	Type of private household
Unterbringungsformen	Housing arrangements
Wohnungen	**Dwellings**
Anzahl der Räume in der Wohnung	Number of rooms
Vorhandensein eines Badezimmers oder einer Duschecke	Bathing facilities
Art der Beheizung der Wohnung	Type of heating
Belagsdichte der Wohnung	Density standard (floor space)
Anzahl der Räume je Bewohner mit Hauptwohnsitz	Density standard (number of rooms)
Gemeindekennziffer der Nutzungseinheit	Geographical area
Zahl der Bewohner	Number of occupants
Typ der Nutzungseinheit	Type of Dwelling
Nutzfläche	Useful floor space
Eigentumstyp der Nutzungseinheit	Legal form of the dwelling
Art der Unterkunft	Type of living quarter
Vorhandensein eines WCs	Toilet facilities
Objekte	**Buildings**
Anzahl der oberirdischen Geschoße	Number of floors
Bauperiode	Period of construction
Art der Beheizung	Type of heating
Gebäudeeigentümertyp	Type of ownership
Nutzfläche des Gebäudes	Useful floor space
Gemeindekennziffer	Geographical area
Gebäudestatus	Status of the building
Gebäudetyp	Type of building
Unternehmen	**Enterprises**
Gemeindekennziffer des Unternehmens	Location place of work (enterprise)
ÖNACE des Unternehmens	Industry (branch of economic activity)
Rechtsform des Unternehmens	Legal form of the dwelling
Anzahl der Selbstständigen im Unternehmen	Self-employed at the enterprise
Anzahl Unselbstständige im Unternehmen	Employees at the enterprise