Quality Assessment of the Austrian register–based Census 2011

Combination of Evidence from Multiple Administrative Data Sources

Christopher Berka*, Stefan Humer*, Manuela Lenk**, Mathias Moser*, Henrik Rechta**, Eliane Schwerer**

Abstract: This paper investigates the quality of register data in the context of a standardised quality framework. The special focus of this work lies on the assessment of Census data and how to deal with uncertainty that arises from multiple sources (registers). To take the uncertainty associated with support and conflict between several registers into account, Dempster-Shafer's Theory of Evidence is applied. This 'fuzzy' approach allows us to investigate the quality of databases with multiple underlying sources for a single attribute and provide both quality measures and plausibility intervals.

Keywords: Dempster-Shafer Theory, Fuzzy Logic, Quality Assessment.

1 Introduction

The importance of register–based information for statistical analyses has increased strongly in recent years. It is obvious that the use of such already recorded information reduces the respondent's burden significantly, as opposed to survey data. Therefore utilizing these data sources becomes increasingly popular among NSIs. In addition, recent changes in legislation urge Statistics Austria to intensify the use of register data in order to minimize the costs of collecting data.

Accordingly Statistics Austria will rely on administrative data for the Austrian Census in 2011. In contrast to the Test–Census in 2006 the Census itself will not be accompanied by supplementary questionnaires. This shift from a traditional census (using surveys) towards a register–based one also challenges the process of quality measurement. For surveys a widely agreed methodology exists in terms of quality assessment, such as sampling errors. In contrast to surveys, there is no such toolkit for the evaluation of register–based data.

As a consequence various methods have been established to assess the quality of this kind of data. First and foremost it is common to double–check register data by issuing questionnaires in order to measure and enhance the quality of administrative records. As outlined above, for Austria a Test–Census in 2006 was accompanied by such a evaluation survey which provided valuable information on the quality of Austrian registers (see Lenk, 2008). However, an exhaustive examination of registers through surveys is a long–lasting

*Vienna University of Economics and Business (WU), Augasse 2-6, 1090 Vienna, Austria.
**Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. *manuela.lenk@statistik.gv.at*

process and only feasible for countries which have a long lasting experience in the usage of registers and strongly depend on them (see e.g. Hokka & Nieminen, 2008). Since the use of administrative records for statistical purposes is a relatively new approach in Austria, a different methodology for the quality measurement is needed.

In general the quality assessment of register data has to fulfill several properties, e.g. transparency, accuracy and feasibility. To achieve these goals we set up a general framework, which allows to evaluate the quality of registers with regard to all the information available. This also includes metadata, which is an enhanced approach apposed to classical data evaluation methods. With respect to census applications the assessment is a three step process: raw data (Registers), combined dataset (Census Database) and imputed dataset (Final Database). In this context we refer to databases as registers which are created by rulesets based on the rawdata. Therefore they are not administrative sources themselves but merely derived from administrative register information. At each of these levels changes in quality are monitored in order to give quality measures for all attributes in all registers and databases.

The quality information on the rawdata level is derived from three hyperdimensions: Documentation (HD^D), Pre–processing (HD^P) and External Source (HD^E). HD^D includes all quality related aspects prior to seeing the data and condenses them to the degree of confidence we put in the data owner. These are for example plausibility checks, data collection methods or legal enforcements of data recording. The second aspect on the rawdata level focuses on the data pre–processing (HD^P) at Statistics Austria. It measures the formal correctness by checking for implausibilities, errors or missing items in the data. In a third step we investigate the accuracy of the data by comparing it to an external source (HD^E). This is primarily done using existing surveys (i.e. the Austrian Microcensus). If an attribute is not found in the Microcensus we rely on expert opinions.

In a next step these three hyperdimensions can be combined to form a quality indicator per attribute and register which comprises all available information. Furthermore we can use this quality measures to assess the quality of databases, such as the Census Database. Most important, this process is independent from data processing to avoid endogeneity problems. Further details on the general assessment and process flow can be found in Berka et al., 2010.

If we focus on the rating of the Census Database (CDB or Ψ), the census itself can be seen as a process which assigns several socio–economic attributes to every relevant person. In case of multiple sources this is realised based on a predefined ruleset, which picks the most appropriate information from the underlying registers. To measure the quality of this process, it is necessary to take the support or conflict that arises when we compare multiple sources with the Census Database into account. Followingly we compare the values that the ruleset chooses for each person in the database with the entries in the rawdata registers. If the entries match, we consider the ruleset to be 'good' and the quality indicator will be high. In the terminology of our framework this is equivalent to the degree of confidence we have in the Census Database. The sources that are used for comparison are the single registers on the rawdata level (see figure 1, left–hand side).

When we compare the CDB to the rawdata registers we can basically distinguish three cases: a) a single comparison register is available (see figure 1, attribute C), b) multiple registers to compare with (see figure 1, attribute A) and c) no rawdata registers with a

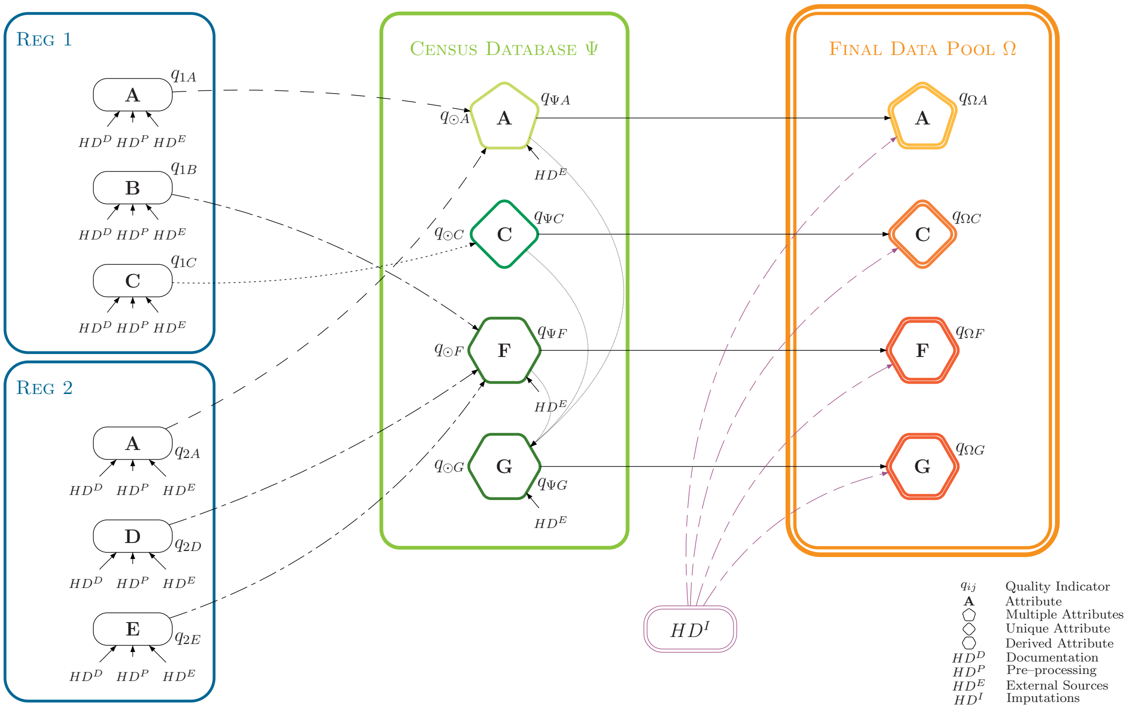

Figure 1: Quality Framework

similar attribute are disposable (see figure 1, attributes F and G). Case a) is trivial to assess since the confidence we put in the CDB is simply q_{ij}, where q_{ij} is the quality indicator for the specific attribute j in register i. Case c) is more complicated to deal with, since the attribute in the CDB is derived from other attributes. Therefore we can not use the quality indicators of the base registers directly. To create a consistent and comparable quality measure it is necessary to take the amount of information flow from base attributes to the derived attribute into account. This is subject to further research.

In this paper we focus on the assessment of case b). Applying weighted averages is problematic since in this standard approach we ignore the uncertainty that is associated with conflict among registers. More specifically a register that disagrees with the CDB does not necessarily imply that the latter is wrong. However it may express a degree of uncertainty on the value of the database. Subsequently we will apply a specific form of this fuzzy logic, namely the Dempster-Shafer Theory. In this paper we focus on the application of this theory on the register-based census. However, since the framework was composed for generic process flows this is only one specific application area.

In the next section we will give a brief overview on fuzzy logic and fuzzy sets in general. Furthermore we provide more detailed information on the Dempster-Shafer Theory and give calculation examples. Section 3 introduces Dempster-Shafer theory in our quality framework and outlines why and how it is applied. Finally section 4 derives preliminary quality measures for selected attributes of the Census.

2 Probability and Uncertainty

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 (see Helton, 1997):

Stochastic Uncertainty results from the fact that systems can behave in different ways. 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.

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 consensus 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 function exists 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 an hospital. Some doctors may have different beliefs about the true reason for the sickness. One doctor might consider a malfunction of the liver as being the reason and has a degree of belief of 90%. It could be that an other doctor thinks of some other reasons than a malfunction of the liver and therefore beliefs that the likelihood of the malfunction being the true reason for the illness is just 7%. An easy way 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).

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 (see Dempster, 1967). It can be regarded as a generalization of Bayes' rule and conflates multiple belief functions by aggregating their *basic probability assignment* functions. Dempsters Rule is a strictly conjunctive procedure and accents agreement of multiple sources of information.

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

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

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

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.

3 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 origin 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 (case b), see Introduction). 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 with $n \leq i$) 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.

However, it could be that a lot of register contain information about the same attribute. If the numbers of registers become 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 for each case (i.e. from REG_1 to REG_n) a look-up table that 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 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.

$$Bel = \frac{1 - \lightning - \omega - U^n}{1 - \lightning} \qquad \xi = \frac{U^n}{1 - \lightning}$$

Symbol	Name
Bel_{REG}	Degree of belief
ξ	Normalisation of uncertainty
$\lightning$	Logical impossibility
ω	Conflict
U^n	Uncertainty

This general equation is applied to each observation for each attribute. The CDB marks the actual belief for a certain observation and therefore defines which belief is calculated. 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 moment 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 deliver a more sophisticated picture of the quality assessment for an attribute in the CDB.

4 Results

Table 1: 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 1). 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 are no specific rational behind this weighting. One could apply sensitivity analyses to get an idea of the impact of the weights. However this will be of interest if more quality indicators for a greater number of attributes are available and is thus subjected to future research. For details on the calculations of these indicators we refer to (Berka et al., 2010).

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 2.

Table 2: Possible Combinations of Certainty and Uncertainty for four Registers

Constellation	*bpa*	Constellation	*bpa*
UUUU	0.10591	CUUU	0.00001
UUUC	0.00001	CUUC	0.00014
UUCU	0.00174	CUCU	0.01773
UUCC	0.02240	CUCC	0.22889
UCUU	0.28500	CCUU	0.00003
UCUC	0.00004	CCUC	0.00038
UCCU	0.00467	CCCU	0.04770
UCCC	0.06029	CCCC	0.61597

Following this short example we will present first results of the application of the Dempster-Shafer theorem on the quality framework for the Austrian census. We will again focus on the attribute sex for reasons of simplifications. Table 3 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.

Table 3: Results of the Dempster-Shafer Application for the attribute sex

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

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 indicators are very high. For the 5% quantile the quality indicator is already very close to 1. This concentration is also supported by a very low standard deviation (σ). 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.

5 Conclusion

This paper investigated the application of the Dempster–Shafer theory on the quality assessment of register data. With special focus on the Austrian Census in 2011 we applied this fuzzy logic approach to evaluate rulesets based on register information. In this context we calculated the beliefs that the ruleset's choice for an attribute is the *true* value based on the underlying data.

This was done for the case of four registers which have different opinions over the true value of the attribute sex. The degree of belief we put in each register is derived from several hyperdimensions which try to condense all the available quality–related information in each register. Using these indicators we derive all possible combinations of certainty and uncertainty between the four base registers. In a next step we use combinations rules from the Dempster–Shafer framework to aggregate this (un-)certainty values.

We found that the support for the values chosen in the CDB is relatively high at 0.9987 for the attribute sex. From this follows that the ruleset (on which the CDB is created) is rather good. Accordingly for this attribute it assigns the most probable values to each statistical unit, according to the base registers.

Further research involves the investigation of case c) (see Introduction), where the CDB attribute is derived from various other variables. In this case the raw register do not have an opinion on the attribute in the CDB itself. They only provide quality indicators for a part of the information that enters the according attribute in the CDB.

References

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

Census Team. (2010, 02). *Draft commissin regulation on quality reporting for the 2011 censuses of population and housing: Some aspects of the reporting on the coverage and on the imputation and deletion of data records.* (Eurostat, F1)

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).

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.

Dempster, A. (2008). Upper and lower probabilities induced by a multivalued mapping. *Classic Works of the Dempster-Shafer Theory of Belief Functions*, 57–72.

Eurostat. (2003a). Item 4.2: Methodological Documents - Definition of Quality in Statistics. In *Working group assessment of quality in statistics.*

Eurostat. (2003b). Quality assessment of administrative data for statistical purposes. In *Assessment of quality in statistics.*

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.

Hokka, P., & Nieminen, M. (2008). Measuring the Quality of the Finnish Population Register with a Survey. Special Focus on Non–Response. In *European conference on quality in official statistics.* Eurostat.

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.

Klir, G., & Wierman, M. (1998). *Uncertainty-based information: Elements of generalized information theory*. Physica-Verlag, Heidelberg.

Lenk, M. (2008). *Methods of register-based census in austria* (Tech. Rep.). Statistics Austria, Vienna.

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

Sevastianov, P., & Dymova, L. (2009). Synthesis of fuzzy logic and dempster-shafer theory for the simulation of the decision-making process in stock trading systems. *Mathematics and Computers in Simulation*, *80*(3), 506–521.

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

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.