A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011

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

Published in:
Austrian Journal of Statistics

Published: 01/01/2010

Document Version:
Publisher's PDF, also known as Version of record

AUSTRIAN JOURNAL OF STATISTICS
Volume 39 (2010), Number 4, 299–308

A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011

Christopher Berka[1], Stefan Humer[1], Manuela Lenk[2], Mathias Moser[1], Henrik Rechta[2], Eliane Schwerer[2]
[1]Vienna University of Economics and Business, Vienna
[2]Statistics Austria, Vienna

Abstract: Along with the implementation of a register-based census we develop a methodological framework to assess administrative data sources for statistical use. Key aspects for the quality of these data are identified in the context of hyperdimensions and embedded into a process flow. Based on this approach we develop a structural quality framework and suggest a concept for quality assessment and several quality measures.

Zusammenfassung: Mit speziellem Fokus auf die Registerzählung 2011 in Österreich wird ein Framework zur Qualitätsbewertung von Registerdaten erstellt. Dabei werden jene Faktoren untersucht, die potentiell die Qualität von Administrativdaten beeinflussen können. Mit Hilfe von Maßzahlen werden Verfahren vorgestellt, die zur Bewertung von Registern, dem Datenaufbereitungsprozess und dem authentischen Datenbestand verwendet werden können. In diesem Beitrag werden die Inhalte dieses Frameworks vorgestellt.

Keywords: Administrative data, Register-based Census, Quality assessment.

1 Introduction

Administrative records have become more important for statistical analyses in statistical institutions and social sciences in recent years. The use of administrative data sources has a long tradition in Scandinavian countries and is applied extensively for statistical purposes, one of which is a register-based census. These data have several advantages over standard surveys (Bruhn, 2001). First of all, administrative data deliver more information on a certain part of the population than most survey data. Furthermore, they are already recorded and reduce the statistical burden of respondents significantly. On the contrary, the quality of administrative data heavily depends on the data-source keeper. In general the national statistical institutions (NSI) have little information on the accuracy and reliability of these data.

Since Austria, amongst other countries, will carry out its first register-based census in 2011, it is central to assess administrative registers and to evaluate their quality prior to this task.

However, there is few literature which deals with quality assessment of administrative data sources. The approach of Finland focuses on the comparison of administrative and survey data (Ruotsalainen, 2008). Other countries, such as the Netherlands take a more structural approach. Their aim is to cover the quality of different registers in a quality framework using different dimensions to assess data quality and accuracy. Daas, Ossen,

Vis-Visschers, and Arends-Tóth (2009) developed a checklist for the quality evaluation of administrative data sources which is structured in three different hyperdimensions of quality aspects.

Our approach contributes a framework for administrative data to the field of quality research, which evaluates these records structurally. It is an extension of the model suggested by Daas et al. (2009), who were the first to outline such a process in a comprehensive way. This allows both the NSI and external researchers to assess the data sources they use.

2 Data Sources

For the register-based census seven base registers and several comparison registers are merged. The base registers determine the attribute totals like the number of buildings and dwellings, the number of enterprises or the number of persons with main residence in Austria. They also provide additional information on the core topics needed for the census. The 'backbones' of the census are the Central Population Register (CPR) and the Central Social Security Register (CSSR). Other base registers are the Tax Register (TR), the Unemployment Register (UR), the Register of Educational Attainment (EAR), Business Register of Enterprises including their local units (BR) and Housing Register of buildings and dwellings (HR). All these registers can be linked with unique keys.

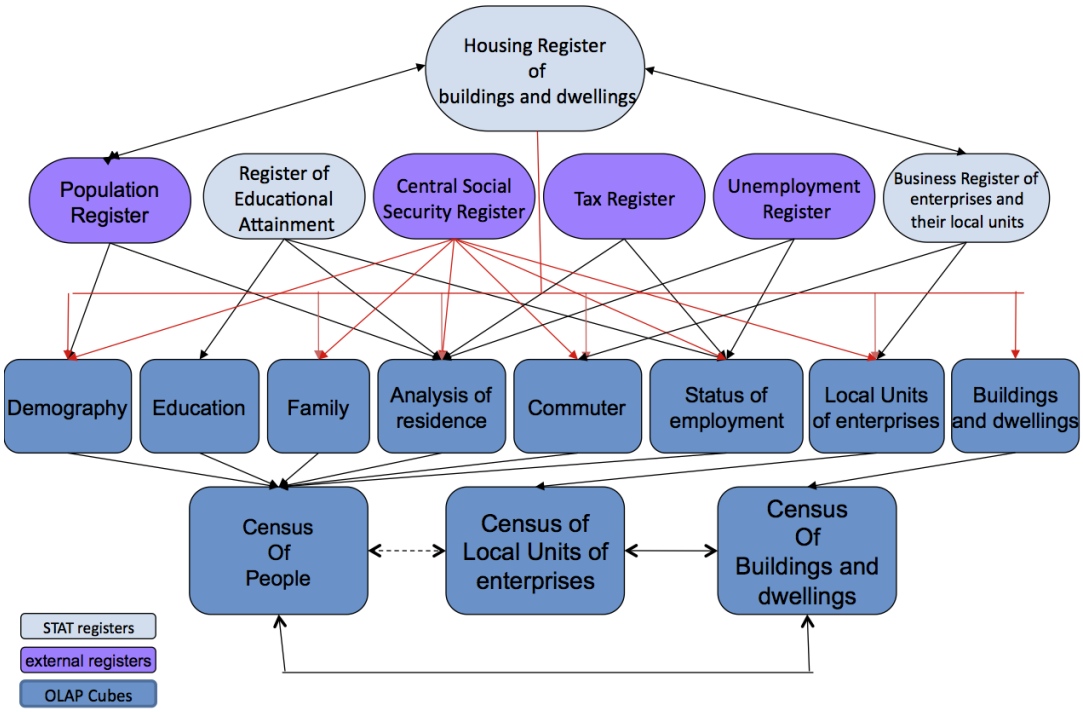

Figure 1: Registers and Topics

The comparison registers are mainly used for cross checks and to add information not or only partly included in the base registers. Due to the fact that these registers are

recorded independently the attribute's values can differ. Therefore we use the principle of redundancy, i.e. collect all attribute information from different registers and check for inconsistencies.

Figure 1 gives an overview of all fields of analysis. The top of the figure gives an overview of the base registers and their connection to different census fields. These eight main fields are the base for three final cubes which represent the Census of People, the Census of Local Units and the Census of Buildings and Dwellings.

3 Quality Framework for the Census

Statistical data quality can be covered by several dimensions (see Eurostat, 2003a). This also applies to administrative data as has been stressed by Eurostat (2003b). In the framework we focus on data accuracy, since this is the most challenging dimension. Moreover, accuracy is essential for the quality of the register-based census and is at the same time a major unknown property of register data. Quantification of data accuracy is realised by a framework which is closely tied to the data flow yet independent from data processing. This is necessary since results of the quality assessment must not influence but evaluate the processing procedure. Whether low quality ratings lead to a revision of the data processing steps has to be determined for each statistical application independently. Experience from the test census suggests that this is not a major concern for the Austrian Census, since data quality is expected to be fairly high.

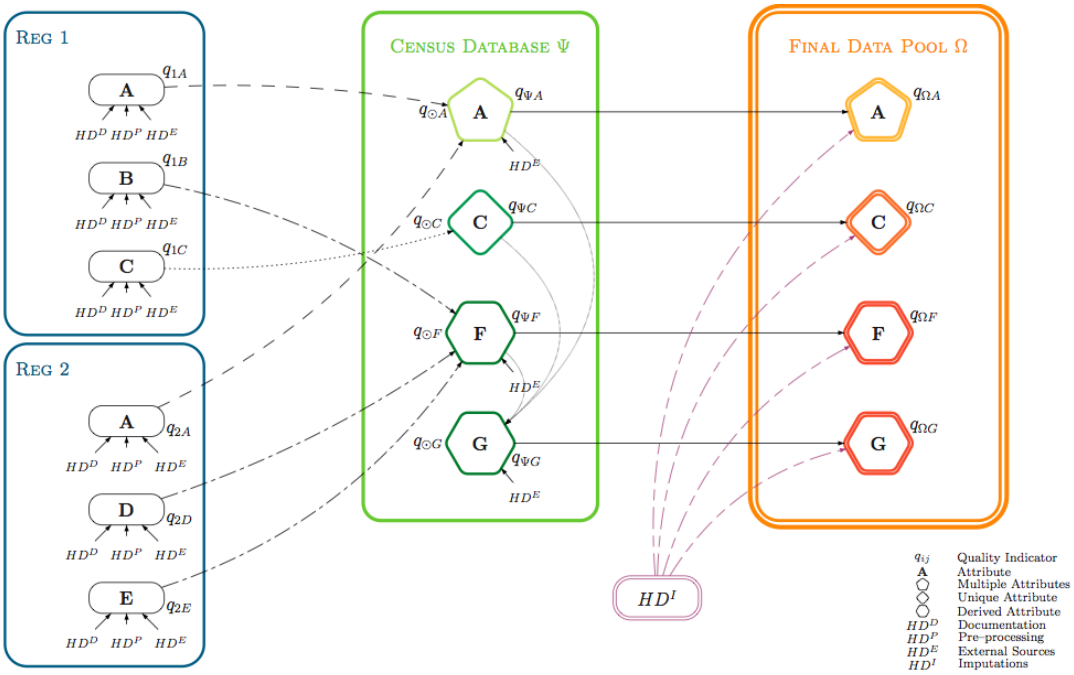

Figure 2: Quality Framework

The quality framework results in exactly one accuracy quality indicator for each attribute in each register or data pool. It is linked to the data flow on three different levels.

1. Raw data (registers)
2. Census Databases (CDB)
3. Final Databases (FDB)

In a first step Statistics Austria receives the raw data (henceforth *registers*, see boxes on the left-hand side in Figure 2). Secondly these different sources are combined to data cubes, the Census Databases (CDB), by using unique IDs only. These cubes only include information available from the registers (raw data). Finally we enrich the CDB with imputations of item non-response. These steps result in Final Databases (FDB), which consist of both real and estimated values. In each of these three steps (Register, CDB and FDB) the data flow is linked to the quality assessment, so that changes can be monitored from a quality perspective.

4 Assessment of Registers

The quality assessment starts on the administrative register level. The quality measure q_{ij} for an attribute is defined as a value between 0 and 1, the latter being the highest possible value. This measure is derived from a set of sources (*hyperdimensions*, HD) which should cover all quality information available for this attribute. Every hyperdimension delivers one (aggregated) quality indicator for each attribute. In order to combine these different quality aspects we use weights (v) for each hyperdimension, which accordingly sum up to 1 (see Figure 3). The weighted average of the hyperdimensions are the quality indicators q_{ij} per register and attribute (see equation (1))

$$q_{ij} = v^D \cdot hd_{ij}^D + v^P \cdot hd_{ij}^P + v^E \cdot hd_{ij}^E = \sum_{k \in D,P,E} v^k \cdot hd_{ij}^k \,. \tag{1}$$

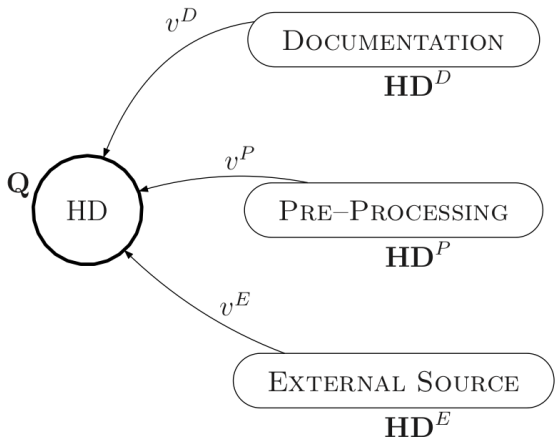

Figure 3: Framework for Quality Assessment of Registers

The three hyperdimensions on raw data level are *Documentation* (HD^D), *Pre-processing* (HD^P) and *External Source* (HD^E), as shown by Figure 3. This differentiation assures

that all available information on different quality aspects is utilised. Every single hyperdimension gives a quality measure which is then weighted and combined to form an overall quality measure for each attribute in each register.

4.1 Documentation (HD^D)

This hyperdimension describes quality-related processes at the register authority as well as the documentation of the data (metadata). In other words, the reliability of the data owner is checked. For this purpose we set up a questionnaire, which contains 16 open-ended and nine scored questions. Open questions give the possibility to record information that might be necessary for data users in general, e.g. which attributes are included in the data set. Additionally scored questions provide specific quality-related information (see Table 1).

Table 1: Scored Questions – HD Documentation

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

The questionnaire is filled in for each attribute in each register. This task is carried out by experts of the NSI who have intense knowledge of the data and register authority. All the answers are weighted according to their importance for the census (accuracy). The relative importance of the different questions is rated by experts of the NSI. In a next step the sum of scores for each attribute is compared with the theoretical maximum score. This comparison results in assessment of the attribute's quality at the register authority, which is hd_{ij}^D.

4.2 Pre-processing (HD^P)

The second hyperdimension is concerned with formal errors in the raw data. Thus it checks for definition and range errors, as well as missing primary keys and item non-response. The end result from the HD *Pre-processing* is given by the ratio of usable records to the total number of records. Again, this procedure is carried out for each

Table 2: HD Pre-processing

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

attribute in each register. The quality measure hd_{ij}^{P} shows the proportion of usable records for each attribute in each data source. If the proportion of useable records for an attribute in a certain register is smaller than that of the same attribute within another register, the quality measure will accordingly be lower.

4.3 External Source (HD^{E})

In a last step the registers are checked against a benchmark. Common benchmarks are representative surveys as for example the Austrian Microcensus (see also Hokka and Nieminen, 2008). However, such a benchmark does not exist for a number of attributes. In this case we use local expert opinion on the data source (expert interview).

Microcensus The Austrian Microcensus is a rotating panel and includes survey information on population subsamples. Accordingly it can be used to benchmark the accuracy of registers. Despite the shortcomings of surveys in general, the Microcensus is the best source for comparison available. We assume that the Microcensus provides the true values on the unit level which implies that the marginal distributions for the corresponding attributes in the Microcensus are correct. The errors from this assumption seem to be neglectable on higher aggregation levels. Furthermore we have to account for differing reference dates between the Microcensus and the registers. For rather static variables (e.g. sex, date of birth) this is not an issue. However other attributes are a subject to frequent change (e.g. employment status) which means that the Microcensus will be restricted to the relevant periods. The quality framework distinguishes two different methods to detect data deviation, using the Microcensus.

One way is to use the Microcensus as a sample and to check for errors on the item level. Since the data include unique identifiers, we are able to link the register data to the Microcensus. Therefore the accuracy of the attributes for each administrative source can be evaluated, if the corresponding unit can be found in the Microcensus. An example for this procedure is given in Table 3. For the attribute *Marital Status* we assume that 5000 units could hypothetically be linked from a register to the Microcensus. In a next step we check whether the value of the attribute is the same in both sources. In the case of the marital status 'single', 17 persons are single according to the register but not to the Microcensus. Applying this procedure for each attribute value (single, married, divorced, widowed) we can measure the percentage of agreement between register and Microcensus.

Table 3: Microcensus Benchmarking – marital status

	Single	Married	Divorced	Widowed	
Error	17	43	24	16	100
$hd^E_{i,MaritalStatus}$					0.98

In this completely virtual example 100 errors (2%) are identified. This would indicate that 98% of the value's of the register's attribute *marital status* are correct. Another approach is to evaluate whether the grossed up marginal distribution of the observed population for a certain attribute is the same in both the administrative source (i.e. the register, 'REG') and the external source (Microcensus, 'MC'). This comparison is illustrated by Table 4.

Table 4: Comparison of Marginal Distribution – marital status

	Single	Married	Divorced	Widowed
MC	0.43	0.30	0.23	0.04
REG	0.46	0.31	0.19	0.03
REG/MZ	1.08	1.05	0.83	0.94

Expert Interview Some register attributes are not included in the external source. Consequently it is not possible to benchmark them, e.g. by using the Microcensus. In this case the external source comparison is replaced by an expert interview. For such attributes the expert is asked on his/her assessment regarding the data accuracy of the corresponding attribute. This approach is very subjective and therefore only carried out if no external data source for the benchmark approach is available. To account for the subjective nature of such an expert interview, the weight for HD^E will be adapted.

5 Linking of Registers (CDB)

The CDB includes every attribute needed for the census and is created using a fixed ruleset based on the registers, which we evaluated in chapter 4. To rate the CDB, it is necessary to distinguish three types of attributes.

- Unique Attributes: An attribute exists in exactly one register and is taken from the Census Database, e.g. type of heating in a dwelling (see attribute C in Figure 2).
- Multiple Attributes: An attribute shows up in several registers. The various information from these sources are combined (e.g. majority principle) in order to derive a valid attribute for the Census Database, e.g. sex (see attribute A in Figure 2).
- Derived Attributes: A new attribute in the Census Database is created based on data from different attributes, e.g. household status (see attribute F and G in Figure 2).

For **unique attributes** the quality assessment is simple. If attribute C can take the values 0 or 1 and is only available in one register, REG1, it will be directly tansferred to the CDB. Suppose attribute C has a quality indicator of $q_{1C} = 0.97$ in the origin register (which is a belief, according to the Dempster-Shafer Theory, see e.g. Shafer, 1992). Since the attribute is the same in REG1 as well as in the CDB, we know that the attribute's quality will be the same in both sources, hence $q_{1C} = q_{\odot C} = q_{\Psi C}$.

Table 5: Linking of CDB – Unique Attributes

PIN	REG 1 ($q_{11} = 0.97$)	Ψ	$q_{\odot C} = q_{\Psi C}$
⋮	⋮	⋮	⋮
ID3456	0	0	0.97
ID3457	0	0	0.97
ID3458	1	1	0.97
ID3459	0	0	0.97
ID3460	1	1	0.97
ID3461	1	1	0.97
⋮	⋮	⋮	⋮
			$\mu = 0.97$

In the case of **multiple attributes** the procedure mentioned below is one option out of a set of alternatives (e.g. Bayesian approaches). Furthermore it applies only to discrete attributes, which is not a problem for the census. To illustrate this approach for multiple attributes, suppose attribute A could take the values 0 or 1 and it existed in two registers, REG1 and REG2. On the data level attribute A will be included in the CDB based on a fixed ruleset. On the quality level we know that the quality of REG1 and REG2 is $q_{1A} = 0.99$ and $q_{2A} = 0.8$ respectively. As a result of this we know the plausibility of the values in the CDB, depending on whether the two registers agree or disagree. For ID3457 in Table 6 this process can be formally written as $(0.99 + (1 - 0.8))/2 = 0.595$ because Ψ uses the value "1", which is correct according to REG 1 & 2 with a probability of 0.99 and (1 - 0.8) respectively.

Derived attributes need more complex procedures. If the derived attribute is based on separate registers, which do not overlap each other, the procedures illustrated above are applied. The only difference is that we choose different q_{ij} for the values, depending on their origin registers. However, several derived attributes are created independently based on other information and characteristics. For example the attributes sex, marital status and date of birth are used to create the attribute household status. Since these rulesets can become rather extensive, we need to weight quality indicators based on the ruleset itself, i.e. based on the rule of combination on the data level. The procedure for this will be experimentally developed. Furthermore it is necessary to assess the errors of the derivation process itself. For this type of attribute, it can be helpful to check the validity with an external source (HD^E, see subsection 4.3).

Finally we end up with exactly one quality indicator for each attribute in the Census Database, which is based on all the information available from the used registers. So

Table 6: Linking of CDB – Multiple Attributes

PIN	REG 1 ($q_{11} = 0.99$)	REG 2 ($q_{21} = 0.8$)	Ψ	$q_{\odot A}$
⋮	⋮	⋮	⋮	⋮
ID3456	0	0	0	0.895
ID3457	1	0	1	0.595
ID3458	0	1	1	0.405
ID3459	1	1	1	0.895
ID3460	1	1	0	0.105
ID3461	0	0	0	0.895
⋮	⋮	⋮	⋮	⋮
				$\mu(q_{\odot A})$

far this only includes real data entries, while the CDB still faces the problem of item non-response.

6 Conclusion and next Steps

In this paper we presented a structural approach for the quality assessment of administrative data. Taking a special focus on the Austrian census in 2011, we distinguish between three stages of data processing. These are the raw data level, the linked dataset (CDB) and the linked imputed dataset (FDB). Each of these steps is linked to the quality assessment. For the raw data (registers) we check the reliability of the register authority (HD Documentation), the formal correctness (HD Pre-processing) and the accuracy with respect to data consistency (HD External Source). The CDB is evaluated based on the merging procedures and the quality measures we derived on the raw data level. The quality framework can be applied to other statistical projects and is therefore also of use to external scientific researchers.

Our current research focuses on the last step, the Final Database, which will be the data pool used for the census. The FDB corresponds to the CDB after imputations are applied. Therefore the amount of item non-response is effectively reduced. Since we already know the quality indicators per (non-imputed) attribute from chapter 5 we only need to account for the imputation process itself (see Figure 2). This is realised by using information from the hyperdimension Imputation (HD^I). We use weighted averages for the combination of the quality indicator of the CDB and the quality indicator of HD^I, that take the proportion of imputed items per attribute into account. The evaluation of different imputation methods and a suitable scoring for these are ongoing research.

References

Bruhn, A. (2001). *The next population and housing census in sweden is planned for 2005 - it will be totally register-based* (Tech. Rep.). Statistics Sweden.

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

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.*

Hokka, P., and 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.

Ruotsalainen, K. (2008). *Finnish register-based census system* (Tech. Rep.). Statistics Finnland.

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

Authors' addresses:

Christopher Berka, Stefan Humer, Mathias Moser
Department of Economics
Vienna University of Economics and Business
Augasse 2-6
A-1090 Vienna

Manuela Lenk, Henrik Rechta, Eliane Schwerer
Directorate Population Statistics
Statistics Austria
Guglgasse 13
A-1110 Vienna