QUALITY INDICATORS FOR THE INDIVIDUAL LEVEL – POTENTIAL FOR THE ASSESSMENT OF SUBGROUPS

E.-M. Asamer[1], H. Rechta[2], C. Waldner[3]

[1] *Statistics Austria, Vienna, Austria; eva-maria.asamer@statistik.gv.at*
[2] *Statistics Austria, Vienna, Austria; henrik.rechta@statistik.gv.at*
[3] *Statistics Austria, Vienna, Austria; christoph.waldner@statistik.gv.at*

Abstract
In the scope of the transformation to a register-based census 2011 in Austria, a quality framework for statistical data based on administrative sources was developed. Now, this quality framework is used annually to evaluate the quality of the register-based labour market statistics. These quality indicators offer a wide range of possibilities to analyse the attributes on their own but also in combinations. This paper gives a short overview of the quality framework and its three stages (raw data level, census data base, final data pool) as well as the different types of attributes (simple, multiple, derived). In the second part of the paper we present an approach for analysing different subgroups, thus showing the full potential of the quality framework. Crossing specific values of an attribute with other attributes or source registers, offers the possibility to analyse strengths and weaknesses of register-based statistics for these subgroups. Then the approach is applied on real examples from the register based labour market statistics 2013.

Keywords: Register based Statistics, Quality Indicators, Quality of Census

1. Introduction

At the reference day 31 October 2011, Statistics Austria carried out Austria's first register-based census. This means that the population and housing census was conducted by using administrative data sources. Such register-based statistics have a long tradition in the Nordic countries (see United Nations, 2007) and hold several advantages in comparison to classical surveys. For example, such a procedure is very cost efficient and there is no respondent burden anymore. To assess the quality of the census, Statistics Austria developed a general

quality framework for statistical data based on administrative sources (see Lenk, 2012). Now this measure is used to evaluate the annual register-based labour market statistics. The aim of this paper is to describe certain topics of the labour market statistics 2013 in a quality-related view. As preparation, we briefly recall in Section 2 the main parts of the quality framework. In Section 3, we focus on the quality results – mainly for multiple attributes. In general, the overall quality in the labour market statistics is very well, so we investigate critical subgroups of around 400 people and below. Investigations on selected questions (e.g. quality in relation to the place of usual residence) lead to disclose hidden weaknesses. In Section 4, we finish with a short discussion on the quality assessment and some closing remarks.

2. Data sources and quality assessment

Appropriate data sources are crucial for register-based statistics. For the register-based labour market statistics, we use 7 base registers as the backbone and 8 comparison registers from more than 50 data holders. 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 the quality of data (see Lenk, 2009). The data owners are responsible for the data maintenances of their data bases.
Hence, the importance of quality assessment for register-based statistics has to be emphasised. The data processing for the Austrian register-based labour market statistic is divided into three levels that have to be considered in the quality assessment: the raw data (i.e. the registers i), the combined dataset (central data base CDB) and the final 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 1 illustrates the data processing, beginning with the raw data from the various administrative data holders. The individual data lines are matched via a unique personal key (branch-specific personal identification number for official statistics bPIN OS) and merged to data cubes in the CDB. Finally, missing values in the CDB are imputed in the FDP where every attribute for every statistical unit obtains a certain quality indicator.

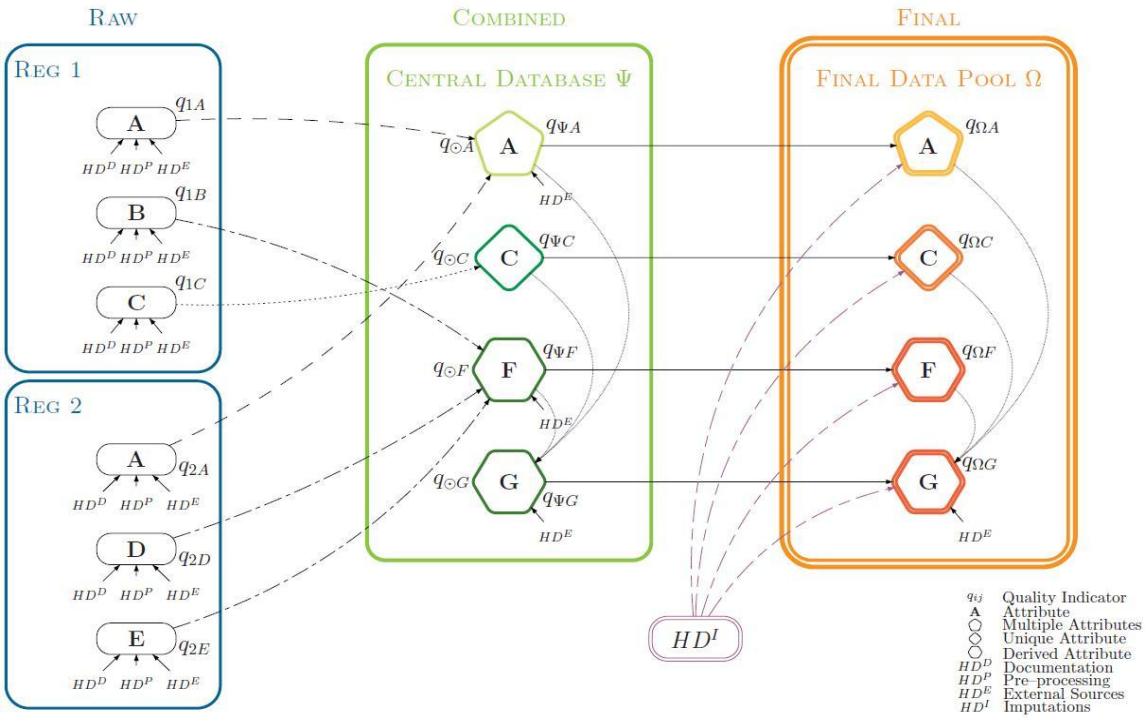

Figure 1: Quality framework for the register-based labour market statistics

2.1. The Raw Data Level

At the raw data level (blue boxes in Fig. 1) we assign to each attribute in each data source three hyperdimensions: Documentation (HD^D), Pre-processing (HD^P) and External Source (HD^E). HD^D describes quality-related processes as well as the documentation of the data (metadata) for the administrative authorities. HD^P is based on the share of usable records on all records. HD^E for raw-data level assesses the data-quality of the source registers in comparison to an external source, in our case, the Austrian microcensus. Given these three quality measures, an overall quality indicator q_{ij} for each attribute *j* in a register *i* on register-level can be derived as an average value.

2.2. The Central Data Base CDB

The entire information from the registers is combined in the Central Database (CDB, green box in Fig. 1) which covers all attributes of statistical interest. 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 Fig. 1). For this reason, the measure of quality in the CDB is the same as in the raw data.

Multiple attributes show up in several registers, e.g. legal marital status (cf. attribute A in Fig 1). Since there are multiple data sources providing a certain attribute, a predefined ruleset picks the most appropriate value for the CDB according to the constellation in the source registers. To assess the validity of this chosen value, all the available information is taken into account. The Dempster-Shafer Theory (DST) for the combination of evidence is applied to derive a quality measure for these attributes for each statistical unit.

The quality measures on the raw data level are considered as beliefs in the correctness of the value. DST for the combination of evidence takes into account all available evidence from the registers to form a quality-indicator $q_{\odot}^n$ on the CDB-level for each statistical unit *n*.

Derived attributes are based on different attributes, e.g. SIE (Status in employment) (cf. attributes F and G in Fig. 1). The registers do not contain any information for these attributes in the required specification, but related information. The quality measure $q_{\odot}^n$ for each statistical unit *n* is the average of the qualities of the input attributes regarding *n*.

A (optional) further comparing to an external source HD^E yields the last CDB-quality indicator q_{Ψ}^n . Note that the quality indicator for a missing value is set to zero.

2.3. The Final Data Pool FDP

The data generation process is completed after the imputation of missing values in the CDB. The result is the FDP. To assess the quality q_{Ω} of the FDP a fourth Hyperdimension HD^I is

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

computed. It is based on the quality of the input and the quality of the imputation model. For a detailed explanation of the quality assessments see Schnetzer et al. (2015).

3. Results

The general results of the quality assessment for the Austrian Census 2011 as well as for the register-based labour market statistics are available on www.statistik.at[3], more precisely they are part of the Methodeninventar §14. There are results for each (unique, multiple) attribute on raw data level and for each attribute on CDB and FDP level.

3.1 Quality indicators - A first look

Since the quality indicators are computed on the individual level we can evaluate the average quality by arbitrary selected attributes. A first impression can be obtained simply by choosing the multiple attributes COC, POB, AGE, LMS, SEX[4] grouped by the *place of usual residence* (GEO) on Laender level (federal province), see Table 1.

Table 1: Average quality for multiple attributes per place of usual residence

GEO	$\bar{q}_{\Omega\ \mathrm{AGE}}$	$\bar{q}_{\Omega\ \mathrm{SEX}}$	$\bar{q}_{\Omega\ \mathrm{LMS}}$	$\bar{q}_{\Omega\ \mathrm{COC}}$	$\bar{q}_{\Omega\ \mathrm{POB}}$
Austria	**0.999**	**1.000**	**0.952**	**0.991**	**0.991**
Burgenland	1.000	1.000	0.954	0.995	0.993
Carinthia	1.000	0.999	0.952	0.992	0.992
Lower Austria	1.000	1.000	0.955	0.993	0.989
Upper Austria	0.999	1.000	0.962	0.992	0.991
Salzburg	0.999	0.999	0.953	0.992	0.988
Styria	0.999	1.000	0.956	0.993	0.992
Tyrol	1.000	0.999	0.953	0.992	0.988
Vorarlberg	0.999	1.000	0.960	0.991	0.990
Vienna	0.999	1.000	0.937	0.986	0.993

[3]http://www.statistik.at/wcm/idc/idcplg?IdcService=GET_PDF_FILE&RevisionSelectionMethod=LatestReleased&dDocName=104555

[4] COC: country of citizenship, POB: place of birth, AGE: age, LMS: legal marital status, SEX: sex

For example one can see that the worst average quality $\bar{q}_{\Omega\,\mathrm{LMS}}$ for the LMS of all Laender is measured for Vienna. There are two main reasons for this:

1. There are more divorced people in Vienna (10.27%) than the Austrian average value (7.77%). Data of divorcees generally have a worse average quality (0,779) for many reasons (e.g. caused by statistic rules).
2. The central population register (CPR) is (beside the Tax register (TR)) the main data source for LMS. Therefore it is not surprising that the average quality per federal province correlates with the coverage in percent of the CPR (see Fig. 2).

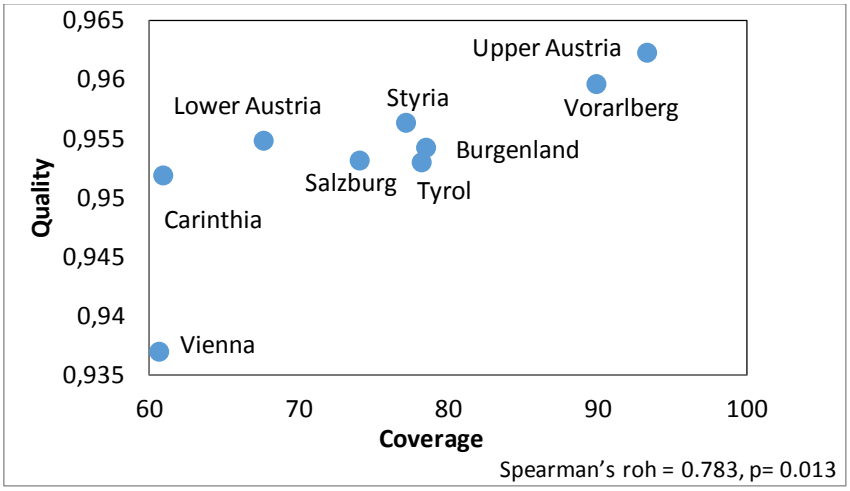

Figure 2: Average quality and coverage rate for LMS of the CPR per Laender

The CPR covers the LMS in Vienna only to 60.66%. Carinthia has a similar coverage rate, but all other Laender have much higher rates. This is caused by the fact that the attribute LMS has only been registered in the CPR since 2006. Thus, item-non response can be improved by migration. But much more important for the improvement is the maintenance by the municipalities.

The quality of LMS in Vienna is worse by comparison, but still very good. This is due to the fact that 96.17% of the population of Vienna has a LMS from at least one source - in other words, the *principle of redundancy* is crucial.

3.2 Redundancy and Quality

In general, the *principle of redundancy* (cf. Lenk, 2009) improves the quality for the multiple attributes (i.e. for AGE, SEX, LMS, POB and COC). Beside the number and the accuracy of the source registers, the information for a reference day plays a role for the output quality.

AGE is delivered from eleven registers. Since the birthday is invariant from the date of excerpt, the average quality grows by the number of sources (see Fig. 3).

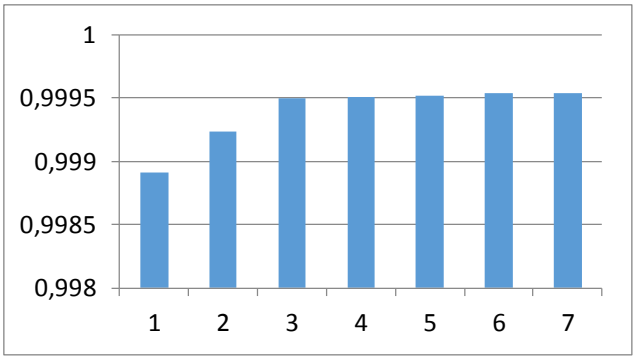

Figure 3: Average quality of AGE per the number of sources

Data for the LMS are obtained in potentially eleven source registers too. For many people, this attribute changes in the course of their lives. Hence, the accuracy (i.e. the up-to-dateness) of the sources and the date of database excerpt are crucial[5]. Therefore, the average quality decreases if there are more than two sources, as is shown in Fig. 4.

[5] Such considerations are involved in the raw data assessment HD^D by asking a questionnaire to the data holders containing questions like "Is the information available for the cut-off date?" or "Is the attribute relevant for the data source keeper?".

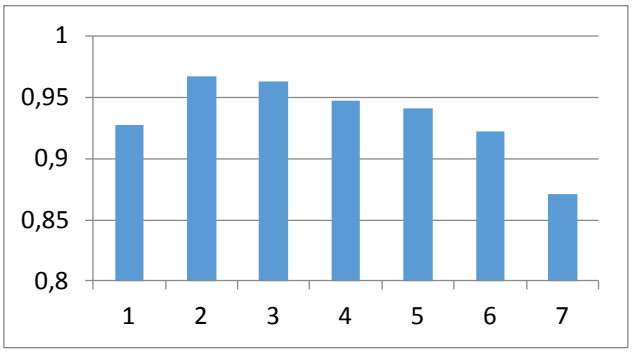

Figure 4: Average quality of LMS per number of sources

As we have seen above, redundancy plays a major role for the output quality. Hence, it is of interest to know "how redundant" an attribute actually is. This is measured by the *redundancy rate (for an attribute A)*, which is defined as the sum of the coverage rates (for the attribute *A)* of the data sources, i.e. the average number of sources (for *A*). For the multiple attributes the rate depends on AGE, mainly because of two big registers, TR and FAR (Family Allowance Register). Their intersection generates two peaks of "high redundant" range, the first between 15 and 25 years, the second between 26 and 60 years (cf. Fig. 5).

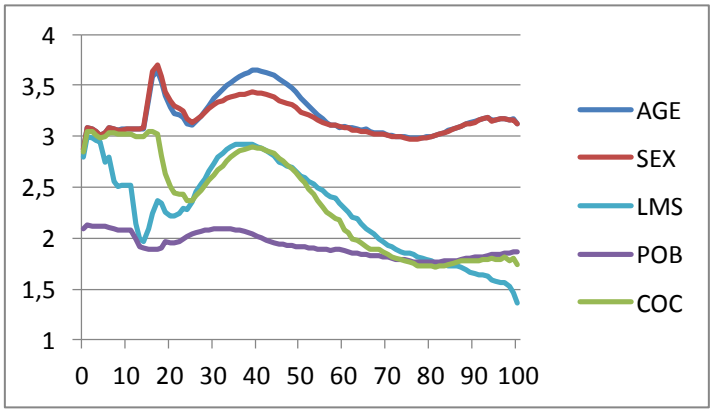

Figure 5: Redundancy rate per attribute and AGE

Note that, if AGE correlates with an arbitrary attribute *A* and the peak-to-valley value of the redundancy rate of a multiple attribute *M* is big enough, then the redundancy rate can affect the quality of *M* grouped by *A*. An example is *A*=CAS (Current activity status) and *M*=COC.

3.3 Assessment for selected subgroups

Now we analyse some attributes, in particular we are investigating the weaknesses of certain subgroups.

AGE, SEX: Evaluating the quality of AGE and SEX (Fig. 6) and zooming into the scale shows that there is a negative peak at the age of 12 years for the attribute SEX. The cause of it is that there are around 400 children, born in 2000 or 2001 with consistent information in the CSSR (Central Social Security Register) and FAR, but different information in the CPR. It seems that there were some mistakes at the initial phase of the CPR in 2001. Since it is not allowed for Statistics Austria to report back possible data errors to the data owners, the data of those children cannot be adjusted at the source. However, the differences decrease annually by around 40 persons, as information is updated when people move or request a certificate of registration from the CPR for other reasons (e.g. passport application).

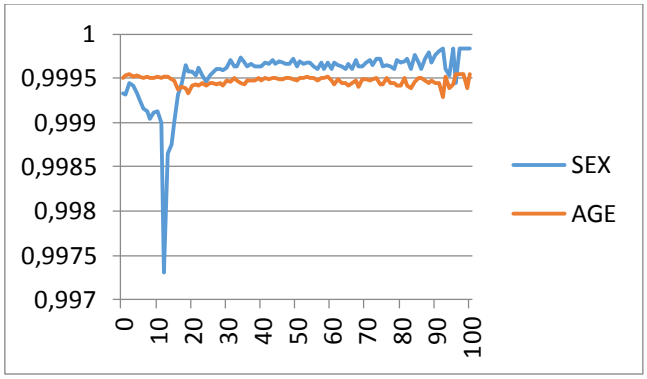

Figure 6: Average quality for AGE and SEX per age

POB: Confusions by name, controversial international legal circumstances or interpretations can lead to different values in registers for subgroups concerned.

Table 2: Average quality for POB and percentage of Austrians population

POB	$\bar{q}_{\Omega\ \mathrm{POB}}$	% of the Austrian population
Total	0.991	100.00
Republic of the Congo	0.991	<0.01
People's Republic of China	0.988	0.17
Democratic Republic of the Congo	0.871	0.01
Republic of China	0.902	0.02

- *Democratic Republic of the Congo*: The name is similar to *Republic of the Congo*. Hence, if this attribute is not relevant for a data holder, there is a substantial risk of confusion.
- *Republic of China*: For the register-based census 2011, the Commission Regulation (EC) No 1201/2009 (2009: L329/46), stipulates the distinction between *Republic of China* and *People's Republic of China*. The reasons for the low quality are twofold. First the name is similar to *People's Republic of China*. Second, the political and legal status is confusing, international not consistent and currently not legitimate as a matter of international law.

Of course, these countries are only marginally represented, but it is not impossible that clusters of them can affect bias in certain municipalities.

4. Conclusion

Since the quality indicator is computed on statistical unit level, the framework enables to assess attributes on macro as well as on micro level. This allows the analysis of different subgroups. Cross tabulating specific values of an attribute with others can disclose the behavior of the data. This approach is applied to the register-based labour market statistics 2013. The investigations for small marginal groups demonstrate that one can detect weaknesses for very small subgroups, and (ideally) can explain them. In summary, the quality measure is confirmed as a very useful tool for the assessment of subgroups. However, this does not exhaust all the possibilities which arise through the framework (e.g. annually monitoring of raw data quality, comparing and analysing raw data quality, assessing the usability of new data sources,…).

5. References

Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H., & Schwerer, E. (2010). A Quality Framework for Statistics based on Administrative Data Sources using the Example of the Austrian Census 2011. Austrian Journal of Statistics, Volume 39, Number 4, 299-308.

Berka, C., Humer, S., Lenk, M., Moser, M., Rechta, H., & Schwerer, E. (2012). Combination of evidence from multiple administrative data sources: quality assessment of the Austrian register-based census 2011. Statistica Neerlandica, Volume 66, Issue 1, 18-33.

Commission Regulation (EC) No 1201/2009 of 30 November 2009 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns

Lenk M. (2009). Methods of the Register-based Census in Austria. Contributed Paper to the Seminar on Innovations in Official Statistics, United Nations, New York.

Lenk M. (2012) Quality assessment of register-based census data in Austria, Conference contribution to the UNECE-Eurostat Expert Group Meeting on Censuses Using Registers, Geneva, 22-23 May 2012

Schnetzer, M., Astleithner, F., Cetkovic, P., Humer, S., Lenk, M., and Moser, M. (2015), Quality Assessment of Imputations in Administrative Data, Journal of Official Statistics, Vol. 31, No. 2, pp. 231–247

United Nations. (2007). Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics. New York and Geneva: United Nations Publication.