Quality Assessment of Register-Based Statistics – Preliminary Results for the Austrian Census 2011

Predrag Ćetković[1], Stefan Humer[1], Manuela Lenk[2], Mathias Moser[1],
Henrik Rechta[2], Matthias Schnetzer[1], Eliane Schwerer[2]

[1] Vienna University of Economics and Business (WU), Augasse 2-6, 1090 Vienna, Austria.

[2] Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. *manuela.lenk@statistik.gv.at*

Abstract The present paper investigates the quality of register data in the context of a standardized quality framework. The focus lies on the assessment of the quality of derived attributes. Such attributes are of high importance for the register-based census in Austria. In order to get a quality measure for the necessary attributes of the census, we have to check the accuracy of the register data. Among other things, the congruency of data between the registers and a comparison data source have to be examined. This may lead to complications in the case of derived attributes, since there may be no data available, which could be used directly for comparison with the register data. Therefore, we have to consider alternative methods in applying our quality framework for derived attributes.

Keywords: administrative data, register-based census, derived attributes.

1. Introduction

Administrative records have become more important for statistical analyses in recent years. The use of administrative data sources has a long tradition in Scandinavian countries and is applied extensively for statistical purposes. One major application is, for example, the register-based census. Administrative data have several advantages over standard surveys. For example, 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 provider. In general, national statistical institutions (NSIs) have little information on the accuracy and reliability of these data. Since Austria, among other countries, will carry out its first register-based census in 2011, it is a central task to assess administrative registers and to evaluate their quality.

Quality assessment of register data has to fulfill several properties like transparency, accuracy or feasibility. To achieve these goals, we set up a general framework, which makes it possible to evaluate the quality of registers with regard to all available information. The present paper deals with the application of this quality framework for the case of derived attributes. These attributes are of high importance, because it is possible that none of the available registers contains an attribute, which is necessary for the register-based census. In this case, related attributes, which could be used for the derivation of the relevant attribute, have to be found. Since a relevant attribute may be derived from several raw data attributes, we would have to check the accuracy of all raw data information. Thus, appropriate comparison data for each raw data attribute should be available in order to check for congruency of data between the registers and a comparison data source. If there is no such comparison data available, we would rely on expert opinions. Since expert opinions may be associated with problems of subjectivity, we consider an alternative method, where only the congruency between the derived attribute itself (data in the Census Database) and the comparison data is checked. The derived attribute we have analyzed in this paper, is the *current activity status*. For this attribute it is also possible to check the congruency between registers, which were used in the derivation process, and the comparison source. Thus, we are able to compare the results of both methods in order to check for possible discrepancies between the two alternatives.

The remainder of the paper is structured as follows. Section 2 gives a general overview of the quality framework and explains its most important elements. The application of the quality framework for analyzing derived attributes is explained in detail in section 3. Section 4 then shows the results of the quality assessment of the attribute *current activity status*. The last section concludes.

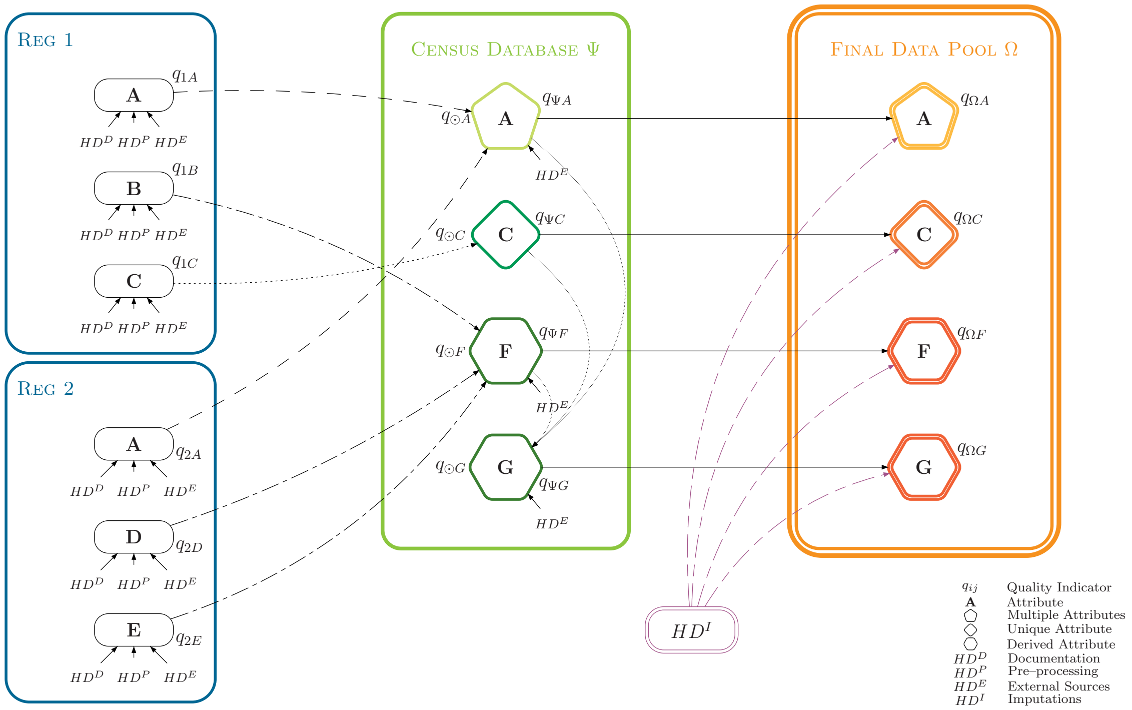
2. Quality Framework

Statistical data quality can be covered by several dimensions like timeliness or accuracy (Eurostat, 2003a). This also applies to administrative data as has been stressed by Eurostat (2003b). There is only few literature which deals with quality assessment of administrative data sources. Some national statistical institutions, like Statistics Finland, focus on the comparison between administrative and survey data (Ruotsalainen, 2008). Other countries, for example the Netherlands, take a more structural approach (Daas *et al.*, 2009). Their aim is to cover the quality of different registers in a framework using different dimensions to assess data quality and accuracy. They developed a checklist for the quality evaluation of administrative data sources, which is structured in three different hyperdimensions of quality aspects. Our approach is an extension of the framework proposed by Daas *et al.* (2009) and it contributes a framework for structural assessment of administrative data to the field of quality research. This allows both the NSI and external researchers to assess the data sources they use.

In our quality 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 realized by a framework, which is closely tied to the data flow, but 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 (Lenk, 2008).

The quality framework, which is shown in Figure 1, is linked to the data flow on three different levels. In a first step, Statistics Austria receives the raw data (henceforth registers, see boxes on the left-hand side in Figure 1). In the next step, these different sources are combined to data cubes, the Census Database (CDB), by using unique IDs. These cubes solely include information available from the registers (raw data). Finally, we enrich the CDB with imputations of item non-responses. These steps result in a Final Data Pool (FDP), which consists of both real and estimated values. In each of these three steps (Registers, CDB and FDP) the data flow is linked to the quality assessment, so that changes can be monitored from a quality perspective. As a result, exactly one quality indicator for each attribute in each register or data pool is calculated (q_{ij} in Figure 1).

Figure 1: *Quality Framework*



The quality assessment of the registers consists of three hyperdimensions: *Documentation (HD^D)*, *Pre-processing (HD^P)* and *External Source (HD^E)*. The first hyperdimension, HD^D, includes all quality related aspects prior to seeing the data. Such aspects are, for example, plausibility checks, data collection methods or legal enforcements of data recording by the provider of the administrative data. Thus, it is a measure of the degree of confidence we put in the data provider. HD^D is realized through a questionnaire which is filled out in accordance with the register authority. For each question there is a maximum score that can be obtained. Summing up the score for each question and comparing this sum to the maximum score leads to the quality indicator

$$HD^D : \frac{obtained\ score}{maximum\ score} \tag{1}$$

The second aspect of the quality framework, HD^P, 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-responses. Usable records are therefore calculated by subtracting all incorrect entries from the total number of observations. The quality measure for the hyperdimesion Pre-Processing is given by:

$$HD^P : \frac{number\ of\ usable\ records}{total\ number\ of\ records} \tag{2}$$

In the last step we then investigate the congruency of the data by comparing it to an external source (HD^E). This is primarily done using existing surveys (i.e. the Austrian

Microcensus). The Microcensus is an appropriate comparison source, because we can link its data via a unique key with the data in the registers or the CDB in order to compare the values and check for consistency on the unit level. As a result, we get the quality measure

$$HD^E : \frac{number\ of\ consistent\ values}{total\ number\ of\ linked\ records} \tag{3}$$

If an attribute is not found in the Microcensus, we rely on expert opinions. The expert is a person at Statistics Austria, who is responsible for the administrative register and therefore has experience with the quality of the data. For further information on the three hyperdimensions see Berka *et al.* (2010).

The quality indicator q_{ij} on register level results from a weighted combination of the three hyperdimensions. Thus, appropriate weights, which resemble the relative importance of each hyperdimension, have to be chosen. In a further step, we can use the quality indicators to assess the quality of the data in the Census Database.

In comparing the CDB with the raw data registers, we can generally 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 raw data register with a similar attribute is 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}, which is the quality indicator for the specific attribute j in register i. A unique attribute is, for example, the *level of education*. For multiple attributes (e.g. *sex*), a specific method must be applied in order to deal with quality indicators from different data sources. This is most important in cases where the information differ between these data sources. In this case, the *Dempster-Shafer theory* is an appropriate method to assess the quality of the data (Dempster, 1968; Shafer, 1992). A detailed investigation of the quality of multiple attributes is provided in Berka *et al.* (forthcoming). The case of derived attributes (e.g. *current activity status*) is subject to the present paper and will be explained in detail in the following sections.

3. Quality assessment of derived attributes

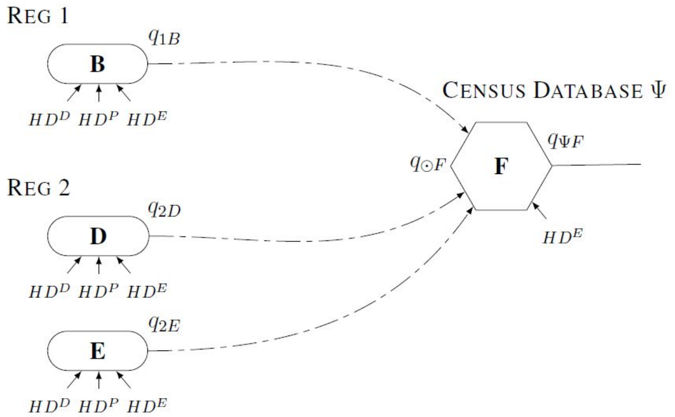
Derived attributes are such, for which the registers do not contain any information in the required specification. However, if the raw data contain attributes, which are related to those we are looking for, they could be used for the derivation of the latter (e.g. attribute F in Figure 1). Such an attribute is, for example, the *current activity status*, which can be derived from various registers, like the *Unemployment Register* or the *Central Social Security Register*. Moreover, a relevant attribute may also be derived from an attribute in the CDB (e.g. attribute G in Figure 1). This may be necessary, if there is no information on raw data level, which could be used directly for the derivation of the relevant attribute. An example for this type of attribute is the *occupation*, which is derived on CDB level (among other) from the *current activity status*, which is a derived attribute itself.

As has been mentioned, more than one register may be used for the derivation of a specific attribute. Thus, if the number of used registers gets large, we would have to assess the quality for a high number of attributes used in the derivation process. Apart from the extended number of applications, no further problems will arise for the hyperdimensions

Documentation and Pre-processing. By contrast, the hyperdimension External Source may lead to further complications, since the congruency of data between all used registers and the comparison source has to be checked. Particularly cases, where no appropriate information for each raw data attribute can be found in the primary comparison source (Austrian Microcensus), will be associated with additional problems and will require other external sources. Alternative external sources are, for example, expert opinions. However, since expert opinions may suffer from subjectivity, the reliability of this type of external source could be questioned. Additionally, such expert interviews would be associated with an increased work effort. In order to deal with these shortcomings, we consider an alternative method, which only differs to the first one with respect to the application of the hyperdimension External Source.

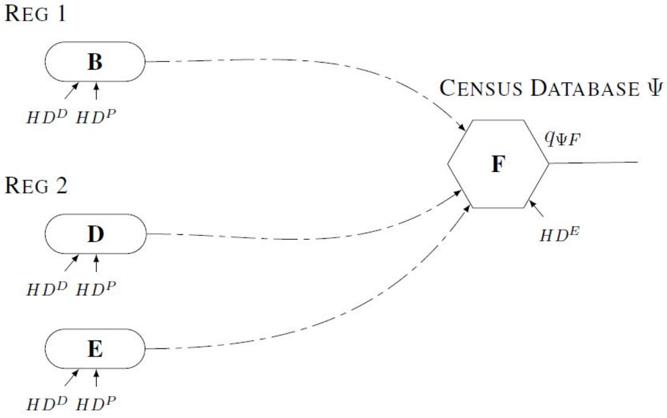
The first method of assessing the quality measure for derived attributes is shown in Figure 2. The three hyperdimensions are all applied on raw data. This results in the quality indicators q_{1B}, q_{2D} and q_{2E} for the attributes B, D and E respectively. A weighted combination of the three quality indicators will then lead to the quality indicator for the derived attribute ($q_{\odot F}$). It may also be necessary to assess the errors of the derivation process itself. Therefore, it can be helpful to check the validity of the derived attribute using an external source (HD^E for the attribute F in Figure 2). A combination of $q_{\odot F}$ with the hyperdimension HD^E on CDB level leads to the quality indicator $q_{\Psi F}$. However, since for the purpose of this paper we are interested in the quality measure $q_{\odot F}$, the indicator $q_{\Psi F}$ has not been calculated for the first method.

Figure 2: *Derived attributes, method a*



In the second method, only the hyperdimensions Documentation and Pre-processing are done on raw data level, while the hyperdimension External Source is applied on CDB level (Figure 3). Since there is no direct measure of HD^E for raw data, the quality indicators for the attributes on raw data level, q_{1B}, q_{2D} and q_{2E}, can not be assessed. This is due to the possibility that different attributes in the CDB may be derived from the same raw data attribute. As the hyperdimension External Source is applied on CDB level, a variety of HD^E-quality indicators would be available for the same raw data attribute. Thus, an assignment of the calculated quality measures to raw data would lead to ambiguous results. However, it is possible to assess the quality indicator for the derived attribute ($q_{\Psi F}$), which is calculated by a weighted combination of HD^D and HD^P on raw data level with HD^E on CDB level.

Figure 3: *Derived attributes, method b*



4. Results

For the Austrian register-based census, many attributes are of the nature of derived attributes. The first attribute of this type we deal with, is the *current activity status*. For the derivation of this attribute, we use several registers. Since none of these registers contain the current activity status in the required specification, related attributes have to be found. Additionally, the specification of these related attributes differs from register to register, so that we end up with 8 different attributes, each of which is included in a separate register. The used registers are the *Central Social Security Register (CSSR)*, the *Unemployment Register (UR)*, the *Register of Social Welfare Recipients (RSWR)*, the *Data of the Federal Chambers (FC)*, the *Registers of Public Servants of the Federal State and the Laender (RPS)*, the *Conscription Register (CR)*, the *Tax Register (TR)* and the *Register of Enrolled Pupils and Students (REPS)*.

As has been mentioned in Section 3, the application of the hyperdimension HD^E may lead to complications for the case of derived attributes. This is due to the possibility that the primary comparison source, the Austrian Microcensus, may either not contain all attributes, which are necessary for comparison or the specification in the Microcensus does not fit with the specification of raw data. For the current activity status, it was possible to find an attribute in the Microcensus (*activity status*), which could be used for comparison with the current activity status on CDB level and all attributes on raw data level except the data from the Register of Enrolled Pupils and Students. The latter was therefore compared with an other attribute in the Microcensus (*participation in education*). However, it was necessary to respecify the relevant raw data attributes as well as the current activity status, so that they could be compared with Microcensus data.[1] The eventual categories for the REPS are: *currently in education* and *currently not in education*. All other raw data attributes as well as the current activity status itself have been classified into the following specifications: *employed*, *unemployed*, *not*

[1] The Register of Enrolled Pupils and Students contains only persons currently in education. The attribute participation in education from the Microcensus is classified into the categories *currently in education* and *not currently in education*. The other raw data attributes have in sum more than 1,100 specifications. By applying a ruleset, Statistics Austria reduces these different categories to about 40. In a further step, we reduce these 40 classes to 5 in order to make a comparison with Microcensus data possible.

economically active, *military and civil servants* and *persons under 15 years*.[2]

The whole population in the Central Database consists of all unique entries in the Central Population Register (CPR). The current activity status in the CDB is derived by using a predefined ruleset, where each applied register contributes to a different degree in the derivation process. The applied ruleset is in accordance with international standards. In order to get an overall quality measure for the attribute current activity status, the quality indicators of the relevant raw data attributes have to be weighted by their contribution to the derivation of the current activity status. These contribution shares are shown in Table 1, where it can be seen that the current activity status has been derived in most cases from the Central Social Security Register (77.18% of all CDB entries). Because of a lack of data in other registers, 5.02% of the entries in the CDB have been derived from data in the Central Population Register (last column in Table 1).

Table 1: *Shares of registers in the derivation of the current activity status in per cent*

	$CSSR$	UR	$RSWR$	FC	RPS	CR	TR	$REPS$	CPR
w_i	77.18	3.15	0.79	0.09	0.18	0.18	0.24	13.17	5.02

The results of the first method, where all three hyperdimensions are applied on raw data level, is shown in Table 2. As can be seen, the hyperdimension Documentation shows a high variability between the registers. It should be mentioned here that this hyperdimension has been hitherto conducted only for the Central Social Security Register, the Unemployment Register and the Register of Enrolled Pupils and Students. The values for the remaining registers are therefore approximated values. The hyperdimension Pre-Processing assigns a high quality to all attributes. Because there are in general only a few items, which do not have an unique ID, the measure for HD^P is in most cases slightly less than one. By contrast, the raw data do not suffer from item non-responses or out of range-values. According to the hyperdimension External Source, raw data is in most cases consistent with data in the Microcensus. However, with a value of 0.38, the attribute from the Unemployment Register has a very low quality when it is applied for the derivation of the current activity status. This is probably due to the different definition of unemployment between the Unemployment Register and the Microcensus.[3]

As the Central Population register does not contain any information regarding the current activity status, those entries, which have been derived from the CPR have been defined as not economically active and their quality indicator has been set to 0. The three hyperdimensions have been equally weighted by 1/3. The combination of the quality

[2] Persons under 15 years are not directly surveyed in the Austrian Microcensus regarding the attribute activity status. Thus, for the application of the hyperdimension External Source, these persons have been dropped out of Microcensus data. As persons under 15 years are not highly represented in most registers, dropping out this group will not really influence the results.

[3] In comparing the Unemployment Register with the Microcensus, 1,539 persons could be linked for the 4th quarter 2009. In the Unemployment Register, 1,216 out of these 1,539 cases have the status unemployed. The remaining cases are mostly persons, which participate in job-training courses and thus are not counted as unemployed. From the 1,216 cases, which are unemployed according to the Unemployment Register, only 481 are also declared as unemployed in the Microcensus, whereas 354 persons are considered as employed and 381 as not economically active.

Table 2: *Results, method a*

Register	HD_i^D	HD_i^P	HD_i^E	q_{ij}
$CSSR$	0.86	0.97	0.92	0.92
UR	0.62	1.00	0.38	0.67
$RSWR$	0.93	0.99	0.91	0.94
FC	0.38	0.98	0.95	0.77
RPS	1.00	0.98	0.97	0.98
CR	0.88	1.00	0.77	0.88
TR	0.79	0.96	0.95	0.90
$REPS$	0.86	0.98	0.83	0.89

indicators q_{ij} of the raw data attributes (by using weights w_i) results in the quality measure for the attribute current activity status ($q_{\odot current\ activity\ status}$), which has a value of 0.862.

$$q_{\odot F} = \sum(q_{ij} * w_i) = \sum[(\frac{1}{3}HD_i^D + \frac{1}{3}HD_i^P + \frac{1}{3}HD_i^E) * w_i] = 0.862 \qquad (4)$$

Table 3 shows the results for the second method. The values for the hyperdimensions Documentation and Pre-Processing as well as the weights for the three hyperdimensions are the same as in the first method. The hyperdimension External Source, which has been assessed for the attribute current activity status in the Central Database, has a high quality. As a consequence, the quality indicator of four registers would be improved in comparison to the first alternative. This is particularly true for the Unemployment Register, where the quality indicator now would be 0.85, compared to 0.67 in the first method. However, as the hyperdimension has been done on CDB level, we can not really assign these quality indicators to the registers (see Section 3 for an explanation). The weighted quality measure for the current activity status ($q_{\Psi current\ activity\ status}$) now has a value of 0.872, which is slightly higher than in the first method.

Table 3: *Results, method b*

Register	HD_i^D	HD_i^P	HD_Ψ^E	(q_{ij})
$CSSR$	0.86	0.97	0.92	0.92
UR	0.62	1.00	0.92	0.85
$RSWR$	0.93	0.99	0.92	0.95
FC	0.38	0.98	0.92	0.76
RPS	1.00	0.98	0.92	0.97
CR	0.88	1.00	0.92	0.93
TR	0.79	0.96	0.92	0.89
$REPS$	0.86	0.98	0.92	0.92

$$q_{\Psi F} = \sum[(\frac{1}{3}HD_i^D + \frac{1}{3}HD_i^P + \frac{1}{3}HD_\Psi^E) * w_i] = 0.872 \qquad (5)$$

5. Conclusion

In this paper we investigated the quality of administrative data for the purpose of applying these data for the register-based census. A general quality framework was adapted in order to deal with derived attributes, which are of high importance for the census. For this purpose, two different methods have been carried out. The first method does the whole quality assessment (all three hyperdimensions) on raw data, whereas the second method shifts the hyperdimension External Source to the data in the Census Database.

The first derived attribute we have dealt with is the current activity status. In order to get this attribute, related attributes from 8 different registers have been used. The results for the first method show that most of the used raw data attributes have a high quality measure. Thus, the overall quality indicator for the current activity status is 0.862, which is fairly high. If the second method is applied, the quality indicator for the current activity status increases slightly to 0.872.

The similarity of the results of the two alternatives indicates that there are no problems in applying the second method for the quality assessment of the attribute current activity status. This is a positive finding, because the hyperdimension External Source has to be done only for the derived attribute and not for all raw data attributes. Thus, complications associated with non-availability of comparison data or subjectivity of potential expert opinions are reduced. However, the positive result for the current activity status does not guarantee that the application of the two alternative methods would lead to the same conclusion for other derived attributes.

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.

Berka C., Humer S., Lenk M., Moser M., Rechta H., Schwerer E. (forthcoming) Combination of evidence from mulitiple administrative data sources – Quality assessment of the Austrian register-based census 2011, *Statistica Neerlandica*.

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

Dempster A. (1968) A generalization of bayesian inference, *Journal of the Royal Statistical Society. Series B (Methodological)*, 30, 205–247.

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*.

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

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

Shafer G. (1992) Dempster-Shafer Theory, in: *Encyclopedia of artificial intelligence*, Shapiro S (Ed.), Wiley, 330–331.