Distr. GENERAL

8 October 2012

WP. 11
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Seminar on New Frontiers for Statistical Data Collection
(Geneva, Switzerland, 31 October-2 November 2012)

Topic (ii): New methods and technologies

A QUALITY MONITORING SYSTEM FOR STATISTICS BASED ON ADMINISTRATIVE DATA

Contributed Paper

Prepared by Predrag Ćetković*, Stefan Humer*, Manuela Lenk**, Mathias Moser*, Matthias Schnetzer*, Eliane Schwerer**

***Statistics Austria, Unit Register-based Census, *Vienna University of Economics and Business*

I. Introduction

1. Administrative data sources have significantly gained importance for statistical purposes in the European Union. The processing of data which has already been recorded by administrative authorities offers numerous advantages compared to survey data such as diminishing costs, removed burden for respondents or the prompt availability of the data.

2. The increasing importance of administrative data raises the question of the quality of these data sources. Quality of administrative data can be covered by several dimensions like, for example, timeliness or accuracy [5]. In our quality framework, we focus on data accuracy, since this is the most challenging dimension. Moreover, the accuracy of the administrative source is essential for the quality of register-based statistics. The quality assessment is realized by a framework, which is closely tied to the data flow, but independent from data processing. Because of this separation, the quality assessment can evaluate the processing procedure without influencing it. The predominant feature of the quality monitoring system is its applicability for other register-based statistics.

3. The paper proceeds as follows. The next section shows an overview of the quality framework. It is followed by an explanation of the quality assessment on raw data level. Then the quality evaluation for a merged data cube (Central Database) and an imputation-enriched data pool (Final Data Pool) is explained for three different types of attributes (unique, multiple and derived attributes). The last section comprises a short summary of the most important findings and a conclusion.

II. Quality framework

4. Quality assessment of administrative data has to fulfill several properties like accuracy or feasibility. To achieve these goals, we set up a general framework (see Figure 1).

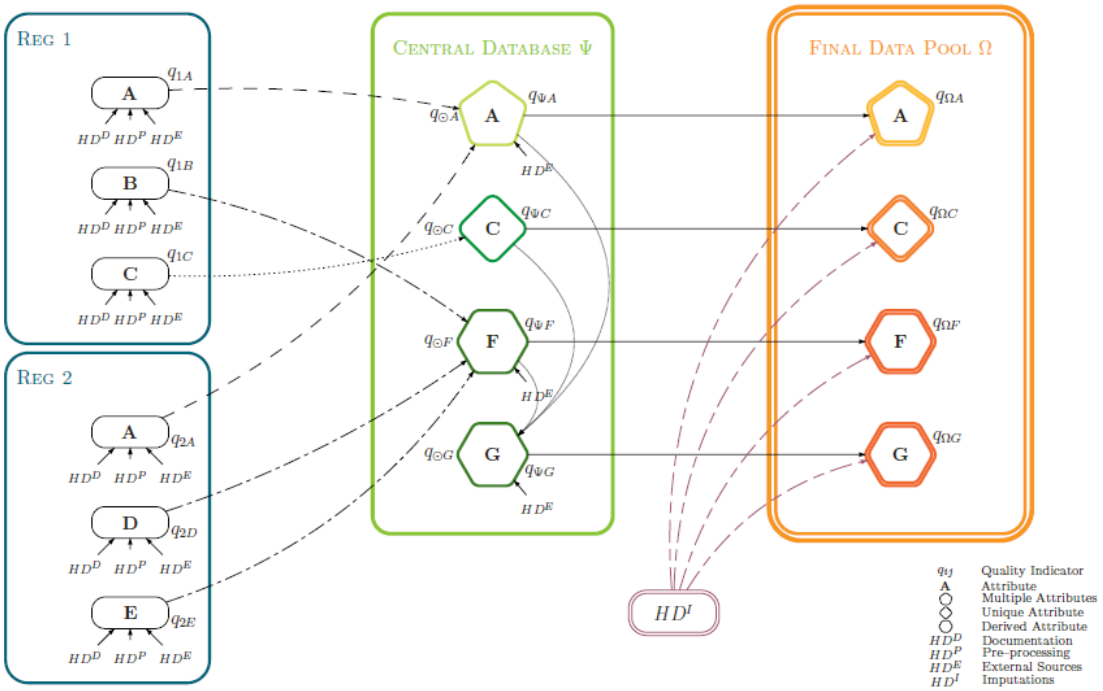

Figure 1: *Quality framework for register-based statistics*

5. The data flow shown in Figure 1 consists of three levels: raw data (i.e. registers), combined data set (*Central Database*, henceforth CDB) and imputed dataset (*Final Data Pool*, henceforth FDP). The data process starts with the receipt of the raw data from the administrative data authorities (*registers REG 1* and *REG 2* on the left-hand side in Figure 1). In the next step, these different sources are combined via a unique key and merged to the CDB (box in the middle of Figure 1). Finally, the FDP is constructed by the enrichment of the CDB with imputations for item non-response (right-hand side in Figure 1). Thus, the FDP 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.

III. Quality assessment on register level

6. The quality assessment of the registers consists of three hyperdimensions: *Documentation (HD^D), Pre-processing (HD^P) and External Source (HD^E).* Prior to seeing the data, HD^D describes quality-related processes at the register authority as well as the documentation of the data (i.e. metadata). The degree of confidence and reliability of the data holder is monitored by the use of a questionnaire containing nine scored questions (see Table 1).

Can we detect data changes over time?	
Is the information available for the reference date?	
DEFINITIONS	
Are the 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?	

Table 1. *Scored Questions – HD Documentation*

7. The national statistical office (NSO) is therefore able to check for data collection methods or legal enforcements of data recording which may significantly influence the quality of the data. The questionnaire is filled out in accordance with the register authority and should thus deliver convincing results. For each question there is a maximum score that can be obtained. The quality indicator is a simple ratio between the *obtained score* and the *maximum obtainable score*. This evaluation is carried out for each attribute in each register.

8. The second hyperdimension, HD^P, deals with formal errors in the raw data, i.e. it checks for definition and range errors, as well as missing primary keys and item non-response. Therefore, usable records are calculated by subtracting all these incorrect and missing entries from the total number of observations. The relation of *usable records* to the *total number of records* defines the quality for HD^P. Again, this procedure is carried out for each attribute in each register.

9. In the last step we then investigate the congruency of the register data by comparing it with an external source (HD^E). This is primarily accomplished by the comparison with already existing representative surveys. An appropriate comparison data source is a survey that can be merged with the register data via a unique key, in order to compare the values of the two different data sources and check for consistency on a unit level. The quality measure for HD^E is given by the ratio of the *number of consistent values* to the *total number of linked records*. Of course a survey only includes a subset of the register data, but for a meaningful quality assessment it is sufficient to link only those entries, which are also included in the survey.

10. If such a benchmark does not exist, we suggest relying on expert opinion (expert interview). The expert is a person at the NSO, 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 [1].

11. Given these three quality measures, an overall quality indicator q_{ij} for each attribute and register can be derived as a weighted average. Thus, appropriate weights, which resemble the relative importance of each hyperdimension, have to be chosen. The resulting quality indicator ranges between 0 and 1 – the closer to one, the better the quality of the assessed attribute. An important advantage of this measurement is its simplicity, which offers the possibility for an uncomplicated comparison of the quality of attributes within and between the registers.

12. Table 2 shows some results for the attributes *sex*, *full-* or *part-time employment* and the *highest level of education* in five registers for data of the register-based labor market statistics 2008.

Here, we are working with equal weights for the three hyperdimensions. It can be seen that the quality of the attribute sex is very good in all of these registers. However, there are some notable differences between the hyperdimensions. For example, Register 3 has a very low value for the hyperdimension Documentation (HD^D), due to the fact that the attribute sex is not relevant for this specific register authority. The indicators for the hyperdimension Pre-processing (HD^P) are mostly influenced by missing primary keys, while range and definition errors only play a minor role. For the hyperdimension External Source (HD^E), the framework returns very high quality measures, which means that there is a high degree of agreement between the register data and survey data.

Register	Attribute	HD^D	HD^P	HD^E	$q^{(33,33,33)}$
REG 1	SEX	1.000	1.000	0.998	0.999
REG 2	SEX	0.792	0.942	0.999	0.911
REG 3	SEX	0.444	0.746	0.997	0.729
REG 4	SEX	0.792	0.993	1.000	0.928
REG 3	FT/PT	0.381	0.698	0.847	0.642
REG 5	EDU	0.928	0.950	0.800	0.891

Table 2. *Quality measures for data of 2008*

IV. Quality assessment of the Central Database and the Final Data Pool

13. The entire information of the registers is linked to a new data cube, the Central Database (CDB), which covers all attributes of interest for the register-based statistics. Concerning the evaluation of the quality of the attributes on CDB level, we distinguish three types of attributes that occur in this linking process:

    (a) *Unique attributes* exist in exactly one register (see Figure 1, attribute C).

    (b) *Multiple attributes* show up in several registers (see Figure 1, attribute A).

    (c) *Derived attributes* are created based on other attributes, because the registers do not contain the required attribute (see Figure 1, attributes F and G).

14. Current research is focused on the calculation of quality indicators in the Final Data Pool (FDP), which corresponds to the Census Database after imputations are applied. The amount of item non–response is effectively reduced by imputations. However, the imputation process itself has to be monitored. This is done by using information from the hyperdimension Imputation (HD^I), which is an ongoing task.[1]

15. The following subsections illustrate the quality assessment of these different kinds of attributes on CDB and FDP level. For the sake of completeness, artificial quality measures for imputed values are supposed.

A. Unique attributes

[1] Included in this hyperdimension are imputation, estimation, record linking without unique key and if necessary statistical matching.

16. Figure 2 shows the quality assessment of a unique attribute, which is illustrated with the attribute *highest level of education (EDU)*.

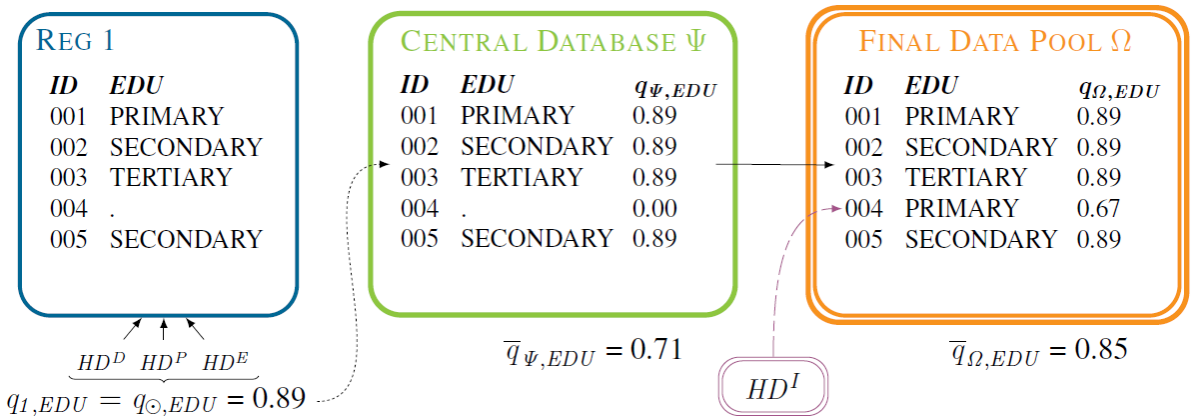

Figure 2: *Quality assessment of unique attributes*

17. The quality assessment starts with the application of the three hyperdimensions on register level (described in section 2) and is assumed to be 0.89. In the case of unique attributes, the information from the register is directly transferred to the CDB. As a consequence, the quality indicator is equal to that of the register level. The quality for item non-response is set to 0 on CDB level (see ID 004 in the middle of Figure 2) because the quality of this entry will be updated with the quality of the imputed value. The quality measure for this value is related to the imputation process itself and is assumed to be 0.67 (see ID 004 on the right-hand side of Figure 2). Those entries, which are non-imputed, will have the same quality in the FDP as in the CDB. Calculating an average of the quality measures on FDP level will result in the final quality indicator of 0.85 for the assessed attribute. Thus, the quality has risen in comparison to the average quality on CDB level (quality of 0.71).

B. Multiple attributes

18. An illustration of the quality assessment for the multiple attribute *sex*, which is assumed to be included in two different registers, is shown in Figure 3. In the case of multiple attributes the information of the registers is combined to one valid value for the CDB by certain decision rules.[2]

[2] The decision rules have to be developed by the NSO and can be based on the experience with the data quality so far.

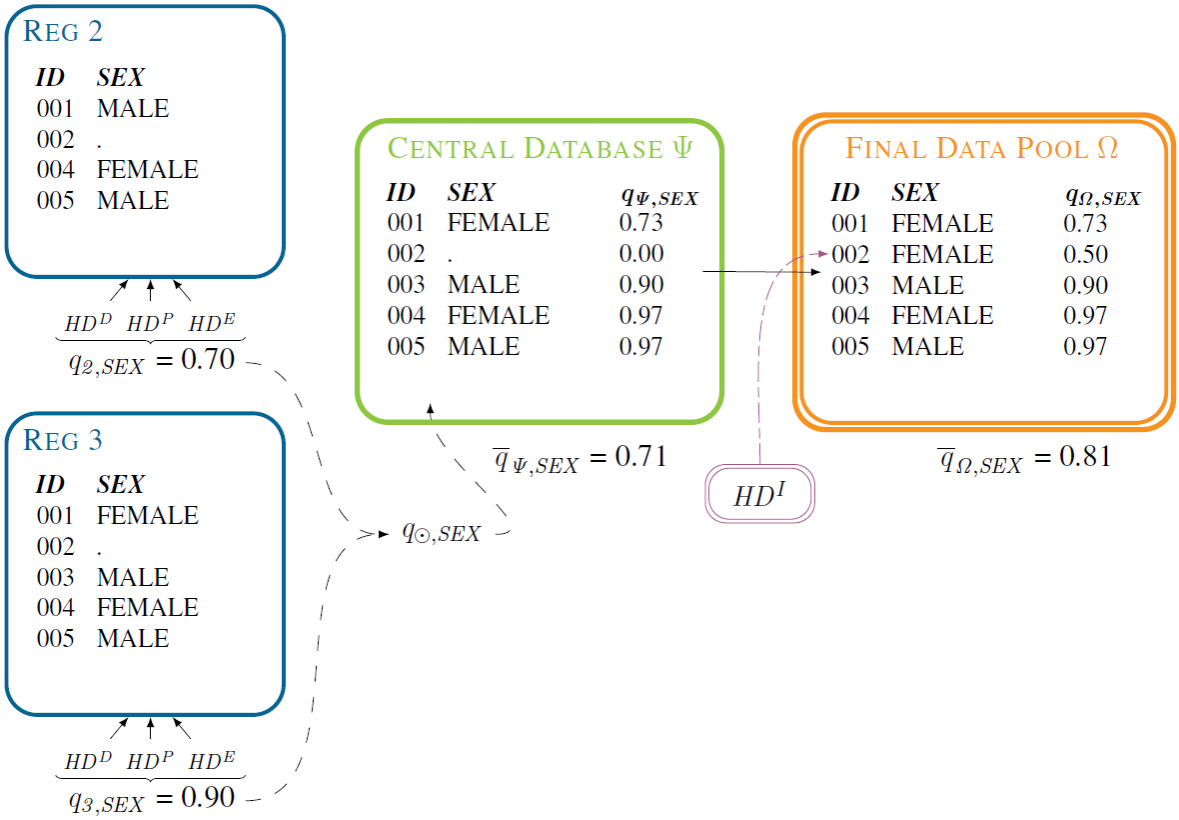

Figure 3: *Quality assessment of multiple attributes*

19. Suppose that the quality indicator for the attribute sex in register two is 0.7 and the quality indicator for the same attribute in register three is 0.9. While the value for the attribute sex of the IDs 004 and 005 is the same in the two registers, the information of ID 001 differs between the two data sources (male in REG 2 and female in REG 3). If registers provide contradictory information, it may not be clear which register is wrong. Applying simple weighted averages and neglecting coinciding and opposing evidence could lead to delusive conclusions, whereas more sophisticated methods like the *Dempster-Shafer theory*[3] are used to evaluate the quality on CDB level. It allows combining information from different registers while the degree of belief in the data source is taken into account. Thus, the quality indicator increases if there is consistency between the sources (see ID 004 and 005). However, the indicator decreases when a conflict occurs (see ID 001).[4] A detailed application of the Dempster-Shafer theory on multiple attributes is given by [2].

20. ID 003 gets the quality indicator of register 3, since this person is included only in that register. ID 002 has a missing item and thus the quality of the attribute sex of this person is set to 0 in the CDB. The average quality of the attribute sex in the CDB is 0.71.

21. The enrichment of the CDB by imputations for the sex of ID 002 leads to the Final Data Pool. The quality of the imputed value for ID 002 is assumed to be 0.50, whereas all non-imputed values get the quality indicator of CDB level. The final quality indicator for the attribute sex is the average of the quality measures in the FDP and is given by 0.81. Thus, the quality improves from the CDB to the FDP.

C. Derived attributes

[3] For detailed information about the Dempster-Shafer theory see [4, 6].

[4] A simple mean would yield the same quality measure for conflict and congruence between the two registers.

22. This section refers to the quality assessment of an attribute, which is derived from an attribute in the CDB (attribute G in Figure 1). For a detailed description of the quality assessment of derived attributes see [3]. Figure 4 shows the quality assessment for the attribute *type of commuter (COM)*, which is derived from the *current activity status (CAS)*.

23. The derivation process is as follows: Employed persons are defined as economically active commuters, pupils/students are commuting to their educational institution and unemployed people or homemakers do not commute.[5] For ID 004, no information about the current activity status is available, so the type of commuting is also unknown for this person. Since the values for the attribute COM are directly derived from the current activity status, the quality is the same as that for the attribute CAS (assumed to be 0.91). ID 004 has a quality of 0 in the CDB, due to its item non-response for the attribute-value of CAS and COM.

24. The Final Data Pool is filled with the data of the CDB and contains also the imputed value for the attribute CAS of ID 004. The quality of CAS is assumed to be 0.73. The attribute COM has the same quality, because it is directly based on the value of CAS. The quality measures in the FDP for the other IDs are the same as in the CDB. Again the final quality indicator for the attribute type of commuter is given by the average of the quality measures on FDP level and has a value of 0.87. Again, this means an increase in the quality in comparison to the CDB.

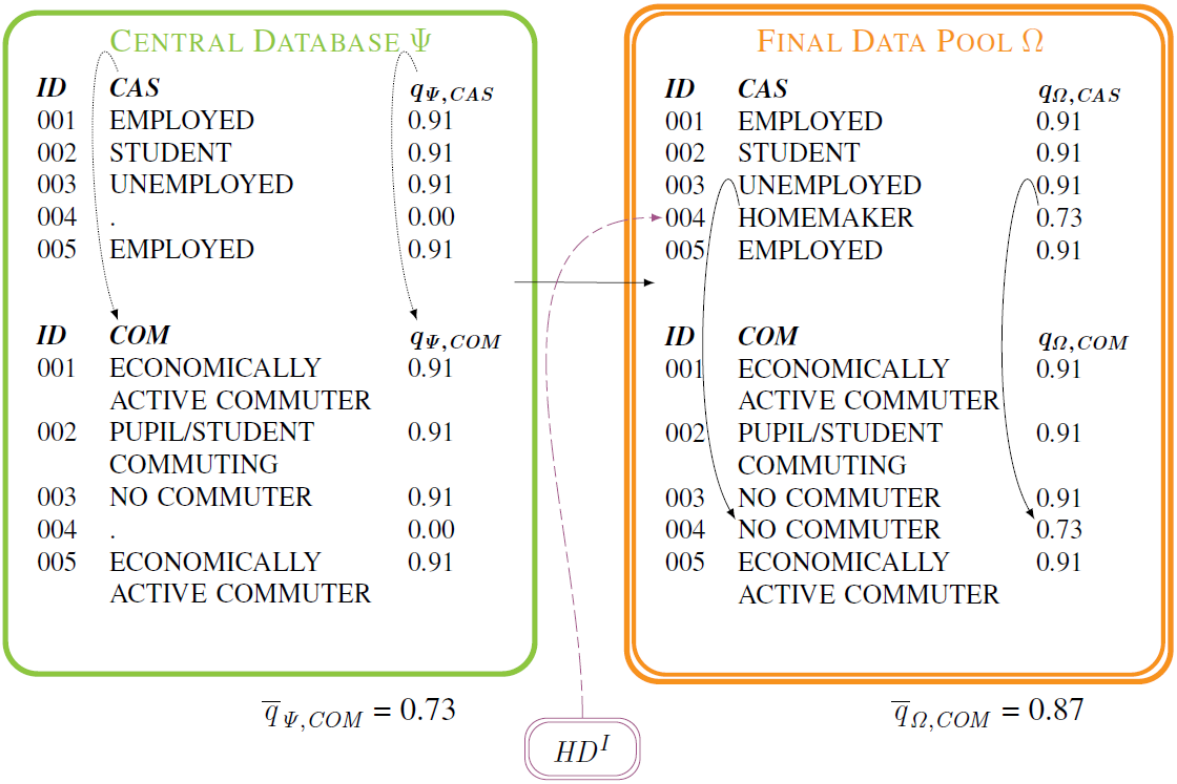

Figure 4: *Quality assessment of derived attributes*

V. Conclusion

[5] Employed persons, who work at home, should not be considered as commuters. For the case of simplicity, we abstract from such cases.

25. This paper presents a structural approach for the quality assessment of administrative data. The quality monitoring is based on three different data levels: 1. the raw data, 2. the Central Database and 3. the Final Data Pool. The most comprehensive part of quality monitoring refers to raw data, where three different quality assessment steps are applied (HD Documentation, HD Pre-Processing and HD External Source). The resulting quality measure is then updated for the data in the CDB and FDP. To guarantee the applicability of the quality framework for the register–based statistics, the procedure was tested with register-based labor market data from 2008.

26. The application of the quality framework has different implications for the three different types of attributes, which are considered to be relevant for the usage of administrative data for statistical purposes. First, unique attributes get their quality directly from the register. Second, the quality of multiple attributes has to be calculated by a combination of the qualities from the different registers. We have suggested the application of the Dempster-Shafer theory, since this theory is able to handle conflicting information from various registers. Finally, derived attributes get their quality from the related attributes.

27. Accordingly, the difference between the three types of attributes lies mainly in the transformation of the register quality to the quality on CDB level. We expect that the proposed methods for unique and multiple attributes are also applicable for other attributes than those, which have been presented in this paper. However, different derived attributes could also require a different treatment in their quality assessment.

28. A decisive advantage of the quality framework at hand is the independence of quality assessment and data processing. The separation from the processing procedure is required to evaluate the process without exerting influence on it. This offers the possibility to apply the methods on various register–based data sets. Moreover, the cooperation between the NSO and the register authorities is intensified because the data holder is integrated in the quality assessment process.

VI. References

[1] 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*, Vol. 39, No. 4, 299--308.

[2] 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*, Vol. 66, No. 1, 18--33.

[3] Ćetković, P., Humer, S., Lenk, M., Moser, M., Rechta, H., Schnetzer, M. & Schwerer, E. (2011), *Quality Assessment of Register-Based Statistics - Preliminary Results for the Austrian Census 2011*. Conference contribution at ESSnet on Data Integration, Madrid.

[4] Dempster, A. (1968). A generalization of bayesian inference. Journal of the Royal Statistical Society. Series B (Methodological), 30(2), 205--247.

[5] Eurostat. (2003). Quality assessment of administrative data for statistical purposes. In: *Assessment of quality in statistics*. Eurostat, Luxemburg.

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