28th Meeting of the Wiesbaden Group on Business Registers
The Hague, 02 – 06th October 2023

Simon Rommelspacher, Adrian Urban
Federal Statistical Office, Germany (Destatis)

Session 3: Globalisation and Large Case Units

Similarity metric for comparison of enterprise groups

Abstract

Data on enterprise groups are stored and maintained in the statistical business registers at national, European and global level. Enterprise groups are structures of legal units with control relationships between them. The maintenance of these complex structures requires to compare them at different points in time or from different sources.

Destatis has developed a similarity metric for the comparison of enterprise groups from different sources or at different points in time. This metric is intended to provide an indication of whether there is similarity in terms of the associated units and their economic characteristics between two enterprise groups. The similarity metric can have values between 0 and 1. The lower the similarity metric, the lower the similarity. Identical enterprise groups have a similarity metric of 1. For enterprise groups without any overlap of units, a similarity metric of 0 is output.

In this paper, the motivation, development and calculation of the similarity metric will be presented using theoretical examples. Then we will describe use cases for the metric with data from the German statistical business register and the EuroGroups Register. Based on this, we can discuss further possible use cases.

The idea of the similarity metric of enterprise groups can be used for deciding about continuity, or for time series data analyses to study the economic and structural changes of enterprise groups. A metric for other statistical units or a case-by-case analysis of the similarity of certain complex units - for example enterprise groups which are competing in the market - could also be developed with the similarity metric.

Keywords: Enterprise Groups, Data Sources, Globalisation, Business Register

1 Introduction

According to EU-Regulation 2019/2152 of the European Parliament and of the European Council the national statistical business registers shall cover all enterprise groups to which the enterprises and legal units included in the register belong. This applies to all EU Member States. The EuroGroups Register covers all multinational enterprise groups which include at least one enterprise in an EU member state.

Enterprise groups are defined from the fact that legal units can be controlled by another legal unit. Control exists, for example, when more than 50 % of the voting rights are held by another legal unit. This controlling legal unit could also be a natural person. All legal units that are directly or indirectly controlled by a top-level legal unit – the global group head – together form an enterprise group.

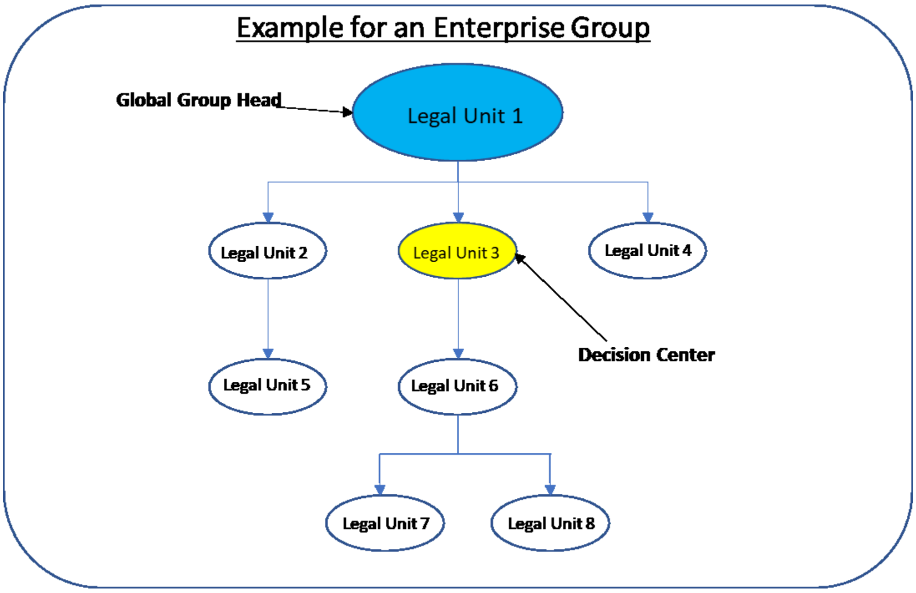


Figure 1. Example for an enterprise group with 8 legal units and the roles global group head and decision center.

An enterprise group in a statistical business register often has some specific characteristics and attributes, e.g. turnover, employees, NACE-Code and so on. These attributes may differ on a global, regional or national level. There can also be specific roles within the enterprise group like the global group head or a global, regional or national decision centre, which is often the legal unit to which the head office of the group or the region belongs. In the landscape of statistical business registers, there are different specifications for the need of attributes and roles.

Both in the production processes and in the assessment processes, e.g. for publications, it is necessary to compare the data on enterprise groups at different points in time and from different registers. In this paper, the comparison of two enterprise groups is first discussed theoretically in chapter two and based on this the new similarity metric for comparison of enterprise groups will be defined. Chapter three will present two use cases of the similarity metric in Germany. Further possible use cases are indicated in the conclusion and outlook in chapter four and should be discussed further in the 28th Meeting of the Wiesbaden Group on Business Registers in The Hague.

2 Comparison of two enterprise groups

For the comparison of thousands or millions of enterprise groups it is easy to identify enterprise groups, which have exactly the same structure and number of the same legal units. In practice there are often not exactly the same structures, because the structures have actually changed or the data sources are of different quality. For combining different data sources or different points in time, rules must be defined as to which enterprise groups are supposed to be the same and which are not. Earlier approaches worked by simple hypotheses like the identity of the global group head and/or the size of the groups. We worked out more sophisticated ways to compare enterprise groups.

Figure 2 shows two enterprise groups from different data sources. In enterprise group B, the structure of five of the legal units is the same as in enterprise group A, but three units are missing compared to enterprise group A. Is this still the same enterprise group?

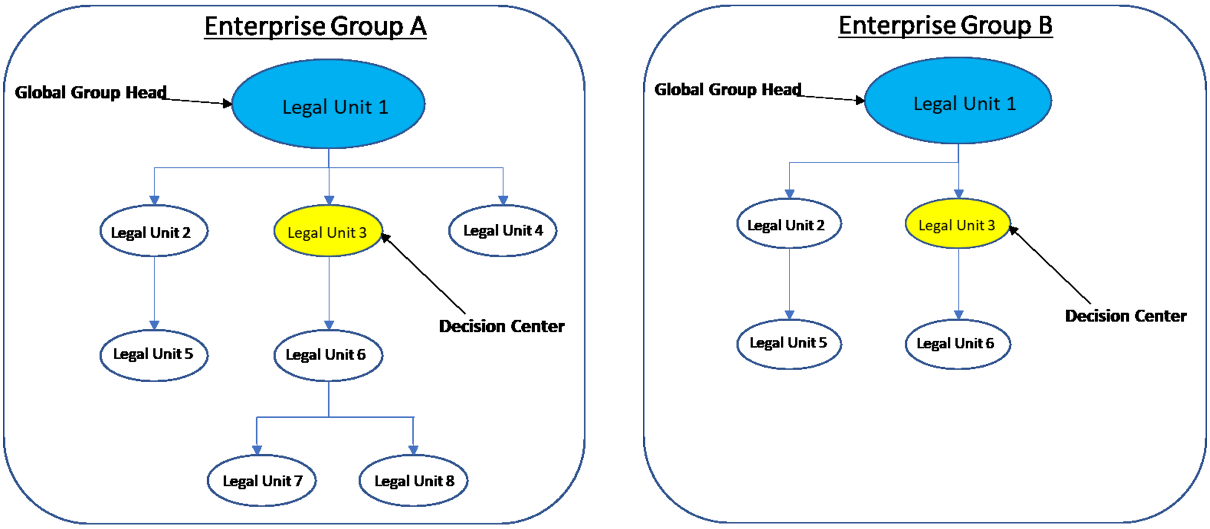


Figure 2: Example for two enterprise groups with similar legal units

To answer this question in a rule-based way, it is necessary to describe and evaluate the similarity of these groups.

For this purpose, various possible solutions are first described and then the idea of a similarity metric is described in a theoretic manner to define the mathematic formula for the similarity metric by Rommelspacher and Urban.

2.1 Challenges and possibilities

In order to describe the possibilities of comparing data at different times and/or in different registers, it is always necessary first to look at which characteristics are available in both sets of data. It is very helpful if unique identifiers for legal units are available to easily match the legal units. In our use cases we always had these unique identifiers. Otherwise, additional process steps must be taken to match the legal units as good as possible. This matching is necessary in order to use our similarity metric for comparing enterprise groups, as it is the only way to compare the legal units belonging to a group.

Economic characteristics can be helpful in the comparison to better reflect the economic importance of the units within the group. Statistical business registers often contain data on employment and turnover. These values can be very different for legal units in a large enterprise group.

The direct relationships between the legal units and the correspondence of all relationships within a group could also be used to calculate similarity. Another possibility would be to compare the roles that some legal units fulfil within the group like the global group head or the global or national decision centre. Therefore, it could be assumed that the similarity of two groups is higher if they have the same global group head. All of these characteristics could be compared, and based on this we considered how much similarity there is between two groups. Calculating and comparing this was the big goal in developing a similarity metric.

2.2 Idea of a similarity metric

Many ideas and use cases for similarity and distance metrics exist in the literature e.g. for the similarity of character strings or whole documents.[1] However, these measures can also be used to compare more complex objects, such as faces, plants or genes, and to quantify and compare (in)similarity.[2]

Typically, distance metrics describe the dissimilarity between two objects and can be easily converted into a similarity metric. A distance metric d(x, y) tries to measure a distance, i.e. how far you have to go from x to y, or how much you have to change on object x to get y. The distance measure is 0 for two identical objects and the greater the difference, the greater the calculated distance. A non-normalised distance metric has no maximum value. A distance metric d(x, y) is a normalised distance metric if $d(x, y) \leq 1$.[3]

A normalised similarity metric s(x, y) is also always between the values 0 and 1, such that $0 \geq s(x, y) \geq 1$. This is done by taking the intersection of the objects x and y or 1-d(x, y) and dividing this by the total number of feature descriptions. The greater the similarity measure s(x, y), the more similarities the objects x and y have. If the objects differ in all characteristics under consideration, the similarity measure should be 0. If the considered characteristics of the objects x and y are exactly identical, s(x, y)=1.

2.3 Similarity metric for comparison of enterprise groups by Rommelspacher and Urban (RUMS)

To quantify the similarity of enterprise groups in different registers or at different points in time, the aim was to define a normalised similarity metric to make the similarities between several enterprise groups comparable. We started to compare the legal units (LEU) in enterprise group A with the legal units in enterprise group B:

$$Similarity(A(LEU), B(LEU)) = \frac{|A(LEU) \cap B(LEU)|}{|A(LEU)|}$$

In this way, examples 1 and 2 in figure 3 show that legal units that are not in group B but in group A are considered in the calculation, but legal units that are in group B but not in group A are not considered in the calculation.

[1] Härdle (2003).

[2] Rogers (1960).

[3] Chen (2009).

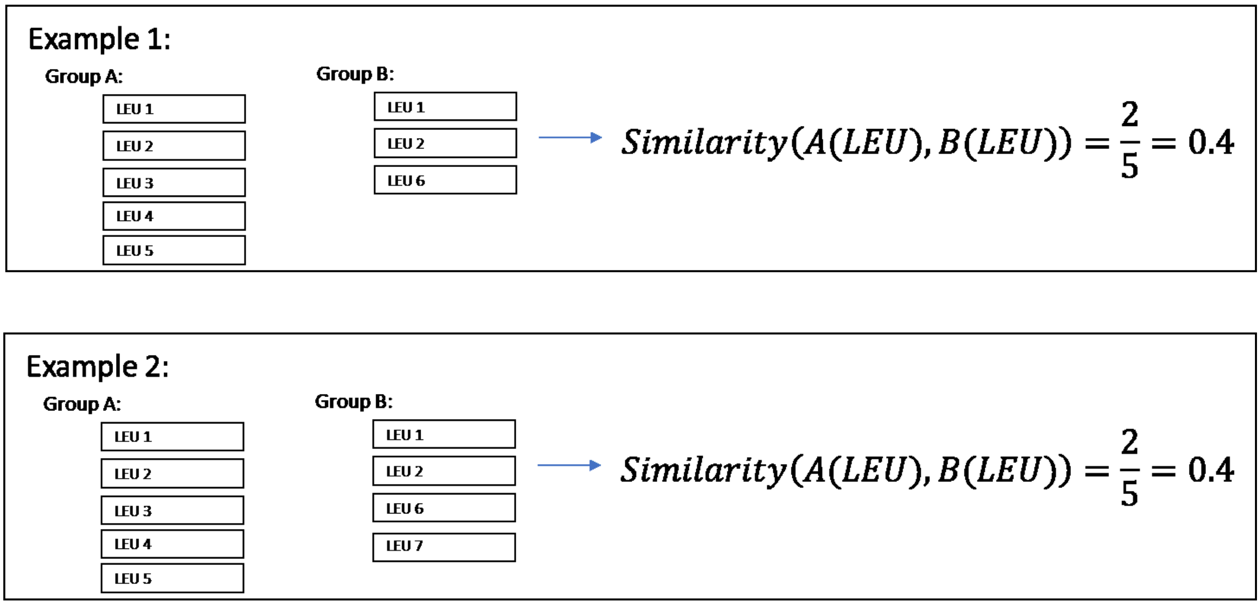


Figure 3. Examples for the first similarity metric

The difference between group A and B is actually greater in example 2, yet the similarity measure is the same. In order to also take into account the units that are in B but not in A when calculating similarity, we need to add another term in which the intersection is divided by the total amount of units in group B. The two terms are first weighted equally with the factor 0.5, so that the result can again only be between 0 and 1.

$$Similarity2(A(LEU), B(LEU)) = 0.5 * \frac{|A(LEU) \cap B(LEU)|}{|A(LEU)|} + 0.5 * \frac{|A(LEU) \cap B(LEU)|}{|B(LEU)|}$$

Figure 4 shows that the similarity metric is now greater for example 1 than for example 2 because of the difference in group B with LEU 7.

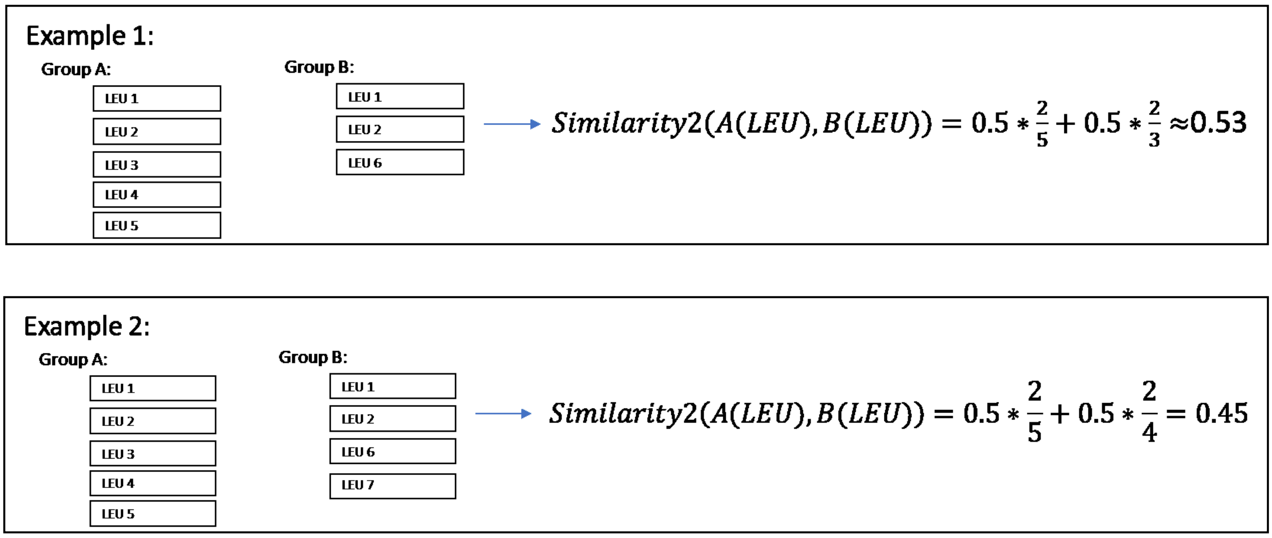
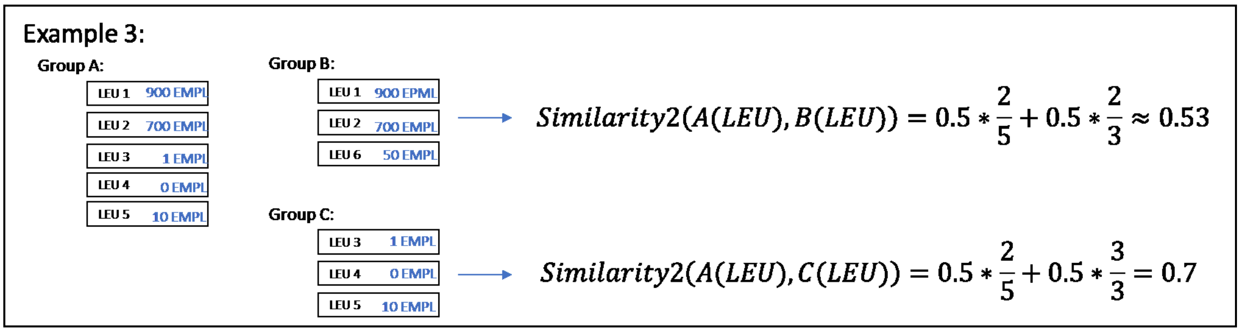


Figure 4. Examples for the second similarity metric

With this similarity metric, any number of enterprise groups from two registers or points in time can be compared with each other with regard to the overlap of legal units within the groups. Assuming that a source is considered to be of better quality, it can be rated higher by setting the factors before the two terms other than 0.5. However, the sum of the factors must add up to 1.

With this approach, however, it can happen that a small value is calculated when comparing two enterprise groups, although the economically most relevant legal units overlap. In Figure 5, the similarity between groups A and C is higher than between A and B, although the most relevant legal units 1 and 2 overlap between groups A and B and not C.



$$Similarity2(A(LEU),B(LEU)) = 0.5 * \frac{2}{5} + 0.5 * \frac{2}{3} \approx 0.53$$

$$Similarity2(A(LEU),C(LEU)) = 0.5 * \frac{2}{5} + 0.5 * \frac{3}{3} = 0.7$$

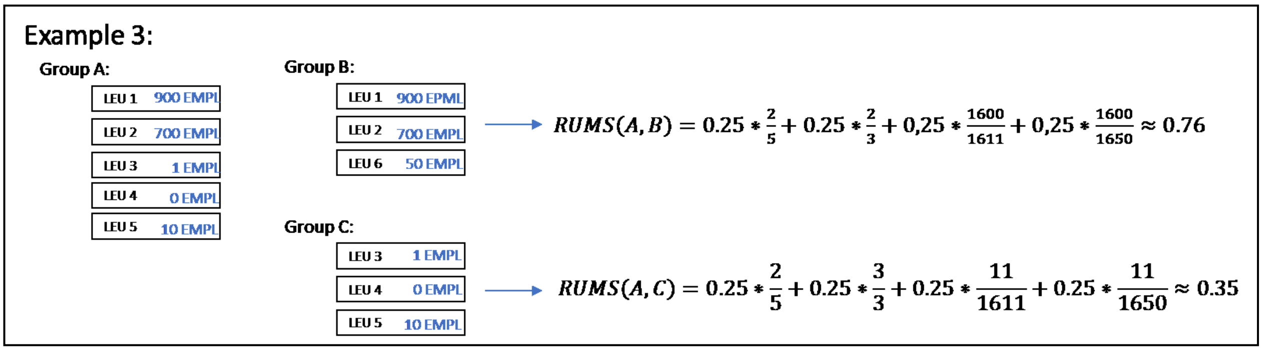
Figure 5. Example of the second similarity metric with three enterprise groups

To take this economic relevance in terms of employment into account, we have included the employees of the overlapping legal units in the formula for calculating the similarity metric for comparing enterprise groups. We call it RUMS ("**R**ommelspacher-**U**rban **M**etric for the **S**imilarity") of two enterprise groups in statistical business registers:

$$RUMS(A,B) = \begin{cases} a * \frac{|A(LEU) \cap B(LEU)|}{|A(LEU)|} + b * \frac{|A(LEU) \cap B(LEU)|}{|B(LEU)|} + c * \frac{\sum Employees(A(LEU) \cap B(LEU))}{\sum Employees(A)} + d * \frac{\sum Employees(A(LEU) \cap B(LEU))}{\sum Employees(B)} & , \sum Employees(A) > 0 \text{ and} \sum Employees(B) > 0 \\ 2a * \frac{|A(LEU) \cap B(LEU)|}{|A(LEU)|} + 2b * \frac{|A(LEU) \cap B(LEU)|}{|B(LEU)|}, & \sum Employees(A) = 0 \text{ or} \sum Employees(B) = 0 \end{cases}$$

$$\text{Under the constraints: } a + b + c + d = 1 \text{ und } \quad 2a + 2b = 1$$

Parameters a, b, c and d can be used to flexibly adjust both the weighting of sources or points in time and the weighting of employment information. With equal weighting, the following RUMS values result, for example 3.



$$RUMS(A,B) = 0.25 * \frac{2}{5} + 0.25 * \frac{2}{3} + 0{,}25 * \frac{1600}{1611} + 0{,}25 * \frac{1600}{1650} \approx 0.76$$

$$RUMS(A,C) = 0.25 * \frac{2}{5} + 0.25 * \frac{3}{3} + 0.25 * \frac{11}{1611} + 0.25 * \frac{11}{1650} \approx 0.35$$

Figure 6.Example of the RUMS-Values with three enterprise groups

As the employees of the overlapping legal units are considered in the RUMS in figure 6, the similarity between enterprise groups A and B is rated significantly higher than that between A and C. Since the data quality and consistency in terms of direct relationship information and roles was considered to be lower in the use cases presented in Chapter 3, these characteristics were not initially considered in RUMS.

3 Use cases

In the German business register, data on enterprise groups are processed and updated annually for a reference year. Thus, when comparing enterprise groups from one year to the next, the question arises as to which group is to be matched to which group from the previous year, and when comparing the national register and the EGR, the question arises as to which group in the EGR is to be matched to which group in the national register.

3.1 Comparison of enterprise groups at different points in time

For the annual processing of data on enterprise groups, a decision must be made for each existing enterprise group in the business register as to whether this enterprise group is to be continued and with which new information. For this purpose, the RUMS value is calculated for all enterprise groups from the two points in time in order to determine the similarity between all groups. The parameters a, b, c and d can be set individually. We first started with an equal weighting. The enterprise groups at the two different points in time can now be divided into three populations.

Enterprise groups that are exactly identical at both points in time have the RUMS-value=1 and here the groups are always clearly in continuity. In figure 7, this is the light red part. The yellow part of figure 7 shows enterprise groups that have no similarity to any of the enterprise groups in the other reporting year, i.e. the RUMS-value is 0. These are groups that are either completely new or no longer exist in the current reference year.

The tricky enterprise groups are in the blue area. These have a varying degree of similarity with one or more enterprise groups from the previous reference year, the RUMS gives values between 0 and 1. There is also the question of how great a similarity must be for allowing the assumption that this is the same enterprise group and that the group from the previous year should be continued. In order to better resolve these two challenges, we have taken an iterative approach. For example, in the first step, only those enterprise groups with a RUMS>0.9 were considered and continued. If there are similarities of one group to more than one other group, the higher RUMS-value is used. At each step, the limit for the RUMS-value was lowered and random checks were made to ensure that the results were still appropriate, and we ended at RUMS>0.3. For all enterprise groups with a RUMS-value lower than 0.3, we decided that these were new groups and not a continuation of the previous year's groups.

With the help of the RUMS, the annual processing of new information could be automated to a large extent and many enterprise groups could be continued in a reasonable way, meaning that they keep the same identification number over time in the German business register. Furthermore, the similarity metric can be an important indicator of either real changes between different reference years or differences in the quality of the data between different reference years. To find out the real causes, the manual treatment of enterprise groups[4] can be focused on these most important constellations in order to efficiently improve data quality.

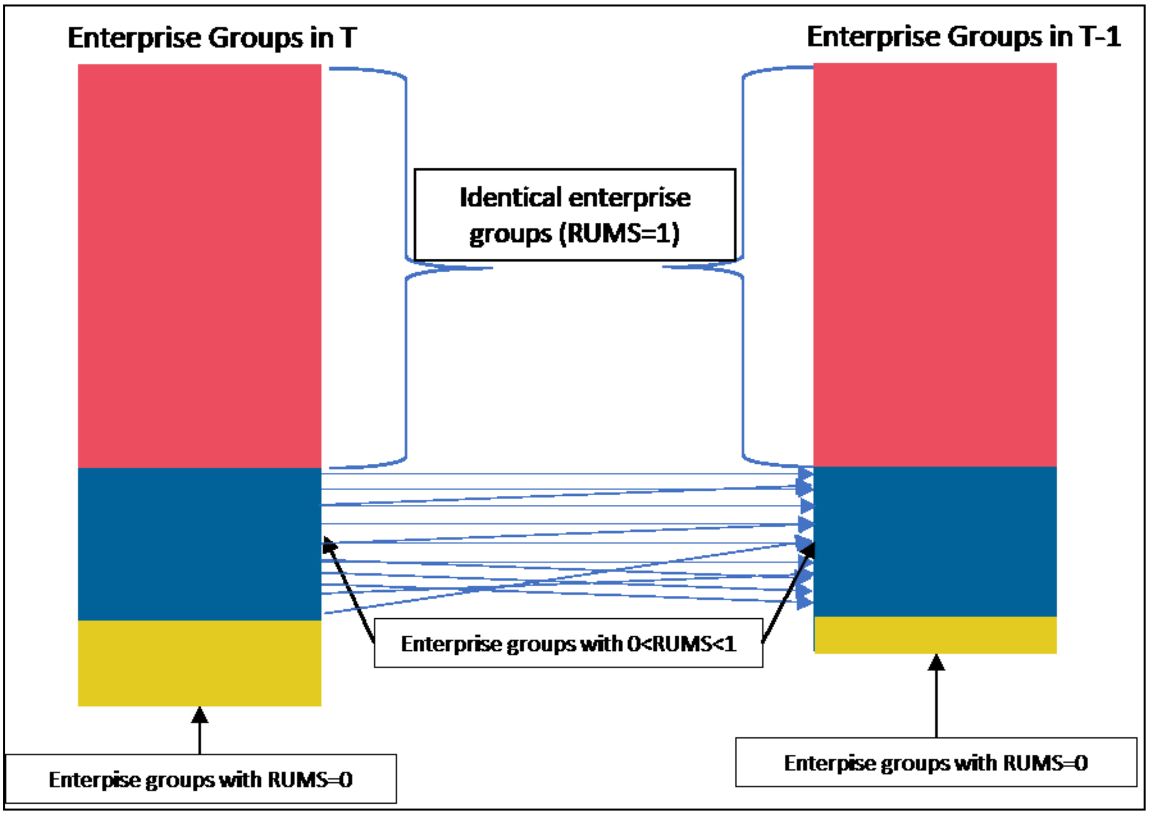

Enterprise Groups in T
Enterprise Groups in T-1
Identical enterprise groups (RUMS=1)
Enterprise groups with 0<RUMS<1
Enterpise groups with RUMS=0
Enterpise groups with RUMS=0

Figure 7. Comparison of enterprise groups at different points in time

3.2 Comparison of enterprise groups from different data sources

In a second use case, we compared the final frame of the EuroGroups Register with the data on enterprise groups in the national statistical business register (NSBR) in Germany of the same reference year. We filtered both registers to multinational enterprise groups with at least one legal unit in Germany. Again, we have equally weighted the parameters a, b, c and d and calculated a similarity for all enterprise groups from the NSBR to all enterprise groups in the EGR.

As a result, we have a large number of enterprise groups that are identical in the EGR and the NSBR. In addition, there are enterprise groups in both the NSBR and EGR that do not exist in the other register, so the RUMS=0. And there are enterprise groups that are in the NSBR and look similar in the EGR. Figure 8 shows an example of these populations in the colours red, yellow and blue.

[4] Depire et al. (2023).

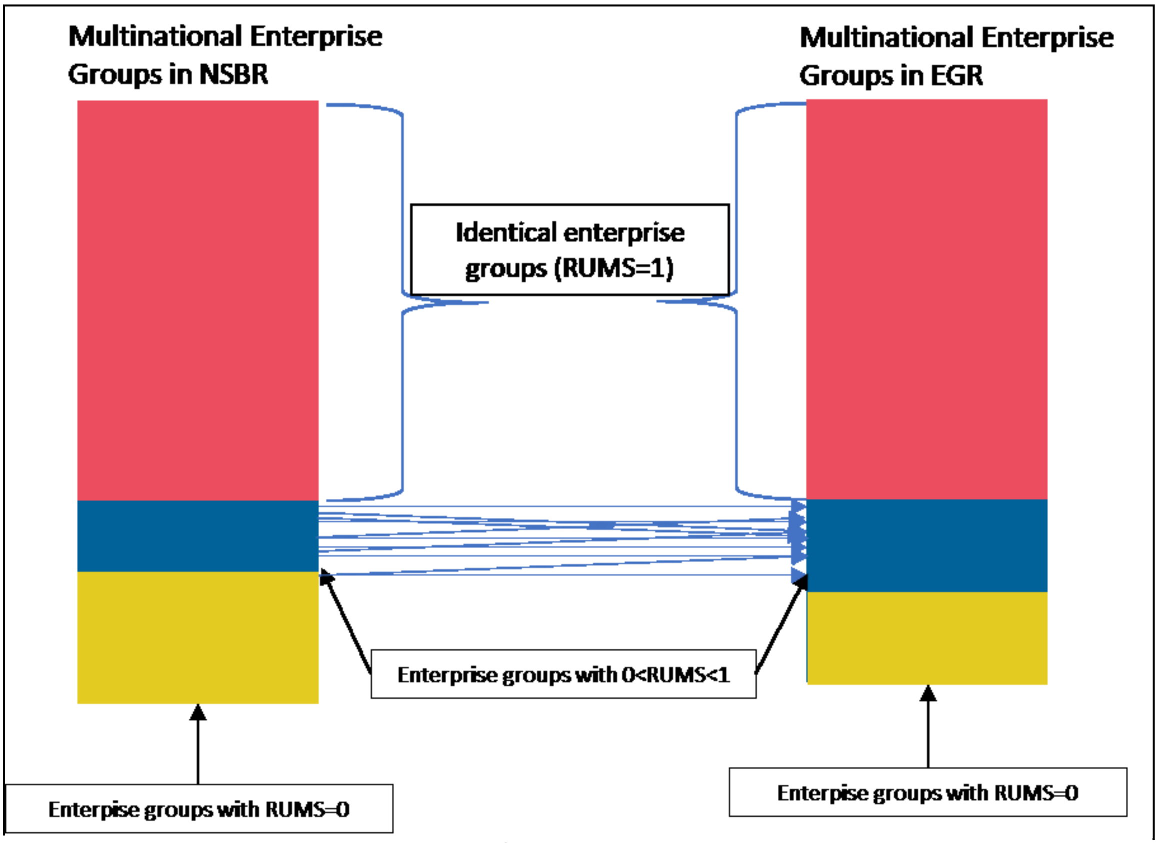

Multinational Enterprise Groups in NSBR
Multinational Enterprise Groups in EGR
Identical enterprise groups (RUMS=1)
Enterprise groups with 0<RUMS<1
Enterpise groups with RUMS=0
Enterpise groups with RUMS=0

Figure 8. Comparison of enterprise groups in NSBR and EGR

The enterprise groups in the blue and yellow areas are to be analysed more closely, as they contain different information for which there may be various reasons. Here, methodological differences as well as technical limitations and also different reference times of the information can be a reason. In a perfect world, the information in the EGR and NSBR would always be identical at the end of a working cycle, so all groups would have a RUMS-value of 1. The RUMS can help to identify and evaluate these differences so that the causes can be found and, if possible, remedied on the way to a perfect register world.

4 Outlook

The two use cases showed that the RUMS can be used as a similarity metric for comparing enterprise groups in statistical business registers and can be useful for analysing very large populations of enterprise groups. Both to establish rules for automated data processing and to highlight critical cases that require special attention.

The RUMS model can be further developed so that more characteristics of enterprise groups, such as the roles in the enterprise groups, the direct relationships between the legal units or additional economic aspects like turnover could be considered in the calculation of similarity. An experimental approach could also be taken to develop similarity metrics that do not refer to the identical units in enterprise groups, but compare the similarity of other characteristics. For example, industry comparisons or patterns in certain types of enterprise groups could be identified.

Whether this is helpful, and what weighting of the different characteristics is most appropriate, depends on the data available and also needs to be tried and evaluated through more use cases. At the 28th meeting of the Wiesbaden Group in The Hague, we would like to raise the question whether other National Registers have use cases or also other approaches to compare data on enterprise groups.

References

[1] Härdle, Wolfgang; Simar, Léopold (2003). Applied Multivariate Statistical Analysis. Berlin; Springer Verlag, p. 381.

[2] Rogers, David, J.; Tanimoto, Taffee. T. (1960). A Computer Program for Classifying Plants. American Association for the Advancement of Science, p.1115-1118.

[3] Chen, Shiyen; Ma, Bin; Zhang, Kaizhong (2009). On the similarity metric and the distance metric. Ontario, Canada; Theoretical Computer Science Vol. 410, p. 2365-2376.

[4] Depire, Alexandre; Sopranidis, Ioannis; Rommelspacher, Simon (2023): A Top-tier approach for the largest Multinational Enterprise Groups in the EU; 28th Meeting of the Wiesbaden Group on Business Registers, 02 – 06 October 2023, The Hague, Session Profiling complex Statistical Units.