ECONOMETRICS. EKONOMETRIA
Advances in Applied Data Analysis
Year 2019, Vol. 23, No. 3
ISSN 1507-3866; e-ISSN 2449-9994

INTRA-COMMUNITY TRADE ASYMMETRIES-BASED CLUSTERING AND LINEAR ORDERING OF COMBINED NOMENCLATURE CHAPTERS USING GENERALIZED DISTANCE MEASURE (GDM)

Iwona Markowicz
University of Szczecin, Szczecin, Poland
e-mail: iwona.markowicz@usz.edu.pl

ORCID: 0000-0003-1119-0789

Paweł Baran
University of Szczecin, Szczecin, Poland
e-mail: pawel.baran@usz.edu.pl

ORCID: 0000-0002-7687-4041

DOI: 10.15611/eada.2019.3.05
JEL Classification: C10, C19, F14

Abstract: Statistical data on foreign trade are collected in all EU member states separately and then passed on to Eurostat where the data are aggregated. Continuous actions are to ensure that all datasets collected at national level are fully comparable. The aim of the paper is to provide a classification as well as an ordering of CN chapters (2-digit codes) according to the quality of data on intra-Community trade of goods. Data were taken from Eurostat's COMEXT database. In ordering the chapters, we utilized the distance from the ideal solution with GDM as the distance measure. The study reveals a structure of goods subject to intra-Community trade that is supplementary to the official nomenclature. In addition, we provided CN chapters ordering according to the overall level of irregularities in reported mirror values of ICS and ICA. The results we obtained are of practical value for both researchers and authorities interested in foreign trade.

Keywords: trade in goods within the EU, Intrastat, GDM, linear ordering, CN chapters.

1. Introduction

The necessity of introducing a new system of public statistics for intra-Community trade became clear right after the creation of the common European market, which involved revoking all customs activities and reporting in trade between the EU member states. Since data on foreign trade was no longer reported through SAD's (single administrative documents, i.e. customs declarations), the EU and its member states needed another source of information on foreign trade between them. Intrastat as a system of public statistics started operating on January 1st, 1993 (in Poland it was first introduced on May 1st, 2004, i.e. the day Poland accessed the EU). Statistical data on foreign trade are collected in all member states separately and then passed on to Eurostat where the data are aggregated [Baran, Markowicz 2018]. Data declared by both sides of an intra-Community transaction should be reflected in their respective national datasets as an intra-Community supply on one hand and an intra-Community acquisition on the other, hence called mirror data. However there are numerous reasons resulting in the differences between mirror data. Some of the reasons are plain errors, like the misclassification of goods, some others are the results of certain characteristics of methodology applied and are hard to avoid, e.g. implementing statistical thresholds, simplified reporting, and exchange rates influencing the reported values. The above mentioned differences are often referred to as asymmetries in official statistics. Continuous actions are undertaken in order to assess the quality of data and to ensure that all datasets collected at national level are fully comparable and that asymmetries are minimized. These actions include data monitoring conducted both by member states' statistical offices as well as by Eurostat with the use of the MOD (mirror outlier detection) method. There also exists a system of bilateral reconciliation studies (rounds) conducted by countries for which huge inequalities in mirror data are detected [Eurostat 2017; GUS 2018]. Analyzing mirror data in international trade has been scope of numerous research papers recently [Carrère, Grigoriou 2014; Javorsek 2016; Markowicz, Baran 2019], but the use of multivariate statistical methods is rare. Thus the scope of the presented work is new to this subject.

The aim of the paper is to provide a classification as well as an ordering of combined nomenclature (CN) chapters (2-digit codes containing wide groups of goods) according to the quality of data on intra-Community trade of goods. Data on intra-Community trade in 2016 were taken from Eurostat's COMEXT database. We consider intra-Community supplies' (ICS) and intra-Community acquisitions' (ICA) values in distinct CN chapters as separate variables recorded by country. Data on intra-Community supplies of a certain country was compared with the mirror data aggregated from partner countries' statistics on intra-Community acquisitions from that country. The same procedure was applied to intra-Community acquisitions and their mirror data. In ordering the chapters, we utilized the distance from the ideal solution with GDM as the distance measure. We obtained groups of CN chapters with similar quality of data measured with scaled asymmetry. The study reveals a structure of goods subject to intra-Community trade that is supplementary to the official terminology. In addition, we provided CN chapters ordering according to the overall level of irregularities in reported mirror values of ICS and ICA measured with aggregate asymmetry. The results we obtained are of great practical value for those searching for and analyzing anomalies in foreign trade; they indicate markets similar to one another according to irregularities. They can also help authorities to prioritize areas of data control.

2. Data and methodology

We used data from the COMEXT database provided by Eurostat. The data contained all intra-Community trade aggregated by CN chapter and country divided into parts containing intra-Community supplies and intra-Community acquisitions from and to all member states in 2016. Such raw data were transformed into the differences between ICS and their aggregated mirror ICA and vice-versa (asymmetry measures) for all chapters in every member state. Then we normalized the differences with a quotient transformation – for clustering we preserved the signs of individual differences in formula (1) and for ordering we stripped them (2), as follows

$$z_{ij} = \frac{x_{ij}}{\max\limits_i \left| x_{ij} \right|} \tag{1}$$

and

$$z_{ij} = \frac{\left| x_{ij} \right|}{\max\limits_i \left| x_{ij} \right|}, \tag{2}$$

where: x_{ij} – difference between ICS and mirror ICA (difference between ICA and mirror ICS in the second part of the survey); i – number of CN chapter; j – number of EU member state.

The sign of the quotient matters in cases of clustering, differences of the same magnitude but differing in sign are by no means similar, thus we chose formula (1) in that case. On the other hand, the ordering of chapters is based on aggregated differences so in that case we need them to be only positive (otherwise the sum of the differences would not be calculated properly) and that is the reason for choosing formula (2) .

Next we used such prepared data to classify CN chapters and to prepare the ranking (ordering). Classification was carried out using a hierarchical algorithm. We used Ward's method of clustering and the similarity measure was GDM1 provided by the clusterSim package in R [Walesiak, Dudek 2019], given by [Walesiak 2016]

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{m}\left(z_{ij} - z_{kj}\right)\left(z_{kj} - z_{ij}\right) + \sum_{j=1}^{m}\sum_{l=1, l\neq i,k}^{n}(z_{ij} - z_{lj})(z_{kj} - z_{lj})}{2\left[\sum_{j=1}^{m}\sum_{i=1}^{n}\left(z_{ij} - z_{lj}\right)^{2} \cdot \sum_{j=1}^{m}\sum_{l=1}^{n}\left(z_{kj} - z_{lj}\right)^{2}\right]^{1/2}}, \tag{3}$$

where: x_{ij} (x_{kj}, x_{lj}) – are i-th, k-th, and l-th observations (here: chapters) of j-th variable (country).

Later we provided a ranking (linear ordering) of CN chapters data prepared as in (2). We did this by finding the distances from all chapters to the ideal chapter (the ideal solution i.e. a vector containing the least differences between individual pairs of countries) with the use of `pattern.GDM1` function from the clusterSim package.

We used variables measured on the interval scale (the differences between mirror data) and we scaled them in a manner similar to ratio transformation (producing ratios of differences is generally permissible). According to Walesiak it is possible to choose between Euclidean, city, Chebyshev, and GDM1 distance measures in our case [Walesiak 2016]. We chose GDM1 from the clusterSim package.

3. Results

The results of the survey are divided into four parts: classification and ordering of chapters according to ICS data quality, then classification and ordering of chapters according to ICA data quality.

3.1. Clustering CN chapters (ICS-based)

We performed hierarchical clustering of CN chapters according to ICS data quality. The agglomerative clustering was performed with Ward's method and we used GDM1 as the distance measure.

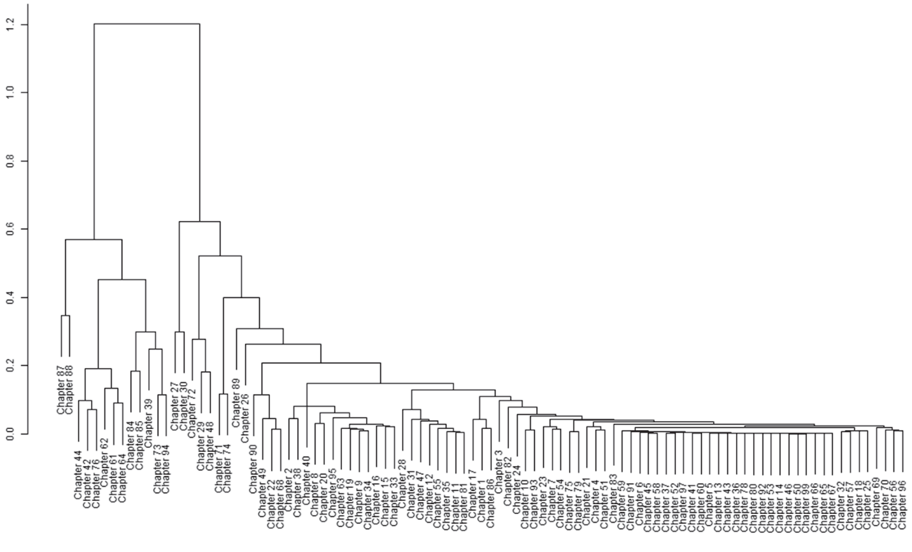

Fig. 1. Clustering of CN chapters based on ICS data quality

Source: own elaboration.

The most natural way of clustering nomenclature chapters seems to be dividing them into two groups, however such a clustering does not look very interesting for a researcher. Instead, exercising our survey's goal we can point out four to six groups that slightly differ from one another. The grouping with six distinct clusters obtained is presented in Table 1.

Table 1. Nomenclature chapters divided into six groups according to ICS data quality

Group I	
Chapter 87	Vehicles other than railway or tramway rolling stock, and parts (…) thereof
Chapter 88	Aircraft, spacecraft, and parts thereof
Group II	
Chapter 42	Articles ofeather
Chapter 44	Wood and articles of wood; wood charcoal
Chapter 61	Articles of apparel and clothing accessories, knitted or crocheted
Chapter 62	Articles of apparel and clothing accessories, not knitted or crocheted
Chapter 64	Footwear,gaiters and the like; parts of such articles
Chapter 76	Aluminium and articles thereof
Group III	
Chapter 39	Plastics and articles thereof
Chapter 73	Articles of iron or steel
Chapter 84	Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof
Chapter 85	Electrical machinery and equipment and parts thereof
Chapter 94	Furniture; bedding, mattresses, mattress supports, cushions and similar stuffed furnishings; lamps and lighting fittings
Group IV	
Chapter 27	Mineral fuels, mineral oils and products of their distillation
Chapter 30	Pharmaceutical products
Group V	
Chapter 29	Organic chemicals
Chapter 48	Paper and paperboard; articles of paper pulp, of paper or of paperboard
Chapter 72	Iron and steel
Group VI – remaining 79 CN chapters	

Source: own elaboration.

Analysing the clusters we observe that similar structures according to data quality are often characteristic to CN chapters containing goods of a similar nature. This could mean that the markets for these products are also similar and they clearly differ from markets for different product groups. It is however easy to spot that there exist intriguing deviations from that general rule, e.g. we find chapter 76 (articles of aluminium) among leather, wood and clothing or chapter 72 (iron, steel) among chemicals and paper. We need to mention that probably due to the huge number of attributes and objects (which are aggregates and do not form a homogenous group) many chapters are finally similar to others in the sense that they construct one large cluster containing elements of random data quality.

3.2. Ranking CN chapters (ICS-based)

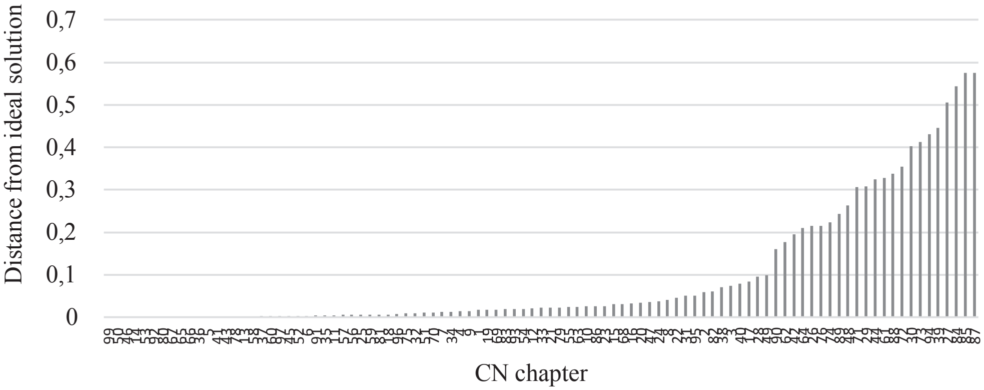

Fig. 2. Ranking of CN chapters based on ICS data quality

Source: own elaboration.

Most of the CN chapters can be described as having high quality mirror data regarding trade in goods between the EU member states. However there exists a large group of chapters, including those with a great turnover, in which data quality is not satisfactory. These are such chapters as: 27 (mineral fuels and oils), 84 (machinery and mechanical appliances), 85 (electrical machinery and equipment), 87 (vehicles, e.g. cars).

Public statistics services collecting data on international trade within the EU should focus on the coordination of data in the above mentioned chapters. For a market analyst this is most interesting part of the whole dataset (some questions are arising: What are the causes of discrepancies? Are they systematic? What are the consequences for the macroeconomic indices?). So these chapters should be examined in the first place although the results of such analysis must be interpreted with caution.

3.3. Clustering CN chapters (ICA-based)

The situation within intra-Community acquisition data is slightly different. The structure of the CN chapters has its own characteristics here as well.

The division into three groups is natural here. We assume that it is adequate according to the aim of the study. Members of the groups are listed in Table 2.

It is interesting (and perhaps symptomatic for some general rule) that also in the analysis of intra-Community acquisitions data we can still observe that some of the chapters containing industrial products are a distinct group that really differs from other chapters, although there is no common wider category to which they belong. Another group observed in both parts of the survey is the group containing chapters 26 and 30, i.e. products of petrochemical industry and pharmaceuticals.

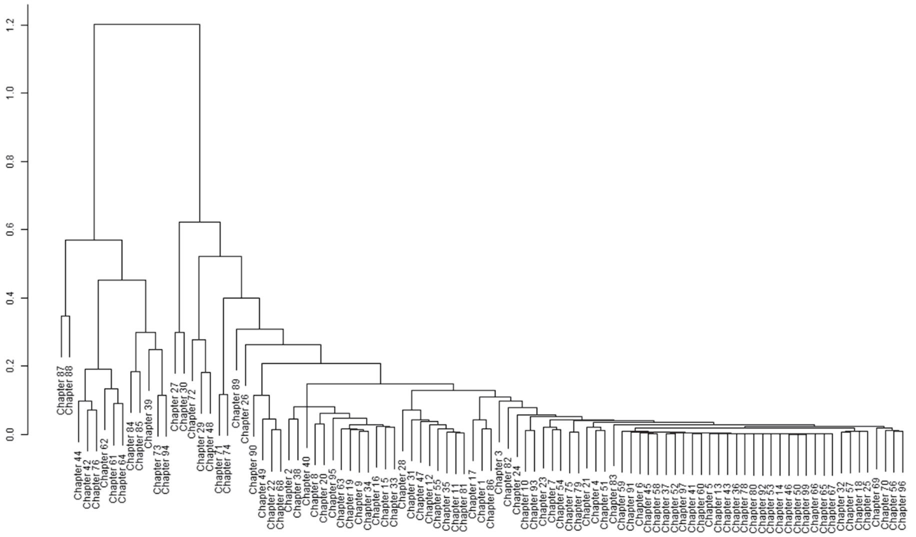

Fig. 3. Clustering of CN chapters based on ICA data quality

Source: own elaboration.

Table 2. Nomenclature chapters divided into three groups according to ICA data quality

Group I	
Chapter 39	Plastics and articles thereof
Chapter 62	Articles of apparel and clothing accessories, not knitted or crocheted
Chapter 73	Articles of iron or steel
Chapter 84	Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof
Chapter 85	Electrical machinery and equipment and parts thereof
Chapter 87	Vehicles other than railway or tramway rolling stock, and parts (…) thereof
Chapter 88	Aircraft, spacecraft, and parts thereof
Chapter 89	Ships, boats and floating structures
Chapter 90	Optical, photographic, cinematographic, measuring, checking, precision, medical or surgical instruments and apparatus; parts and accessories thereof
Chapter 94	Furniture; bedding, mattresses, mattress supports, cushions and similar stuffed furnishings; lamps and lighting fittings
Group II	
Chapter 27	Mineral fuels, mineral oils and products of their distillation
Chapter 30	Pharmaceutical products
Group III – remaining 85 CN chapters	

Source: own elaboration.

3.4. Ranking CN chapters (ICA-based)

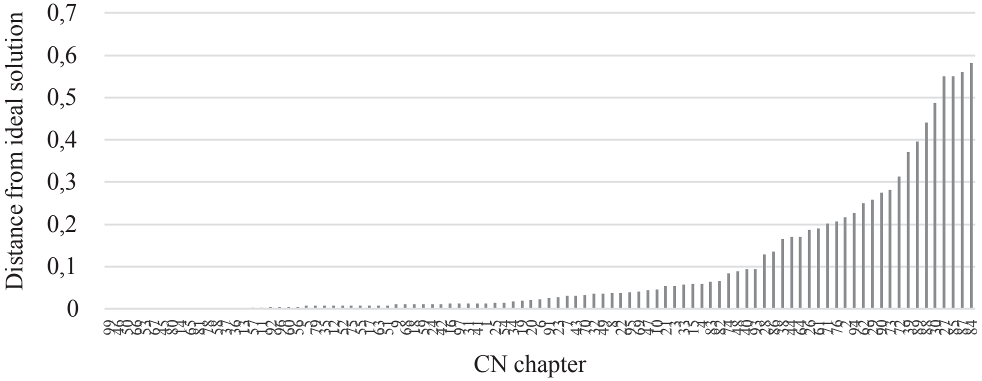

Fig. 4. Ranking of CN chapters based on ICA data quality

Source: own elaboration.

By ordering the chapters according to distance from the ideal solution we can observe that again most of the chapters are characterized by the generally good quality of data and that chapters of low data quality are relatively few. These are mostly the same chapters that we have listed above in the part concerning ICS (the four chapters on the far right are the same as in point 3.2 above: fuels, mechanical and electrical machinery, cars).

4. Conclusions

The classification for both ICS and ICA suggests which CN chapters are most similar according to the structure of differences between aggregated supplies and mirror acquisitions and vice-versa. Clusters obtained in the survey are characterized by their similar influence on data quality in different EU member states.

The ranking of CN chapters shows which chapters have the greatest shares in generating overall asymmetry between declarations and their mirror data. This means that these chapters should be subject to the specific attention of statistical offices collecting foreign trade data in European countries.

The results we obtained are also valuable for practical reasons – they can be useful in anomaly detection in foreign trade data analysis. They help in spotting markets of similar goods according to supposed anomalies (e.g. late reporting or no reporting). They can also help to determine priority areas of control.

Bibliography

Baran P., Markowicz I., 2018, *Analysis of intra-Community supply of goods shipped from Poland*, Socio-Economic Modelling and Forecasting No. 1, The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings, M. Papież, S. Śmiech (eds.), Zakopane, pp. 12-21.

Carrère C., Grigoriou Ch., 2014, *Can mirror data help to capture informal international trade?,* Policy issues in international trade and commodities research study series No. 65, UNCTAD, New York.

Eurostat, 2017, *Compilers guide on European statistics on international trade in goods*, 2017 edition, Manuals and Guidelines, Publications Office of the European Union, Luxembourg.

GUS, 2018, *Handel zagraniczny. Statystyka lustrzana i statystyka asymetrii. Foreign trade. Mirror and Asymmetry Statistics*, Warszawa.

Javorsek M., 2016, *Asymmetries in International Merchandise Trade Statistics: a Case Study of selected countries in Asia-Pacific*, UN ESCAP Statistics Division, Working Paper Series SD/WP/02/ April 2016.

Markowicz I., Baran P., 2019, *Jakość danych dotyczących wewnątrzunijnej wymiany towarowej*, Wiadomości Statystyczne. The Polish Statistician, 6(64), pp. 5-15.

Walesiak M., 2016, *Uogólniona miara odległości GDM w statystycznej analizie wielowymiarowej z wykorzystaniem programu R*, Wydawnictwo UE we Wrocławiu, Wrocław.

Walesiak M., Dudek A., 2019, *clusterSim*: *Searching for Optimal Clustering Procedure for a Data Set*., R package version 0.47-4. https://CRAN.R-project.org/package=clusterSim.

KLASYFIKACJA I PORZĄDKOWANIE LINIOWE DZIAŁÓW NOMENKLATURY SCALONEJ NA PODSTAWIE ASYMETRII W HANDLU WEWNĄTRZWSPÓLNOTOWYM Z WYKORZYSTANIEM UOGÓLNIONEJ MIARY ODLEGŁOŚCI (GDM)

Streszczenie: Dane statystyczne handlu zagranicznego zbierane są w każdym z krajów członkowskich Unii osobno, a następnie przekazywane do Eurostatu, gdzie są agregowane. Wciąż prowadzone są działania mające na celu zapewnienie pełnej porównywalności danych gromadzonych na poziomie krajowym.Celem artykułu jest przeprowadzenie klasyfikacji oraz porządkowania działów nomenklatury scalonej (CN, klasyfikacja towarowa, kody 2-znakowe) według jakości danych nt. wewnątrzwspólnotowej wymiany towarowej. Dane pobrano z udostępnianej przez Eurostat bazy COMEXT. Do uporządkowania działów użyto odległości od wzorca, danej miarą GDM. Badanie ujawniło strukturę towarową w handlu wewnątrzwspólnotowym, którą można uznać za uzupełniającą wobec oficjalnej klasyfikacji. Dodatkowo uporządkowano działy CN według łącznego poziomu nierównomierności w deklarowanych lustrzanych wartościach WDT i WNT. Uzyskane wyniki mają praktyczne znaczenie dla badaczy, ale również dla władz zajmujących się handlem zagranicznym.

Słowa kluczowe: handel wewnątrzwspólnotowy, Intrastat, GDM, porządkowanie liniowe.