

Baumgartner, Christoph; Srhoj, Stjepan; Walde, Janette

Article

Harmonization of product classifications: A consistent time series of economic trade activities

Journal of Economics and Statistics

Provided in Cooperation with:

De Gruyter Brill

Suggested Citation: Baumgartner, Christoph; Srhoj, Stjepan; Walde, Janette (2023) : Harmonization of product classifications: A consistent time series of economic trade activities, Journal of Economics and Statistics, ISSN 2366-049X, De Gruyter Oldenbourg, Berlin, Vol. 243, Iss. 6, pp. 643-662, <https://doi.org/10.1515/jbnst-2022-0034>

This Version is available at:

<https://hdl.handle.net/10419/333263>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<https://creativecommons.org/licenses/by/4.0/>

Christoph Baumgartner*, Stjepan Srhoj and Janette Walde

Harmonization of Product Classifications: A Consistent Time Series of Economic Trade Activities

<https://doi.org/10.1515/jbnst-2022-0034>

Received May 21, 2022; accepted June 20, 2023

Abstract: Firm-product data provide information for various research questions in international trade, industrial or innovation economics. However, working with these data requires harmonizing product classifications consistently over time to avoid internal validity issues. Harmonization is required because classification systems like the EU classifications combined nomenclature (CN) for goods or the Prodcom for the production of manufactured goods undergo several changes. We have addressed this problem and developed an approach to harmonize product codes. This approach tracks product codes from 1995 to 2022 for CN and 2001 to 2021 for Prodcom. Additional years can be conveniently added. We provide the harmonized product codes for CN and Prodcom in the selected period's last (or first) year. Our approach is summarized in an open-source R package so that researchers can consistently track product codes for their selected period. We demonstrate the importance of harmonization using the micro-level trade data for Croatia as a case study. Our approach facilitates working with firm-product data, allowing the analysis of important research questions.

Keywords: product code; concordance; CN; HS; Prodcom; BEC

JEL Classification: C81; F14; O12; C55

1 Introduction

Innovation and international trade are considered a major engine of economic growth (Aghion et al. 2021; Wagner 2019). Innovation can take a variety of forms,

***Corresponding author: Christoph Baumgartner**, Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Innsbruck, Austria, E-mail: christoph.baumgartner@uibk.ac.at. <https://orcid.org/0000-0002-2681-071X>

Stjepan Srhoj, Department of Economics, Faculty of Economics, Business and Tourism, University of Split, Split, Croatia, E-mail: ssrhoj@efst.hr. <https://orcid.org/0000-0002-7105-8088>

Janette Walde, Department of Statistics, Faculty of Economics and Statistics, Universität Innsbruck, Innsbruck, Austria, E-mail: janette.walde@uibk.ac.at. <https://orcid.org/0000-0002-1241-5292>

including product, production or delivery, organizational, marketing or communication innovation (Gault 2018). When an economist thinks of innovation, product innovations, like (autonomous) automobiles, computers, smartphones, robots, or drones, come first to mind. We focus on enabling investigations concerning product innovation and firm-product relations in general. For the purpose of analyzing product innovation, studies mostly use the Community Innovation Survey (CIS)¹ (e.g. Schubert and Tavassoli 2020). CIS is considered one of the richest sources of firm innovation data. Researchers use CIS data to investigate characteristics of firms introducing product innovations, like research and development (R&D) expenditures, firm size, firm age, industry, or profitability (e.g. Audretsch and Belitski 2020; Coad et al. 2016; Griffith et al. 2006; Mairesse and Mohnen 2010). CIS data is also used to conduct policy evaluations of large public R&D grants or public procurement for innovation on the firms' innovation output (Mairesse and Mohnen 2002; Stojčić et al. 2020). Although the availability of CIS data has had a major impact on the field of innovation economics, CIS data suffers from at least two weaknesses. Firstly, only a small sample of relatively larger firms is collected each cycle, and secondly, variables are collected via survey, thus, they are self-reported by the firms and can be imprecise or suffer from a sample selection bias. There are also other weaknesses, such as the documented difficulties in merging CIS microdata with regional identifiers (Hauser et al. 2018).

Transaction data from customs (Combined Nomenclature, CN) and the Prodcom data can be used to investigate product innovation or firm-product relations in general (European Commission 2019, 2021). Research questions concerning determinants of firms' changes in productivity over time, firms' product mix strategy, firm markups, R&D grants impact on the number of newly exported products, or extensive and intensive margins of trade are just few examples that can be analyzed with firm-product data.² These data provide the value and quantity of imported, exported, or sold products at a very detailed product level and enable the analysis of large firm samples. In addition, Prodcom data include a substantial part of all domestically traded products. Questions related to industry-level importance of innovative products can also be investigated using transaction and Prodcom data. Identified new products can be classified into new-to-firm and new-to-market products, combined with the value of these sold or exported products providing additional insights into the magnitude of innovation. Also, these data enable to

1 Link: <https://ec.europa.eu/eurostat/web/microdata/community-innovation-survey> [Accessed May 23, 2021].

2 Customs data (CN) includes information at the firm-product-market level, while Prodcom includes information at the firm-product level, we use the term "firm-product data" to describe both data, although CN adds another dimension.

consider the distance between exporter and destination market (Mayer and Zignago 2011), the type of exports (Wagner 2019), and frequency of simultaneous exports and imports (Bernard et al. 2018). Castellani and Fassio (2019) is one of the few innovation studies using firm-product-market level data.

Consistent time series of product codes are necessary for working with transaction and Prodcom data. However, several product codes are updated annually. Such updates include new product codes, dropped product codes, or product codes merged into a common code. Considering such changes harmonizing the product codes for a given period is mandatory. We contribute to the literature on the importance of harmonizing data to mitigate internal data validity issues (Bellert and Fauceglia 2019; Duprez and Magerman 2022; Liao et al. 2020; Pierce and Schott 2009, 2012; Van Beveren et al. 2012).³ Such harmonization is not only an important topic for product codes. Link (2020) demonstrated the necessity of harmonization for ifo Business Survey (IBS) data. Previous approaches to harmonize product codes are based on the idea of Pierce and Schott (2009, 2012), implemented in Stata or R, and are more or less flexible in terms of selecting the period and classification systems (and also with respect to software and programming skills required for users). We briefly explain the development of the preliminary work to better delineate our approach. Pierce and Schott (2009, 2012) suggested an approach to harmonize the international harmonized system product codes (HS) and provided Stata code to enable researchers to prepare the data consistently for further analyses. Van Beveren et al. (2012) picked up their idea and extended their Stata code from concording HS to CN8 and Prodcom codes. Duprez and Magerman (2022) updated the Stata do-files available from 1988 until 2010 up to 2014.

When using programming code in the form of do files or script files, adjustments to other periods, classification systems and program environments require additional programming effort and this may be error prone. Bellert and Fauceglia (2019) were the first to provide a Stata ado-file for various classification schemes without necessary internal code changes. Not having to intervene in the program structure for updates or change requests was also our objective for the implementation in R. Our approach offers new features and add-ons in several respects. In addition to the harmonized product codes, the display of the development of all product codes over time is the core of our approach. This allows the entire history of merged, split, dropped and new product codes to be easily tracked and interpreted. Also, special cases are marked, such as product codes that remained the same during the observation period, but to which other product codes have been added. We provide

³ Liao et al. (2020) provide a dictionary for product codes to get the appropriate description of the codes in different classification schemes. Their approach is therefore not concerned with the harmonization of product codes in the strict sense. It is implemented as an R package and is under development.

so-called *utilize functions* to count the number of added, kept, and dropped products of each firm from one year to the next. We provide the broad economic categories (BEC) (United Nations 2016) and the System of National Accounts (SNA) classification (European Communities et al. 2009). All necessary calculations are included in an open-source R package that is easy to use and can be adapted for different periods (or updated product code change lists) without internal code changes. The approach shows the links between different product codes over time and allows working with consistent time series for the selected period. The importance of harmonization is demonstrated by applying our approach on Croatian data. We obtain a similar substantial discrepancy between the raw time series and the harmonized time series as Van Beveren et al. (2012) for Belgian firm-product data.

In the remainder of the manuscript, we describe the classification systems, present the idea of harmonization, and demonstrate the importance of harmonizing product codes by applying the approach to firm-product data from Croatia. Finally, we summarize the main features of the approach.

2 Harmonization of Product Codes

Harmonization is based on classification lists and concordance tables provided by Eurostat. There are several classification systems used in the European Union (EU) for trade and production. Although the systems are designed to be similar, there are important differences between the two systems, both at a given time and across several years.

2.1 Combined Nomenclature

EU international trade statistics records the value and quantity of products traded between EU member states, and from EU member states to non-EU countries. At the customs, products must be classified according to the combined nomenclature, which has 8-digits and thus denoted CN8. Its first six digits match the classification system of the (international) harmonized system (HS), which is maintained by the World Customs Organization (WCO 2022). The European classification system CN8 is an extension of the HS6 classification system, analogous to the 10-digit extensions (HS10) used in the USA. Table 1 provides an example for the different product codes and their number of digits. The CN classification was developed to meet the requirements of both the Common Customs Tariffs and the EU's external trade statistics. In intra-EU trade statistics the CN is also used (Directorate-General for Taxation and Customs Union 2022).

Table 1: Example of a CN8 code for 2020.

Classification system	Digits	Code	Description
HS	2	84	Chapter 84 – Nuclear reactors, boilers, machinery and mechanical appliances; parts thereof
HS	4	8421	Centrifuges, including centrifugal dryers; filtering or purifying machinery and apparatus, for liquids or gases
HS	6	842129	Machinery and apparatus for filtering or purifying liquids (excl. such machinery and apparatus for water and other beverages, oil or petrol-filters for internal combustion engines and artificial kidneys)
CN	8	84212920	Made of fluoropolymers and with filter or purifier membrane thickness not exceeding 140 µm

Source: European Commission (2021).

The HS undergoes periodic revisions. Between 1988 and 2022, it was updated five times (in 1996, 2002, 2007, 2012, and 2017). Revision years for HS6 also tend to be years of substantial changes in the CN8 classification. Moreover, there are annual updates of the CN8 classification, such that from one year to the next a product may have a different CN8 code. Such updates are motivated by policy, development of technology, or statistical requirements.

2.2 Prodcom Classification

Firms within the EU must report their industrial production and services in the Prodcom survey, which are specified on the Prodcom list. Although Prodcom regulation is EU-based, this firm-product data is obtained by the National Statistics Institutes of the member states. If the member state considers Prodcom reporting as an administrative burden for the firms, they are able to alleviate the reporting requirement. However, the information in Prodcom covers at least 90 % of national production within each sector defined by the four digits from the statistical classification of economic activities in the European Community (NACE) code (European Communities 2008).

The Prodcom list was developed to measure production in the EU Member States and to allow a comparison between production and external trade statistics (Eurostat 2023). Therefore, Prodcom classification is closely related to the combined nomenclature classification. Similar to the CN classification, also the Prodcom system is an extension of other systems. The first four digits are taken from the NACE code, while the digits 5 and 6 are in line with the Classification of Products by Activity (CPA). The last two digits then make up the actual Prodcom code (cf. Table 2).

Table 2: Example of a Prodcom code for 2020.

Classification system	Digits	Code	Description
NACE	4	0710	Mining of iron ores
CPA	6	071010	Iron ores
Prodcom	8	07101020	Iron ores and concentrates; agglomerated (excluding roasted iron pyrites)

Source: European Commission (2019).

As with CN, the Prodcom codes are subject to annual changes. These changes also include newly established Prodcom codes or definitely dropped codes. Thus, the coverage may change. Products may be covered by Prodcom codes in one year but not covered by any Prodcom code in another year. It is impossible to keep track of the amount or value of these codes over time, so they need to be dropped from the data when harmonizing.

2.3 Classification by Broad Economic Categories

The broad economic categories (BEC) system classifies products into broad economic categories used internationally. This classification was introduced in 1971 and has undergone several revisions, with the latest revision, number 5, in 2016. Today, its development and maintenance are done by the United Nations Statistics Division and is used worldwide. Since the most recent revision, BEC codes have eight categories and contain up to six digits instead of three as they used to. The leading digit indicates the main economic class of the product. The second digit distinguishes between goods and services within the eight main categories. Moreover, BEC codes can be classified into three basic classes defined by the System of National Accounts (SNA), which focus on the end-use of the product. The classes are called “capital goods”, “intermediate goods”, and “consumption goods” in revision 4 of BEC, while they are called “intermediate consumption”, “gross fixed capital formation”, and “final consumption” since revision 5. The SNA classification is the third digit in the BEC codes. The last three digits are the “processing dimension”, the “specification dimension” and finally, the “durability dimension” (for an example, see Table 3).

Concordance files between HS6 and BEC Rev. 4 exist for 1996, 2002, and 2007. For 2012 and 2017, there exists a concordance between HS6 and BEC Rev. 5. Therefore, we provide BEC codes from Rev. 4 until 2011 and BEC codes from Rev. 5 thereafter. Figure 1 illustrates the connection as well as the availability of concordance files

Table 3: Example of a BEC code for Rev. 5.

Classification system	Digits	Code	Description
BEC	1	6	ICT, media, computers, business and financial services
BEC	2	61	Goods
BEC	3	611	Intermediate consumption
BEC	4	6112	Processed
BEC	6	611220	Specified

Source: United Nations (2016).

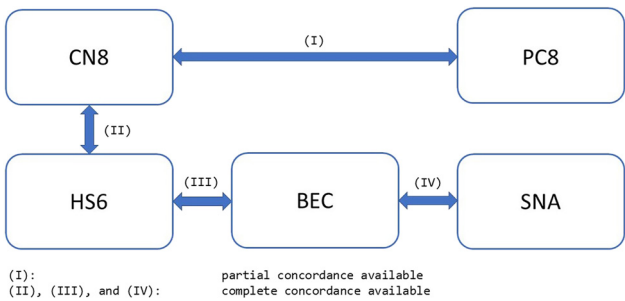


Figure 1: Connections between the different classification systems.

between the classification systems described in this paper.⁴ Recently (simultaneously to our approach⁵) Duprez and Magerman (2022) provide concordance files for CN8 and PC8 on github for 2002–2014.⁶ By default, we use the original concordance files provided at the RAMON server.

2.4 Harmonization

The basic idea behind harmonization is to keep track of every product code during a specific period. In the simplest case, a code does not change during the examined period, i.e. no harmonization is needed. In any other case, all changes associated with a specific code must be considered. There are different product code changes 1-to-1,

⁴ A list of all changes is annually provided by Eurostat on the EU server RAMON: <https://ec.europa.eu/eurostat/ramon>.

⁵ We published our R package *harmonizer* for the first time December 8, 2021; Link: <https://cran.r-project.org/src/contrib/Archive/harmonizer/>.

⁶ Link: https://github.com/glennmagerman/Concordances_CN_PC [Accessed May 18, 2022]. These concordance files can conveniently be included by researchers in our R package if requested.

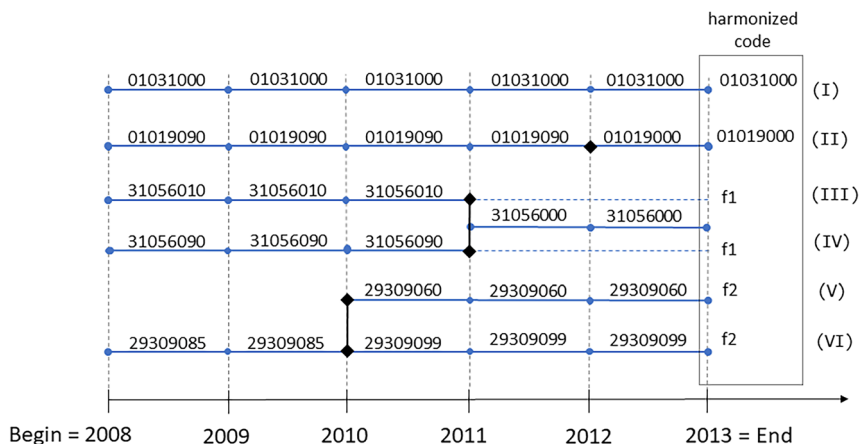


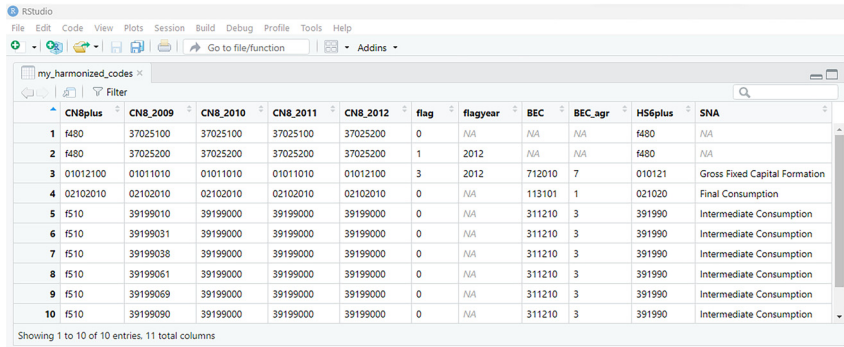
Figure 2: Example for real CN8 code changes for the period 2008–2013.

1-to-many, many-to-1, many-to-many, 1-to-none, and none-to-1. The last two changes indicate that a code was dropped or a new code was created. A 1-to-many or many-to-1 change can occur if two or more codes are split or merged.

Figure 2 visualizes possible product code histories. Diamonds in the figure indicate that product codes change in a given year. In case (I) code 01031000 does not change, and thus this CN8 code is the harmonized code too. Case (II) shows a simple code change, i.e. a renaming of the product code, and thus the former product code 01019090 becomes simply 01019000 (= harmonized code). The remaining lines show more complex changes. Cases (III) and (IV) demonstrate a merge of product codes 31056010 and 31056090 in 2011. They both are assigned at the end of the period the family code f1 (= harmonized code). The family code f2 in the cases (V) and (VI) results from a split of code 29309085 into 29309060 and 29309099 in 2010. One should note that a mixture of the cases (II) until (VI) is possible, resulting in another family code. The final artificial family code is denoted as “f” with a running number. Therefore, the largest number indicates the number of family codes.

All changes above occur for CN8, HS6, and Prodcom product codes’ changes over time. Prodcom codes can be generated or dropped within the period of interest. If changes in these classifications are not taken into account, erroneous entries and exits of products, price bias, or incorrect price and quantity indices result. The harmonized product code for all product codes is provided in a separate column (CN8plus, PC8plus, and HS6plus, cf. Figure 3).

In contrast to the approaches that provide harmonized product codes (family codes) as output (Bellert and Fauceglia 2019; Duprez and Magerman 2022; Pierce and Schott 2009, 2012; Van Beveren et al. 2012), our approach provides, in addition, the entire history of each



	CN8plus	CN8_2009	CN8_2010	CN8_2011	CN8_2012	flag	flagyear	BEC	BEC_agr	HS6plus	SNA
1	4480	37025100	37025100	37025100	37025200	0	NA	NA	NA	4480	NA
2	4480	37025200	37025200	37025200	37025200	1	2012	NA	NA	4480	NA
3	01012100	01011010	01011010	01011010	01012100	3	2012	712010	7	010121	Gross Fixed Capital Formation
4	02102010	02102010	02102010	02102010	02102010	0	NA	113101	1	021020	Final Consumption
5	1510	39199010	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption
6	1510	39199031	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption
7	1510	39199038	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption
8	1510	39199061	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption
9	1510	39199069	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption
10	1510	39199090	39199000	39199000	39199000	0	NA	311210	3	391990	Intermediate Consumption

Figure 3: Harmonized CN8 product code (CN8plus) and history of product codes from 2009 until 2012 as output by the package.

product code (CN8, Prodcod) in concordance with the HS6, BEC, and SNA classifications. The history matrix enables an easy-to-understand output (Figure 3) and allows the researcher to track splits and mergers and to understand the substantive implications.

In comparison to the previous product harmonization approaches (Bellert and Fauceglia 2019; Duprez and Magerman 2022; Pierce and Schott 2009, 2012; Van Beveren et al. 2012), our approach has further add-ons. (i) A product code can merge with another product code but remains the same product code afterward. These merged product codes belong by default to the same family. They are marked by a flag equal to 1 (cf. Figure 3) that allows the researcher to investigate the history and to decide whether to use the original product code or the family code. This choice has consequences, for example regarding the possibility to link HS6, BEC, or SNA classifications. The “flagyear” column has values only for flag values equal to 1, 2 or 3 and indicates the year in which the first change occurred. In this case, there was no change to this specific product code, so the value of the “flagyear” column corresponds to the last year of the period. (ii) New or dropped product codes are indicated by a flag equal to 2. A dropped product code, for example, may be linked to other product codes (due to splits or mergers) and is therefore assigned a family code. At the firm level, the drop of such a family code is interpreted as a product dropped by the firm, although it may be due to a product code dropped by Eurostat or the national Bureau of Statistics. The indicator (flag = 2) therefore makes it possible to remove the dropped (or also the new) product code from the family code. This prevents misinterpretations. The feature is particularly advantageous for Prodcod codes. (iii) Simple code changes, i.e. product codes have received unique but different product codes over time, are not marked as a family (in contrast to Bellert and Fauceglia 2019) but indicated by a flag equal to 3. (iv) Eurostat reports lists of product

Table 4: Comparison of the approaches dependent on the classification systems, the programming environment, and output.

Approach	Pierce and Scott (2009, 2012)	Van Beveren et al. (2012)	Bellert and Fauceglia (2019)	Duprez and Magermann (2022)	Our approach
Classification systems					
CN8	✗	✓	✓	✓	✓
PC8	✗	✓	✓	✓	✓
HS6	✓	✓	✓	✗	✓
HS10	✓	✗	✓	✗	✗
BEC	✗	✗	✗	✗	✓
SNA	✗	✗	✗	✗	✓
Programming environment					
Software	Stata	Stata	Stata	Stata	R
R-package/Stata-ado-file	✗	✗	✓	✗	✓
Output					
Harmonized product codes	✓	✓	✓	✓	✓
History matrix	✗	✗	✗	✗	✓

codes that changed as well as lists of product codes that existed in each year. Our method employs both lists. Thus, the complete information is used to identify especially new or deleted product codes. (v) Each approach deals with specific classification systems, reports different results, and is written for different programming environments. We summarize these features in Table 4.

The complete harmonization approach is available as an open-source R package, called *harmonizer*.⁷ By default harmonization of CN8 codes is possible from 1995 to 2022 and Prodcom from 2008 to 2021, respectively. However, updated data is added regularly to the package by the maintainer and can even be conveniently added by the users themselves.⁸ Also, usage of product codes and product code change files provided by national Bureau of Statistics is feasible. Thus, harmonization is possible for these products too. In Appendix B the commands for computing the harmonized product codes and the history matrix with the R package *harmonizer* are provided.

⁷ Downloadable via CRAN: <https://CRAN.R-project.org/package=harmonizer>.

⁸ See the R package's *vignette* for more (technical) documentation about how the harmonization is implemented and about all package features and functions <https://cran.r-project.org/web/packages/harmonizer/vignettes/harmonizer.html>.

3 Application

We apply harmonization on firm-product-market data of Croatia in the period from 2008 to 2016. For demonstration purpose Table 5 shows one possible (artificial) entry of raw data at firm level.

Table 6 provides the export value of all goods captured by CN8 codes. Columns (2) and (3) show the result without harmonization, while columns (4)–(7) take the harmonization into account. Almost 20 % of the total export value averaged across all years is associated with family product codes. If we do not harmonize the product codes, amount and value of products will be incorrectly attributed to individual product codes, while actually, the codes should be interpreted in the context of a family of product codes (cf. Figure 2). A direct consequence of this misinterpretation can be seen in Table 7.

Table 5: Artifical example of firm-product-market level data, classified by the CN8.

FirmID	Year	CN8	Market	Unit type	Unit amount	Export value
10356	2020	84181080	AT	Number of items	17	15,000

Table 6: Exports of firms in Croatia.

Years	Without harmonization		With harmonization			
	Export value	# of codes	Export value in families	Export value in families [%]	# of families	# of final CN8plus codes
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2008	9448.5	6434	1584.5	16.77	458	5547
2009	7565.1	6214	1333.0	17.62	448	5378
2010	8786.7	6106	1695.8	19.30	443	5347
2011	9459.5	5994	1904.4	20.13	427	5338
2012	9564.3	6235	2067.9	21.62	436	5557
2013	9389.9	6388	2062.6	21.97	457	5630
2014	10,112.2	6681	2095.7	20.72	471	5789
2015	11,266.7	6845	2205.6	19.58	482	5889
2016	11,896.9	6941	2421.1	20.35	485	5924
Average	9721.1	6426.4	1930.1	19.8	456.3	5599.9

Note: Columns (2) and (3) represent the export value and the number (#) of product codes if no harmonization is applied. Columns (4)–(7) summarize harmonized data. Columns (4) and (5) show the absolute or the relative value, respectively, by all product codes that are associated with a family of product codes, while column (6) shows the number of different families of product codes. Column (7) represents the final number of different harmonized codes in each year. All values are in million €.

Table 7: Added and dropped export products.

Year	Without harmonization					With harmonization				
	# of firms (2)	Added products (3)	Value of added products (4)	Dropped products (5)	Value of dropped products (6)	# of firms (7)	Added products (8)	Value of added products (9)	Dropped products (10)	Value of dropped products (11)
2008–2009	5922	29,353 43.90 %	418.7 4.50 %	36,346 54.3 %	628.9 6.70 %	5791	27,875 42.6 %	321.0 3.4 %	34,619 52.90 %	518.1 5.6 %
2009–2010	6478	31,901 52.70 %	1214.9 16.30 %	31,762 52.5 %	673.6 9.10 %	6321	30,214 51.2 %	822.4 11.10 %	29,874 50.70 %	374.5 5.0 %
2010–2011	6748	31,947 51.90 %	874.4 10.20 %	32,664 53.1 %	815.8 9.50 %	6586	30,326 50.5 %	657.5 7.70 %	30,919 51.50 %	557.4 6.5 %
2011–2012	7288	44,599 74.40 %	1835.0 20.00 %	30,187 50.4 %	2039.8 22.30 %	7105	41,938 71.7 %	415.5 4.50 %	27,841 47.60 %	808.1 8.80 %
2012–2013	6988	37,582 52.10 %	684.4 7.30 %	37,893 52.6 %	615.6 6.50 %	6822	36,107 51.3 %	504.1 5.30 %	36,755 52.20 %	512.5 5.40 %
2013–2014	5891	38,975 57.80 %	646.7 7.00 %	34,787 51.6 %	619.1 6.70 %	5738	36,785 56.1 %	569.1 6.20 %	33,491 51.10 %	573.2 6.20 %
2014–2015	6296	46,749 62.20 %	687.1 6.90 %	32,621 43.4 %	397.0 4.00 %	6109	44,116 61.2 %	644.2 6.50 %	31,291 43.40 %	366.6 3.7 %
2015–2016	6638	41,776 46.00 %	666.0 6.00 %	40,426 44.5 %	764.1 6.90 %	6432	39,736 45.8 %	623.3 5.70 %	38,058 43.80 %	693.5 6.30 %
Average	6531.1	37,860.3 55.1 %	878.4 9.8 %	34,585.8 50.3 %	819.2 8.96 %	6363	35,887.1 53.8 %	569.6 6.3 %	32,856 49.2 %	550.5 5.9 %

Note: Columns (2)–(6) show analyses using no harmonization. Every firm that exists in both years of the period of column (1) is included. Columns (3) and (4) show the number (#) and the corresponding value of added products. The percentage below refers to the number of all existing products and values, respectively. Columns (5) and (6) follow the same idea but using dropped products. Columns (7)–(11) can be interpreted analogously, but correct for the issue of harmonization. All values are in million €.

Table 7 shows all added and dropped products in terms of export, i.e. firms started or stopped trading specific products. From 2008 to 2016, added or dropped products are counted from one year to the following. Thus, the firm had to exist in both years. Columns (2)–(5) represent the trading data without harmonization, while columns (7)–(11) summarize harmonized data. The percentages below the values denote the relative amount of added or dropped products for all products in the market. Comparing the value of added products columns (4) and (9), reveal the strong influence of the missing harmonization. One notices that the average amount sold of new-to-firm products without harmonization is 1.5 times higher than the value with harmonization. Without harmonizing the data, firms are interpreted as introducing more products than they do or being more innovative than they are. A similar pattern occurs if one looks at the value of dropped products, i.e. columns (6) and (11). By looking at non-harmonized data, one would overestimate the amount of dropped products of a firm even though some were not dropped at all. Similar findings are obtained for imports (cf. Tables A1 and A2).

We also investigated the firm sales data from products classified by Prodcom codes in Croatia. Table 8 shows the difference between harmonized and non-harmonized data. The raw data without any harmonization, which can be found

Table 8: Products classified by Prodcom in Croatia.

Years	Without harmonization		With harmonization			
	Sales value	# of codes	Value in families	Value in families [%]	# of families	# of final PC8plus codes
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2008	18,349.5	1746	699.5	3.81	60	1607
2009	16,005.5	1704	560.5	3.50	61	1583
2010	16,467.2	1651	558.1	3.39	59	1536
2011	17,238.8	1647	622.4	3.61	60	1532
2012	17,286.5	1615	579.1	3.35	57	1513
2013	16,110.2	1603	528.0	3.28	58	1510
2014	15,837.7	1637	523.4	3.31	61	1540
2015	16,178.9	1623	533.2	3.30	62	1528
2016	15,940.7	1581	539.5	3.38	61	1490
Average	16,601.7	1645.2	571.5	3.4	59.9	1537.7

Note: Columns (2) and (3) represent the sales value and the number (#) of product codes if no harmonization is applied. Columns (4)–(7) present a harmonization for internationally used PC8 codes (i.e. national codes excluded). Columns (4) and (5) show the absolute respectively the relative value of all codes that are associated with a family, while column (6) shows the number of families. Column (7) gives the final number of harmonized codes in a certain year. All values are in million €.

in columns (2) and (3) of Table 8, may also include possible typos and national codes. Within the Prodcom classification system, every country can include codes that are only valid for the nation itself. These codes are not used internationally and can therefore only be harmonized with national concordance files. However, if concordance files are available, it is possible to use them within our package since the use of custom trade data as well as of custom concordance files is possible. Nevertheless, the harmonized results shown here do not include these national codes, since we have not yet had access to the national concordance files of Croatia.

Similar findings to Table 7 providing dropped and added products for Prodcom codes can be found in the Appendix in Table A3. Valid identification of the changes in the firms' product mix allows developing innovation measures and therewith enhancing the second wave of innovation and firm literature. Identified new products can be classified into new-to-firm and new-to-market products, and can also be combined with the value of these products sold or exported – thus providing additional insights into the magnitude of innovation. Researchers can work with the harmonized data to not only identify new-to-firm products but combining these data with firm-level geographical data and identify new-to-firm-region or new-to-firm-national market products. In addition, with the export or sales value, the importance of these new products can be quantified.

4 Conclusions

Similar to research on firms and international trade (Wagner 2019), research on firms and innovation has begun to use transaction data at the level of firms and products or even at the firm-product-market level (Castellani and Fassio 2019). This type of data can be used for many purposes, such as developing innovation measures at different regional levels or identifying characteristics of innovative or high-growth firms. However, such data must be harmonized; otherwise, the validity of the results is compromised. If changes in these classifications are not taken into account, erroneous entries and exits of products, price bias, or incorrect price and quantity indices result.

The open-source R package *harmonizer* comprises all our harmonization efforts. The main output consists of the harmonized product code and our unique history matrix of product codes of various classification systems (CN8 and Prodcom together with HS6, BEC, and SNA). The history matrix reflects the complex temporal

links between product codes that are not readily apparent in the change lists, product code lists, and concordance files. We provide not only the necessary links between product codes but also information to identify special cases, such as deleted or new products in a family of linked product codes. Due to this information, product code changes or drops are not misinterpreted as products discontinued by firms. Furthermore, our R package allows for computing the number of added, kept, or dropped products for each firm each year. In addition, updating the harmonization period is conveniently possible as new concordance files can be provided as input files.

The application of our approach to Croatian firm-product-market data demonstrated the importance of product code harmonization. In our application, the average export value of new products in the firms' product mix without harmonization is 1.5 times higher than the value using harmonized data. The extent of the significance of harmonizing is in line with the application of Van Beveren et al. (2012) on the data of Belgium.

Establishing a standardized harmonization procedure for trade (CN8) and sales (Prodcom) data is a desirable goal. For this, different researchers across the EU countries have to contribute to obtaining common knowledge about the sensitivity of handling product code families differently. The goal of our approach is a step towards this standardization, as it enables researchers to investigate the sensitivity with respect to this issue. However, further research is still needed to establish a standardized harmonization procedure for these rich data.

With the harmonization of trade and sales data, new and dropped products can be counted reliably, changes in export and import values can be computed, price indices can be derived, and new products can be distinguished between new-to-firm products or new-to-market products. The *harmonizer* package can be conveniently applied. Our contribution may push forward the research in international trade and innovation economics by enabling research with consistent firm-product data.

Acknowledgments: The authors thank the staff of the Croatian Bureau of Statistics for data access in the safe room and for their data assistance. This work was supported by the research platform Empirical and Experimental Economics of the Universität Innsbruck and by the Croatian Science Foundation within the project IP-CORONA-2020-12-1064. Last but not least the authors would like to thank the Editor and two anonymous reviewers for their valuable comments.

Appendix A

Table A1: Imports of firms in Croatia.

Years	Without harmonization		With harmonization			
	Import value	# of codes	Import value in families	Import value in families [%]	# of families	# of final CN8+ codes
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2008	20,527.8	8222	3147.3	15.33	543	6914
2009	15,295.4	8100	2320.3	15.17	540	6869
2010	14,934.9	7970	2406.0	16.11	527	6827
2011	16,073.0	7829	2830.3	17.61	529	6835
2012	16,073.6	7977	2599.6	16.17	536	6931
2013	16,276.8	8090	2587.2	15.90	543	6981
2014	16,906.4	8075	2943.1	17.41	531	6938
2015	18,260.1	8080	3067.6	16.80	529	6916
2016	18,995.4	8175	3012.4	15.86	532	6941
Average	17,038.1	8057.56	2768.2	16.26	534.44	6905.78

Note: Columns (2) and (3) represent the import value and the number (#) of product codes if no harmonization is applied. Columns (4)–(7) summarize harmonized data. Columns (4) and (5) show the absolute or the relative value, respectively, by all product codes that are associated with a family of product codes, while column (6) shows the number of different families of product codes. Column (7) represents the final number of different harmonized codes in each year. All values are in million €.

Table A2: Added and dropped import products.

Year	Without harmonization					With harmonization				
	# of firms	Added products	Value of added products	Dropped products	Value of dropped products	# of firms	Added products	Value of added products	Dropped products	Value of dropped products
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
2008–2009	15,881	167,909	1360.2	207,749	2234.4	15,662	159,089	1188.3	197,906	1999.3
2009–2010		38.00 %	6.70 %	47.10 %	11.00 %		36.60 %	5.90 %	45.50 %	9.9 %
2009–2010	16,403	171,341	2099.1	188,505	2547.2	16,171	160,777	1366.6	176,265	1756.4
2010–2011		41.90 %	13.90 %	46.10 %	16.90 %		40.00 %	9.10 %	43.80 %	11.6 %
2010–2011	16,645	178,434	1405.0	175,674	1556	16,421	167,633	1083.3	163,835	1203.3
		45.20 %	9.50 %	44.50 %	10.60 %		43.20 %	7.40 %	42.20 %	8.2 %
2011–2012	17,827	260,135	2617.5	169,092	2470.7	17,572	245,834	1377.9	156,074	1089.3
2012		65.30 %	16.70 %	42.40 %	15.70 %		62.70 %	8.80 %	39.80 %	6.9 %
2012–2013	17,082	178,683	1429.3	230,035	1411	16,842	172,798	1230.8	224,343	1217.4
2013		36.10 %	9.00 %	46.50 %	8.90 %		35.50 %	7.80 %	46.10 %	7.7 %
2013–2014	10,409	139,553	1448.9	181,060	1273.9	10,274	133,469	1222.4	175,524.0	1119.2
2014		36.10 %	9.10 %	46.80 %	8.00 %		35.20 %	7.70 %	46.30 %	7.1 %
2014–2015	10,647	160,713	1242.5	134,545.0	1083.9	10,514	151,742.0	1080.6	129,748.0	973.7
2015		45.80 %	7.40 %	38.30 %	6.50 %		44.30 %	6.50 %	37.90 %	5.8 %
2015–2016	11,100	140,076	1412.2	145,586.0	1171.7	10,944	134,651.0	1302.1	138,448.0	1074.0
2016		37.40 %	7.90 %	38.90 %	6.50 %		37.10 %	7.20 %	38.20 %	6.0 %
Average	14,499.3	174,605.5	1626.8	179,030.8	1718.6	14,300	165,749.1	1231.5	170,267.9	1304.1
		43.2 %	10.0 %	43.83 %	10.51 %		41.8 %	7.6 %	42.48 %	7.9 %

Note: Columns (2)–(6) show analyses using no harmonization. Every firm that exists in both years of the period of column (1) is included. Columns (3) and (4) show the number (#) and the corresponding value of added products. The percentage below refers to the number of all existing products and value, respectively. Column (5) and (6) follow the same idea but using dropped products. Columns (7)–(11) can be interpreted analogously, but correct for the issue of harmonization. All values are in million €.

Table A3: Added and dropped Prodcom products.

Year	Without harmonization					With harmonization				
	# of firms (2)	Added products (3)	Value of added products (4)	Dropped products (5)	Value of dropped products (6)	# of firms (7)	Added products (8)	Value of added products (9)	Dropped products (10)	Value of dropped products (11)
2008–2009	3147	878 9.70 %	966.2 5.40 %	945 10.40 %	795.1 4.50 %	2927	783 9.1 %	842.1 4.7 %	822 9.60 %	612.2 3.40 %
2009–2010	3010	549 6.40 %	630.7 4.10 %	648 7.60 %	446.7 2.90 %	2794	415 5.1 %	505.2 3.3 %	488 6.00 %	330.5 2.10 %
2010–2011	3003	493 5.90 %	855.6 5.30 %	559 6.60 %	697.6 4.30 %	2787	403 5.0 %	550.6 3.4 %	444 5.60 %	456.7 2.80 %
2011–2012	2973	669 8.20 %	1206.8 7.20 %	662 8.10 %	698.3 4.20 %	2763	417 5.4 %	960.9 5.8 %	414 5.40 %	420.1 2.50 %
2012–2013	2797	531 6.90 %	1006.0 5.90 %	580 7.50 %	574.2 3.40 %	2592	475 6.5 %	928.2 5.5 %	485 6.60 %	464.0 2.70 %
2013–2014	2873	569 7.30 %	892.7 5.60 %	555 7.20 %	1091.7 6.80 %	2659	529 7.2 %	792.5 5.0 %	496 6.70 %	991.9 6.20 %
2014–2015	3007	546 6.90 %	1110.7 7.10 %	569 7.20 %	1000.5 6.40 %	2759	457 6.1 %	500.6 3.2 %	493 6.60 %	737.2 4.70 %
2015–2016	2995	852 10.80 %	1335.3 8.30 %	916 11.60 %	1061.8 6.60 %	2737	450 6.0 %	311.9 1.9 %	492 6.60 %	356.3 2.20 %
Average	2975.6	635.9 7.76 %	1000.5 6.1 %	679.3 8.3 %	795.7 4.9 %	2752.3	491.1 6.3 %	674.0 4.1 %	516.8 6.6 %	546.1 3.3 %

Note: Columns (2)–(6) show analyses using no harmonization. Every firm that exists in both years of the period of column (1) is included. Columns (3) and (4) show the number (#) and the corresponding value of added products. The percentage below refers to the number of all existing products and value, respectively. Columns (5) and (6) follow the same idea, but use dropped products. Columns (7)–(11) can be interpreted analogously, but correct for the issue of harmonization. All values are in million €.

Appendix B

R commands for harmonizing product codes

```
# Install package (only needed the first time)
install.packages("harmonizer")
# Load package into the library
library("harmonizer")

# Create harmonized CNS product codes, where 'b' is the first and
# 'e' is the last year of interest
my_harmonized_codes <- harmonize_cn8(b = 2009, e = 2014)

# Save as .csv-file
write.csv(my_harmonized_codes, file = ("yourfilepath/outputfilename.csv"))
```

References

- Aghion, P., Antonin, C., and Bunel, S. (2021). *The power of creative destruction: economic upheaval and the wealth of nations*. Harvard University Press, Cambridge, Massachusetts.
- Audretsch, D.B. and Belitski, M. (2020). The role of R&D and knowledge spillovers in innovation and productivity. *Eur. Econ. Rev.* 123: 103391.
- Bellert, N. and Fauceglia, D. (2019). A practical routine to harmonize product classifications over time. *Int. Econ.* 160: 84–89.
- Bernard, A.B., Blanchard, E.J., Van Beveren, I., and Vandenbussche, H. (2018). Carry-along trade. *Rev. Econ. Stud.* 86: 526–563.
- Castellani, D. and Fassio, C. (2019). From new imported inputs to new exported products. Firm-level evidence from Sweden. *Res. Pol.* 48: 322–338.
- Coad, A., Segarra, A., and Teruel, M. (2016). Innovation and firm growth: does firm age play a role? *Res. Pol.* 45: 387–400.
- Directorate-General for Taxation and Customs Union (2022). The Combined Nomenclature, Available at: https://ec.europa.eu/taxation_customs/business/calculation-customs-duties/customs-tariff/combined-nomenclature_en (Accessed 20 May 2022).
- Duprez, C. and Magerman, G. (2022). Correspondences of eu product classifications, Working paper. Available at: https://static1.squarespace.com/static/55e85d72e4b0146280523def/t/62388fb92cba1d26fa8408f6/1647873979840/concordances_live.pdf.
- European Commission (2019). Prodcom list. Official Journal of the European Union, Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019R1933&qid=1653055165284>.
- European Commission (2021). Combined Nomenclature. Official Journal of the European Union 64, Available at: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L:2021:414:FULL&from=EN>.
- European Communities (2008). System of national accounts 2008, Available at: <https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF> (Accessed 23 January 2023).

- European Communities, International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations and World Bank (2009). System of national accounts 2008, Available at: <https://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf> (Accessed 18 January 2023).
- Eurostat (2023). *Industrial production statistics introduced – PRODCOM*. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Industrial_production_statistics_introduced__PRODCOM (Accessed 22 August 2023).
- Gault, F. (2018). Defining and measuring innovation in all sectors of the economy. *Res. Pol.* 47: 617–622.
- Griffith, R., Huergo, E., Mairesse, J., and Peters, B. (2006). Innovation and productivity across four European countries. *Oxf. Rev. Econ. Pol.* 22: 483–498.
- Hauser, C., Siller, M., Schatzer, T., Walde, J., and Tappeiner, G. (2018). Measuring regional innovation: a critical inspection of the ability of single indicators to shape technological change. *Technol. Forecast. Soc. Change* 129: 43–55.
- Liao, S., Kim, I.S., Miyano, S., and Zhu, F. (2020). Concordance: product concordance. R package version 2.0.0, Available at: <https://CRAN.R-project.org/package=concordance>.
- Link, S. (2020). Harmonization of the ifo business survey's micro data. *Jahrb. Natl. Stat.* 240: 543–555.
- Mairesse, J. and Mohnen, P. (2002). Accounting for innovation and measuring innovativeness: an illustrative framework and an application. *Am. Econ. Rev.* 92: 226–230.
- Mairesse, J. and Mohnen, P. (2010). Chapter 26 – using innovation surveys for econometric analysis. In: Hall, B.H. and Rosenberg, N. (Eds.). *Handbook of the economics of innovation*, 2. North-Holland, Amsterdam, Netherlands, pp. 1129–1155.
- Mayer, T. and Zignago, S. (2011). Notes on cepii's distances measures: the geodist database. Technical report, Available at: <https://doi.org/10.2139/ssrn.1994531>.
- Pierce, J.R. and Schott, P.K. (2009). ConCORDING U.S. harmonized system categories over time. Working Paper 14837. National Bureau of Economic Research, Available at: <https://www.nber.org/papers/w14837>.
- Pierce, J.R. and Schott, P.K. (2012). A concordance between ten-digit U.S. harmonized system codes and SIC/NAICS product classes and industries. *J. Econ. Soc. Meas.* 37: 61–96.
- Schubert, T. and Tavassoli, S. (2020). Product innovation and educational diversity in top and middle management teams. *Acad. Manag. J.* 63: 272–294.
- Stojčić, N., Srhoj, S., and Coad, A. (2020). Innovation procurement as capability-building: evaluating innovation policies in eight central and eastern European countries. *Eur. Econ. Rev.* 121: 103330.
- United Nations (2016). Classification by broad economic categories Rev.5, Available at: <https://unstats.un.org/unsd/trade/classifications/Manual%20of%20the%20> (Accessed 20 May 2022).
- Van Beveren, I., Bernard, A.B., and Vandenbussche, H. (2012). ConCORDING EU trade and production data over time. Working Paper 18604. National Bureau of Economic Research, Available at: <http://www.nber.org/papers/w18604>.
- Wagner, J. (2019). *International trade in goods: Evidence from transaction data*. World Scientific Publishing Company Pte. Limited, Hackensack, New Jersey.
- WCO (2022). Harmonized system – frequently asked questions, Available at: http://www.wcoomd.org/en/topics/nomenclature/overview/harmonized_system_faq.aspx (Accessed 20 May 2022).