Data Observer

Philipp Leppert*

The Micro Data Linking-Panel: A Combined Firm Dataset

https://doi.org/10.1515/jbnst-2019-0013

Abstract: The following article informs about the Micro Data Linking-Panel (hereafter MDL-Panel), a new firm[1] dataset, available since May 2018 via the Research Data Centre of the Federal Statistical Office and the Research Data Centre of the Statistical Offices of the Federal States (RDC) (Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (2018a, 2018b). The dataset originates from two projects funded by Eurostat over the period 2014 to 2016 with the title "Micro data linking of structural business statistics and other business statistics" (MDL). The projects combined data of various business statistics for the reference years 2008 to 2013 and were aimed at offering new perspectives on business statistics topics, which are not addressed in the ordinary publication program. Furthermore, MDL served to gain insights about the involved methodological challenges of linked data in Official Statistics. The first chapter provides background information on the projects funded by Eurostat. Chapter 2 contrasts the contents of the MDL-Panel to other products available via the RDC. The third chapter describes the scope of the dataset and its content in more detail. The article closes with future prospects for the MDL-Panel.

Keywords: combined firm data, panel data, Official Statistics

1 Project description

Since 2005, § 13a of the Federal Statistics Law allows Official Statistics to link micro data from business statistics and environmental statistics, if the final statistical product does not include additional response burden for the reporting firms. Starting in 2014, the Federal Statistical Office and eight other national

1 In German business statistics the statistical unit, or firm, is currently referred to as the smallest legally independent unit which keeps accounts for commercial and/or tax purposes.

***Corresponding author: Philipp Leppert,** Statistisches Bundesamt (Destatis), Wiesbaden, Germany, E-mail: philipp.leppert@destatis.de

statistical institutions participated in two MDL projects to assess the analytical potential of linked data from different sources of business statistics. In the first project the target objective was to create linked cross-sectional micro datasets from all structural business statistics surveys for the reference years 2008 to 2012. These datasets can be considered as the "backbone" of the MDL-Panel. In order to complement them, they were linked with other business statistics like the business register, the Community survey on ICT usage and e-Commerce, the Inward Foreign AffiliaTes Statistics (IFATS) and the international trade in goods statistics. The data linkage was carried out via a unique identifier available from the business register except for the international trade in goods statistics, where the data can only be linked via the tax id. The follow-up project added data from reference year 2013 as well as variables from the business demography, which is based on the business register. After completing the project, we decided to make the created dataset available for research via the RDC.[2]

2 Comparing the MDL-Panel to other RDC products

The MDL-Panel includes data sources from different statistical domains, which together form a new unique dataset. Currently, there are no published tables for MDL by the Federal Statistical Office or Eurostat and in contrast to other products available via the RDC the researcher cannot match his or her calculated results.[3] All statistics in Table 1 were linked and are hence components of the presented MDL-Panel. They are described in more detail in their corresponding quality reports published by the Federal Statistical Office. Please refer to the metadata report of the MDL-Panel to receive more information on each included statistic.[4]

Some of the linked statistics are already available as individual datasets via the RDC or are even included in other linked datasets like the AFiD-Panel (Malchin/Voshage 2009). The MDL project focused *only on a limited number*

2 Note that due to methodological considerations linked data from the Community survey on ICT usage and e-Commerce as well as IFATS are not available in the RDC version of the MDL-Panel.

3 Published tables are in general available from https://www-genesis.destatis.de/genesis/online.

4 The metadata report of the MDL-Panel is available from https://www.forschungsdatenzentrum.de/de/10-21242-42231-2013-00-05-1-1-0.

Table 1: Business statistics included in the MDL-Panel.

NACE section[1]	Statistical domain	Statistic (EVAS)[2]	Type of survey	Reference year	Standard supply of the RDC?
B, C	Structural business statistics	42231	Complete survey with cut-off limit	2008–2013	Yes
B, C		42251	Sample survey	2008–2013	Yes
B, C		42252	Sample survey	2008–2013	No
D, E		43211	Complete survey with cut-off limit	2008–2013	Yes
D, E		43221	Complete survey with cut-off limit	2008–2013	Yes
F		44211	Complete survey with cut-off limit	2008–2013	No
F		44221	Complete survey with cut-off limit	2008–2013	No
F		44252	Sample survey	2008–2013	No
F		44253	Sample survey	2008–2013	No
F		44254	Sample survey	2008–2013	No
G		45341	Sample survey	2008–2013	No
I		45342	Sample survey	2008–2013	No
H, J, L, M, N, S 95[3]		47415	Sample survey	2008–2013	Yes
B-J, L-N, S 95	International trade in goods statistics[4]	51141	Complete survey with cut-off limit	2009–2013	No
B-J, L-N, S 95		51231	Evaluation of customs declarations	2009–2013	No
B-J, L-N, S 95	Business register	52111	Register-based evaluation	2008–2013	Yes

[1] NACE Revision 2. Statistical classification of economic activities in the European Community.
[2] For information about the statistic behind the EVAS see https://erhebungsdatenbank.estatistik.de/eid/evas.jsp.
[3] Data for NACE division S 95 is only available from reference year 2010 onwards for all concerned statistics.
[4] Note that due to a limited storage period the available data from the international trade in goods statistics was incomplete for reference year 2008 and the related variables are only available for the following reference years. For reference years 2009 and 2010 the breakdown of imports and exports by partner country is restricted due to availability of data.

of variables available across all structural business statistics surveys, which are among others turnover, value added at factor cost, total purchases of goods and services, wages and salaries, gross investment in tangible goods and the number of employees. This might be a *drawback* for very specific research questions and the researcher may rather consult the individual version of the needed statistic, if available. In addition to this, the researcher might consider linking the MDL-Panel with available individual statistics.

The major advantage of the MDL-Panel, however, is the coverage of surveyed firms of all structural business surveys, including the economic areas mining and quarrying, manufacturing, energy and water supply, construction, wholesale and retail trade as well as the service sector. In other words the whole surveyed population of what Official Statistics refers to as "the non-financial business economy in Germany" is present in this dataset. Moreover, we included selected variables of the business register like the firm's group affiliation and of the business demography like indicators of a firm's growth. Finally, we are able to provide firm level data on international trade in goods between member states of the European Union (EU) and with countries outside the EU derived from the international trade in goods statistics. Note that due to the data preparation process of the MDL-Panel the reproducibility of any related figure published by the Federal Statistical Office is *not* ensured.

3 Scope of the MDL-Panel

3.1 Basic information

Column 3 of Table 2 shows a total of 246.556 surveyed firms that are available in the MDL-Panel for reference year 2013. For example in NACE section B and C, that is mining/quarrying and manufacturing, three structural business surveys were combined to a total of 22.517 firms. This population consists of 16.625 firms with at least 20 employed persons originating from a complete survey with a cut-off limit regarding a firm's investments and an associated sample survey regarding other structural variables. The sample survey with a reduced questionnaire for 5.892 firms with less than 20 employed persons was added to complement the population. This allows, among others, for an analysis of smaller firms compared to larger firms, which is at the moment not possible with the available individual statistics of the RDC.

Table 2: MDL-Panel basic information (broken down for reference year 2013).

NACE section	Covered Population	Surveyed linked firms	Included structural business statistics surveys (EVAS)	Included other business statistics (EVAS)
B, C	See quality report of included structural business statistics surveys	22.517	42231 42251 42252	51141 51231 52111
D, E		7.017	43211 43221	
F		10.279	44211 44221 44252 44253 44254	
G		46.795	45341	
I		9.825	45342	
H, J, L, M, N, S95		150.123	47415	
The above	The above	246.556	The above	The above

Table 3: MDL-Panel reference years 2008–2013.

Industry code	Reference year	Surveyed linked firms	Surveyed linked firms available in all reference years	Population total (extrapolated)
B, C, D, E, F, G, H, I, J, L, M, N, S95	2008	206.734	47.106	1.886.300
	2009	222.974	47.106	2.027.720
	2010	228.800	47.106	2.073.681
	2011	234.248	47.106	2.157.591
	2012	241.815	47.106	2.189.519
	2013	246.556	47.106	2.192.778

Table 3 shows that in the time dimension there are 47.106 firms which were surveyed continuously in each reference year from 2008 to 2013. The share of these firms, compared to all surveyed firms, is roughly one fifth per reference year of the MDL-Panel because most sample surveys rotate parts of their covered population after a fixed or variable amount of time. This is done in order to smooth the response burden among the reporting firms. Note that sample survey designs are commonly optimized with respect to sampling variance, which gives a very large firm a higher probability to be observed in each reference year compared to a very small firm.

3.2 Data integration and validation

As in Germany the structural business statistics are organized domain-specific with distinct laws for each target population, linking the related surveys to create the "backbone" of the MDL-Panel was not without issues. Each survey is conducted based on a specific law which imposes restrictions on the contents of the survey. It is obvious that some structural variables relevant in a certain industry like manufacturing are not necessarily relevant for the service sector as well and hence the respective laws restrict the scope of the survey to the variables most meaningful for covering the target population. This is also a reason for the limited number of variables available in MDL across different industries. Also we faced the problem that due to time inconsistencies it was not always possible to link the firms originating from the structural business statistics surveys to the other business statistics, which resulted in the loss of a few firms compared to the actual number of surveyed firms in the individual statistic. Even though the collected survey data is cleaned by means of automatic and manual plausibility checks within the scope of each statistical domain, we conducted further validation checks considering the time dimension like attrition, changing industry affiliation or demographic events to ensure the validity of the data.

3.3 International trade in goods data at the level of the firm

A key feature of the MDL-Panel is the availability of a firm's import and/or export values, which are derived from the international trade in goods statistics. However, in Germany the firm is not necessarily the reporting unit in this statistic and certain assumptions have to be made to make this data applicable for a firm level analysis at all. First, the system of data collection varies between goods traded with members of the EU and goods traded with third countries outside the EU. While in the latter case the data is compiled via customs declaration, reporting obligations for exports and/or imports within the EU arise for an entrepreneur by being subject to value-added tax in Germany.[5] Both systems, however, share the registration of all trade in goods transactions via the firm's tax id. Now, this means that tax schemes like group taxation ("Organschaft") become an issue for Official Statistics as they comprise multiple firms in order to form a single tax entity ("VAT group"). In the following, the head of the VAT group ("Organträger") is referred to as the

5 For more information see (Allafi 2012).

parent and any other firm within the VAT group is referred to as a subsidiary ("Organgesellschaft").

In the statistical domain, the parent's tax id appears by customs declaration for traded goods with third countries outside the EU. Regarding the international trade in goods with EU members, the parent usually obeys the reporting obligations itself. As a VAT group consists of at least two entities, we do not know whether the parent itself caused a goods movement within its firm or any other subsidiaries with their firms as they all share the tax id of the parent. Most parents are located in NACE group M 70.1 (Activities of head offices) while frequently having subsidiaries in manufacturing or wholesale and retail trade. We assume that firms in the service sector are not predominantly involved in trading goods but rather trading services. If we do not adjust the firm level data to this assumption, a large amount of the actual traded goods would be attributed to these firms.[6] To evade this data collection issue, we reallocate the exports within a VAT group from a parent in the service sector to its subsidiaries in NACE sections B to G, if there are any. After this reallocation only 4 % of the total exported goods volume remains in the service sector (Kaus/Leppert 2017). Regarding imports, the redistribution within a VAT group is also carried out but regardless of the industry affiliation of the parent's subsidiaries. The redistribution is done via a key variable available from the business register, which represents the estimated turnover for each member of the VAT group (Wagner 2004). Hence, the firm's export and import values are interconnected with turnover, which is, given the available data source, the best measure to provide trade in goods data at the firm level. Our redistribution is conceptually comparable with the official approach carried out by the trade by enterprise characteristics statistics (TEC) (Allafi 2012). We used the business register to link the international trade data to the structural business statistics surveys via the tax id.

Note that the problem to attribute exports and imports within a VAT group to the respective firms extends to the available breakdown by country of destination (exports) and country of origin (imports). Given the data, it is not clear which firms within a VAT group do trade with the countries reported by the parent's tax id and we have to assume that all eligible subsidiaries trade with each partner country according to their relative share of the VAT group's turnover. The reallocation approach described above remains unaffected and tends to lead to a higher number of firms engaged in foreign trade and a higher number of listed partner countries respectively.

---

6 Note that there is also the international trade in services statistics, which is compiled by the German Central Bank and was not part of the MDL project.

There is a possibility to improve the allocation of exported goods with EU members within a VAT group by using VIES data.[7] This data source allows to breakdown the export shares of the parent and each subsidiary within the VAT group and has already been used for crafts statistics with promising results (Feuerhake/Giebenhain 2017). Not only does this warrant more exact export figures at the firm level but also provides the actual countries of destination for each subsidiary. However, for goods traded with third countries outside the EU there is currently no remedy to circumvent the described allocation issue.

3.4 Extrapolation of linked survey data

As the MDL-Panel includes the sampling weights of each respective sample survey, the firms can in general be extrapolated to the survey's target population. However, related figures published by the Federal Statistical Office cannot be exactly reproduced with these sampling weights. For example, a complete and sample survey with the same target population is extrapolated such that the sampling weights of the sample survey are tied to a certain variable population total of the complete survey. Hence, it is not possible to produce the exact figures with the sampling weights of the sample survey that we linked to the complete survey. Moreover, the sampling weights are designed only for extrapolating variables from the structural business statistics surveys and not for variables originating from the other linked business statistics (see Table 1). For example the population total value of imports and exports, originating from the international trade in goods statistics, cannot be reproduced with the sampling weights originating from the structural business statistics surveys. In the first project we tried to solve this issue with adjustment factors to improve the consistency of MDL figures with published figures like IFATS (Jung/Käuser 2016). Due to the variety of statistical domains and different survey methodologies this approach proved unfeasible. In the second project we considered the General Regression Estimator (Deville/Särndal 1992) which can ensure consistent extrapolated figures across different target populations. However, this approach yielded only reliable results for highly aggregate figures. On the other hand, the trend of the total extrapolated import and export values from 2009 to 2013 can be captured with the sampling weights and is also in line with figures published by the Federal

7 VIES (VAT Information Exchange System (European Commission 2019).

Statistical Office, although underestimating the total import and export values by 2% and 4% respectively (Kaus and Leppert 2017).

3.5 Selected applications of the MDL-Panel

Kaus and Leppert (2017) use the MDL-Panel to assess the contribution of firms which are involved in international trade regarding structural variables like value added, wages and salaries or gross investment in tangible goods. Furthermore, they demonstrate the heterogeneous nature of firms which are involved in international trade by characterizing them with respect to structural variables as well as firm characteristics like type of trader, group affiliation or the number of countries traded with. The relationship between a firm's productivity and its type of international trade activity (i. e. being an importer, exporter or both) and number of countries traded with is examined as well.

4 Future prospects

Even though the MDL-Panel was finalized with reference year 2013, it is designed to be complemented and continued with the already processed data sources. However, this is only meaningful if we can guarantee the comparability of its content as well as variable definitions and underlying survey designs over time. For the currently available variables of structural business statistics surveys this is only a small issue since variable definitions marginally changed in the past. Regarding the international trade in goods statistics, however, the compilation concepts were altered with reference year 2015. There, the concept of partner countries regarding imported goods was modified, which leads to a shift between a firm's reported countries in the time dimension. Hence, it is challenging to warrant the comparability of linked data from different sources for a long period.

With reference year 2018 even the concept of the firm will change due to the introduction of the *enterprise*[8] as statistical unit in business statistics, which deviates considerably from the currently applied legal unit (Opfermann/Beck 2018).

8 "The enterprise is the smallest combination of legal units that is an organizational unit producing goods or services, which benefits from a certain degree of autonomy in decision-making, especially of the allocation of its current resources". Council Regulation (EEC) No 696/93 of March 15, 1993 on the statistical units for the observation and analysis of the production system in the Community (Official Journal of the European Communities No L 76, page 5).

Moreover, with the introduction of FRIBS (Framework Regulation Integrating Business Statistics), laws concerning different industries or economic areas will be harmonized and allow for more comparative survey designs and survey scope as well as a higher comparability of variable definitions between statistical domains (Waldmüller/Weisbrod 2015). Depending on the researchers' demand, a relaunch of the MDL-Panel adjusted to the recent developments in business statistics might be possible.

5 Data access

Access to the micro data of the MDL-Panel can be applied for research purposes via the Research Data Centres of Official Statistics. The MDL-Panel can be analyzed via remote execution or in one of the PC workplaces for guest researchers (safe centre), which are located in almost 20 cities in Germany. An analysis of the data is possible with the statistical software Stata or SAS as well as partly with R or SPSS. For access to the MDL-Panel 250 Euro will be charged. PhD students and students can request a discount. For additional information on the access to the MDL-Panel see the following link: https://www.forschungsdaten zentrum.de/en/access.

References

Allafi, S. (2012), Außenhandelsergebnisse nach Wirtschaftszweigen 2010. WISTA Wirtschaft und Statistik Ausgabe 9: 764–765.

Deville, J.-C., C.-E. Särndal (1992), Calibration Estimators In Survey Sampling. Journal of the American Statistical Association 87: 376–382.

European Commission (2019), VAT Information Exchange System. Available from: Retrieved on 2019-02-04: https://ec.europa.eu/taxation_customs/business/vat/eu-vat-rules-topic/vies-vat-information-exchange-system-enquiries_de.

Feuerhake, J., M. Giebenhain (2017), Innergemeinschaftliche Warenexporte im Handwerk. WISTA Wirtschaft und Statistik Ausgabe 3: 39–52.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (2018a), Metadatenreport. Teil I: Allgemeine und methodische Informationen zum Micro Data Linking-Panel (MDL-Panel), Berichtsjahre 2008 bis 2013 für die On-Site-Nutzung. Version 1.

Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder (2018b), Metadatenreport. Teil II: Produktspezifische Informationen zum Micro Data Linking-Panel (MDL-Panel) 2008 bis 2013 für die On-Site-Nutzung. Version 2. DOI: 10.21242/48121.2013.00.05.1.1.0.

Jung, S., S. Käuser (2016), Herausforderungen und Potenziale der Einzeldatenverknüpfung in der Unternehmensstatistik. WISTA Wirtschaft und Statistik Ausgabe 2: 95–106.

Kaus, W., P. Leppert (2017), Aussenhandelsaktive Unternehmen in Deutschland: Neue Perspektiven Durch Micro Data Linking. WISTA Wirtschaft und Statistik Ausgabe 3: 22–38.
Malchin, A., R. Voshage (2009), Offical Firm Data for Germany. Schmollers Jahrbuch 129: 501–513.
Opfermann, R., M. Beck (2018), Einführung des EU-Unternehmensbegriffs. WISTA Wirtschaft und Statistik Ausgabe 1: 63–75.
Wagner, I. (2004), Schätzung fehlender Umsatzangaben für Organschaften im Unternehmensregister. WISTA Wirtschaft und Statistik Ausgabe 9: 1001–1008.
Waldmüller, B., J. Weisbrod (2015), Neuere Entwicklungen in den Unternehmensstatistiken. WISTA Wirtschaft und Statistik Ausgabe 5: 33–48.