RESEARCH DATA CENTRE (FDZ)
of the German Federal Employment Agency (BA)
at the Institute for Employment Research (IAB)

FDZ-METHODENREPORT

Methodological aspects of labour market data

01|2022 EN Update: Identifying mothers in administrative data

Dana Müller, Andreas Filser, Corinna Frodermann

Bundesagentur für Arbeit

Update: Identifying mothers in administrative data

Dana Müller (IAB), Andreas Filser (IAB), Corinna Frodermann (IAB)

Die FDZ-Methodenreporte befassen sich mit den methodischen Aspekten der Daten des FDZ und helfen somit Nutzerinnen und Nutzern bei der Analyse der Daten. Nutzerinnen und Nutzer können hierzu in dieser Reihe zitationsfähig publizieren und stellen sich der öffentlichen Diskussion.

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

List of Figures

Zusammenfassung

Die administrativen Daten der Bundesagentur für Arbeit bieten eine wichtige Datenbasis für die Arbeitsmarktforschung. Welche Informationen gesammelt werden, ist über die Aufgaben der Bundesagentur für Arbeit definiert. Daher sind nicht alle Informationen in den Daten enthalten, die für verschiedene Forschungsfragen relevant sind. Das betrifft zum Beispiel Informationen zu der Geburt von Kindern, die wichtig für die Analyse der Erwerbsbiografien von Frauen sein können. Nach wie vor unterbrechen insbesondere Mütter ihre Erwerbstätigkeit, um sich der Kinderbetreuung zu widmen. Diese Erwerbsunterbrechungen können unterschiedliche Effekte auf die Erwerbsbiografien von Müttern haben, wie z.B. Lohneinbußen, Karrierenachteile oder vermehrte Teilzeitbeschäftigung. Der FDZ-Methodenreport 13/2017 (Müller/Strauch 2017) zeigte eine Möglichkeit auf, familienbedingte Erwerbsunterbrechungen mit Hilfe indirekter Identifikatoren in den administrativen Daten zu ermitteln. Mit dem vorliegenden FDZ-Methodenreport wurde diese Identifikationsstrategie aktualisiert und an neue Datensatzversionen angepasst. Wir validieren unsere Identifikationsstrategie mit Hilfe offizieller Geburtsstatistiken. Der Programmcode wird als Anhang zur Verfügung gestellt und kann nach Bedarf angepasst werden.

Abstract

Administrative data of the Federal Employment Agency provide an important data base for labour market research. However, the Federal Employment Agency's tasks define what kind of information is recorded. Therefore, some information is missing that is relevant for various research purposes. One example is information on childbirth, which can be important when analysing female employment biographies. It is still mainly mothers who interrupt their employment to care for their children. These employment interruptions could have different effects on the career of mothers, e.g. wage reductions, reduced career opportunities or an increase in part-time employment. The FDZ-Methodenreport 13/2017 (Müller/Strauch 2017) presented one way to identify family-related employment breaks using indirect identifiers in the administrative data. With this FDZ-Methodenreport, we updated the identification strategy and adapted it to new dataset versions. We use official statistics on births to validate our identification strategy. We make our program code available as an appendix for the research community. Users can adapt when necessary.

Keywords

childbirth, mothers, administrative data

Acknowledgements

The authors would like to thank Katharina Glanert-Strauch for her contribution to the previous version of the DoFile and FDZ-Methodenreport (Müller/Strauch 2017). Moreover, we are grateful to Ann-Christine Bächmann, Johanna Eberle, Katharina Schneider, Philipp vom Berge and Jonas Jessen for helpful comments and suggestions as well as to all users and IAB-colleagues who gave feedback of a preliminary version of the program code. Furthermore, we would like to thank Florian Zimmermann and Benedikt Zapf for editing the Excel sheet and Dagmar Theune for editing the layout.

1 Introduction

Administrative data of the German Federal Employment Agency provide a rich resource for labour market research. They include information on employment, unemployment benefits, participation in labour market programs and registered job search as well as sociodemographic and establishment information. In comparison to survey data, those administrative data have various advantages. Large sample sizes, mandatory notifications and a longitudinal observation period on a daily basis that spans several decades make sure that panel mortality and selective responses do not influence empirical results. Furthermore, the importance of the notifications for social security contributions and other administrative purposes usually ensure a high data quality.[1]

However, the Federal Employment Agency's tasks define what kind of information is recorded. Therefore, some information is missing that is relevant for various research purposes. One example is information on childbirth, which can be important when analysing female employment biographies. It is still mainly mothers who interrupt their employment to take care of their children. These employment interruptions could have different effects on the career of mothers, e.g. wage penalties or an increase in part-time employment.

Information on childbirth is not directly available in the administrative data of the Federal Employment Agency, yet other information helps to identify childbirth indirectly. One possible strategy is to identify employment interruptions due to family formation as described by Schönberg (2008). She uses the meanwhile outdated IAB Employment Sample with an observation period from 1975 until 2001. This data contains a variable ("btyp") that indicates employment interruptions. Furthermore, Schönberg (2008) tests her results with linked pension register data that include children's date of births. Her results show that it is possible to identify nearly 90% of childbirths in West Germany using the administrative data if the age of mothers falls into a specific range. However, the identification of first childbirth during un/employment episodes is easier because further pregnancies often take place during parental leave or non-employment periods, which are not observable in the data. Nevertheless, the "btyp" variable does no longer exist in the follow up data versions of the SIAB ("Sample of Integrated Labour Market Biographies", see Frodermann et al. 2021).

In the previous FDZ-Methodenreport, Müller and Strauch (2017) presented another way to identify family-related employment breaks. While Schönberg (2008) uses employment interruptions to identify mothers, they used an indirect identifier specifying the reason of cancellation/notification/termination of an employment or unemployment spell in the variable 'grund'. Official statistics on births in Germany were used to validate the identification strategy. In this FDZ-Methodenreport, we provide an updated identification strategy suitable for the new SIAB_7519 version as well as for all ADIAB-Datasets of the IAB since version 7519. The paper is structured into three parts. The following chapter describes the information we use from the administrative data for our procedure. Afterwards, we describe the identification of mothers and

[1] This is not always the case for information that is not directly relevant for the calculation of pensions or benefits, for example information on schooling.

expected childbirths in general (details are included in the available program code). In Chapter 4, we discuss limitations. We sum up our identification strategy in the last chapter.

2 Background on relevant administrative information

There are two main ways to identify maternity in the administrative data, one using employment notifications and one using process data of the Federal Employment Agency about unemployment and benefits. Status updates from both sources are compiled in the variable 'grund', the reason of cancellation/notification/termination of a spell.

In the employment notifications, mothers are identifiable because employers deregister mothers-to-be from employment when they enter maternity leave. In Germany, mothers are legally entitled to fourteen weeks' maternity leave (at least six weeks before and eight weeks after childbirth). Maternity leave notifications are filed under "entitlement to other compensation by the statutory health insurance provider" (value 151 of 'grund', value 51 in previous SIAB versions), because statutory health insurance providers pay the maternity allowance. Unfortunately, notifications with *grund=151* are not exclusively caused by maternity leave, but might also indicate deregistration due to long-term illness. When a person is absent due to illness for more than six weeks, health insurance steps in to pay the sickness allowance. This transfer of payment obligations also causes an employment deregistration with *grund=151.* Although we cannot clearly differentiate between long-term illness and maternity leave in the administrative data, both causes for *grund=151* cluster in different age groups. Childbearing typically occurs during young and middle adulthood (see Excel sheet in the online appendix). In contrast, long-term illness-related absence is rare in these age stages, but is more prevalent among older workers (Meyer et al. 2015). Our identification strategy therefore rests on the assumption that women in a specific age bracket do not face employment interruptions due to long-term sickness.

In the administrative data on non-employed individuals, mothers are identifiable because the Federal Employment Agency deregisters them from unemployment at the start of maternal protection. This deregistration takes place when women are unemployed and receive unemployment benefits according to Social Code Books II or III. Unemployed women who received unemployment benefits I before maternity leave receive maternity allowance in the amount of the previous benefits. The allowance is paid by the statutory health insurance provider instead of the Federal Employment Agency because a pregnant mother is no longer available to the labour market due to the employment ban during maternity leave. Women who work and are additionally dependent on unemployment benefits II also receive maternity allowance. Women who depend solely on unemployment benefits II (i.e. welfare) do not receive maternity allowance.

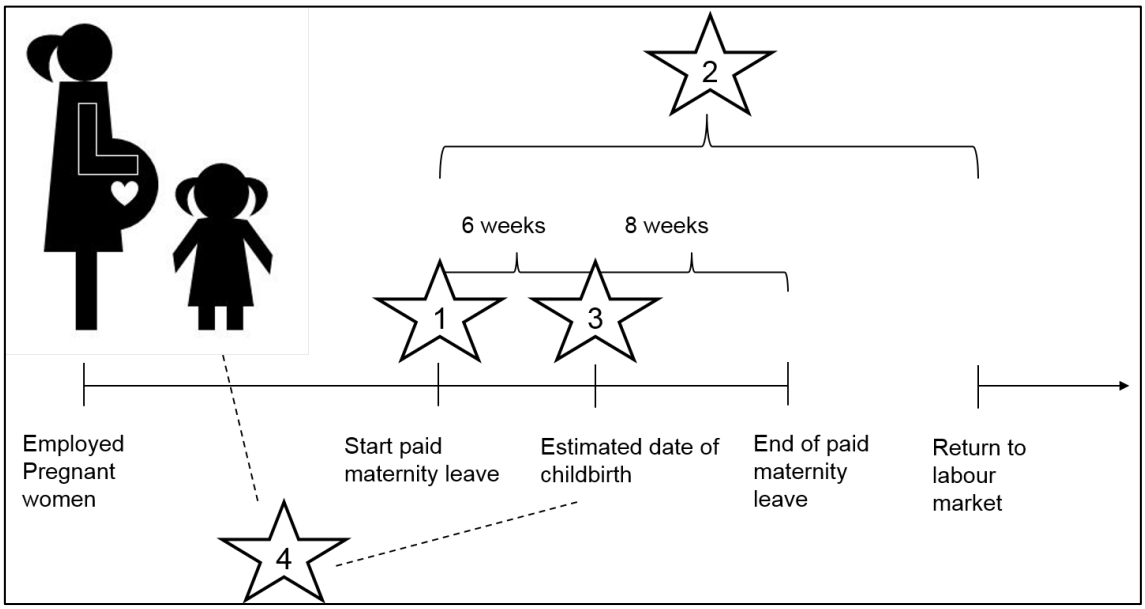
The timing of paid maternity leave is determined by the estimated birth date from maternal logbook. The employer informs the employee about the starting date of maternity protection. Women insured at the statutory health insurance receive up to 13 Euro per day as maternity allowance. The employer makes an additional payment so the total compensation equals the previous net wage. Pregnant women who are privately insured can apply for a one-time maternity

benefit of 210 euros. In either case, both pregnant women and their employers are exempt from social insurance contributions during paid maternity leave.
The availability of the data and the logic of the notification procedures imply caveats that researchers need to keep in mind when using the proposed algorithm. The limitations are described in Chapter 4.

3 Identification strategy

Figure 1 illustrates our four-step-approach to identify mothers and (expected) childbirths in the administrative individual data maintained by the Research Data Centre. Beyond the description below, further information is included in the program code available in the appendix of this report. The program code is suitable for all datasets building on administrative information, such as the Sample of Integrated Employment Biographies (SIAB). However, the program code is only applicable for data sets from versions 7519 onwards, like the SIAB_7519 and other ADIAB data products. For earlier data versions, please refer to the previous code of Müller/Strauch (2017). Furthermore, the identification strategy requires information that is only available in the weakly anonymous versions accessible via on-site use or remote execution. The factually anonymous versions (for example the SIAB Regional File) contain only aggregated deregistration information in the variable 'grund', which is unsuitable for an accurate identification of mothers. We tested the program code with the most recent version of the Sample of Integrated Employment Biographies (SIAB_7519_v1) using Stata 16.

Figure 1: Steps of mother's respectively (expected) childbirth's identification



The first step (star 1 in Figure 1) identifies the beginning of paid maternity leave. The entry into paid maternity leave is recorded in the variable 'grund' as mentioned above. We mark all

deregistrations with the value 151 (51 in previous SIAB versions), which come from the notification procedure of the social security system and are filed 'due to other entitlements from statutory health insurance'. Value 151 (51 in previous versions) is included in the source 'Employment History'. Furthermore, we mark all deregistrations into maternity leave and maternity allowance with the values 1114 (2015 in previous versions), 502 (3002 in previous versions) and 310 (6010 in previous versions), which come from process data of the Federal Employment Agency and are included in the sources 'Benefit Recipient History', 'Unemployment benefit II Recipient History', and 'Jobseeker History'. We do not include the value 152 (52 in previous versions) of 'grund' which includes the notification 'childcare leave'. Employers file this notification when mothers take the last year of their parental leave at a later date, which is possible until the child reaches its 8th birthday. At this stage, we consulted childbirth statistics of the Federal Statistical Office in order to restrict the plausible childbearing age stages. The number of identified age-specific childbirths in our data needs to be lower than the number of age-specific childbirths of the official statistics because not all women are recorded in the administrative data of the Federal Employment Agency. Using an iterative process, we determined the maximum childbearing age for the identification of second and higher births to be 40. However, we set the age restriction for the *first* childbirth at the 38th birthday of employed women after comparing our results with the childbirth statistics of the Federal Statistic Office. For women older than 38, the probability of long-term illness increases at a certain age (Meyer et al. 2020), thus increasing the probability of misspecificating motherhood.

In the second step (star 2 in Figure 1), we calculate the duration of employment interruptions to avoid misspecification due to long-term sickness. Therefore, each interruption needs to be longer than 97 days (duration of maternity leave). We also take into account women who do not re-enter the labour market. In this case, deregistration is on the last observed spell. Here, we generate the employment interruption using the end of the observation period in the data. Furthermore, we create a separate variable to highlight these spells. All notifications followed by an interruption shorter than 98 days are ignored.

The third step (star 3) includes the computation of the expected date of delivery. Only 4% of all childbirths are exactly at the expected date of delivery. Researchers need to be aware that an inexact date of childbirth may affect their research results. We use the original end date (endorig) and add 42 days (maternity leave starts six weeks before estimated childbirth) to compute childbirth.[2]

In step 4 (star 4), we check the gap between siblings to avoid misinterpreting multiple deregistration as independent births. We do not use a hard limitation of 40 weeks (281 days) of pregnancy because the expected delivery is only correct for 4% of all childbirths. Furthermore, the preterm rate was 9.2% in 2010 for Germany (March of Dimes et al. 2012). Therefore, we use a more realistic measure and use a minimum gap between siblings of at least 32 weeks (224 days)[3]. With our identification strategy, we generate 324,456 (expected) childbirths and 234,239 mothers, respectively in the SIAB 7519.

Finally, we prepare some variables that might be useful for research purposes including the number of children per episode, the number of children and a dummy variable for being mother

[2] We cannot use the end date of the episode (endepi) because overlapping periods of employment or non-employment have different durations and time specific start and end dates after episode splitting.

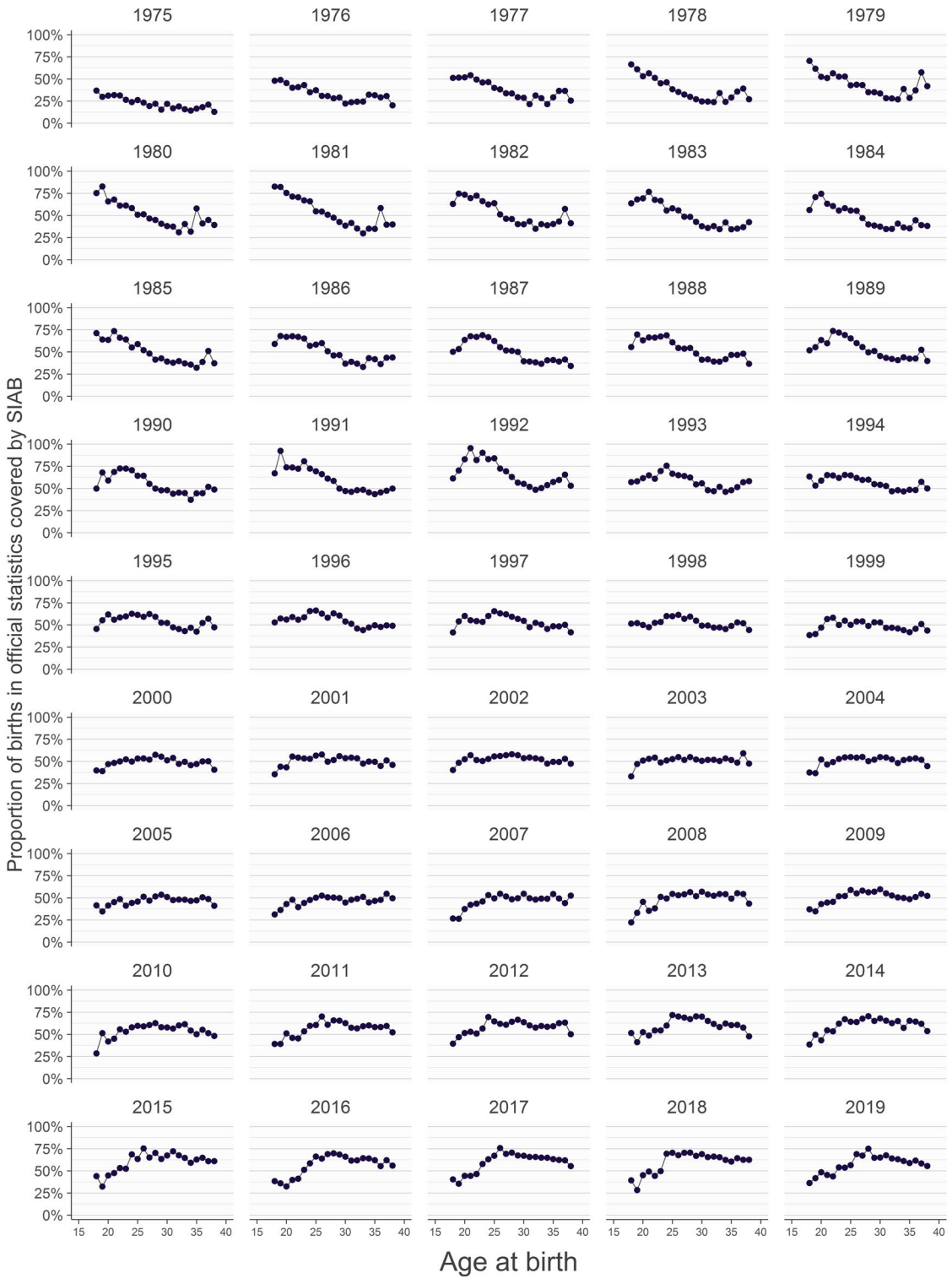
[3] In the program, we produce both measures of 281 days and 224 days to show how many deliveries are affected.

or childless. Researchers need to be aware that we underestimate the number of children for the reasons mentioned in chapter 2.

We compare our identified (expected) childbirth with the official statistics of the Statistical Office of Germany using annual statistics on childbirth by age of mothers. We multiply the numbers of (expected) childbirth with a factor of 50 because the SIAB 7519 is a 2% subsample of the Integrated Employment Biographies. Figure 2 summarizes the proportion of annual age-specific birth numbers in the official birth statistics covered in the SIAB. Detailed results are presented in an Excel sheet (see online appendix[4]). The results show that the number of childbirths identified in the SIAB data is smaller than the number of childbirths in the official statistics. This is not surprising because we do not have any information on civil servants and self-employed. Furthermore, we can only identify a childbirth if the mother has a record in the administrative data sources.

[4] The Excel sheet cannot be designed freely accessible at the moment due to the disproportionate financial effort involved. However, the information it contains can be found in the following paragraph as well as in the description of Figure 2. We are happy to provide individual support upon request.

Figure 2: Proportion of births in official statistics covered by mother identification in SIAB_7519_v1, age cohorts 18-40, see Appendix for detailed table.



4 Limitations

As mentioned before, there are several limitations that researchers must consider when using the proposed algorithm:

1. We can identify (expected) childbirths only if women have a record in the administrative data sources. Consequently, we do not capture maternity leave of self-employed or public servants. Furthermore, we are not able to identify childbirths before women enter the labour market.
2. It is not possible to differentiate between live birth and stillbirth.
3. The algorithm is limited to first births up to the 38th birthday of employed women to avoid misspecification at older ages.
4. The calculated date of childbirth might differ from the actual date of childbirth. Reason is that we compute childbirth based on the end of the spell before maternity leave and add 42 days (see step 3 in Chapter 3). Researchers have to be aware that deviations resulting from preterm and late birth might occur.
5. The number of children per woman can be underestimated for two reasons: (i) twin births are not identifiable and (ii) consecutive births can only be identified if the mother is employed between the birth of two children or has another record in the administrative data sources. This can have different consequences: For example, there is evidence that less births can be identified among low-skilled individuals. The reason might be that they are less likely to have an (employment) record between two births (Huber, 2019).

5 Conclusion

This FDZ-Methodenreport is an updated version of the FDZ-Methodenreport of Müller/Strauch (2017). We describe the strategy to identify mothers respectively (expected) childbirths in the administrative data of the Federal Employment Agency by using indirect identifiers. We validate our results with official statistics on childbirths. Although the identification strategy has some limitations, it is a helpful strategy for researchers who want to use the administrative data of the Research Data Centre for their research purposes in the context of female employment biographies.

6 References

Frodermann, C.; Schmucker, A.; Seth, S.; Vom Berge, P. (2021): Sample of Integrated Labour Market Biographies (SIAB) 1975-2019. (FDZ-Datenreport 01/2021 (en)), Nürnberg, 76 S.

Huber, K. (2019): The role of the career costs of children for the effect of public child care on fertility and maternal employment, BGPE Discussion Paper No. 185

March of Dimes, PMNCH, Save the Children, WHO (2012): Born Too Soon: The Global Action Report on Preterm Birth. Eds CP Howson, MV Kinney, JE Lawn. World Health Organization. Geneva. http://www.who.int/pmnch/media/news/2012/201204_borntoosoon-report.pdf (accessed on October 31, 2017).

Meyer M., Wiegand S., Schenkel A. (2020): Krankheitsbedingte Fehlzeiten in der deutschen Wirtschaft im Jahr 2019. In: Badura B., Ducki A., Schröder H., Klose J., Meyer M. (eds) Fehlzeiten-Report 2020. Fehlzeiten-Report, vol 2020. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-61524-9_23

Schönberg, U. (2008): Does the IABS reliably identify maternity leave taking? FDZ-Methodenreport, 03/2008 (en).

Appendix

Download of the Stata do-file:

- https://doku.iab.de/fdz/reporte/2022/MR_01-22_EN_programs.zip

Download of the Excel sheet with the comparison of official statistics on childbirth by age of mother and childbirth by age of mother generated with the SIAB_7519_v1:

- https://doku.iab.de/fdz/reporte/2022/MR_01-22_comparison_of_official_statistics.zip

Imprint

FDZ-Methodenreport 01|2022 EN

Date of publication
17 January 2022

Publisher
Research Data Centre (FDZ)
of the Federal Employment Agency (BA)
in the Institute for Employment Research (IAB)
Regensburger Str. 104
D-90478 Nuremberg

Download
https://doku.iab.de/fdz/reporte/2022/MR_01-22_EN.pdf

Documentation version
DOI: 10.5164/IAB.FDZM.2201.en.v1

All publications in the series "FDZ-Methodenreport" can be downloaded from
https://fdz.iab.de/en/FDZ_Publications/FDZ_Publication_Series/FDZ-Methodenreporte.aspx

Website
https://fdz.iab.de/

Corresponding author
Dana Müller
Phone: +49 911 179-2409
Email Dana.Mueller@iab.de