Schmollers Jahrbuch 121 (2001), 639 – 648
Duncker & Humblot, Berlin

European Data Watch

This section will offer descriptions as well as discussions of data sources that may be of interest to social scientists engaged in empirical research or teaching courses that include empirical investigations performed by students. The purpose is to describe the information in the data source, to give examples of questions tackled with the data and to tell how to access the data for research and teaching. We will start with data from German speaking countries that allow international comparative research. While most of the data will be at the micro level (individuals, households, or firms), more aggregate data and meta data (for regions, industries, or nations) will be included, too. Suggestions for data sources to be described in future columns (or comments on past columns) should be send to: Joachim Wagner, University of Lueneburg, Institute of Economics, Campus 4.210, 21332 Lueneburg, Germany, or e-mailed to ⟨wagner@uni-lueneburg.de⟩.

Individual tax statistics data and their evaluation possibilities for the scientific community

By Markus Zwick[1]

1. Preliminary remarks

Today there is intensive competition on the market for information services. Driven by the enormous technical progress in data processing, customer requirements are today increasingly wide ranging and, above all, specialised. Here the scientific community in particular places great demands on those providing data. As a flexible, innovative user of data, it demands individually customised products from the official statistics authorities. Apart from the use of aggregated information, this above all includes the use of microdata.

[1] I would like to thank ORR Roland Sturm for information and suggestions, especially with regard to chapter 3.

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

The official statistics authorities gather individual pieces of information and collate these into statistical results. The Bundesstatistikgesetz – BStatG (Federal Statistics Law) stipulates that these statistics are only allowed to be published if the individual values of variables cannot be allocated to the individual statistical units. The right of informational self-determination – as defined by the Federal Constitutional Court in its reason for its judgement on the population census – means that part of the information is not allowed to be made public by the official statistics authorities. This societal self-restriction is particularly meaningful in tax statistics. The legal taxpayer, who has to provide tax authorities with a detailed explanation of how he earns his income, generally has an interest in preventing third parties from gaining access to these data. The official statistics authorities are responsible for protecting these individual rights, on the one hand, but also, on the other hand, for seeking possibilities for making available for utilisation the great information potential of available individual data, especially for making it available to the scientific community. The following explanations from the area of tax statistics illustrate that this balancing act is possible.

The work begins by outlining the existing individual data of the different tax statistics. In the course of this introduction, not only the available characteristics but also the individual statistics are briefly explained here. In the third part the possibilities for external scientists to use individual data from official statistics are presented. The fourth chapter describes the practical implementation of this procedure in the area of tax statistics. A brief outlook rounds off the work.

2. The individual tax statistics data[2]

Since the modification of the tax statistics law (StStatG) by the annual tax law of 1996 it has been possible for the official statistics authorities to keep the individual data of the different tax statistics on a centralised basis[3]. Until then the statistical evaluation of the different tax statistics was almost exclusively carried out within the framework of previously defined tables in accordance to different combinations of characteristics. Any changed questions which only arose after the processing could only be answered – if at all – in a time-consuming process involving a great amount of personnel, since the data were not available on a centralised basis. For this reason very great reserves of tables were sometimes produced and kept in stock, in

[2] You can also find further information on the following statistics under www.destatis.de.

[3] See von der Lippe (1997) and Zwick (1998).

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

order to be able to react reasonably flexibly to additional requests for data despite this problem.

Thanks to the centralised storage of the data records, it is now possible to carry out relatively up-to-date evaluations. Up to now the prime focus has been on the *wage and income tax statistics*. These statistics, which are compiled every three years, in each case examine for a complete year the accrual of the different types of fiscal revenues and the development of the ratio between the total income and the taxable income of each individual legal taxpayer. In addition, statistics on the tax liability and tax rebates[4] are compiled. This information can be used as the basis for working out how the tax liability arises for every legal taxpayer in each individual case. In addition the data records also contain socio-economic characteristics, such as marital status, age, sex or religious affiliation. These data records can also be sorted according to regional categories, right down to local district level. This information is available for approximately 30 million legal taxpayers. Since by definition in the case of tax splitting between a married couple the legal taxpayer comprises two persons, the data pool contains approximately 38 million German citizens, for whom up to 400 characteristics are documented. If requested, a complete data record description can be sent, and this also applies to the following statistics. The data records of these statistics which are compiled every three years are available for the years 1992 and 1995. In spring 2002 data for the 1998 survey year are expected to be available. The relatively great time lag until the statistics are available is above all due to the long time allowed for taxpayers to make their tax declaration. Since in the most important cases full use is generally made of the time allowed for making a declaration, an earlier processing of the data would lead to considerable distortions. Currently, however, it appears that in the future these statistics could be made more up-to-date through shortening the periodicity. There is a further reference to this in section 5.

In addition to the income information of the legal taxpayers, the data of the wage and income tax statistics contain information on the related *Personengesellschaften* (partnerships). Here, amongst other things there is a categorisation according to legal form of the companies, industrial sector and regional affiliation.

The statistical data on the Körperschaftssteuer (corporation tax) is collected in the same three-yearly rhythm as the above-mentioned statistics. For this purpose in some cases more than 400 characteristics are recorded for around 550,000 legal persons. Apart from the source of income of the legal persons, the individual data contain information on their legal form, industrial sector, as well as regional information.

[4] See Rosinus (2000) for the latest results and further methodical information.

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

In addition to cross section analyses, dynamic analyses are also possible in the *turnover tax statistics* thanks to the annual surveys. These statistics collect data in a very differentiated way on the turnover of the companies which are subject to turnover tax, for instance according to industrial sector, legal form and regional criteria. Since 1996 approximately 2.8 million data records, each containing about 100 characteristics are available for every year and the latest statistics available are for 1999. Currently the turnover tax statistics are the only statistics which enable a deeper insight into the tertiary sector in Germany, although information about the turnover development of the independent professions is not available. Since the data on turnover tax statistics are an integral part of the future business register, currently intensive work is being carried out on temporal linkage methods and a matching with the information on independent professions from wage and income tax statistics.

For the first time the individual data are also available for the re-introduced *trade tax statistics* for 1995. These statistics, which are now carried out every three years, contain about 2.3 million legal taxpayers, each with approximately 100 characteristics. Apart from the legal form, the industrial sector and regional information, these statistics contain information on the tentative tax, as well as other information, such as data on profit and loss development of the business establishments.

3. Procedure of using individual data of the official statistics

In 1998 and 1999 the statistical offices of the Federation and the Länder conducted intensive investigations to find out ways of improving the utilisation of different individual data by external scientists. Within a working group operating at a Federal and Länder level, with the title "Wirtschaftsstatistische Einzeldaten für die Wissenschaft" (Individual Economic Statistical Data for the Scientific Community), the widest range of possibilities was discussed. The work culminated in a compilation of procedures, describing the possible cooperation between official statistics and the scientific community.

The procedures,

1. Implementation of special evaluations by the statistical offices on behalf of the scientific community,
2. Scientists make their own analysing programs available,
3. Scientists receive individual de facto anonymised data and
4. Scientists analyse individual non-anonymised data in the respective statistical office,

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

enable different ways of using the information potential of individual data.

The first procedure is a form of cooperation which has been practised for a long time between the statistical offices and the scientific community. A mutually-specified set of questions is evaluated within the statistical offices using the individual data, the results are checked for cases where confidentiality has to be maintained and they are then delivered to the data consumers. In this procedure only the statistical offices have access to the individual material, while the scientists receive the results which have been tested to ensure maintenance of confidentiality.

The second procedure represents a relatively new form of cooperation. Here, in comparison to procedure 1 an elementary working step is transferred to the level of the data user. The user formulates an evaluation question using program syntax. The statistical office takes over the program code and uses it on individual data material which has not been made anonymous. With this type of cooperation particularly great requirements are demanded from the specialist statistician with regard to maintenance of confidentiality. Apart from the "normal" checking of the evaluated results, which takes place before the data is passed on, he is also obligated to control the program syntax. Naturally it can be assumed that a scientist interested in the data would not risk spoiling his reputation by, for instance, programming "Trojan horses" which "smuggle" the information which is meant to be kept confidential past the confidentiality maintenance. However, it is necessary that the specialist statistician tests the functioning mechanism of transformation steps which are generated by external program syntax. Since the market for information-processing program packages is now constantly growing, for reasons of practicability it is necessary to limit the syntax to just a few program packages. Up to now SAS and SPSS have been used for such work.

In the third procedure scientists receive individual de facto anonymised data, for their own research purposes. This scientific "privilege" of Section 16, sub-section 6 of the Federal Statistics Law (BStatG) allows the researchers to work directly with the individual official data[5]. According to the Federal Statistics Law (BstatG), individual de facto anonymised data are those data which can only be allocated to a statistical unit with a relatively great amount of time, costs and manpower. With regard to the compilation of de facto anonymised data a differentiation can be made between two cases. In the first instance the statistical offices compile the data on a case-by-case basis, performing a special evaluation to respond to a specific inquiry. In the second instance, when there is great demand in individual statistical

[5] See also Köhler (1999).

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

areas, the statistical offices provide de facto anonymised microdata files (Scientific Use Files) as standard products. Due to the anonymising measures, de facto anonymised data files have a reduced amount of information compared to the original individual data.

The information services outlined here go beyond the provision of an informational infrastructure. Each of the procedures which have been presented has its own advantages, but also procedure-related disadvantages. For this reason there can be no standard answer as to which of these presented approaches is suitable for which type of research[6]. Therefore, during the preparation for a scientific project an intensive exchange should take place between scientists and specialist statisticians. This will ensure that the service of providing the data preparation, which involves costs, can be individually tailored to the object of the research. Due to the limited capacities in the statistical offices, it is possible in individual cases that despite ensured financing there may be individual incidences where it is only possible to implement such a cooperation to a limited extent or not at all. However, here the official statistics authorities are trying to adapt their capacities to the demand.

If certain preconditions are met, narrowly defined wording of questions can be used to enable external scientists to work with the original individual data according to a fourth approach. These are questions of official statistics which external scientists cooperate in formulating. In such a framework scientists carry out research work on behalf of a statistical office on a contractual basis and under obligation to keep the statistical confidentiality in the closed area of the official statistics. The responsibility for the work is in the hands of the statistical offices; hence they alone have use of the research results. The scientists can obtain a restricted right of utilisation regarding the results of the cooperation which are not required to be kept confidential.

Restrictions exist within each procedure, which limit a completely free research being carried out on the individual data of the official statistics. However, as the following chapter illustrates, a mix of these procedures allows the processing of all questions, at least in the sphere of tax statistics.

---

[6] The procedures described are conceived for specific research use; in the scientific sphere these procedures should tend to be only used in terms of their results – also due to cost reasons.

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

4. Utilisation of individual data in the tax statistics[7]

The main user of the individual data of the different tax statistics is the Federal Ministry of Finance (BMF). Here special evaluations in accordance with procedure 1 are generally carried out. Changing questions, particularly within the framework of tax reform projects, are programmed within the Federal Statistical Office, the program is used on individual data and the results are made available to the BMF. For instance, last year work was carried out on loss calculations, special depreciation allowances or extraordinary income. Samples are taken for the BMF from the material contained in all wage and income tax statistics. Here, in particular, one can cite a 1% sample drawn for a micro-simulation model used by the BMF. In addition, a 10% sample containing approximately 3 million individual data records is being kept in reserve. The random samples, which are drawn according to the "principle of the comparable precision for disaggregated results", also enable really complicated calculations to be carried out in a short time. Calculations based on the entire material can, under certain circumstances, take several days. These samples are also available for work outside the BMF.

However, specially prepared material from the Federal Statistical Office is also in demand from research institutes, university research, associations and other users. Hence, for instance, the data from the wage and income tax statistics of 1995 were, amongst other things, used for preparing in-depth structural analyses on church taxpayers for the two big German churches. In the area of university research special evaluations which were paid for have been used and are being used for preparing dissertations and professorial theses.

Within the course of a work on the income distribution of income in the case of the self-employed the second procedure was used[8]. With the help of different evaluation programs written in SPSS, in the protected area of official statistics the data of wage and income tax statistics from 1992 and 1995 were evaluated. In addition to descriptive evaluations, in particular the programs determined different inequality and disparity parameters. The statisticians checked the contents of the syntax, applied it and sent their results to the data user after ensuring that the obligation to maintain confidentiality was respected. This procedure has several advantages: with this type of cooperation complicated and time-consuming programming work can be carried out by the data user himself. In particular in university research this

[7] With regard to this chapter, see also Zwick (1999); further questions about utilisation can be directly put to the author under markus.zwick@destatis.de

[8] With regard to the results see Merz (2000)

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

is an advantage which should not be underestimated, since information services normally obtained through the official statistics, which entail costs, can be substituted by own research time. A further advantage is the flexibility. Program adaptations were able to be carried out within the shortest of time by the data consumers. In this way it was possible to carry out several evaluation runs, each adapted to the respective purpose, within the course of one day. Programs and results were exchanged by e-mail.

Within the course of two research projects external scientists were able to use de facto anonymised individual data from wage and income tax statistics. In both cases a data file adequate for handling the problems was created on the basis of a specific question. The production of a general Scientific Use File with the data of the wage and income tax statistics is an unsatisfactory approach, because of the great number of characteristics which exist. In order to ensure the data security, such a data file would have to be sufficiently rough and would only be able to have a sub-section of the existing characteristics. As previous experience shows, the rapidly changing questions which are to be answered by the data material of the tax statistics, frequently require another set of data. Each general, previous stipulation of the data would mean a restriction of the possible research projects. For this reason at this moment in time it does not appear worthwhile creating a Scientific Use File of the wage and income tax statistics.

The wage and income tax statistics is the only statistical source of information on high income in Germany[9]. In voluntary sample surveys of households, as well as in surveys where the provision of information is mandatory, this sub-population cannot be substantiated at all or only very inaccurately due to methodical reasons. Therefore, despite some restrictions due to tax terminology and design possibilities, the wage and income tax statistics provided an important basis for describing high income in the German Federal Government's report on poverty and wealth. The commissioned report "Hohe Einkommen, ihre Struktur und Verteilung" [High income, its structure and distribution] (which the Federal Ministry of Labour and Social Affairs plans to publish) received intensive support from the Federal Statistical Office in the form of a procedure mix. In his capacity as acting expert, Prof. Dr. Merz from the University of Lüneburg had the possibility of basing his work on special preparations, an own program syntax and a de facto anonymised microdata file which had been created just for his work. Within the course of the special evaluation an estimation of the missing sub-population of rich households in the "Einkommens- und Verbrauchstichprobe" – EVS[10] (sample survey of income and expenditure) was carried out on the

9 See, amongst others, Bach / Bartholmai (2000)

10 Cut-off limit for the 1998 EVS: household net income of DM 420,000

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

basis of the 10% sample of the wage and income tax statistics for 1995. An SPSS syntax module was used for calculating the distribution of wealth according to the most varied of concepts and demarcations on the basis of a sample involving 3 million cases. With the aid of the de facto anonymised microdata file it was possible for the scientists to independently carry out a PROBIT analysis of wealth.

The fourth cited form of analysis of individual data by scientists only takes place in a few cases. In this type of cooperation the researcher in question must provide his services without charge, and in return also only receives a limited utilisation right to the results of his work. On the other hand, the statistical offices must integrate an external scientist in the working operation. A pending work within the tax statistics is intended to illustrate that these adversities for both sides can produce a positive aspect.

In cooperation with the Humboldt University of Berlin the Federal Statistical Office would like to develop a tax simulation model based on microdata. This model is then intended to be used by Destatis for the most varied of questions. A further project which would enable scientists to work with individual data of the tax statistics has been earmarked in the research and development plan of the Federal Statistical Office under the name Integrated Microdata File. This branch of research, which has been discussed since the 1960s, attempts to integrate different data records (e.g. wage and income tax, sample survey of income and expenditure, microcensus) in one data file. Plans envisage dividing this project into different sub-projects and then, if necessary, also putting them up for tender to external scientists.

5. Outlook

With the German Federal Government's report on poverty and wealth a new expanded requirement has been placed on the data of official statistics. Various works have shown that the existing data is not adequate for certain questions. Apart from other improvements, the sustained shortening of the periodicity of the wage and income tax statistics is also called for. The evaluations of the participating ministries are also positive with regard to the advantages of annual wage and income tax statistics. A politically motivated expansion of the data on which this is based also expands the data material available for scientific projects. However, apart from methodical and technical questions, the financing of such statistics has to be finally clarified.

Annual wage and income tax statistics would considerably expand the analysis possibilities. In particular dynamic analyses with a very broad pa-

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/

nel would be able to make interesting research work possible. Furthermore, the research geared towards an "integrated microdata file" could lead to a considerable expansion of the questions capable of being analysed[11].

Finally it must be pointed out that there could be a considerable change in the way individual data are dealt with in Germany. With regard to this, in particular developments in Europe can be cited. For instance, various regulations of the European Community contain attitudes towards maintaining confidentiality which deviate from German law. Here developments could arise which are not foreseeable for Germany. Hence, for instance, a "Center for Research of economic Micro-data" was created in the Netherlands in 1998 and is undergoing a trial period. At EUROSTAT a similar institution is envisaged.

References

Bach, Stefan / *Bartholmai,* Bernd (2000): Möglichkeiten zur Modellierung hoher Einkommen auf Grundlage der Einkommensteuerstatistik, Deutsches Institut für Wirtschaftsforschung, Diskussionspapier 212, Berlin

Bork, Christhart (2000): Steuern, Transfer und private Haushalte, Fiananzwissenschaftliche Schriften, Band 99, Peter Lang GmbH, Europäischer Verlag der Wissenschaften

Köhler, Sabine (1999): Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, 133 – 148, in Band 31 der Schriftenreihe Forum der Bundesstatistik ‚Methoden zur Sicherung statistischer Geheimhaltung', Herausgegeben vom Statistischen Bundesamt, Metzler- Poeschel, Stuttgart

Merz, Joachim (2000): The Distribution of Income of Self-Employed, Professions and Employees, in: R. Hauser und I. Becker (publisher), The Personal Distribution of Income in an International Perspective, Berlin / Heidelberg / New York, Springer Verlag

Rosinus, Wolfgang (2000): Die steuerliche Einkommensverteilung, Wirtschaft und Statistik, 6, 456 – 463

Spahn, P. B. / *Galler,* H. P. / *Kaiser,* H. / *Kassella,* Th. / *Merz,* J.(1992): Mikrosimulation in der Steuerpolitik, Wirtschaftswissenschaftliche Beiträge, 66, Physica-Verlag, Heidelberg

von der Lippe, Peter (1997): Änderung des Gesetzes über Steuerstatistiken, Steuer & Studium, 6, 265 – 268

Zwick, Markus (1998): Einzeldatenmaterial und Stichproben innerhalb der Steuerstatistiken, Wirtschaft und Statistik, 7, 566 – 572

– (1999): Steuerstatistische Einzeldaten und ihre Auswertungsmöglichkeiten für die Wissenschaft, Allgemeines Statistisches Archiv, 83, 248 – 253

11 With regard to specific work in the area of microsimulation, see also Bork (2000) and in particular Spahn / Galler / Kaiser / Kassella / Merz (1992)

DOI https://doi.org/10.3790/schm.121.4.639 | Generated on 2025-08-17 20:41:43
OPEN ACCESS | Licensed under CC BY 4.0 | https://creativecommons.org/about/cclicenses/