Wagner, Joachim

Working Paper

Improvements and Future Challenges for the Research Infrastructure: Firm Level Data

RatSWD Working Paper, No. 93

Provided in Cooperation with:
German Data Forum (RatSWD)

Suggested Citation: Wagner, Joachim (2009) : Improvements and Future Challenges for the Research Infrastructure: Firm Level Data, RatSWD Working Paper, No. 93, Rat für Sozial- und Wirtschaftsdaten (RatSWD), Berlin

This Version is available at:
https://hdl.handle.net/10419/186201

German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Working Paper Series

Working Paper No. 93

Improvements and Future Challenges for the Research Infrastructure: Firm Level Data

Joachim Wagner

July 2009

Federal Ministry
of Education
and Research

Working Paper Series of the Council for Social and Economic Data (RatSWD)

The *RatSWD Working Papers* series was launched at the end of 2007. Since 2009, the series has been publishing exclusively conceptual and historical works dealing with the organization of the German statistical infrastructure and research infrastructure in the social, behavioral, and economic sciences. Papers that have appeared in the series deal primarily with the organization of Germany's official statistical system, government agency research, and academic research infrastructure, as well as directly with the work of the RatSWD. Papers addressing the aforementioned topics in other countries as well as supranational aspects are particularly welcome.

RatSWD Working Papers are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Working Papers* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Working Papers* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

Starting in 2009, some of the empirical research papers that originally appeared in the *RatSWD Working Papers* series will be published in the series *RatSWD Research Notes*.

The views expressed in the *RatSWD Working Papers* are exclusively the opinions of their authors and not those of the RatSWD.

Improvements and Future Challenges for the Research Infrastructure: Firm Level Data

Joachim Wagner

Leuphana University Lueneburg (wagner[at]leuphana.de)

Abstract

This article discusses the use of enterprise- and establishment-level data from official statistics to document stylized facts, to motivate assumptions used in formal theoretical models, to test hypotheses derived from theoretical models, and to evaluate policy measures. It shows how these data can be accessed by researchers in Germany today and reports on recent developments that will offer new and improved datasets that combine data collected in separate surveys and by different agencies. The paper makes three recommendations for future developments in this area: (1) change the law to make the combination of data collected by different producers easier, (2) combine firm-level data across national borders and make these data available for researchers, and (3) find ways to enable researchers in Germany to work with confidential firm-level data via remote access 24 hours a day and 365 days per year.

Keywords: firm level data, Germany, FiDASt, KombiFiD, AFiD

1. What are firm-level data?

Firm-level data are data collected at, or related to local production units (establishments) or legal units (enterprises). The technical term used to describe this kind of data in official statistics is *wirtschaftsstatistische Einzeldaten*, or micro data for production units. This kind of data can either be collected in a survey (administered by a statistical office or by other institutions such as an opinion research institute or by a researcher at a university), or produced during a process that is related to administrative issues (for example, collection of taxes on sales or reporting to the social security system), resulting in what is named process-produced data.[1]

Usually, firm-level data are confidential - either by law (if they are collected in surveys from official statistics or are the outcome of administrative processes) or by an agreement between the (private, non-governmental) collector of the data and the firms that delivered the data. The reasons for confidentiality are manifold, including the fact that information delivered by firms that are required to report to surveys administered by official statistics has to be protected against competitors, and also that firms usually are only willing to respond to a survey voluntarily if they can be sure that any information considered to be "sensitive" will not be disseminated.

Confidentiality of firm-level data is a crucial issue for researchers who want to use micro data for production units in scientific studies. Although researchers are not at all interested in any of the establishments or enterprises per se, they need to access the data at the micro level to perform their statistical analyses and econometric estimations and thereby to uncover patterns of firm behavior and test theoretical hypotheses. This paper discusses issues related to the use of confidential firm-level data by independent researchers (i.e., those who are not working for the data producers). It begins with a review of what firm-level data are good for (in section 2), who produces firm-level data in Germany, and how researchers can gain access to these data today (in section 3). In section 4, new and ongoing developments are discussed that are leading to new products - new types of firm-level data will considerably enhance the research potential available to researchers in the near future. Section 5 concludes with a wish list.

1 For a discussion of other organizational data (for example data for organizations within the educational system) and publicly funded non-official organizational data collected by researchers, see the contribution by Stefan Liebig (2008) in this volume.

2. What are firm-level data good for?

Researchers use firm-level data in a wide range of areas in economics for four (not mutually exclusive) tasks, namely

- to document stylized facts that can not be uncovered by looking at aggregate data for industries or regions,
- to motivate assumptions used in formal theoretical models,
- to test hypotheses derived from theoretical models, and
- to evaluate policy measures.

The following three examples from different areas of economics - firm demography, job creation and destruction, and international firm activities - illustrate the need for, and the research potential of, the use of firm-level data:

1. Hopenhayn (1992) considers long run equilibrium in an industry with many price-taking firms producing a homogeneous good. Output is a function of inputs and a random variable that models a firm-specific productivity shock. These shocks are independent across firms and are the reason for the heterogeneity of firms. There are sunk costs to be paid upon entry and entrants do not know their specific shock in advance. Incumbents can choose between exiting or staying in the market. The model leads to three testable hypotheses, namely that firms that exit in year t were in t-1 less productive than firms that continue to produce in t, that firms that enter in year t are less productive than incumbent firms in year t, and that surviving firms from an entry cohort were more productive than non-surviving firms from this cohort in the start year. Wagner (2007a) uses a panel dataset for all manufacturing plants from Germany (1995-2002) to test these hypotheses econometrically, and finds that all three hypotheses are supported empirically.

2. It is often argued that in Germany jobs are mostly created in small- and medium-sized firms, while large firms generally tend to destroy jobs. The *Mittelstand*, or middle class, is considered the engine of job creation. Using panel data for manufacturing firms, Wagner (2007b) demonstrates that this simple view is wrong. Growing and shrinking firms, entries and exits can all be found in substantial amounts in all size classes within each time period considered. Economic policy measures with a special focus on firms from different size classes, therefore, cannot be justified by pointing to an extraordinary large contribution of these firms to job creation.

3. A large number of empirical studies for many countries (surveyed in Wagner 2007c) demonstrate that exporting firms are more productive than non-exporting firms of the same size from the same narrowly defined industry. This stylized fact motivated Melitz (2003) to set aside the standard assumption of homogeneous firms and to develop a model with heterogeneous firms where only the more productive firms in an industry export. This model has become the workhorse of a flourishing body of literature dealing with international firm activities. Using unique, recently released, and nationally representative high-quality longitudinal data at the plant level, Wagner (2007d) presents the first comprehensive evidence on the relationship between exports and productivity for Germany, a leading actor on the world market for manufactured goods. He documents that the positive productivity differential of exporters compared to non-exporters is statistically significant and substantial, even when observed firm characteristics and unobserved firm specific effects are controlled for.

All three examples demonstrate that using firm-level data is not only useful but indispensable for both sound empirical research (including the evaluation of policy measures and the derivation of policy recommendations) and crafting theoretical models that are relevant outside academic journals. In his Nobel lecture, James Heckman (2001, 674) named "the evidence on the pervasiveness of heterogeneity and diversity in economic life" the most important empirical discovery from econometric analyses using micro data. Everybody who ever worked with plant- or enterprise-level data will agree - there is no such thing as a representative firm, not even in 4-digit industries. We would not know this, and would be unable to base our theoretical models and the policy implications derived from these models on this knowledge if firm-level data was not accessible to researchers. Fortunately, such access is possible, as the next section discusses in greater detail.

3. Who produces firm-level data, and how can they be accessed by researchers today?

In Germany, the data for establishments and enterprises are collected or constructed by a number of institutions. The most important among them include:

- the Federal Statistical Office (*Statistisches Bundesamt*, or *Destatis*) and the Statistical Offices of the German *Länder* (*Statistische Ämter der Länder*), which administer a

large number of surveys as well as secondary statistics;

- the Federal Employment Agency (BA: *Bundesagentur für Arbeit*) and its research institute, the Institute for Employment Research (IAB: *Institut für Arbeitsmarkt- und Berufsforschung*), which uses information on employees covered by social security to construct establishment-level information on the number of employees and their average characteristics, and collects information on a wide range of issues for a panel of establishments in annual surveys for the IAB Establishment Panel;
- the German Central Bank (*Deutsche Bundesbank*) has a database with information from balance sheets and data for foreign direct investments of German firms.

Furthermore, firm-level data are collected on a large scale by research institutes (including the Ifo Institute for Economic Research in Munich and the Center for European Economic Research in Mannheim) and by the KfW (*Kreditanstalt für Wiederaufbau*), a bank that is closely related to the German state.

It should be noted that some of these firm-level data include information on the employees working in the firms, leading to what is named linked employer-employee (LEE) data. LEE data for Germany are the salary and wage structure surveys (*Gehalts- und Lohnstrukturerhebungen*) from official statistics, and the LIAB, which combines information from the IAB Establishment Panel with employee information from social insurance records.

More information on the firm-level data for Germany, and references to papers describing their information content, are given in Kaiser and Wagner (2008).

In the past, some of the data producers provided access to confidential firm-level data for researchers on the basis of individual contracts and contacts. For example, various statistical offices of the *Länder* allowed researchers to work with firm-level data either via remote data access (i.e., by sending programs to the office the output of which was checked for violation of data protection rules and then sent to the researchers) or by giving them a special status as an unpaid employee that makes it feasible for researchers to work with the micro data inside the office, strictly in accordance with all relevant data protection rules. Projects that pursued this type of access formed the network FiDASt - an acronym for Firm-Level Data from Official Statistics (*FirmenDaten aus der Amtlichen Statistik*). Results from these projects are documented in various contributions to professional journals and in three workshop volumes (see Schasse and Wagner 1999; 2001; Pohl et al. 2003). Furthermore, the IAB offered researchers the option of using the data from the IAB Establishment Panel via remote data access and the so-called *Schalterstelle*, a contact person in charge of running the programs and checking the output afterwards (see Kölling 2000).

In recent years, following the suggestions of the KVI (*Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik* 2001), most of the important producers of firm-level data - including the Federal Statistical Office and the Statistical Offices of the *Länder*, the IAB, and the Deutsche Bundesbank - established Research Data Centers (or *Forschungsdatenzentren*) that offer researchers convenient ways to work with confidential data via remote data access or by working in-house (see Zühlke et al. 2004; Kohlmann 2005; Lipponer 2003). Furthermore, Scientific Use Files (SUFs) were produced for several datasets that can be used by researchers on their own PCs in the office, as well as Public Use Files (PUF), which can be used by anybody, including students during courses (see Zwick 2007). Other data producers (like the KfW) offer researchers the opportunity to use the confidential firm-level data in joint projects with employees of the producers, including access to the data while working in-house. A survey of who offers what to whom and how is given in Kaiser and Wagner (2008).

Most recently, further progress on the way to a less restrictive access to confidential data was made by locating a Research Data Center outside the data producing institution and inside the institution where the researchers are. The statistical office of Berlin and Brandenburg opened a Research Data Center in the building of the German Institute for Economic Research (*Deutsches Institut für Wirtschaftsforschung, DIW*), making the work with the micro data from German official statistics much more convenient for DIW researchers (and for researchers working in the universities nearby).

Compared to twenty, ten, or even five years ago, things have improved a lot for researchers with regard to access to confidential micro data for establishments and enterprises. As the next section will demonstrate, there is more to come.

4. What will the near future bring? New products in the pipeline

Compared to firm-level data collected by research institutes, data from official surveys have several advantages: they often cover the whole population of targeted firms (not merely a small sample) and the firms are required to answer and answer correctly (there are no missing cases, no missing values, and - it is to be hoped - no wrong answers). Furthermore, the surveys are usually repeated periodically, and the data from various waves can be combined to build panel data sets. The extra costs associated with preparing data from official surveys for scientific research are not zero, but they are only a tiny fraction of what it would cost to collect data in a new survey. That said, there is one disadvantage of these data from official

statistics. Usually, they cover only a small number of items, often fixed by law. This leads to severe limitations with regard to the potential use of these data for scientific analyses.

A promising way to increase the research potential of data from the surveys of official statistics would be to combine the information collected for a unit (enterprise or establishment) in different surveys. This is technically feasible if each unit has a unique identifier (a unit number) that is used in different surveys. Fortunately, this is the case with firms surveyed by the Federal Statistical Office and the Statistical Offices of the *Länder*. Given that the law allows for matching data from various surveys administered by statistical offices, combined information from these surveys can be used in a single empirical investigation. The following example illustrates how such combinations can increase the research potential of firm-level data from official statistics.

Cost structure surveys collect information on, among other things, turnover and various categories of costs. From these data a rate of return can be computed to proxy the profit situation of the firm. How is this rate of return related to export activities of the firm? This question can not be answered using these data alone, because no information on exports is collected in the cost structure surveys. Information on exports, however, is available in another survey - a report covering the activities of manufacturing firms, which does not however contain any information itself about the profit situation of the firm. Combining data from these two different surveys leads to a dataset that makes it possible to investigate the role of exports for profitability (see Fryges and Wagner 2008).

Matched data from surveys collected by the statistical offices have been used in a number of studies recently. The datasets for these studies have been tailor-made by the Research Data Centers to suit the purposes of each respective study. This is both expensive and time consuming. In the AFiD project (where AFiD is an acronym for *Amtliche Firmendaten für Deutschland*, or official firm-level data for Germany) several standardized datasets are prepared that are combinations of data from various surveys (for details see Malchin and Voshage 2009). These combined data are available to researchers via the Research Data Centers of the statistical offices.

Datasets from the AFiD project will offer a convenient way for researchers to investigate questions that could not be answered using data from only one survey. Furthermore, the content of datasets prepared in the AFiD project can be enhanced by adding information from other sources. On the one hand, it is both technically feasible and legal to add data collected in special purpose surveys that are administered by the statistical offices only once. A case in point is the survey on international outsourcing activities of firms recently performed by the

German Federal Statistical Office (*Statistisches Bundesamt* 2008). The data from this survey have a limited amount of information, yet combined with all the other data for firms from the AFiD project these data offer the opportunity for exciting empirical research on various topics related to the determinants and consequences of international outsourcing. Note that the extra costs of adding these data to the datasets already available are negligible. On the other hand, in accordance with the law, and given that it is technically feasible, data from publicly available sources can be matched with the AFiD data to further enhance the information content of these datasets. To give an example, information about patents granted to the firms can be added. Augmented datasets of this type - or what might be labeled AFiD*plus* data - will offer attractive opportunities for empirical investigations in innovative fields.

While combining information available for a single firm from various surveys done by official statistics (in addition to publicly available information from other sources) in the AFiD project is an attractive way to build new rich datasets that are worth much more than the sum of their parts to a researcher, even more attractive datasets can be constructed when confidential firm-level micro data from the vaults of different data producers are matched on top of that. To give an example, information on foreign direct investments of firms is not available from any survey done by the statistical offices, but rather from balance sheet data processed by the German Central Bank (*Deutsche Bundesbank*). Combining AFiD data with the data for foreign direct investments leads to a dataset that makes it possible to investigate problems highly relevant for both scientific analysis and policy debates, including the consequences of foreign direct investments for jobs and wages in Germany.

Due to the sometimes tricky problems related to the definition of economic units, and the different identifiers used for firms by different data producers, this matching can be technically demanding. Furthermore, this is only legally allowed (in Germany, in 2008) if each firm explicitly declares in a written statement which of the data it delivers to the different data producers can be used for the matching. This leads to a fairly high bar set for any project trying to observe this procedure. Recently, the German Federal Ministry for Education and Research (BMBF: *Bundesministerium für Bildung und Forschung*) funded the research project KombiFiD (an acronym for *Kombinierte Firmendaten für Deutschland*, or combined firm-level data for Germany), a feasibility study in which a large number of firms is asked to agree to match their data and in which the technical problems of matching data across the boundaries of data producers are examined. The data from this feasibility study will be available at the Research Data Centers of the data producers involved in KombiFiD - hopefully beginning in the summer of 2009. More information and up-to-date news on the

project can be found at the website: *www.kombifid.de.*

5. A Firmpanelholic's Wish-List

Even considering all the recent progress that has been made in the way that firm-level data are prepared and made available for the use of independent researchers, and even with all the datasets currently under construction in the projects described above, there are still several wishes left unfulfilled. If a good fairy granted me three wishes related to firm-level data, this is what I would ask for:

1. Change the German law so that it is possible to match micro data for firms across the boundaries of data producers without requiring a written consent to this matching from the firms. The reason for this wish is obvious from the discussion presented in section 4.

2. Find ways to combine firm panel data across national borders, and to give researchers access to these data (see the International Public Use Microdata Series Project[2] that collects census data for persons and households from all over the world for a role model dealing with individual level data). The main reason for this wish is that we live in a time of increasing globalization. If the objects of our analysis - the firms - become more and more international, often controlling or being controlled by firms in other countries, the data we use should enable us to learn about the causes and consequences of their behavior by allowing access to micro-level data for all units connected to a firm, legally or otherwise, irrespective of the country these units are located in.

3. Find ways to enable researchers in Germany to work with firm-level micro data via remote access, available 24 hours a day and 365 days per year, rather than requiring them to send programs to the Research Data Centers or to go there in person (see Hundepool and de Wolf (2005) for a description of a pilot project at Statistics Netherlands). The reason for this wish is obvious to any researcher familiar with the conventional ways of working with confidential firm-level data: While it is possible to work using the current means of access, and it is infinitely better to have this opportunity than not to have any opportunity at all - it remains a second-best solution. Time for research is the ultimate constraint faced by researchers and the means of access available today are extremely time consuming. (As an aside, I would like to

2 www.ipums.org/international

add that the Scientific Use Files that can be used on the researchers' own PC are in my view no solution when it comes to firm-level data; see Wagner 2005.) While the space limitations for this report make it impossible to go into detail on this point, the example of Denmark (described in Kaiser and Wagner 2008) clearly demonstrates how such an "easy access" policy can be implemented. Based on an approved research proposal, researchers in Denmark can access the data on the mainframe computers in Statistics Denmark from their office PCs, with extremely high penalties for any misuse. Not that long ago, the kingdom of Denmark began in what is today the northern part of Hamburg, some 40 kilometers north from my office at the Leuphana. Given the high price of beer in Denmark I am not sure that I would wish this still to be the case - yet when I look at the ease of access to all kinds of confidential micro data that my colleagues at Danish universities enjoy, I do feel some regret. So, at the end of the day, I do wish that we would start to learn from the Danish experience.

References:

Fryges, H. and Wagner, J. (2008): Exports and profitability - First evidence for German manufacturing firms. Unpublished manuscript.

Heckman, J.J. (2001): Micro Data, heterogeneity, and the evaluation of public policy: Nobel lecture. Journal of Political Economy 109(4), 673-748.

Hopenhayn, H. (1992): Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60(5), 1127-1150.

Hundepool, A. and de Wolf, P.-P. (2005): OnSite@Home: Remote Access at Statistics Netherlands. http://unece.org/stats/documents/ece/ces/ge.46/2005/ wp.6.e.pdf.

Kaiser, U. and Wagner, J. (2008): Neue Möglichkeiten zur Nutzung vertraulicher amtlicher Personen- und Firmendaten. Perspektiven der Wirtschaftspolitik 9(3), 329-349.

Kölling, A. (2000): The IAB-Establishment Panel. Schmollers Jahrbuch / Journal of Applied Social Science Studies 120(2), 291-300.

Kohlmann, A. (2005): The Research Data Center of the Federal Employment Service in the Institute for Employment Research. Schmollers Jahrbuch / Journal of Applied Social Science Studies 125(3), 437-447.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001): Wege zu einer besseren informationellen Infrastruktur. Baden-Baden.

Liebig, S. (2008): Organizational Data. In this volume.

Lipponer, A. (2003): Deutsche Bundesbank's FDI micro database. Schmollers Jahrbuch / Journal of Applied Social Science Studies 123(4), 593-600.

Malchin, A. and Voshage, R. (2009): Official Firm Data for Germany.Schmollers Jahrbuch / Journal of Applied Social Science Studies 129(3) (in press).

Melitz, M.J. (2003): The impact of trade on intra-industry reallocations and aggregate industry productivity. Econometrica 71(6), 1695-1725.

Pohl, R./Fischer, J./Rockmann, U. and Semlinger, K. (Eds.) (2003): Analysen zur regionalen Industrieentwicklung. Sonderauswertungen einzelbetrieblicher Daten der Amtlichen Statistik. Berlin.

Schasse, U. and Wagner, J. (Eds.) (1999): Entwicklung von Arbeitsplätzen, Exporten und Produktivität im interregionalen Vergleich – Empirische Untersuchungen mit Betriebspaneldaten. Hannover.

Schasse, U. and Wagner, J. (Eds.) (2001): Regionale Wirtschaftsanalysen mit Betriebspaneldaten - Ansätze und Ergebnisse. Hannover.

Statistisches Bundesamt (2008): Verlagerung wirtschaftlicher Aktivitäten. Erste Ergebnisse. Wiesbaden.

Wagner, J. (2005): Anonymized firm data under test: Evidence from a replication study. Jahrbücher für Nationalökonomie und Statistik 225(5), 584-591.

Wagner, J. (2007a): Entry, exit and productivity. Empirical results for German manufacturing industries. University of Lüneburg Working Paper Series in Economics 44, March.

Wagner, J. (2007b): Jobmotor Mittelstand? Arbeitsplatzdynamik und Betriebsgröße in der westdeutschen Industrie. Vierteljahrshefte zur Wirtschaftsforschung 76(3), 76-87.

Wagner, J. (2007c): Exports and productivity: A survey of the evidence from firm-level data. The World Economy 30(1), 60-82.

Wagner, J. (2007d): Exports and productivity in Germany. Applied Economics Quarterly 53(4), 353-373.

Zühlke, S./Zwick, M./Scharnhorst, S. and Wende, Th. (2004): The research data centers of the Federal Statistical Office and the statistical offices of the Länder. Schmollers Jahrbuch / Journal of Applied Social Science Studies 124(4), 567-578.

Zwick, M. (2007): CAMPUS files - Free public use files for teaching purposes. Schmollers Jahrbuch / Journal of Applied Social Science Studies 127(4), 655-668.