FDZ Methodenreport

Methodological aspects of labour market data

No. 01/2008(EN)

The IAB Establishment Panel – from Sample to Survey to Projection

Gabriele Fischer (TNS Infratest Sozialforschung)
Florian Janik (IAB Research Unit Establishments and Employment)
Dana Müller (Research Data Centre of the BA at the IAB)
Alexandra Schmucker (Research Data Centre of the BA at the IAB)

Acknowledgements

We would like to thank Lutz Bellmann, Harald Bielenski, Udo Brixy, Micha Andreas Crimmann, Peter Ellguth, Nicole Gürtzgen, Tanja Hethey, Martina Huber, Peter Jacobebbinghaus, Susanne Kohaut, Johannes Ludsteck and Dieter Urban for their constructive criticism. We thank Claudia Schiener and her team for the translation. The authors remain solely responsible for the content.

Contents

1. Introduction

The IAB Establishment Panel is an annual survey of establishments and is unique in Germany, as it represents all industries and establishment sizes nationwide and can also be analysed on a longitudinal basis. The design of the IAB Establishment Panel was developed in the early 1990s and subjected to a wide range of tests. This process also involved parallel development activities taking place on the *Hanover Firm Panel*, which were carried out on behalf of the *Forschungsstelle Firmenpanel* at the University of Hanover (Gerlach et al.: 1998) and the Institute for Applied Economic Research in Tübingen (IAW).

The survey began in West Germany in 1993, with the aim of building up a representative information system for continuous analysis of labour demand. It has been carried out in East Germany since 1996, making it a nationwide survey. The IAB Establishment Panel is conceived as a longitudinal survey, i.e. a large majority of the same establishments are interviewed every year. Consequently, it enables both analysis of developments across time through comparison of cross-sectional data on different points in time, and also longitudinal studies of individual establishments.

Now in the IAB Establishment Panel approx. 16,000 establishments are surveyed on a large number of employment policy-related subjects, including employment development, business policy and business development, investment activities, innovations in the establishment, public funding, personnel structure, vocational training and apprenticeships, new and exiting personnel, recruitment, wages and salaries, working times in the establishment, further training and general data on the establishment. The survey also includes varying focal topics every year. With the exception of Hamburg, all the German federal states (Bundesländer) currently contribute regional extension samples to the IAB Establishment Panel. This firstly enables evaluations on the federal state level, and secondly results in a total range of samples that significantly widens the evaluation options on the nationwide level.

The IAB Establishment Panel contains high data quality, achieved by means of the high-quality sample, the high exploitation level and the sophisticated process of data monitoring and error correction. The survey is carried out by TNS Infratest Sozialforschung GmbH on behalf of the IAB. A general introduction to the IAB Establishment Panel is contained in German in Bellmann (2002) or in English in Kölling (2000).

The IAB Establishment Panel is based on a complex study design, which also presents challenges for users of the dataset. This paper provides an overview of the methodology of the IAB Establishment Panel. It goes into detail on the design of the samples and survey, the weighting process, and data access at the Research Data Centre (FDZ) of the Federal Employment

Agency (BA) at the Institute for Employment Research (IAB). The most important points are presented at the beginning of each chapter. It is intended for users of the IAB Establishment Panel, firstly as a collection of methodological aspects of the IAB Establishment Panel, and secondly to make it easier for first-time users in particular to start using the data.

This paper is also aimed at users of the IAB Linked-Employer-Employee Dataset (LIAB[1]), in which the IAB Establishment Panel is an important component.

FDZ Datenreport 6/2006 (Städele/Müller 2006) contains a description of the individual variables and their frequencies for the years 1993 to 2005. The frequencies of all variables from the 2006 wave are contained in an additional document at: http://doku.iab.de/fdz/iabb/Fallzahlen_IAB Establishment Panel2006.zip. The list of variables from 1993 to 2006 provides an overview of any changes (http://doku.iab.de/fdz/iabb/Variablenliste1993_2006.xls).

2. Sample Design

The Key Aspects in Brief

The population of the IAB Establishment Panel consists of all establishments with at least one employee liable to social security as of 30 June of the previous year.

The basis for sampling is the Federal Employment Agency establishment file. This is the only data source in Germany that covers all industries and establishment sizes. Establishment numbers for the IAB Establishment Panel sample are drawn from the establishment file. These establishment numbers form the basic survey units for the IAB Establishment Panel. Not every establishment number, however, represents a suitable unit for surveying.

The IAB Establishment Panel is based on a disproportionately stratified sample according to establishment size, industry and federal state (Bundesland). Weighted data must be used to make representative descriptive statements on the population.

In the case of industry analyses, users should bear in mind that the change in the system of classification of economic activities (in particular from WS73 to WZ93) means that data can only be compared over time to a limited extent. Comparative industry analyses can be carried out from 1993 to 1999 and from 2000 up to the most recent available data.

The foundation of new establishments is reflected in the IAB Establishment Panel by drawing "new" establishment numbers. These are not necessarily newly founded establishments, however.

[1] This dataset contains both the IAB Establishment Panel data and the process-produced data on individuals of the Federal Employment Agency (BA). This enables a simultaneous analysis of the supply and demand sides of the labour market (Alda/Bender/Gartner 2005).

The population of the IAB Establishment Panel consists of all establishments with at least one employee liable to social security as of 30 June of the previous year.[2] The basis for sampling is the BA establishment file, which is created on a quarterly basis and contains some two million establishments. The establishment file contains all establishments that notify the social security agencies of their employees as required. The establishments receive an establishment number for these notifications from the respective Employment Agency responsible for the establishment, as of 2007 from the BA's central establishment number service. These establishment numbers are compiled centrally in the BA establishment file. The establishment number is the relevant unit for the sampling and weighting processes.

2.1 The Establishment Number

An establishment number is initially a necessary indicator for the administration process. Every establishment is allocated an establishment number as part of the notification procedure for social security. An "establishment" in this sense denotes a regionally and economically separate unit, in which employees liable to social security work. Specific rules govern the allocation of establishment numbers, and are applied in the individual Employment Agencies. New establishment numbers are generally allocated when:

- an establishment is newly founded,
- a change of establishment owner[3] takes place,
- the economic activity classification of the establishment changes, or
- the employer submits an application, e.g. to combine several branches.

No new establishment number is allocated if an establishment moves to a different local authority district or a different employment agency district, unless the employer or health insurer places an application. Establishments in possession of an establishment number but not currently employing any employees liable to social security are referred to as "dormant". The old establishment number can be used again, provided the employer and health insurer agree. If the employer opens a new establishment, it may return to the former establishment number under certain circumstances.

[2] As of 2004, private households (previously from at least five employees as of 30 June of the previous year) and exterritorial units (e.g. embassies) have no longer been surveyed. These types of units are not widely spread. In addition, the survey was not designed for questioning private households. Exterritorial units were only very rarely willing to participate in the survey. The population has thus included approx. 30,000 fewer establishments since 2004, including 1,000 establishment numbers on which the establishment file contains no data.

[3] This does not apply to handovers of establishments to family members, provided the health insurer does not request a new establishment number.

In most cases, the establishment number is a suitable basis for identifying an establishment as a survey unit. However, this is not always the case. The above criteria for allocating establishment numbers imply, for example, that a single company may hold several establishment numbers if its units are in different employment agency districts or constitute differing, separate economic units. The reverse case is also possible: several branches of the same company in a single district are compiled into a single establishment in the establishment number statistics, for example head office and several branches, provided they operate in the same economic activity area and the employer and health insurer agree.

It is ultimately the interviewer who establishes to what extent the establishment number drawn represents a suitable unit for surveying, that is a unit for which the information required for the questionnaire is available in the establishment. The IAB Establishment Panel study design allows for certain deviations in identifying the establishment unit to be surveyed. However, there must always be a relation to the establishment number drawn. Chapter 3.3 contains further details on the establishment number and the identification of the establishments to be surveyed.

2.2 The Sampling Process

The sample for the IAB Establishment Panel is drawn from the establishment file[4] as of 30 June of the previous year, because the information on social security notifications is only available approximately six months after the respective reference date. However, the precise number of employees liable to social security is not available for all establishments as of the reference date. For approximately 5% of employees liable to social security, the information is unavailable as of the reference date, due to differing notification dates (Bundesagentur für Arbeit 2005). In these cases, information is adopted on the basis of the most recent available notification.

The sample is stratified according to establishment size[5], industry and federal state. The sample is multiply disproportional. The target degree of completeness of the cells in the stratification matrix is determined firstly by the number of extension samples (for an overview of the extension samples in the individual German federal states, see Table E in the appendix) and secondly according to the “principle of optimum stratification” proportional to the number of employees by establishment size classes (Buttler/Fickel 2002, p. 147f). As a result, large establishments, small federal states and small industries and the manufacturing industry in East

[4] For further information on the establishment file, see Brixy/Fritsch (2002).
[5] For the sample and weighting of the IAB Establishment Panel, the term “establishment size” denotes the number of employees liable to social security in the establishment.

Germany[6] are overrepresented (see Table A in the appendix). This disproportionate stratification has consequences for analyses, which are described in Chapters 4 and 5.

The stratification matrix has altered over time for the industries, as a result of changes in the system of economic activity classification (from WS73[7] to WZ93[8]) and modified aims with relation to content.

At the beginning of the panel in 1993, the stratification matrix consisted of 16 economic sectors, increasing to 20 in 2000 and falling to 17 in 2004 (see Tables B to D in the appendix). Users should bear these changes in mind when observing specific industries over time. Particularly the shift from WS73 to WZ93 in 2000 resulted in significant changes in the classification of establishments, meaning that comparisons across the entire survey period are only possible to a limited extent. Nor can a more aggregated classification achieve comparability.

Up to 2006, separate samples were drawn for each federal state (Bundesland) and for West and East Berlin. It is, however, no longer possible to allocate establishment numbers geographically precisely to West or East Berlin in the establishment file, and Berlin has thus been treated as a single entity since 2007. This has effects on the definition of West and East Germany. Up to and including 2006, East Germany included East Berlin and West Germany included West Berlin. As of 2007, West Germany consists of all West German federal states apart from Berlin, with the entire city of Berlin defined as part of East Germany.[9] This applies for the sampling process and thus also for the projection.

For the purpose of drawing the sample, the target values for the net sample are established in dependence on the financed interviews on the federal state level and for the manufacturing industry in East Germany. For these 17 partial samples (since 2003), a target degree of completeness is set for every cell of the stratification matrix (17 industries and 10 establishment size classes) according to the “principle of optimum stratification”. The size of the first interviewee sample is then calculated on the basis of the various expected response rates of the individual samples and survey methods.

---

[6] This extension sample is financed by the Institute for Economic Research Halle (IW Halle).

[7] WS73 is a classification of economic activities produced especially by the Federal Employment Agency (Bundesanstalt für Arbeit 1973), which differs from the classification of economic sectors used by the Federal Statistical Office (WZ79) (Bender et al. 1996: 16). Since the introduction of WZ93, the Federal Employment Agency and the Federal Statistical Office have used the same classifications.

[8] The change in the system of classification for economic sectors from WZ93 to WZ2003 (Federal Statistical Office: 2003) does not restrict comparability of industries, as the changes took place below the level of classification used in the IAB Establishment Panel.

[9] The extension sample for the manufacturing industry in East Germany forms an exception in this case. At the request of the Institute for Economic Research Halle (IWH), West Berlin was not added to East Germany, but the establishments from East Berlin were no longer included. The extension sample for the manufacturing industry thus refers to all East German federal states apart from Berlin from 2007 on.

2.3 The IAB Establishment Panel Partial Samples

The longitudinal character of the IAB Establishment Panel is also reflected in the sample. Firstly, the IAB attempts to survey as many establishments as possible over an extended period. Secondly, the IAB Establishment Panel sample must also depict the dynamics of establishment closures and "new" establishments. The annual gross sample thus consists of four respective partial samples:

1. responding establishments from the previous year ("continuers sample"),
2. non-respondents from the previous year willing to being surveyed again,
3. "new" establishment numbers,
4. extension sample.

These partial samples are necessary to depict continuity and change in the establishment population. Sample 1, the responding establishments from the previous year, consists of the establishments that are part of the existing stock of establishments from one year to the next. This sample ensures the longitudinal character of the IAB Establishment Panel. The non-

respondents from the previous year (sample 2) raise the number of cases in cross-sectional terms. In a survey of establishments over an extended period of time such as the IAB Establishment Panel, a concentration solely on the establishments continuing to exist from one year to the next would lead to selection effects. Establishments that have existed over an extended period differ in many operative characteristics from newly founded establishments. In order to depict this dynamic, "new" establishment numbers (sample 3) are added to the IAB Establishment Panel sample every year. These establishment numbers had at least one employee liable to social security as of the reference date, but not in the previous year. Such an establishment number does not necessarily denote a newly founded establishment – as explained above in the description of establishment number allocation. It can also be a "dormant" establishment number or an establishment that has existed for some time, but has only recently taken on an employee liable to social security. This definition deviates from that of newly founded establishments. These establishments should therefore not be uncritically identified as newly founded establishments. Users should make additional reference to the information from the questionnaire in such cases. The attribute "new" is always written in quotation marks below for these establishment numbers, so as to prevent false conclusions. Despite these difficulties arising from the establishment file system, this procedure is the only feasible option for depicting the activity of founding new establishments within the parameters of the sampling process (cf. footnote 20).

In addition to the samples described above, it is sometimes necessary to add further existing establishments as of the reference date, in order to make up for losses and achieve the required number of cases in the individual federal states (sample 4).

Up to the 2001 wave, all establishment numbers ever contained in the gross sample in any of the waves that did not participate or no longer took part in the survey were blocked for all further waves. That is, they were no longer available for the sampling process. As of the 2002 wave, establishment numbers that have already been included in the gross sample, but became non-respondents in the meantime, can be included again after a certain period of time. The reason for this decision was that the population in certain industries or federal states was almost exhausted in the upper establishment size classes, causing problems in filling the cells in the stratification matrix. The intervening period is determined based on the need for extension establishments. An artificial identification number (idnum) is allocated for processing in the dataset. For organisational reasons, the establishments added back into the gross sample receive a new identification number.[10]

3. Survey Design, Field Work and Editing

The Key Aspects in Brief

The questionnaire consists firstly of one block of questions asked in identical form every year. These questions are supplemented by subject areas that are repeated at regular intervals. For longitudinal analyses, users should pay attention to the precise formulation of the questions, as some deviations occur. There are also questions on specific focus subjects, which vary every year.

The survey is generally carried out in the form of face-to-face interviews, with written postal surveys also taking place in some federal states (Bundesländer). Users should thus bear in mind the survey type variable for analyses of the federal states in question.

The field phase takes place in the third quarter of every year. The interviewed establishments are informed in advance by letter. Before the interview starts, the correct establishment unit is identified. This is mainly done by comparing the number of employees in the interviewed establishment with the data from the previous year's survey or the BA statistics.

All in all, the IAB Establishment Panel has good response rates for surveys of this kind. However, there are differences depending on the partial sample type and the survey method: "continuer" establishments have a higher response rate than establishments included for the first time, and

establishments surveyed in face-to-face interviews have a higher response rate than establishments surveyed by post.

In order to enhance the data quality, the data are checked and errors are corrected in the course of an extensive editing process.

3.1 Questionnaire Design and Pre-Test

As well as the sample, the questionnaire also has to take the longitudinal character of the IAB Establishment Panel into account. The main aim is to gather certain information on a regular basis in order to measure developments. Simultaneously, up-to-date questions relevant for labour market policy also have to be included in the survey.

The IAB Establishment Panel questionnaire contains numerous questions that are asked in every wave, so as to depict changes consistently over time. This basic programme of questions is generally identical over the years. Several questions, however, have had to be changed at some point. Users should bear these changes in mind for longitudinal analyses.[11]

Up to the 2007 wave, this basic programme consisted of the subject blocks Employment Development, Business Policy and Development, Vocational Training, Personnel Structure and Personnel Movements, Investments, Wages & Salaries and Adherence to Collective Agreements. In addition, specific subject blocks are also regularly included in the questionnaire at certain intervals, e.g. subjects such as further training, working time, public funding and innovations. Please also bear in mind that the “general information” in the questionnaire is not gathered every year for every establishment, but usually only for first-time respondent establishments. In the case of continuer respondents, the relevant information is carried over from the previous year.

The design of the questionnaire was changed slightly in 2007. The basic programme was supplemented by basic indicators from the previously multi-year subject blocks Further Training, Innovation and Working Times. These areas will be surveyed annually from the 2008 wave.[12] For all subject blocks regularly surveyed in the IAB Establishment Panel, the basic information is now available in every year, thereby raising the analysis potential, as content-based links are also

[10] It cannot be assumed, however, that an establishment with different identification numbers and identical enterprise numbers in the different waves represents the same economic unit (see Chapters 2.1 and 3).

[11] FDZ Datenreport 2/2006 (Alda et al. 2006) contains a description of adaptations of sample variables to remain consistent over time. Additionally, PanelWhiz, a Stata tool for creating panel datasets quickly and easily, is due for launch shortly (for further information on PanelWhiz: http://www.panelwhiz.eu/).

[12] To implement this concept, several questions were deleted from the basic programme. These included the questions on planned investments and investments in information and communication technology.

possible on a cross-sectional basis. At two-year intervals, these basic indicators are supplemented or extended by additional questions. [13]

Alongside the fixed elements of the questionnaire, varying current focus subjects are included every year. These include, for example, questions on demand for qualified employees, on employment of older workers or securing employment and location.

The design of the IAB Establishment Panel was developed and extensively tested in the early 1990s. Before the survey was extended to cover the East German federal states, test surveys were carried out by TNS Infratest and the Institute for Socio-economic Structural Analysis (SÖSTRA) on behalf of the IAB. Furthermore, newly developed or re-formulated questions for the current focus subjects are tried out in pre-tests. Interviews including these new questions are held in approximately 100 establishments across the country. The specially trained interviewers[14] not only test the real interview situation, but also deal with any problems with understanding or interpretation with the interviewee. The findings from these pre-tests are integrated into the development of the questionnaire.

3.2 Survey Method

The survey is generally carried out in the form of face-to-face interviews in the establishments, by employees of TNS Infratest. Interviewer continuity is a decisive factor for the success of the survey. TNS Infratest therefore aims to send the same person to carry out interviews in each respective establishment. As the questionnaire contains a relatively large number of questions on figures, which the interviewees cannot answer spontaneously, the questionnaire may be left behind in the establishment. This means that data not available during the interview is researched afterwards and added to the questionnaire. This option is mainly used by larger establishments. The majority of establishments, however, are surveyed on an entirely face-to-face basis (2006: 73%).

Written surveys have taken place in some cases since 2000. This became necessary because extension samples had to be realised in certain federal states (Bundesländer) in order to carry out evaluations on the federal state level. The financial means were not sufficient for the large number of person-to-person interviews, however, which is why a large part of the sample is surveyed by post in these federal states. This applies to Bremen from 2000 to 2002 and Hamburg in 2000 and 2001. The establishments from the extension samples have been surveyed by post in Saarland since 2001 and in Schleswig-Holstein since 2002. For the first postal surveys in 2000, tests were carried out to establish to what extent the mixed methods would result in distortion of

[13] This created a regular rhythm for the multi-year modules. Previously, some subject blocks were only surveyed every three years.
[14] These interviewers are project employees of the IAB in selected local employment agencies (ProIAB).

content. In some cases, these tests found significant differences in response behaviour between the face-to-face interviews and the written postal surveys. When analysing the federal states in question for the relevant periods, researchers should therefore take the survey type variable into account.

3.3 Field Phase and Identification of the Correct Establishment Unit

As the surveys take place in the third quarter, the questions on stock figures generally refer to the reference date of 30 June, and the questions on flow figures (e.g. newly recruited or exiting personnel) to the first half of the survey year. This is designed to minimise memory errors. The establishments are informed of the survey by a letter of announcement[15] in mid-June. The interviewers then contact the establishments. The establishments surveyed in writing receive the questionnaire with the letter at the end of June. These establishments are sent reminder letters to motivate them to participate in mid-July and mid-August. Both the face-to-face interviews and the written postal surveys are generally completed by mid-October. In parallel, the surveyed data undergoes checks and errors are eliminated (see Chapter 3.5).

Before the interview starts, the interviewers have to ensure that the interview is carried out for the correct establishment unit. It is essential for longitudinal analyses that the interviews always refer to the same establishment unit. This is the only way to ensure that changes over time (such as employment development) depict actual changes and are not based on deviations in the surveyed establishment unit. The correct establishment unit is identified at the beginning of the interview, on the basis of the establishment name, the number of employees liable to social security on the reference date (30 June of the previous year) and – in the event of deviations – the establishment number.

In the case of first-time respondent establishments, the employee figures are compared with the number of employees liable to social security as of the reference date (30 June of the previous year) according to the establishment file. If it proves impossible to carry out the interview with precisely the unit drawn in the sample because this does not constitute an economically viable unit or there is no specific information available on this unit in the establishment, the survey may deviate from the originally selected unit, in accordance with fixed rules. The interviewed establishment unit must, however, bear some relation to the establishment number drawn. This may be the case if a larger establishment unit is surveyed which contains the establishment number, or if a smaller establishment unit is surveyed for which the establishment number partially applies. These deviations are documented in the address protocol, so that the same (deviating) unit

[15] The establishments receive letters from the Federal Employment Agency (BA) and the Confederation of German Employers' Associations (BDA), requesting that they participate in the survey.

can be surveyed again the following year. If it proves impossible to find a link to the establishment number drawn for first-time respondent establishments, no interview is carried out.

In the case of continuer respondent establishments, the correct establishment unit is determined on the basis of the employee figures from the previous year's questionnaire. If deviations occur, the establishment number is again used to determine whether at least part of the establishment or a larger unit bears the establishment number in question. If this is the case, the survey is carried out and the relationship between the unit surveyed and the establishment number is documented in the address protocol. These establishments can then be evaluated for the cross-sectional survey, but are no longer available for panel analyses.

The establishments surveyed by post receive the relevant information for determining the correct survey unit along with the questionnaire, and are requested to refer all data to this unit.

3.4 Non-Response to Interviews and Questions

The response rate to the surveys has varied between 63% and 73%.[16] The variations in the response rates are mainly due to differing sizes of the extension samples. As the response rates among establishments surveyed for the first time are significantly lower than those of continuer establishments[17], the total response rate is much lower in the years with large extension samples. Alongside differences between first-time and continuer respondents, the response rate also differs according to the survey method. Establishments are far more willing to participate in the face-to-face interview than in the written survey. For example, in 2006 the response rate among first-time establishments contacted by post was 12.9% - significantly lower than the establishments first interviewed orally at 36.3%. Similarly, in the same year the continuer establishments were less willing to participate in the written survey at 61.9% than in the on-site interviews. The response rates for the orally interviewed continuer establishments, however, are stable at between 81% and 84%.

To judge the survey quality, one must look at the non-responses to specific questions as well as the non-responses to the entire survey. Questions with high non-response rates are either hard to understand, hard to answer, or participants frequently refuse to respond to them. These values are registered as "no response" and coded "-9".[18] Across the waves of the IAB Establishment Panel, the sensitive variables such as business volume, total wages & salaries, share of advance

[16] This calculation does not include interviews unsuitable for evaluation.
[17] E.g. the response rate for the new entries sample in 2006 was 30%.
[18] A few questions in the first waves had the additional value "don't know" (-8). As the IAB Establishment Panel surveys almost exclusively facts that are known to the establishment, it does not generally make sense to distinguish between "don't know" and "no response". For questions not asked of certain establishments due to the questionnaire filter system, the dataset contains system missings.

performance and cost of debt in total sales and total investment grants, always have the highest non-response rates, but these are relatively stable.

In the written survey, the “no response” rates are considerably higher than in the oral interviews.

The lower rate of non-response items and the higher response rates in the face-to-face interviews underline the data quality arising from the survey method applied.

3.5 Data Checking (Editing)

In parallel to the field phase, the data are comprehensively checked. As well as monitoring the interviewers, this process mainly optimises the data quality. To do so, the responses to individual questions are checked for consistency and plausibility. This process makes additional use of both questions linked to the question being checked and questions from the previous year, provided these are available. If errors/implausibilities cannot be cleared up on the basis of the questionnaire data, the establishment is contacted by telephone and the problem is solved in conjunction with the interviewee. The editing process also includes another check of the relationship to the correct establishment unit. Should the editing process establish that first-time respondents do not have any relation to the establishment number and the establishment was thus wrongly surveyed, these interviews are excluded.[19] The same applies to interviews with high rates of missing or erroneous responses and questionnaires that arouse suspicion of falsified interviews.

4. Cross-Sectional Evaluations

The Key Aspects in Brief

Cross-sectional evaluations enable content-based analyses for a single wave or time-series comparisons over several waves.

In order to make representative statements for the population, it is necessary to carry out cross-sectional weighting for descriptive evaluations, due to the disproportionate sample approach. This weighting enables regional, industry-based and size-differentiating studies. Below the dimensions of the stratification matrix taken into account for the weighting process, valid statements are only possible to a limited extent.

The basis for cross-sectional evaluations is the cross-sectional cases of the respective wave. These consist of all establishments with at least one employee liable to social security as of the reference date (30 June of the previous year). These cases can be selected with the variable

[19] Deviations are allowed for continuer establishments, as described above. The deviations must be plausible and documented. The interviews are used for the cross-section, but not for the longitudinal analysis.

querxxxx. “xxxx” stands for the respective survey year (e.g. quer2006 for the survey year 2006).

The respective cross-sectional weighting factor is denoted by hrxxxxq (e.g. hr2006q).

As the IAB Establishment Panel is based on a sample, users must take error tolerances into account in their evaluations.

Please also take suitable account of the disproportionate nature of the sample for multivariate studies.

The survey design of the IAB Establishment Panel enables evaluations on a cross-sectional and longitudinal basis. For both evaluation options, users should take care to select the respective correct cases for the study and apply the weighting accordingly. The two types of evaluation are based on different logics: a cross-sectional study looks at the establishments in the survey in the relevant year. In a longitudinal study, however, researchers can look at developments of individual establishments involved in the survey over a longer period. Additionally, the dynamics of foundations and closures[20] can also be integrated into the analysis. Both the logic of the definition of cross-sectional and longitudinal cases and the respective weightings differ from one another. As most analyses are generally either cross-sectional or longitudinal, the two types of evaluation are presented separately below.

4.1 Definition of the Cross-Sectional Cases

The cross-sectional cases represent the population of the respective wave. The criteria for a cross-sectional case are that a valid questionnaire has been completed and that the establishment had at least one employee liable to social security as of the reference date of 30 June of the previous year. In the datasets of the individual waves, the cross-sectional cases can be identified respectively via the string variable querxxxx (xxxx describing the year of the respective wave, e.g. quer2006 for the 2006 wave). These contain the value “Q”[21]. These cases form the basis for the cross-sectional weighting in each wave.

4.2 Cross-Sectional Weighting

The multiply disproportionate structure of the sample is corrected with the aid of a weighting procedure and adapted to the structure of the population. The weighting takes place in the form of a projection onto the number of establishments in the population. This applies for the whole

[20] For a definition of “new” establishments in the IAB Establishment Panel, please see Chapter 2.2. Genuine newly founded establishments can also be identified via the foundation year surveyed in the questionnaire.

[21] In the 1996 wave, users must also include the letter ‘L’ for the definition, which describes one-off additional cases in West Berlin.

of Germany, for East and West Germany, and also for the individual federal states (Bundesländer) and the establishments in the manufacturing industry in East Germany.

The required structures for weighting the cross section are produced from the BA establishment file as of the reference date (30 June of the previous year). This consists of the population per federal state across the industry and size classifications, in accordance with the stratification matrix for the sampling process (from 2003: 170-cell matrix). As a valid projection requires sufficient net cases per weighting cell, not all federal states can be weighted using this 170-cell matrix. In the federal states with a lower number of cases, the projection takes place onto the respective marginal distributions, i.e. the margin of the establishment size classifications and the margin of the industries.[22]

The projection includes all cross-sectional cases with the surveyed actual values on employment, industry allocation and federal state. The establishments are thus allocated to the separate cells of the stratification matrix according to their responses to the questionnaire. The sampling cell and the weighting cell of a single establishment may therefore differ from one another.[23]

In the first step of the cross-sectional weighting process, the “new” establishment numbers from the new entries sample are projected onto the appropriate target values from the establishment file.

In a second stage, all surveyed establishments are projected per federal state onto the target structure according to the BA. As mentioned above, this takes place in the large federal states with sufficient numbers of cases on the basis of the stratification matrix, and in the smaller federal states on the basis of marginal distributions of industries and establishment size. Projecting the establishments onto the target structure according to the BA can present a problem if very few or no cases at all are present in the individual cells in the sample. In order to have the correct number of establishments nevertheless, cells are combined in such cases.

A separate weighting process is carried out for the extension sample in the manufacturing industry in East Germany. The weighting factors are calculated across 10 establishment size classes and 15 branches of the manufacturing industry, without differentiation by federal state.

Next, the marginal distribution of establishments via industry and establishment size class is checked again and adapted as necessary.

[22] This applies to Bremen and Schleswig-Holstein, for example.
[23] It is possible that more establishments are allocated to a cell in the course of the survey than are present in this cell in the population. Up to 1998, such cases resulted in weighting factors below one. From 1999, the lowest possible value of the weighting factor in such cases is one.

The final stage is to adjust the weighting factors to the number of employees liable to social security as of the reference date. In this step, care is taken to preserve the number of establishments across establishment size classes, industries and federal states, despite the adjustment to the number of employees liable to social security. For this reason, the adjustment of the employees liable to social security takes place within the stratification cells. The weighting steps are repeated on an iterative basis.

This process may produce very large weighting factors. This applies particularly to small establishments, for which the numbers in the population are relatively high and the selected sets in the sample are relatively small. In order to limit the resulting problems of outliers, the level of the weighting factor is checked and limited to a maximum of 3,000.

As a result, every establishment is given an individual weighting factor (see below: "Alteration of Weighting Factors"). Subsequent to the weighting process, the weighted data are intensively checked and compared with external sources in cooperation with the IAB and other institutions[24] involved in evaluating the IAB Establishment Panel, in order to achieve maximum validity.

Alteration of Weighting Factors

Due to the nature of the weighting process, it is generally the case that an establishment is allocated different weighting factors in two subsequent waves. This applies to both the cross-sectional weighting process and the longitudinal process. There are various reasons for this, e.g..:

- because the target structure (the number of establishments or employees in the target structure of the respective weighting cell) changes,

- because the number of surveyed establishments in the weighting cell changes due to extensions or non-responses, and/or

- because an establishment changes size, industry or federal state between two waves.

[24] Alongside the IAB, the IAB Establishment Panel is evaluated by further research bodies on behalf of the participating federal states and institutions. In 2006, these bodies were the BAW Regional Economic Research Institute, the Institute for Applied Economic Research (IAW) Tübingen, the Institute for Socio-economic Structural Analysis (SÖSTRA), the Institute for Economics, Labour and Culture Frankfurt (IWAK), the Halle Institute for Economic Research (IWH) and the International Institute for Empirical Socio-Economics (INIFES).

4.3 Notes on Cross-Sectional Analyses

The respective cross-sectional weighting factor is stated in the datasets as hrxxxxq (e.g. hr2006q). All variables surveyed in the questionnaire can be weighted on the basis described above. It is possible – depending on the observation period – to carry out regional (by federal state), industry-specific and size-differentiating analyses[25]. Researchers should, however, bear in mind that no valid statements are possible below the stratification dimensions used in the weighting process (e.g. on the district (Kreis) level). Please also take into account that evaluations on the federal state level are not possible for Hamburg, due to insufficient numbers of cases. Strongly differentiated descriptive analyses should also bear the number of cases in mind. Researchers are recommended to use only analysis groups containing at least 20 establishments in their unweighted state.

As with all sample surveys, researchers must allow for a statistical error tolerance with the results of the IAB Establishment Panel. The data gathered in the sampling process are estimates for the share of absolute size of the corresponding characteristics in the population, and are by nature imprecise to a certain extent, which can be approximated. In general, the greater the dispersion of a characteristic in the population and the smaller the number of (unweighted) cases forming the basis of the estimate, the higher the statistical imprecision becomes. When interpreting the data, it is important to know the level of imprecision.

The error tolerance table (see Table H in the appendix) can be used to estimate the imprecision closely for various numbers of cases and proportions. With proportions and an analysis of 15,000 cases, deviations of 2 percentage points and more can be assumed, as a rule of thumb. Some of the questions have filter questions before them in the questionnaire, which lowers the number of cases. These questions have higher tolerance levels. The above rule of thumb does not apply to absolute numbers either. Your statistics software may offer a simple option for providing confidence areas.[26] If the distribution of an estimator for a statistic is unknown or your software does not allow you to calculate confidence intervals for this statistic, these can also be determined by means of bootstrapping (Green 2003, pp. 924f).

For descriptive evaluations of cross-sectional data, researchers are recommended to round employee figures to whole thousands and to use percentages without decimal points.

[25] One prerequisite – which cannot generally be checked, but is usually regarded as given in this type of study design – is that a selection is made for the characteristic in question, regarding willingness to participate and response behaviour only along the lines of the weighting dimensions.

[26] Stata, for instance, allows estimates of weighted numbers of cases including confidence intervals with the command -svy: total-, while proportions are calculated using -svy: proportion-. Alongside the weighting factor, users also have to provide an identifier for the stratification matrix cells.

Due to the disproportionate sample, which is corrected by the weighting process, the results of weighted and unweighted multivariate estimates may differ. In the cross-section, researchers can use both weighted and unweighted multivariate processes as they see fit. In order to control the effects of the disproportionately stratified sample, users are recommended to adopt the respective stratification variables (industry, establishment size and federal state as of the reserve date in the previous year) as independent control variables in the model.

5. Longitudinal Analyses

The Key Aspects in Brief

Longitudinal analyses allow researchers to look at developments in individual establishments. They also enable the study of foundation and closure dynamics.

Five longitudinal sections are defined in the IAB Establishment Panel, with the starting years 1993, 1996, 2000, 2003 and 2007.

Panel cases consist of establishments already contained in the starting years, “new” establishments in the subsequent waves and establishments going out of operation during the panel period. A key prerequisite for the definition of the panel cases is that the same establishment unit is surveyed as in the previous year. These cases are identifiable in the dataset by the variables panyy_zz, whereby yy stands for the starting year and zz for the last year included in the longitudinal section (e.g. pan93_06 defines the panel cases from the first wave in 1993 up to the wave in 2006).

This definition implies that not all establishments in any cross-section are contained in a longitudinal section that includes the respective cross-sectional year. This applies to the establishments in the reserve sample, which have been drawn from the stock records, and to establishments at which a deviating unit was surveyed. Only in the starting wave of the respective longitudinal section are all establishments also cross-sectional cases.

The disproportionate sample approach from the cross-section is continued in the longitudinal section. For this reason, descriptive longitudinal analyses must also be carried out with weighted data, so as to make representative statements on the population. The respective weighting factor is contained in the variable hryy_zzp (e.g. hr93_06p).

Multivariate or panel-econometric analyses must also take account of the disproportionate nature of the sample, bearing in mind the error tolerances.

On the basis of the wave code, researchers can also define their own longitudinal sections. There are no weighting factors for these, however.

Unlike cross-section analyses, longitudinal evaluations offer the possibility of analysing developments and links between establishment characteristics across time, on the individual establishment level. Both the definition of panel cases and the longitudinal weighting process are designed significantly differently to the respective processes in cross-section. The longitudinal processes have to take into account the dynamics of foundations and closures, as well as changes in the individual establishments over time, such as growth and shrinkage. For an illustration of a weighted longitudinal analysis, we recommend Chapter 6.2 of IAB Forschungsbericht 5/2007 (Fischer et al. 2007, only in German).

5.1 Definition of Longitudinal Cases

Panel cases consist of establishments from the cross-section sample of the respective basis year plus all “new” establishments from the new entries and extension samples of the following years. All panel cases must have either a suitable interview for the identical establishment unit from the respective first up to the latest available survey, or the information that the establishment number is no longer valid. Thus, not only those establishments that have at least one employee liable to social security at the start and end of the longitudinal period are defined as panel cases. Were the panel cases limited to this definition, only partial analysis of the establishment dynamic would be possible across time. The establishment dynamic includes the closure of establishments as well as the arrival of “new” establishments. In order to depict these dynamics, the panel case definition includes establishments no longer in operation, “new” establishments and establishments without employees liable to social security. Establishments that change location from West to East Germany or vice versa are excluded as panel cases.

The starting point for forming a longitudinal section is all cross-sectional cases existing in the starting year. These represent the stock of all existing establishments in the starting year, the development of which shall be observed in the subsequent years. In the subsequent waves, panel cases are then defined as follows:

1. Panel cases from the previous wave that still exist:

This group represents the “surviving establishments”. In order to avoid distortion in the analyses, however, only those surviving establishments are continued as panel cases for which the interview could be carried out with the same establishment unit as in the previous wave. If it is not possible to hold an interview with the same unit as in the previous year, this establishment is still available for the cross-section but not for longitudinal analysis.

2. Panel cases from the previous wave that went out of operation in the previous wave or earlier waves:

These establishments depict establishment closures. No interview can be carried out for these establishments. That means that from the year in which they went out of operation, they are only contained with the information that these establishments were no longer in operation during the relevant period for the panel case definition.

3. “New” establishment numbers from the new entries samples of the subsequent years:

Alongside existing establishments and those no longer in operation, “new” establishments make up the third part of the establishment dynamic. The “new” establishment numbers, which are adopted into the sample in each wave (see Chapter 2), are thus also a component of the panel case definition, provided an interview was carried out with the establishment unit drawn by the BA.[27]

The establishments that are drawn from the stock data in each wave as replacements for non-responders are not included in the panel definition, as the development of the stock establishments is only observed from the respective starting year on.

In the IAB Establishment Panel, various longitudinal sections are defined, beginning with five different starting years:

Longitudinal section 1 (West Germany only) in the period from 1993 to 2006,

Longitudinal section 2 in the period from 1996 to 2006,

Longitudinal section 3 with the starting year 2000 up to the latest available data,

Longitudinal section 4 with the starting year 2003 up to the latest available data,

Longitudinal section 5 with the starting year 2007 up to the latest available data.

The panel cases are identifiable in the respective datasets by the variables panyy_zz, whereby yy denotes the starting year and zz each subsequent year (e.g. pan93_06 defines the panel cases from the first wave from 1993 to 2006). The variables are denoted by “P” if the case is part of the corresponding longitudinal panel. Weighting factors also exist for these defined panel cases. For the analysis, the respective periods for study do not necessarily have to be identical with the longitudinal periods. The analysis period must, however, be completely contained in the respective longitudinal period. For example, the longitudinal panel from 1996 to 2006 can also be used to evaluate the period from 1998 to 2001.

[27] First-time respondent establishments from the partial sample of “new” establishment numbers, for which a deviating unit was surveyed (see Chapter 3.3), are not defined as “new” establishments and thus not taken into account in the construction of the longitudinal section.

5.2 Longitudinal Weighting

Similarly to descriptive evaluations of cross-sectional data, descriptive analyses of longitudinal data also lead to distorted results without weighting, due to the disproportionate sample. When carrying out longitudinal weighting for an establishment survey with a disproportionate sample, various requirements must be borne in mind:

The establishments in the stock data in the respective waves must reproduce the industry and establishment site structures as of the reference date for the respective wave within the longitudinal weighting. That means that these establishments have to be weighted for each individual wave in the longitudinal panel to match the target matrix of the establishment file.[28] As many establishments are contained in the stock data in several waves, this means that these establishments have to be weighted in various waves to match differing populations, but only one factor per establishment is possible for the entire longitudinal period.

Establishments that grow or shrink during the period present a particular challenge for longitudinal weighting, in as far as the changes lead to an alteration in the size class relevant for the weighting process, as the average weighting factors vary strongly depending on the establishment size, due to the disproportionate sample. For new definitions of a longitudinal section, the cross-sectional factor of the basis wave serves as an entry factor for the longitudinal weighting, and for existing longitudinal sections the longitudinal factor of the previous wave is taken. If an establishment switches, e.g., from one size class to the next higher class between two waves, it initially receives an entry factor that is too high – due to the fact that the factor of the previous wave serves as the entry factor – meaning that this establishment would be too highly weighted. Vice versa, if an establishment shrinks from a larger size class to the next lower class in the stratification matrix, this establishment initially receives a factor that is too small, as it had received a smaller factor as a large establishment in the previous year, to balance out the disproportionate sample. These effects are checked, so as to minimise distortions in the evaluation of the respective wave.

The longitudinal weighting process is designed in such a way that an analysis of the panel cases with the longitudinal weighting factors – for comparably defined sub-groups – results in approximately the same distributions for all questions from the previous surveys as in the corresponding cross-section analyses of the wave in question.

Even with such a large sample as in the IAB Establishment Panel, longitudinal weighting theoretically requires considerably more parameters to be checked than is practically possible and sensible. Longitudinal weighting therefore has to concentrate on the key parameters.

[28] Due to the low number of longitudinal cases, the industry structure is compounded into seven industries for the longitudinal weighting process.

Longitudinal weighting takes place in eight subsequent steps, which are explained in detail in the appendix. With the exception of the longitudinal panel from 1993, which contains only West German establishments, all steps are carried out separately for East and West Germany.[29]

The individual steps towards calculating the longitudinal weighting factors take the following aspects into account:

- The structure of the establishments by industry, size class and federal state contained in the latest available data,
- the number of existing, no longer operational and newly founded establishments in every year of the longitudinal period,
- survival and removal of the establishments from the stock data of the starting year and the "new" establishment numbers added across the longitudinal period,
- growth and shrinkage of the establishments from the stock data of the starting year and of the "new" establishment numbers added across the longitudinal period,
- disproportionate non-responses, on the basis of the questionnaires from the previous year, and
- the industry and size structure of the population in each of the waves contained in the longitudinal period.

5.3 Notes on Longitudinal Analysis

The respective weighting factor is contained in the variable hryy_zzp (e.g. hr93_06p). All variables surveyed in the questionnaire can be weighted on the basis described above. It is possible – depending on the observation period – to carry out regional (by federal state), industry-specific and size-differentiating evaluations. Researchers should bear in mind the error tolerances and numbers of cases for differentiated descriptive longitudinal analyses, just as for cross-sectional studies (see Chapter 4.3). Table F in the appendix depicts all panel cases, and the panel cases with valid interviews in each year are shown in Table G.

Researchers should also bear the following in mind for descriptive longitudinal studies: longitudinal weighting is less precise than cross-sectional weighting. This applies particularly to weighted employee figures. We recommend rounding the absolute figures very roughly or applying measures of dispersion. Please bear in mind that the percentages calculated may be imprecise.

[29] As detailed in Chapter 2.2, there has been a new definition of West and East Germany from 2007 on. After this date, West Germany consists only of the West German federal states, not including Berlin. East Germany consists of all East German federal states and Berlin.

Due to the disproportionate sample, which is corrected by the weighting process, the results of weighted and unweighted multivariate estimates may differ. In longitudinal studies, researchers can use both weighted and unweighted multivariate processes, according to their own judgement. Most users use unweighted data. In order to control the effects of the disproportionately stratified sample, we recommend adopting the respective stratification variables (industry, establishment size and federal state as of the reference date of the previous year) in the models as independent control variables.

5.4 The Wave Code: Defining Your Own Longitudinal Sections

Researchers also have the option of creating their own panel case definitions for other time periods. They can do so with the aid of the wave code (wellxxxx; xxxx stands for the survey year). This variable denotes the answers to four central questions for identifying longitudinal cases:

1. Did the establishment have employees liable to social security as of 30 June of the previous year?
2. Is an up-to-date interview available which is suitable for evaluation?
3. Did the establishment take part in the survey the previous year?
4. Was the same establishment unit surveyed as in the previous year?

Every establishment in the gross sample is clearly coded for the respective wave. This string variable contains a total of 11 values (see Table 1). [30]

[30] The following rules apply for issuing wave codes: establishments surveyed for the first time in the respective wave and from which a suitable interview is available are given the code A. In the first wave in 1993, this applies to all cases from the basic sample. From the second wave, code A applies to all establishments from the new entries and extension samples. From the 2002 wave, it is possible that establishments again count as first-time respondents and are allocated code A, if they are drawn from the new entries and extension sample following a period of exclusion. However, in such a case they receive a new identification number. Code A is allocated both to cases from the basic sample West in 1993 and East from 1996, and to all “new” establishment numbers from the new entries samples in the respective waves and the cases from the extension samples. In the first wave, only codes A and H are permitted. From the second wave on, codes B, C, D, H and Y are permitted, provided code A to G is included in the previous year. Codes E and G contain only cases allocated code H in the previous year. No further attempts are made to survey establishments that fail to give an interview in two subsequent waves. These establishments receive code H in the first non-response year and code X in the subsequent year. Establishments definitively out of operation are given code Y, if they went out of operation during the current wave. In the subsequent waves, these cases are given code Z. Establishments with codes W, X, Y and Z are not surveyed again, and are given the same code in the following wave. Only establishments coded Y are contained in the following wave under code Z.

Table 1: Wave Codes

		Code	
		With	Without
		employees liable to social security as of 30 June of previous year	
1.	Cases with interview in curr. wave		
1.1.	First-time respondents	A	Not permitted
1.2.	Continuer respondents		
1.2.1.	with interview in prev. year:		
	same unit as prev. year	B	C
	different unit as prev. year	D	Not permitted
1.2.2.	no interview in year -2		
	same unit as year -2	E	Not permitted
	different unit as year -2	G	Not permitted
2.	Cases with no interview in curr. wave		
2.1.	No response, survey in future	H	
2.2.	Extension cases, no survey in future	W	
2.3.	No response, no survey in future	X	
2.4.	Establishment out of operation:		
2.4.1.	in current wave	Y	
2.4.2.	earlier	Z	

Researchers who wish to define their own longitudinal section[31] on the basis of the logic described in Chapter 5.1 should use the cross-sectional cases of the starting year as their basis. These can be selected via the variable querxxxx. In the subsequent years, only establishments with the wave codes B or C (continuer establishment, same unit as previous year surveyed), Y (no longer operative establishment from the current wave) or Z (no longer operative establishment from the previous wave) should be defined as panel cases. The “new” establishments, identified by the variable neuxxxxr[32] (xxxx stands for the survey year), should be added. When defining their own longitudinal panels, researchers should ensure that the same establishment

[31] The wave code also offers an alternative option for identifying cross-sectional cases. These consist of all establishments with the codes A, B, D, E and G in the respective wave.

[32] This variable can be added to the dataset if required. “New” establishments are coded “N”, however only if no deviating unit was surveyed (see Chapter 5.1). The individual waves also contain the variable neuxxxx, which, however, only denotes a new establishment number. Waves 1996 and 1997 contain establishments that were only surveyed once and cannot therefore be included in longitudinal analyses. This information is contained in the variable zusatz (codes W4 and W5), which can also be added if necessary.

unit was surveyed at all times. For example: a longitudinal panel based on waves 1 to 3 should not contain cases with wave codes D or G (different unit surveyed) in wave 2 or 3. We would like to point out that no weighting factors are available for self-defined longitudinal sections. It is therefore impossible to carry out descriptive analyses that are representative of the population on this basis.

6. Data Access

As the surveyed establishments have been assured that their data will only be published in anonymous form and not passed on to third parties, external researchers only have access to the data via the Research Data Centre (FDZ) of the Federal Employment Agency (BA) at the Institute for Employment Research (IAB). Researchers have the choice between a research visit to the FDZ or controlled remote data access.

To apply for remote data access for a project, researchers need only fill out a brief application form and fax this to the FDZ. The FDZ decides whether to grant data access. As a rule, permission is sent by email within a few days. The researchers should then create evaluation programs in SPSS, Stata or SAS on the basis of test data, and send these to the FDZ by email. The FDZ carries out the analyses with the original data and returns the results after checking for data protection.

For a research visit at the FDZ researchers should make a detailed application for the research project in advance. This is evaluated by the FDZ and the Federal Ministry for Labour and Social Affairs (BMAS). Once the BMAS has granted permission, a contract is signed and an appointment is subsequently made with the FDZ. The visiting researchers work in the FDZ directly with the data from the IAB Establishment Panel at special workplaces. The analysis results are checked for data protection, to ensure that only absolutely anonymous findings leave the building.

Researchers who have previously taken part in a research visit to the FDZ are also free to carry out further analyses for the same project via remote data access. No further application is necessary in such cases.

Researchers have access to all survey waves of the IAB Establishment Panel via the FDZ. The centre provides a separate dataset for each wave containing all data from the questionnaire, the weighting factors, wave codes and other technical variables. In addition to these datasets, there is also an “organisation file” for every year (from 1998). This file contains the identification numbers of all previously surveyed establishments, and various additional information from the BA statistics. Users can obtain some of these characteristics additionally, provided they explain the need for the data in detail. The FDZ provides the following characteristics from the organi-

sation file, on application: district (Kreis) code, local authority (Gemeinde) size, industry classification WZ1993 (for the years 2000-2003) and industry classification WZ2003 (from 2004). These characteristics are already contained in the test data.[33]

The FDZ offers a number of aids for preparing analyses with the IAB Establishment Panel online, e.g. dataset descriptions, test data and questionnaires. The data access routes are also explained on the FDZ homepage, and all application forms are available for download. (http://fdz.iab.de)

For further information on the current survey wave, content-based results and a wide range of literature, please see the IAB Establishment Panel homepage (http://betriebspanel.iab.de).

[33] Additionally, users can obtain the variables zusatz and neuxxxxr from the organisation file without providing an explanation. These characteristics are, however, not contained in the test data.

References

Alda, H./ Bender, S./ Gartner, H. (2005): “The Linked Employer-Employee Dataset created from the IAB Establishment Panel and the Process-Produced Data of the IAB (LIAB)”. In: Schmollers Jahrbuch 125, pp. 327-336.

Alda, H./ Dundler, A.,/ Müller, D./ Spengler, A. (2006): "Aufbereitung eines Paneldatensatzes aus den Querschnittsdaten des IAB-Betriebspanels“. FDZ Datenreport No. 2/2006.

Bender, S./ Hilzendegen, J./ Rohwer, G./ Rudolph, H. (1996): "Die IAB-Beschäftigtenstichprobe 1975-1990“. In: Beiträge zur Arbeitsmarkt- und Berufsforschung 197, Nuremberg.

Bellmann, L. (2002): "Das IAB-Betriebspanel. Konzeption und Anwendungsbereiche." In: Allgemeines Statistisches Archiv 86 (2), pp. 177-188.

Brixy, U./ Fritsch, M. (2002): "Die Betriebsdatei der Beschäftigtenstatistik der Bundesanstalt für Arbeit" In: Fritsch, M./ Grotz, R. (eds.), "Das Gründungsgeschehen in Deutschland. Darstellung und Vergleich der Datenquellen", Heidelberg: Physica Publishers, pp. 55-78.

Bundesagentur für Arbeit (2005): "Statistik der sozialversicherungspflichtigen Beschäftigung und geringfügig entlohnten Beschäftigung"

http://www.pub.arbeitsamt.de/hst/services/statistik/000200/html/qualitaetsberichte/qualitaetsbericht-bst.pdf (as available on 07/08/2007).

Bundesanstalt für Arbeit (1973) (ed.): "Verzeichnis der Wirtschaftszweige für die Statistik der Bundesanstalt für Arbeit", Nuremberg.

Buttler, G./ Fickel, N. (2002): "Statistik mit Stichproben", Reinbek/ Hamburg: Rowohlt Publishers.

Fischer, G./ Wahse, J./ Dahms, V./ Frei, M./ Riedmann, A./ Janik, F. (2007): "Standortbedingungen und Beschäftigung in den Regionen West- und Ostdeutschlands Ergebnisse des IAB-Betriebspanels 2006". IAB Forschungsbericht No. 5/2007.

Gerlach, K./ Hübler, O./ Meyer, W. (1998) (eds.): "Ökonomische Analysen betrieblicher Strukturen und Entwicklungen". The Hannover Firm Panel, Frankfurt/New York: Campus Publishers.

Green, W. H. (2003): "Econometric Analysis". Prentice Hall.

Kölling, Arnd (2000): "The IAB-Establishment Panel." In: Schmollers Jahrbuch 120, pp. 291-300.

Städele, D./ Müller, D. (2006): "Codebuch zum IAB-Betriebspanel, Version 1: Längsschnitt 1993 – 2005" (6th revision). FDZ Datenreport No. 6/2006.

Statistisches Bundesamt (2003) (ed.): "Klassifikation der Wirtschaftszweige", Wiesbaden.

Appendix

Calculating the Longitudinal Weighting Factors

The process described below is used for every individual longitudinal section. Modifications result from the grade of differentiation of the individual parameters, depending on the available numbers of cases and the total period to be depicted.

The first step towards calculating the longitudinal weighting factors is to adapt the latest available cross-sectional cases to the structure of the population (number of establishments and employees liable to social security as of 30 June of the previous year). The entry factor is the previous year's panel weighting factor, or – for new establishment numbers – the cross-sectional weighting factor of the current year. This step corresponds to the cross-sectional weighting process described in Chapter 4.2, but on the basis of a smaller number of cases, as only those cases are included that belong to the respective longitudinal and cross-section. The stratification matrix used in the respective cross-sectional weighting process is the basis for this step.

The next stage is to adapt the basic data on the stock establishments for each individual year included in the longitudinal panel and for the "new" establishment numbers and those no longer in operation between the years. Each year's stock data includes the establishments with at least one employee liable to social security as of the reference date. The "new" establishment numbers consist of those from the stock data that had no employees liable to social security as of the reference date one year previously. The "no longer operational" establishment numbers are those in the stock that no longer have any employees liable to social security as of the reference date one year later[34]. This step depicts the "recovery" or "dying out" of establishment numbers across time.

The third step is to adapt the establishments "surviving" and "no longer operational" contained in the latest available data from the starting waves of the respective longitudinal section up to the current wave and to check the surviving and no longer operational newly founded establishments for every year of the longitudinal panel.

In the fourth and fifth steps, the establishment dynamic (growth and shrinkage of establishments) is taken into account. Changes in establishment size class, however, are only taken into account between the starting wave of the respective longitudinal panel (or the first survey of new establishment numbers) and the respective end year of the panel period. Any changes

[34] Please bear in mind that establishments no longer operational according to their wave code have actually been closed, as far as this can be established by attempting to make contact for interviews. However, "no longer operational" establishment numbers also include establishments that have no employees liable to social security as of the reserve date or for which the establishment numbers are no longer used for other reasons (see Chapter 2.1).

between these dates are not checked in the panel weighting process. Due to extremely low numbers of cases in certain combinations, the theoretically possible combinations are also compounded as follows: establishment is in the same establishment size class as at the beginning as of the latest available dataset, or has grown, or has shrunk. Establishments' developments can therefore only be roughly depicted using this process. In the fourth step, the establishment dynamic is thus taken into account for the stock of the starting wave of the respective longitudinal section. In the fifth step, this is done for the "new" establishment numbers added during the longitudinal period.

Any disproportionate non-responses are then corrected, on the basis of answers to individual questions from the previous year's survey. Multivariate estimates are used to calculate selection probabilities on the basis of the questionnaire responses from the previous wave. These multivariate non-response analyses are carried out for every individual longitudinal period. The analyses – with the exception of the longitudinal panel from 1993 and the longitudinal panel for the manufacturing industry in East Germany – are carried out separately for East and West.

Steps seven and eight are to adapt the cases from the respective cross-section to the requirements of the stratification matrix for every survey year included in the longitudinal section. To limit the number of weighting cells, only the establishment size classes (seventh step) and a simplified industry structure with seven categories (step eight) are taken into account.

The eight weighting steps are repeated on an iterative basis.

In certain cases, using the above parameters for the panel weighting process can lead to extremely high weighting factors. To limit the resulting problems of outliers, only weighting factors between one[35] and a maximum of 4,000 are permitted for panel weighting. As a consequence, the theoretically necessary target numbers for individual characteristic values may not be reached due to the weighting process.

[35] Up to 2001, longitudinal weighting factors of less than one were permitted (on the causes, see Chapter 4.2).

Table A: Sampling Probability by Establishment Size

Number of employees liable to social security	Average sampling probability
1 to 4	0.0011
5 to 9	0.0015
10 to 19	0.0030
20 to 49	0.0089
50 to 99	0.0153
100 to 199	0.0304
200 to 499	0.0862
500 to 999	0.1504
1,000 to 4,999	0.8765
5,000 and over	0.9127
Total	0.0043

Reference: IAB Establishment Panel 1993

Table B: Classification of Economic Activities by 16 Industries for Sampling and Projection of Wave 1993 to 1999

Code	Economic activities	Code acc. to question-naire 1993-1998	Code acc. to question-naire 1999
1	Agriculture, hunting and forestry	1	010
2	Energy supply , mining and quarry-ing	2	020
3	Raw material processing	3-6	030-060
4	Manufacture of capital goods	7-12	071, 072, 073, 080, 091, 092, 100, 110, 120
5	Manufacture of consumer goods	13-16	130-160
6	Construction	17-18	170-180
7	Wholesale and retail trade	19	190
8	Transport, communication	20	200
9	Financial intermediation	21	210
10	Insurance industry	22	220
11	Restaurants, homes	23-25	230-250
12	Education, publishing	26-27	260-270
13	Human health activities	28	280
14	Liberal professions	29-34	290-340
15	Other services	35	350
16	Non-industrial organizations/ public administration	36-41	360-410

For wave 1999 the 41 classified branches were broken down on a more differentiated 3-digit level which was necessary to identify establishments from the producing industries for an extension sample on be-half of the Institute for Economic Research Halle (IWH). Questionnaire codes 7 and 9 were affected most by this differentiation.

Table C: Classification of Economic Activities by 20 Industries for Sampling and Projection of Wave 2000 to 2003

Code	Industries	WZ93 code	Code acc. to questionnaire 2000-2003
1	Agriculture, hunting and forestry, fishing	01, 02, 05	01
2	Mining and quarrying, electricity, gas and water supply	10-14, 40, 41	02
3	Manufacture of food products, beverages and tobacco	15-16	03
4	Manufacture of consumer products	17-19, 21-22, 36	04-05, 18
5	Manufacture of industrial goods	20, 23-27, 37	06-11
6	Manufacture of capital and consumer goods	28-35	12-17
7	Construction	45	19, 20
8	Trade, maintenance and repair of motor vehicles, motor cycles, personal and household goods	50-52	21-23
9	Transport, storage and communication	60-64	24, 25
10	Financial intermediation	65-67	26, 27
11	Hotels and restaurants	55	33
12	Education	80	34
13	Health and social work	85	35
14	Computer and related activities	72	28
15	Research and development	73	29
16	Legal, accounting, book-keeping and auditing activities; tax consultancy; market research and public opinion polling; business and management consultancy; holdings; advertising	741, 744	30
17	Real estate activities	70	31
18	Renting of machinery and equipment without operator and of personal and household goods; other business activities	71, 742-743, 745-748	32
19	Other community, social and personal service activities	90, 92-93	36-38
20	Public administration and defence; compulsory social security; private households with employed persons; activities of membership organizations n.e.c.; extra-territorial organizations and bodies	75, 91, 95, 99	39-41

Table D: Classification of Economic Activities by 17 Industries for Sampling and Projection as of Wave 2004

Code	Industries	WZ 2003 code	Code acc. to questionnaire as of 2004
1	Agriculture, hunting and forestry, fishing	01, 02, 05	01
2	Mining and quarrying, electricity, gas and water supply	10-14, 40, 41	02
3	Manufacture of food products beverages and tobacco	15-16	03
4	Manufacture of consumer products (not including manufacture of wood products)	17-19, 21-22, 36	04-05, 18
5	Manufacture of industrial goods (including manufacture of wood products)	20, 23-27, 37	06-11
6	Manufacture of capital and consumer goods	28-35	12-17
7	Construction	45	19, 20
8	Trade, maintenance and repair of motor vehicles, motor cycles, personal and household goods	50-52	21-23
9	Transport, storage and communication	60-64	24, 25
10	Financial intermediation	65-67	26, 27
11	Hotels and restaurants	55	33
12	Education	80	34
13	Health and social work	85	35
14	Industrial services	70-74	28-32
15	Other services	90, 92-93	36-38
16	Non-industrial organizations	91	39
17	Public administration	75	41

Table E: Regional Extension Samples 1993-2007

	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Schleswig-Holstein										X	X	X	X	X	X
Hamburg								X	X						
Lower Saxony								X	X	X	X	X	X	X	X
Bremen								X	X	X	X	X	X	X	X
North Rhine-Westphalia								X	X	X	X	X	X	X	X
Hesse									X	X	X	X	X	X	X
Rhineland-Palatinate								X	X	X	X	X	X	X	X
Baden-Württemberg								X	X	X	X	X	X	X	X
Bavaria								X	X	X	X	X	X	X	X
Saarland									X	X	X	X	X	X	X
Berlin				X	X	X	X	X	X	X	X	X	X	X	X
Brandenburg				X	X	X	X	X	X	X	X	X	X	X	X
Mecklenburg-Western Pomerania				X	X	X	X	X	X	X	X	X	X	X	X
Saxony				X	X	X	X	X	X	X	X	X	X	X	X
Saxony-Anhalt								X	X	X	X	X	X	X	X
Thuringia				X	X	X	X	X	X	X	X	X	X	X	X
Manufacturing industries East Germany							X	X	X	X	X	X	X	X	X

Table F: Panel cases

	_94	_95	_96	_97	_98	_99	_00	_01	_02	_03	_04	_05	_06
pan93…	3,739	3,614	8,004	3,252	3,221	2,872	2,957	3,243	3,053	3,230	3,501	3,647	3,871
pan96…				8,048	7,537	6,718	6,581	6,842	6,441	6,705	7,086	7,252	7,562
pan00…								12,135	10,528	10,006	9,987	9,773	9,772
pan03…											14,179	13,154	12,448

Table G: Panel cases with valid interviews per year

	_94	_95	_96	_97	_98	_99	_00	_01	_02	_03	_04	_05	_06
pan93…	3,489	2,916	2,381	1,989	1,701	1,412	1,222	1,053	892	762	656	580	509
pan96…				6,658	5,483	4,565	3,943	3,427	2,952	2,576	2,257	1,987	1,740
pan00…								10,840	8,746	7,281	6,347	6,636	4,841
pan03…											12,769	10,761	9,108

Level of confidence table

p = (%) / n =	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
100	(–)	(–)	(–)	(–)	12.2	13.0	13.5	13.9	14.1	14.1	14.1	13.9	13.5	13.0	12.2	11.3	10.1	8.5	6.2
200	(–)	(–)	7.1	8.0	8.7	9.2	9.5	9.8	9.9	10.0	9.9	9.8	9.5	9.2	8.7	8.0	7.1	6.0	4.4
300	(–)	4.9	5.8	6.5	7.1	7.5	7.8	8.0	8.1	8.2	8.1	8.0	7.8	7.5	7.1	6.5	5.8	4.9	3.6
400	(–)	4.2	5.0	5.7	6.1	6.5	6.7	6.9	7.0	7.1	7.0	6.9	6.7	6.5	6.1	5.7	5.0	4.2	3.1
500	(–)	3.8	4.5	5.1	5.5	5.8	6.0	6.2	6.3	6.3	6.3	6.2	6.0	5.8	5.5	5.1	4.5	3.8	2.8
600	(–)	3.5	4.1	4.6	5.0	5.3	5.5	5.7	5.7	5.8	5.7	5.7	5.5	5.3	5.0	4.6	4.1	3.5	2.5
700	2.3	3.2	3.8	4.3	4.6	4.9	5.1	5.2	5.3	5.3	5.3	5.2	5.1	4.9	4.6	4.3	3.8	3.2	2.3
800	2.2	3.0	3.6	4.0	4.3	4.6	4.8	4.9	5.0	5.0	5.0	4.9	4.8	4.6	4.3	4.0	3.6	3.0	2.2
900	2.1	2.8	3.4	3.8	4.1	4.3	4.5	4.6	4.7	4.7	4.7	4.6	4.5	4.3	4.1	3.8	3.4	2.8	2.1
1.000	1.9	2.7	3.2	3.6	3.9	4.1	4.3	4.4	4.4	4.5	4.4	4.4	4.3	4.1	3.9	3.6	3.2	2.7	1.9
1.100	1.9	2.6	3.0	3.4	3.7	3.9	4.1	4.2	4.2	4.3	4.2	4.2	4.1	3.9	3.7	3.4	3.0	2.6	1.9
1.200	1.8	2.4	2.9	3.3	3.5	3.7	3.9	4.0	4.1	4.1	4.1	4.0	3.9	3.7	3.5	3.3	2.9	2.4	1.8
1.300	1.7	2.4	2.8	3.1	3.4	3.6	3.7	3.8	3.9	3.9	3.9	3.8	3.7	3.6	3.4	3.1	2.8	2.4	1.7
1.400	1.6	2.3	2.7	3.0	3.3	3.5	3.6	3.7	3.8	3.8	3.8	3.7	3.6	3.5	3.3	3.0	2.7	2.3	1.6
1.500	1.6	2.2	2.6	2.9	3.2	3.3	3.5	3.6	3.6	3.7	3.6	3.6	3.5	3.3	3.2	2.9	2.6	2.2	1.6
1.600	1.5	2.1	2.5	2.8	3.1	3.2	3.4	3.5	3.5	3.5	3.5	3.5	3.4	3.2	3.1	2.8	2.5	2.1	1.5
1.700	1.5	2.1	2.4	2.7	3.0	3.1	3.3	3.4	3.4	3.4	3.4	3.4	3.3	3.1	3.0	2.7	2.4	2.1	1.5
1.800	1.5	2.0	2.4	2.7	2.9	3.1	3.2	3.3	3.3	3.3	3.3	3.3	3.2	3.1	2.9	2.7	2.4	2.0	1.5
1.900	1.4	1.9	2.3	2.6	2.8	3.0	3.1	3.2	3.2	3.2	3.2	3.2	3.1	3.0	2.8	2.6	2.3	1.9	1.4
2.000	1.4	1.9	2.3	2.5	2.7	**2.9**	3.0	3.1	3.1	3.2	3.1	3.1	3.0	2.9	2.7	2.5	2.3	1.9	1.4
2.500	1.2	1.7	2.0	2.3	2.4	2.6	2.7	2.8	2.8	2.8	2.8	2.8	2.7	2.6	2.4	2.3	2.0	1.7	1.2
3.000	1.1	1.5	1.8	2.1	2.2	2.4	2.5	2.5	2.6	2.6	2.6	2.5	2.5	2.4	2.2	2.1	1.8	1.5	1.1
4.000	1.0	1.3	1.6	1.8	1.9	2.0	2.1	2.2	2.2	2.2	2.2	2.2	2.1	2.0	1.9	1.8	1.6	1.3	1.0
6.000	0.8	1.1	1.3	1.5	1.6	1.7	1.7	1.8	1.8	1.8	1.8	1.8	1.7	1.7	1.6	1.5	1.3	1.1	0.8
8.000	0.7	0.9	1.1	1.3	1.4	1.4	1.5	1.5	1.6	1.6	1.6	1.5	1.5	1.4	1.4	1.3	1.1	0.9	0.7
10.000	0.6	0.8	1.0	1.1	1.2	1.3	1.3	1.4	1.4	1.4	1.4	1.4	1.3	1.3	1.2	1.1	1.0	0.8	0.6
15.000	0.5	0.7	0.8	0.9	1.0	1.1	1.1	1.1	1.1	1.2	1.1	1.1	1.1	1.1	1.0	0.9	0.8	0.7	0.5
20.000	0.4	0.6	0.7	0.8	0.9	0.9	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.9	0.9	0.8	0.7	0.6	0.4
25.000	0.4	0.5	0.6	0.7	0.8	0.8	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.8	0.8	0.7	0.6	0.5	0.4

Example: A sample size where n = 2,000 includes 30 % building society depositors. From the table it can be concluded that the confidence interval is ± 2.9 % (highlighted value). With a probability of 95.5 %, the real percentage value of building society depositors in the population is ± 2.9 % of 30 % i.e. lying somewhere between 27.1 % and 32.9 %.

The low percentage value in small samples, where no level of confidence is given, can only be interpreted in a limited way, as this confidence level accounts for more than half the percentage values (e.g ± 8.5 % where n = 100 and p = 10 %).

$p - t\sigma \leqq p \leqq p + t\sigma$

$\sigma = \sqrt{2}\ \sqrt{\frac{p\,(100-p)}{n}}$

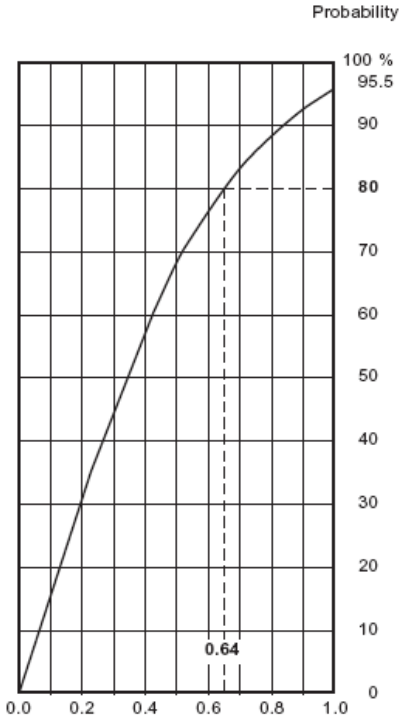
t = 2 ⇒ 95.5 % significance level
P = variable percentage in the population

$\sqrt{2}$ = design factor

p = variable percentage in the sample (in %)
n = sample size
Please note: any analysis of subgroups must, as a rule, be expressed as a percentage of the whole sample size!

tns infratest

Curve of estimation of the confidence level at a reduced level of significance.



Reduction factors according to the confidence interval as shown in the table.

Example:
In the graph a significance level of 80 % corresponds to a reduction factor of 0.64 (see the dotted line). The example shown here of building society depositors reads as follows: with a probability of 80 % the real percentage value is 30 % ± 1.9 % (2.9 % x 0.64 = 1.9 %), i.e. lying between 28.1 % and 31.9 %.

Imprint

FDZ Methodenreport
No. 01/2008 (EN)

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

Editorial staff
Stefan Bender, Dagmar Herrlinger

Technical production
Dagmar Herrlinger

Download
http://doku.iab.de/fdz/reporte/2008/MR_01-08_EN.pdf

Internet
http://fdz.iab.de/

Corresponding author
Dana Müller, Institute for Employment Research,
Regensburger Str. 104, D-90478 Nuremberg
Phone: +49-(0)911/179-2409
E-Mail: dana.mueller@iab.de