VATT-KESKUSTELUALOITTEITA
VATT-DISCUSSION PAPERS

226

Integrated Panel of Finnish Companies and Workers

Ossi Korkeamäki
Tomi Kyyrä

Valtion taloudellinen tutkimuskeskus
Government Institute for Economic Research
Helsinki 2000

ISBN 951-561-324-8

ISSN 0788-5016

Valtion taloudellinen tutkimuskeskus

Government Institute for Economic Research

Hämeentie 3, 00530 Helsinki, Finland

Email: ossi.korkeamaki@vatt.fi, tomi.kyyra@vatt.fi

Yliopistopaino

Helsinki, July 2000

KORKEAMÄKI OSSI AND KYYRÄ TOMI: INTEGRATED PANEL OF FINNISH COMPANIES AND WORKERS. Helsinki, VATT, Valtion taloudellinen tutkimuskeskus, Government Institute for Economic Research, 2000, (C, ISSN 0788-5016, No 226. ISBN 951-561-324-8.

Abstract: In this paper we describe the compiling of the integrated data which comprises longitudinal information on Finnish companies and workers. The company panel of the data is dynamically representative over time and covers a wide range of the economy. The worker panel covers in principle all people who have been employed in the private sector during the period 1988-96. Worker and company records can be easily matched to form linked data sets of companies and their employees. We describe the data sources used and discuss some problems which might occur in using the data as well as the potential of the data in future labour market research. To provide evidence on the quality of the data, we also report some descriptive figures with emphasis on job and worker flow statistics.

Key words: Panel data, matched data, job and worker flows

Tiivistelmä: Tämä paperi sisältää kuvauksen yhdistetystä paneeliaineistosta, joka sisältää seurantatiedot sekä yrityksistä että työntekijöistä. Yrityspaneeli sisältää yrityksiä useilta eri toimialoilta ja se on muodostettu siten, että se pysyy edustavana yli ajan. Työntekijäpaneeli kattaa periaatteessa kaikki yksityisellä sektorilla periodilla 1988-96 työskennelleet henkilöt. Työntekijä- ja yritystietoja voidaan helposti yhdistää toinen toisiinsa muodostettaessa erilaisia linkitettyjä aineistoja. Kuvaamme käytetyt aineistolähteet, jonka lisäksi keskustelemme jonkin verran aineiston tutkimuskäyttöön liittyvistä mahdollisista ongelmista sekä aineiston tarjoamista tutkimusmahdollisuuksista. Aineiston laadun ja luotettavuuden testaamiseksi esitämme myös joitakin aineistosta laskettuja deskriptiivisiä lukuja, jotka painottuvat työpaikka- ja työntekijävirtoihin.

Asiasanat: Paneeliaineisto, yhdistetty aineisto, työpaikka- ja työntekijävirrat

Contents

1. Introduction **1**

2. Company panel **3**

3. Worker panel **7**

4. On the linking of company and worker records **9**

5. Supplementary data **11**

5.1 Supplementary company panel 11
5.2 Supplementary worker panel 13

6. Some tests of the quality of the IP data **15**

6.1 Disentangling the company identifiers – enterprise demography and the reliability of job and worker flow measures 15
6.2 Comparing the employment measures in the worker panel and the company panel 25
6.3 Wages in the worker panel and the company panel 27

7. Conclusion **31**

References **32**

Appendix: Information sources **33**

Appendix: Definitions of flow measures **36**

Appendix: List of variables for the company panel **38**

Appendix: List of variables for the worker panel **41**

1. Introduction

Several authors in applied labour economics have pointed out the importance of data sources which combine information from both worker records and company records. One decade ago this kind of data was not available anywhere. In recent years, however, researchers have put enormous effort in finding and using matched company-worker data sets.[1] Because of the high quality and centralised nature of the Finnish statistical system, the extensive administrative registers existing in Finland provide great potential for creating these kind of data sets. Recognising this the Government Institute for Economic Research and Research Institute of the Finnish Economy, in co-ordination with Statistics Finland, begun a few years ago to find out possibilities for collecting the data which would contain longitudinal information on Finnish companies and workers. In this paper we briefly describe the result of this time-consuming and arduous project, which is also part of The Economic Crisis of the 1990s in Finland Research Programme of the Academy of Finland.

In particular we describe the compiling of the integrated data which consists of a panel of Finnish companies and a large panel of private-sector workers. This data was constructed by combining the existing databases maintained by Statistics Finland, and it is nicknamed the *Integrated Panel of Finnish Companies and Workers* and referred to shortly as the IP data. The company panel of the IP data consists of a large sample of companies from wide sectors of the economy for the period 1989-1995. It was constructed by combining the annual surveys of the Financial Statements Statistics, and it is dynamically representative over time due to the annual rotation of the underlying sample. The worker panel, based on the Employment Statistics, covers in principle all workers who have been employed with a private sector company during the period 1988-1996. The worker and company records can be easily matched to form linked data sets of companies and their employees.

Together the company panel and worker panel of the IP data cover a considerable portion of the economic activity in Finland. This kind of data source obviously enables us to study several issues that are beyond the scope of analysis based on data sets containing information only on either companies or workers. Moreover, the time period covered by the IP data is an exceptional one in Finnish economic history as it covers the whole business cycle from the boom of the late 1980s trough the deep recession of the early 1990s to the years of recovery.

Below we discuss the different administrative data sources used as well as some problems that are likely to occur when linking the worker and company records. Since the practices of Statistics Finland impose some problematic restrictions on

[1] This development is described in detail in Abowd and Kramarz (1999).

the use of the IP data, we also describe the supplementary version of the data that had to be constructed in order to bring some flexibility in the empirical analysis of the data. To assess the quality of the IP data, we further represent some descriptive figures with emphasis on job and worker flow statistics. We first construct ‘synthetic’ companies from the worker panel by identifying workers associated with the same company and correcting for changes in the company identifiers which are due to some administrative reasons. We then calculate several statistics separately from both the worker panel and company panel to be compared with each others. We find both sources of data to produce consistent results. In addition there turns out to be a need to control for different changes in company identifiers when counting labour market flows or detecting changes in the company population.

The rest of the paper is organised as follows. The next two sections describe the company panel and worker panel of the IP data respectively. In Section 4 we briefly discuss issues related to the linking of these two panels, whereas the supplementary data is described in Section 5. Our brief descriptive analysis is presented in Section 6, and the final section concludes the paper.

2. Company panel

The principal data source on companies is the Financial Statements Statistics (FSS). It is a survey that Statistics Finland compiles annually on the basis of corporate income statement and balance sheet data. By combining the annual surveys of the FSS, we have constructed a panel of over 11,700 companies for the period 1989-1995, covering annually 4,000 to 6,000 companies from different sectors of the economy. In 1989 the sampling method, industry classification and information content of the FSS changed so that the data for previous periods are not consistent with the later periods. Moreover, the sampling method changed again in 1995 in the manufacturing and construction sectors and one year later in other sectors. We have therefore used the FSS data only for the period 1989-1995 to construct our company panel.

Over our observation period, the annual surveys of the FSS cover the population of all companies above a certain size threshold and a stratified sample from smaller companies.[2] In the sampling of small and medium-size companies, the company's industry and size of personnel were used as stratification variables. A total of 64 industry groups and 6 size groups were used. The main industry groups are manufacturing, construction, trade, business services and transport.[3] The threshold size for large companies was allowed to vary across the main industry groups. The threshold size for large companies is 100 employees in manufacturing and trade, and 50 employees in construction. In trucking, which is a sub-sector of transport, the threshold size is 50 employees, whilst there are so few companies in the other branches of transport that they were all included in the annual survey. In business services, which is dominated by small companies, the threshold size is lower than in other industries, being 20 employees. The sample was rotated annually by replacing a fraction of the oldest companies in each stratum with new ones. The rotation sampling was mainly used to keep the survey representative in each point of time.

Our panel is organised into five main industry groups, across which the information content slightly varies.[4] There are over 200 variables available in each industry group, though some of them are available only for large companies.[5] Some basic information on the company data is given in Table 2.1. The data for period t are from the accounting periods ending between the 1st of April in year t and the last day of March in year $t + 1$. The accounting period of most companies is the calendar year. In cases where the length of the accounting

[2] During 1994 and 1995 Statistics Finland gradually gave up the stratification sampling. Nowadays it surveys the whole population of companies above a certain size threshold and collects information on smaller companies from the administrative registers of tax authorities.

[3] It should be stressed that transport contains also communications as a sub-sector.

[4] Transport is such a heterogeneous group that information content further varies across sub-sectors.

[5] A variable list containing the most important variables is given in the appendix.

period deviated significantly from 12 months, the data were adjusted to correspond to a normal-length accounting period. Some companies have also changed their accounting period during the observation period.

Table 2.1 Some basic information by industry group

	Number of companies	Observation period	Size threshold (employees)
Manufacturing	2,982	1989-1994	100
Construction	1,491	1989-1994	50
Trade	3,934	1989-1995	100
Business services	2,124	1990-1995	20
Transport[6]	1,187	1989-1995	-

Annual data were converted to correspond to the information content of the last period (1994 or 1995 depending on the industry). At the beginning of 1993 the law on bookkeeping changed which in turn affected the information content of the FSS. It was not possible to convert all variables to be consistent over time, so the meaning of certain variables may vary over time. Moreover, there are some variables that are available for certain cross sections only.

Some pieces of essential information were missing from the annual surveys of the FSS. We have therefore collected some additional variables from the Register of Enterprises and Establishments. These variables contain information on the ending dates and length of the accounting periods, foreign trade, the share of foreign ownership and the country from which the largest foreign investment has been made.

Along with the obvious advantages of the FSS data, there are also some limitations that are worth noting. The complex sampling design results in a large variation in the sampling probabilities (across companies and over time) and, as a consequence, in the corresponding sampling weights.[7] This may in turn magnify response bias and problems resulting from the changes in the sample composition

[6] Trucking differs from other sub-sectors of transport in having a size threshold of 50 employees and records being available only from 1992 onwards.

[7] For each company the fraction of companies sampled from its stratum is known. These fractions are *not* corrected for response bias however, so they may give a somewhat biased guidance for weighting the observations. Unfortunately, the response bias is likely to be associated with company characteristics (e.g. large companies are more likely to return questionnaires appropriately filled than the smaller ones). On the other hand, the FSS survey is statutory and mandatory, and Statistics Finland demands for answers continually from non-respondents and from those companies which did not fill the questionnaire appropriately. Consequently, the fraction of the company records lost because of non-response and rejected questionnaires has been relatively low, being around 15 per cent annually.

available for empirical analysis. From the viewpoint of dynamic panel data analysis, the major drawback of the data is that there are a number of companies for which we have observations only for very few (consecutive) cross sections. Moreover, the companies with too few observations for purposes of estimating the dynamic panel data models which require several consecutive observations on each company are more likely to be the small ones due to the sampling method underlying the FSS. As a result, the composition of the estimation sample may differ significantly from that of the data as a whole.

Table 2.2 Number of companies by entry year and length of observation period

Entry	Number of cross-sectional observations							
year	1	2	3	4	5	6	7	Σ
1989	683	689	616	530	507	725	455	4,205
1990	290	294	220	241	460	306		1,766
1991	261	183	213	492	334			1,438
1992	181	171	446	452				1,250
1993	176	412	435					1,023
1994	587	451						1,038
1995	953							953
Σ	3,131	2,155	1,930	1,715	1,301	1,031	455	11,718

According to the table above, there are 3,131 companies (27% of all companies) in the data for which we have only one cross section available. For example, if at least four cross-sectional observations on each company are required by the estimating technique, we have only 4,502 companies (38% of all companies) potentially available for the estimations. Since some of these companies will be further lost because of additional data restrictions, the estimation sample will cover only a relatively small share of companies in the data. Although the data as a whole should be representative in the cross-sectional dimension over time, the samples available for dynamic panel data estimations are certainly not. Unfortunately, there is perhaps very little one can do to solve this problem. Where several consecutive cross-sectional observations are required, the loss of companies available for estimations is likely to lead to estimation samples that will not include observations from each stratum. This in turn rules out the re-

weighting of observations in the estimation samples as a solution to the change in composition of the data.[8]

Of course, one may restrict the attention purely on large companies which are all surveyed every year and hence avoid problems resulting from the sample rotation and stratification sampling of small and medium-sized companies. This would be a pity in the sense that almost all of the company panels available elsewhere contain only large companies. On the other hand, the threshold sizes for large companies across industries are quite low in the FSS.

In sum, the company panel based on the FSS has several important advantages, though it is not free from drawbacks. Whilst the vast majority of existing company data sets covers only the manufacturing sector, which nowadays accounts for a relatively small and declining share of aggregate production and employment, the company data based on the FSS represents a wide range of the economy. Moreover the data remain representative in the cross-sectional dimension over time, and the information content is quite rich. Due to the sample rotation, the major drawback of the data is a relatively small number of cross-sectional observations available for most of the small and medium-sized companies. Given the large number of observations in the data, there will still be enough observations also for purposes of estimating the models that require several consecutive cross sections, though this must be done at the expense of representativeness of the data.

[8] The re-weighting of observations in the estimation sample will probably further increase the variation in the sampling weights as the companies lost because of too few cross-sectional observations are more likely to be associated with low initial sampling probabilities.

3. Worker panel

Data on workers were drawn from the annual records of Employment Statistics (ES), which is one of the major databases of Statistics Finland. The main objective of the ES database is to produce regional figures on the economic activity of the Finnish population, and it covers effectively all people with a permanent residence in Finland. Since 1987, Statistics Finland has updated the ES database annually by combining information from over 20 administrative registers, and its information content is extensive.

The ES database matches also some employer information to individuals who hold a job at the last week of the year. Amongst others, each individual in the ES database is associated to his or her employer with a company (and establishment) identifier,[9] which provides a link to the company data.[10] For the worker panel of the IP data we have selected all persons from annual records of the ES database with a company or establishment identifier of the business enterprise at least in one of the years between 1988 and 1996.[11] Business enterprises were separated from other companies using the information on the legal status of the employing company.[12] As a result, the worker panel of the IP data covers practically *all* persons who have been employed (at the end of year) with a private sector company during the period 1988-1996. For selected people a set of variables for the period 1987-97 was collected by combining the annual records of the ES database. The total number of persons in the worker panel is slightly below two million; over half of them having been employed with companies covered by our company panel.

The huge number of people in the worker panel together with the relatively long time period it covers makes the data very large, so the number of variables to be

[9] The company and establishment identifiers in the ES database are identical to those used in the Register of Enterprises and Establishments, which in turn serves as the sampling frame for the FSS survey on which our company panel is based. In this sense one may expect that the linking of worker and company records of the IP data is based on a highly reliable procedure.

[10] The additional information available on the employer is the ownership, legal status and industry of the company.

[11] To be specific, the population for the period 1988-1993 was based on the data collected by Mustaniemi (1997) who used some minor but troublesome age and earnings restrictions in composing her data. To be selected in the data, she required that a person had to be associated with the business enterprise at least in one of the years between 1988 and 1993, and in such a year the person had to be aged between 16 and 65 and exhibit a certain positive amount of earnings, which in turn depends on the starting date of his or her job spell. Since the person was selected in the data and longitudinal information was collected over the whole period if the requirements were met in *any* year, these restrictions have probably an insignificant effect but when elderly people are considered one should bear these in mind. We did not impose these restrictions when complementing the data to cover the period 1994-1996, however.

[12] Strictly speaking, pension trusts and foundations, employment pension institutes, unemployment benefit funds, housing corporations, trusts and foundations, non-profit corporations, state enterprises, public bureau and corporations, church and other religious corporations were excluded.

included in the worker data had to be kept in a manageable range. An additional factor limiting the number of variables to be included in the data was Statistics Finland's concern on privacy protection. This is because the possibility of matching individuals in the worker panel to their employers in the company panel makes the identification of statistical units in the both data sets in principle much more easier.

The worker panel contains some basic information, such as age, gender, marital status, education and home municipality. In addition there is annual information on months worked and being unemployed, labour market status at the end of year, annual incomes, dates of employment and unemployment spells, amongst other.[13] Overall, since the main purpose of the worker data is to serve as a complementary source of information for the company data, its information content is likely to be sufficient for a vast range of research purposes.

[13] See the variable list in the appendix.

4. On the linking of company and worker records

The records on companies and workers in the IP data can be matched by using the company identifiers attached to each individual in the worker panel. The only difficulty in linking these data sources is that the observation period is the accounting period in the company data, whilst the company identifiers in the worker panel are attached to workers on the basis of employment relationships in force at the last week of the year. It follows that if the company serves as the statistical unit in the analysis and one likes to obtain additional information from the worker records to be matched to each company, it may be necessary in some cases to exclude companies whose accounting periods deviate significantly from the calendar year.[14] This restriction combined with the fact that several companies in the company panel are observed only over few cross sections may worsen the sample selection problems, where the matched company-worker panel is aimed for use in the estimation of dynamic panel data models.

Overall we believe that the linking of the company and working records available in the IP data to form matched data sets provides new avenues in many branches of applied labour economics. By the use of worker records, one can for example decompose the company's labour force by worker characteristics and observe the wage distribution within the company as well as gross worker flows in and out of the company. It is thus possible to study a number of issues related to labour demand that are beyond the scope of empirical studies based on the company data only.

Since we can follow workers over the period 1988-1996 independently of their labour market state, we can identify the entry channels of the new workers hired by the companies (unemployment, another job, etc.) as well as the exit channels of senior workers (unemployment, retirement, etc.). With the IP data one can thus study the hiring strategies of employers or assess whether the low-wage companies are losing their workers to the companies that offer higher wages.

Of course, one may also match employer information from the company panel to individual workers. This in turn enables us to study several issues concerning wage determination that cannot be considered with data on workers only. With the IP data one can also model the supply and demand sides of the labour market together. It is in fact possible to use this kind of data in estimating richer specifications of the equilibrium search and dynamic matching models than have been estimated until now.

The examples above were only to mention a few. It is further worth noting that the time period covered by the IP data is an interesting one from the economist's

[14] It may be possible in some cases to re-label time periods in the company data to be more consistent with the calendar year.

point of view. In the last years of the 1980s the Finnish economy overheated and finally collapsed in the early 1990s. The recession that hit the economy was exceptionally severe: the annual change in the GNP was negative during the period 1991-1993, and in the worst year of 1991 the GNP decreased over 7 per cent. Large-scale job destruction took place in virtually every sector of the economy, leading to exceptionally high levels of unemployment, the unemployment rate being close to 20 per cent. Although the economy has grown strongly from 1994 onwards, aggregate unemployment has stuck at a high level. The observation period of the IP data thus covers the whole business cycle from the boom trough the deep recession to the years of recovery. This observation combined with the availability of information on both the supply and demand sides of the labour market in the IP data emphasises the potential of the data in future applied research concerning especially labour market issues.

5. Supplementary data

Due to the strict privacy protection regulations followed by Statistics Finland, only researchers employed with Statistics Finland may be given permission to access the IP data described in Chapters 3 and 4. Consequently, additional samples of the company and worker panel of the IP data were constructed for external use. These samples were manipulated in such a way that they fulfil the requirements for use outside Statistics Finland, and they were created in order to bring some additional flexibility in the use of the IP data.

As will become clear below, the supplementary version of the data is *not* suitable for the final analysis. Due to the manipulations required by the statistical authorities, the supplementary version is not representative with respect to the underlying populations. It includes also additional errors in the company data variables that had to be generated for privacy protection reasons. The supplementary version is generally meant to be used in testing the computer programs that the researcher have to run with the original IP data later on. It might also be used in some preliminary analysis of the data in which case one must bear in mind the underlying manipulations when interpreting the results.

5.1 Supplementary company panel

According to the law, Statistics Finland is not allowed to give for external use data sets in which statistical units can be identified. With respect to the company records, Statistics Finland is especially rigorous. This is partly because the population of companies is very heterogeneous and relatively small compared to that of individuals or households. Consequently, it is believed that individual companies can be easily identified indirectly. We were therefore forced to resort to somewhat cumbersome methods to form a sample of company data that could be used outside Statistics Finland.

First of all, information on the company's industry was supposed to assist the identification of individual companies to a great extent. Therefore original industry codes had to be aggregated to contain only five main industry groups of the FSS, i.e. manufacturing, construction, trade, business services and transport. Second, two variables collected from the Register of Enterprises and Establishments for the company data had to be excluded from the supplementary samples because they were deemed to contain delicate information. These variables refer to the share of foreign ownership and the country from which the largest foreign investment was made.

Third, the income statement and balance sheet data themselves were believed to identify individual companies. To prevent this possibility, we were required to add random noise terms to the original variables of the FSS data. However, the

population of large companies was assumed so small and heterogeneous that adding noise terms in the variables is not enough to secure non-identification in their case. All large companies had to be therefore dropped out of the sample. On the other hand, there are so many small companies that their identification indirectly using financial statement information was assumed so troublesome in any case that it was not required to manipulate their data at all. Random noise terms were thus added only to the data of medium-sized companies.

Table 5.1 Threshold size based on personnel by industry

	Small	Medium-sized	Large
Manufacturing	0-99	100-499	500+
Construction	0-49	50-499	500+
Trade	0-99	100-499	500+
Business services	0-19	20-99	100+
Transport	-	0-99	100+
Measure	none	noise	excluded

Since the average company size differs from industry to industry, the threshold size for medium-size companies was allowed to vary across industries according to Table 5.1. The size of the company is defined as the average number of employees over the accounting periods the company is observed in the data.[15] The number of excluded companies was 178 in manufacturing, 21 in construction, 59 in trade, 123 in business services and 83 in transport.

For the medium-sized companies the noise was added to variables by multiplying the original values by random terms generated from the log normal distribution. If we denote the variable *i* of company *j* in period *t* with $x_{ji}(t)$, the new variable is then given by

$$y_{ji}(t) = x_{ji}(t) \times \varepsilon_{ji},$$

where $\ln \varepsilon_{ji} \sim N(0,\sigma^2)$ and time-invariant ε_{ji} are i.i.d. across all *i* and *j*. Moreover, σ^2 was chosen so that

[15] There are few companies whose average size of personnel is below the threshold size but the number of employees in some period exceeds considerably the threshold size. These companies were included in the supplementary data, but the cross sections in which the size of personnel is well above the threshold value had to be deleted.

$$\Pr(-0.15 < \ln \varepsilon_{ji} < 0.15) = 0.95 \approx \Pr(0.86 < \varepsilon_{ji} < 1.16).^{16}$$

In words, each variable value was multiplied by a random term that changed the original value by some 15 per cent or less in absolute terms with a probability of 95 per cent, the expected change being close to zero. There are two points to be emphasised. First, different variables of a given company were multiplied by different (independent) random terms. Second, the random multipliers do not vary over time, so that each variable of a given company was multiplied by the same error term in all cross sections. Thus adding noise to the variable does not change its pattern over time.

The use of the log normal distribution in generating random multipliers has some advantages. First of all, taking logarithm of the new variable $y_{ji}(t)$ yields

$$\ln y_{ji}(t) = \ln x_{ji}(t) + e_{ji},$$

where the normal error term $e_{ji} = \ln \varepsilon_{ji}$ is constant over time and whose variance is known. Taking further first differences we obtain

$$\Delta \ln y_{ji}(t) = \Delta \ln x_{ji}(t),$$

in which case the error term will be eliminated entirely.

5.2 Supplementary worker panel

For the supplementary data on workers, we first selected all workers from the worker panel who have been employed with companies covered by the supplementary company data during the period 1988-1996, i.e. workers employed with small and medium-sized companies of the company data (see Table 5.1). Secondly, we sampled randomly one-third of those workers who can be matched only to the large companies, which had to be dropped out of the supplementary company data.[17] The first group covers some 650,000 workers and the latter one about 150,000 workers, so that combining these groups yields the supplementary worker data of around 800,000 workers.[18]

Information content differs slightly from that of the whole worker data because some of the variables had to be excluded from the supplementary data. The excluded variables are the home municipality and some variables referring to employing company, such as the industry and legal status of the company. These

[16] It follows that $\sigma^2 = (0.15/1.96)^2$, $E(\varepsilon_{ji}) = \exp(\sigma^2/2) \approx 1.0029$ and $Var(\varepsilon_{ji}) = \exp(2\sigma^2) - \exp(\sigma^2) \approx 0.0059$.

[17] This is because we were not allowed to include all workers associated with large companies.

[18] The staff of Employment Statistics Department further required that all persons whose annual income subject to state taxation exceeds 1.5 million FIM in some of the observation years have to be excluded from the supplementary sample. There were 1,015 such individuals.

variables had to be excluded in order to make impossible the identification of individual companies in the supplementary company data by matching worker records to company records. However, the supplementary worker data still contain company and establishment identifiers, so the linking of worker and company records is possible just like in the case of entire data sets.

6. Some tests of the quality of the IP data

In this section we do some cross checking between the company panel and worker panel in order to highlight the quality of the data. The emphasis is on labour force related issues like job and worker flows. We begin with an application of enterprise demography and demonstrate how one can overcome certain problems associated with job and worker flow measures that result from changing company identifiers in the administrative registers. By comparing some aggregate figures computed from the worker and company panels, we show that both sources of data produce consistent results and that there are no major problems in matching the worker and company records in the IP data.

6.1 Disentangling the company identifiers – enterprise demography and the reliability of job and worker flow measures

As was pointed out earlier, the worker panel of the IP data contains information basically on all Finnish individuals that can be matched to a private sector company or establishment in any of the years 1988-1996. The link between individuals and their employers available in the Employment Statistics database – the company and establishment identifiers of the employer – has been previously used in counting job and worker flow measures as well as in augmenting company and establishment data with the age and education measures of personnel.[19]

The problem is that company (and establishment) identifiers in the administrative registers can change for several different reasons. If there is a change in the ownership or industry classification of the company, its identifier is subject to change. Merger and dispersal of companies cause usually some existing company identifiers to disappear and new ones to appear. A new company identifier may enter the administrative register because a new company was established or because some existing company created a subsidiary. On the other hand, the company identifier may disappear from the register because the company exited from the market or because it was taken over and absorbed by some other company. When counting job and worker flow measures, it is crucial to have information on factors behind the observed changes in the company identifiers.[20]

---

[19] Ilmakunnas and Maliranta (2000) contains job and worker flow calculations for trade, hotels and restaurants, finance and insurance, and a very detailed analysis of manufacturing industries. Piekkola and Böckerman (2000) deals with job and worker flows in manufacturing, trade, construction and business services. In both studies establishment identifiers attached to workers have been used instead of company identifiers.

[20] In detecting job changes one may prefer to compare the establishment identifiers attached to workers over time. We focus here on the use of company identifiers because the link between the workers and companies in the IP data is the company identifier attached to each employed worker in the worker panel. The problems resulting from different types of changes in the identifiers are basically the same in both

To illustrate this point, suppose that in the worker data we observe the company identifier attached to a given worker to change from one year to the next one. This change may occur because the worker has changed his or her employer or, alternatively, because the identifier of the employing company has changed for some reason. If one is unable to separate between this kind of events, the job and worker flows cannot be measured accurately.

Furthermore, there is no reasonable way to study to which extent employment and labour market flows are attributed to company closures and births without information on demographic changes in the company identifiers. Tuija Mustaniemi (1997) has particularly studied enterprise demography in the Finnish manufacturing industries and retail trade. By utilising the method of Baldwin et al. (1992) and Persson (1998), she found that of 'administrative' company births only 67 per cent in manufacturing and 76 per cent in retail trade can be interpreted as a creation of a new company.[21] That is, up to one third of the new company identifiers entering the administrative registers are due to the reasons other than company closures and births, such as restructuring business units between the existing companies or changes in the industry classification of the existing companies.

To investigate the data and facilitate future research, we applied the same sort of enterprise demographic analysis as mentioned above. Namely, from the worker panel we formed 'synthetic' companies by grouping and counting workers associated with identical company identifiers.[22] That is, the synthetic company is defined as a unique company identifier in the worker panel and its labour force is equal to the total number of workers associated with that identifier. [23]

To be able to sort 'real' company closures and births out of 'artificial' ones, we computed the worker flow from one company to another between two consecutive years for all possible pairs of the synthetic companies. Let $N_{i,j}$ be the number of persons who worked at the company i at the end of year t but who

---

cases, though the establishment identifiers are somewhat less sensitive to change due to demographic events.

[21] An administrative birth was defined as the appearance of a new company identifier in the Register of Enterprises and Establishments, with an additional restriction that it was possible to find at least one worker from the Employment Statistics database who can be attached to the new company identifier. This additional restriction was made to sort out non-operating companies as it imposes the restriction that the new company employs at least one person (which might be the entrepreneur).

[22] It should be stressed that we are implicitly restricting our attention on companies with a positive labour force, i.e. on companies that can be matched to individuals in the Employment Statistics database. In what follows, the number of synthetic companies in a given industry is clearly less than the corresponding number of companies in the Register of Enterprises and Establishment which contains also a large number of non-operating companies.

[23] It appeared that the industry classification code of the employer in the worker panel occasionally varies across workers attached to the same company. The industry code for the synthetic company is therefore defined as the mode of the industry codes of workers under the same company identifier.

worked at the company j at the end of year $t + 1$. N_i is the total number of employees of the company i at the end of year t and N_j is the total number of employees of the company j at the end of year $t + 1$. Given that the identifier of the company j appears for the first time in the worker panel in year $t + 1$, we define that a 'real' birth of the company j has occurred if

$$\frac{N_{i,j}}{N_i} \leq 0.5 \quad \text{and} \quad \frac{N_{i,j}}{N_j} \leq 0.5 \quad \text{hold for all possible } i.$$

In other words, the entry of a new company identifier is interpreted to be a result of company birth if no existing company transfers over half of its labour force to the new company and if no existing company contributes over half of the labour force of the new company.[24] Otherwise the entry of the new company identifier is considered to be 'artificial' in the sense that it is due to some restructuring between the existing companies, such as a transfer ($N_{i,j} / N_i > 0.5$ and $N_{i,j} / N_j > 0.5$), merger ($N_{i,j} / N_i > 0.5$ and $N_{i,j} / N_j < 0.5$) or dispersal ($N_{i,j} / N_i < 0.5$ and $N_{i,j} / N_j > 0.5$). Demographic events for small companies were treated somewhat differently; see Mustaniemi (1997) for details.[25]

When dealing with the synthetic companies we do not allow for the company identifiers to change for 'artificial' reasons defined above. Namely, when the existing company disperses into new companies or creates subsidiaries, we do not treat the new companies being created as independent units but collect them under the identifier of the preceding company, i.e. we replace forwards the new company identifiers by the old one. In the cases of mergers and take-overs we proceed as the merging companies would have been together forever by replacing their old company identifiers backwards by the new one. This kind of procedure is obviously very rough but it enables us to separate the births and deaths of companies from the events reflecting restructuring between the existing companies.

As an example, suppose that two companies, A and B, decide to merge and form a new company C. The identifiers of companies A and B disappear and the new identifier for company C appears in the administrative registers. Consequently, the company identifiers attached to employees of companies A and B will change in the worker panel in the year of merging to correspond to the new company C. If we are counting job and worker flow measures by comparing the original company identifiers over consecutive years, we will observe an equal number of jobs being destroyed (by companies A and B) and created (by company C). In

[24] An alternative threshold value used in the previous studies is 0.6.

[25] We applied two different ways of counting workers: either to count every worker associated with a company identifier or to count only workers that have been employed at least six months during the year(s) in question. In the text we present the results based on the latter approach as the results do not differ decidedly from the first option but seem to be more reliable.

addition the equal number of job-to-job transitions amongst the workers will occur. In this case, however, no jobs are actually destroyed nor created by these employers. Neither does any worker change his or her job.

When dealing with the synthetic companies, the exits of company identifiers A and B as well as the entry of the new company identifier C resulting from the merging will be classified as 'artificial'. In this particular case all identifiers A and B observed in the worker data will be replaced by identifier C. Thus there will be only one synthetic company C whose labour force prior to merging will be set equal to the sum of employees of companies A and B. If we are using the 'demography-corrected' company identifiers based on the synthetic companies, our job and worker flow measures will not pick up the spurious impact of these two merging companies. By the use of synthetic companies, we can thus control for changes in the company identifiers that do not reflect actual job and worker flows.

Whether companies A and B should be interpreted as having exited from the market or not, or whether the company C should be treated as a new firm entering the market or not, depends on the research agenda. But as long as we are concerned with labour market flows it seems reasonable to separate these events from the company births and closures followed by job creation and job destruction.

Figure 6.1 Births and closures of companies in worker panel counted by original codes and after recoding in 1988-1996, all industries

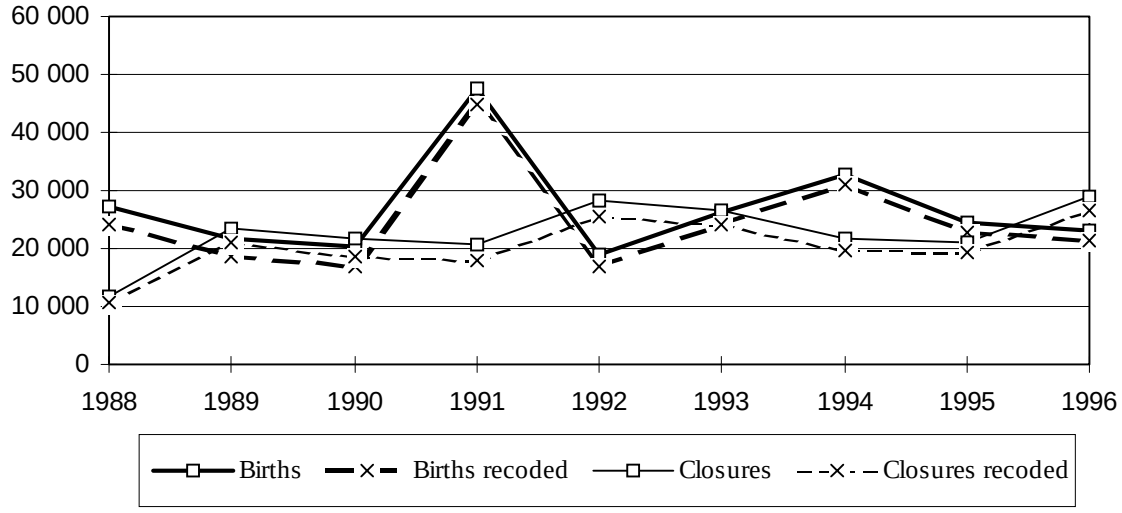
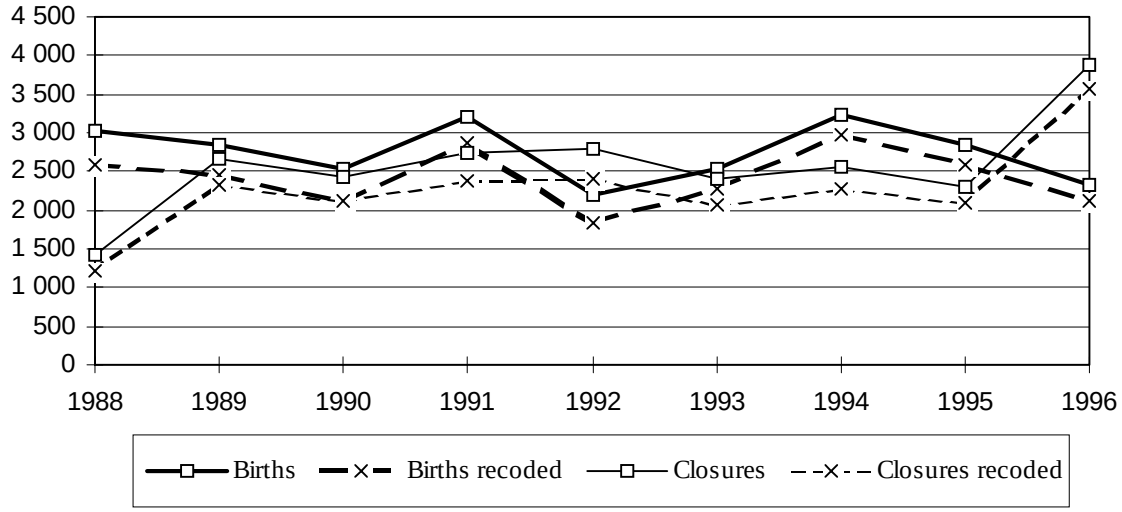


Figure 6.1 shows the number of company births and closures over time as measured by original company identifiers and by demography-corrected company identifiers. The difference between these two measures is quite small and does not call for further investigation. There seems to be a disproportionately

large number of company births in 1991 and 1994, however. This phenomenon is explained by a change in turnover taxation in 1991 and by the introduction of value added taxation in 1994, both of which extended the existing tax base towards smaller companies. During 1991 and 1994 a number of small companies became liable for the business taxes for the first time, and were consequently added to the Register of Enterprises and Establishments.[26]

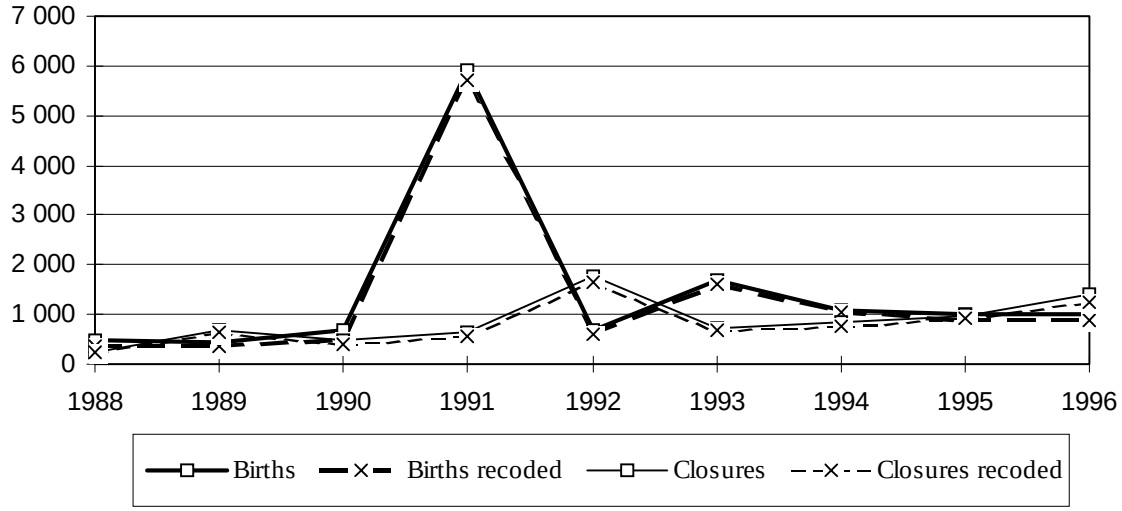
Since the tax reforms affect our results basically through very small companies (often the ones comprising only the entrepreneur) and the average company size varies across industries, one may expect to find inter-industry differences with respect to the company birth figures. This indeed is the case as illustrated in Figures 6.2 and 6.3. Whereas the time series for manufacturing do not exhibit any dramatic patterns, a huge spike in the number of company births occurs in 1991 in transport. Since the average size of the labour force in manufacturing has been traditionally relatively large, the tax reforms did not change notably the tax base in manufacturing.

Figure 6.2 Births and closures of companies in worker panel counted by original codes and after recoding in 1988-1996, manufacturing



[26] Recall that the company identifiers in the worker panel originally came from the Register of Enterprises and Establishments. This register in turn covered basically all registered companies subjected to turnover tax up until 1993, after which the criterion of inclusion has been mainly that the company is subject to value added taxation.

Figure 6.3 Births and closures of companies in worker panel counted by original codes and after recoding in 1988-1996, transport



Turning to job and worker flows, Figures 6.4 and 6.5 depict the flow measures calculated from the worker panel using both the original and the demography-corrected company identifiers.[27] Now it becomes evident that the creation of consistent company coding was a worthwhile exercise. The difference between job and worker flow figures counted by the original and demography-corrected company identifiers is on average some five percentage points, accounting for 20 to 30 per cent of the total flows.

Even though correcting for artificial births and closures did not have a notable effect on the aggregate numbers of company births and closures, the effect on the job and worker flow measures is considerable. This is because the companies experiencing changes in their company identifiers for artificial reasons are much larger on average than companies actually entering and exiting the market. As a result, correcting for spurious changes in the company identifiers affects the job and worker flow measures with a higher weight than it affects the aggregate figures of company births and closures.

[27] We have used formulas by Davis and Haltiwanger (1996) which are reproduced in the appendix. The same method has been applied in other Finnish studies as well. We continue to count only workers employed for six months or more in the year(s) in question.

Figure 6.4 Job creation rate and job destruction rate in worker panel companies counted by original codes and after recoding in 1988-1996, all industries

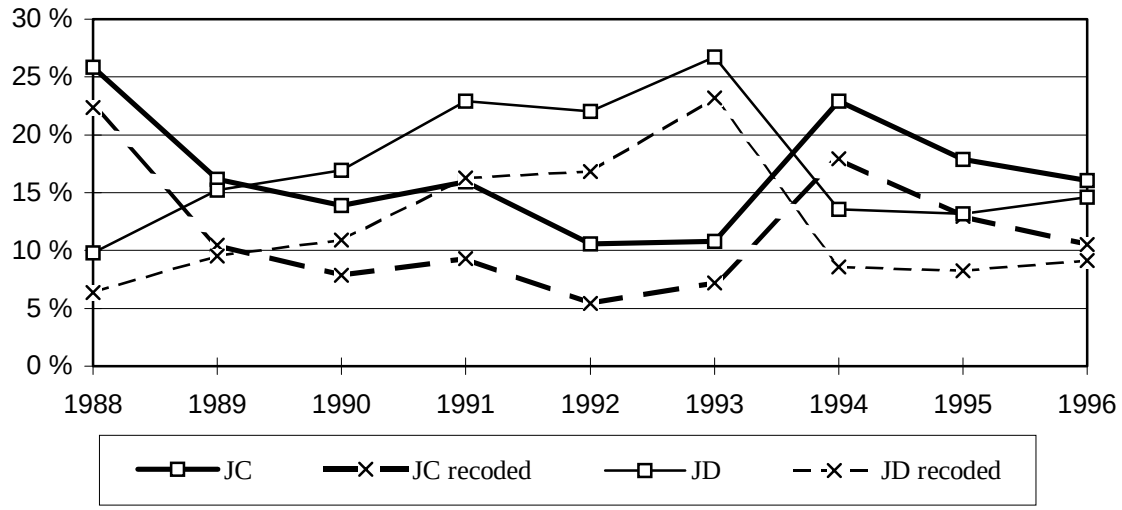


Figure 6.5 Hiring rate and separation rate in worker panel companies counted by original codes and after recoding in 1988-1996, all industries

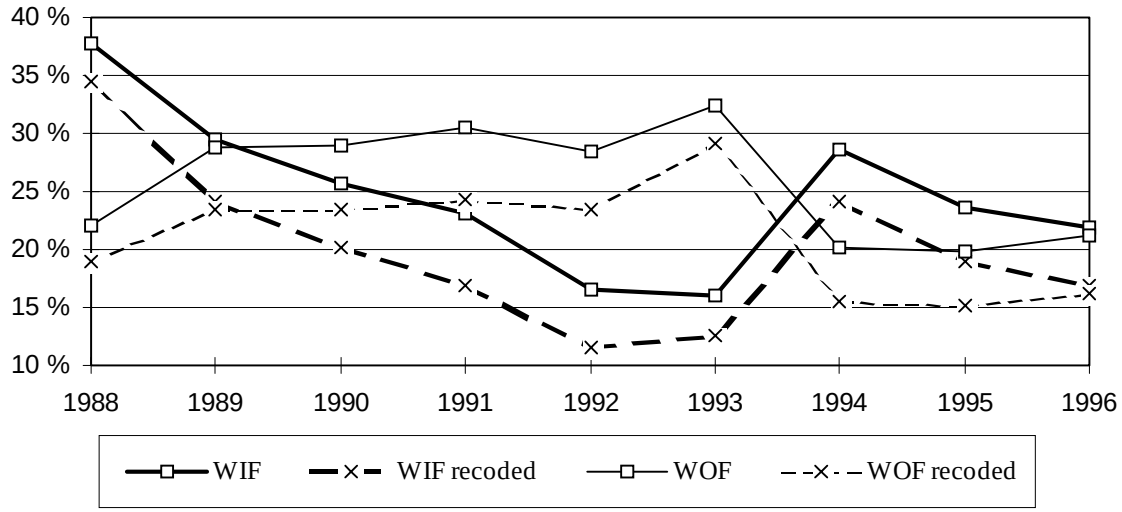


Figure 6.6 Job creation rate and job destruction rate in worker panel companies counted by original codes and after recoding in 1988-1996, manufacturing

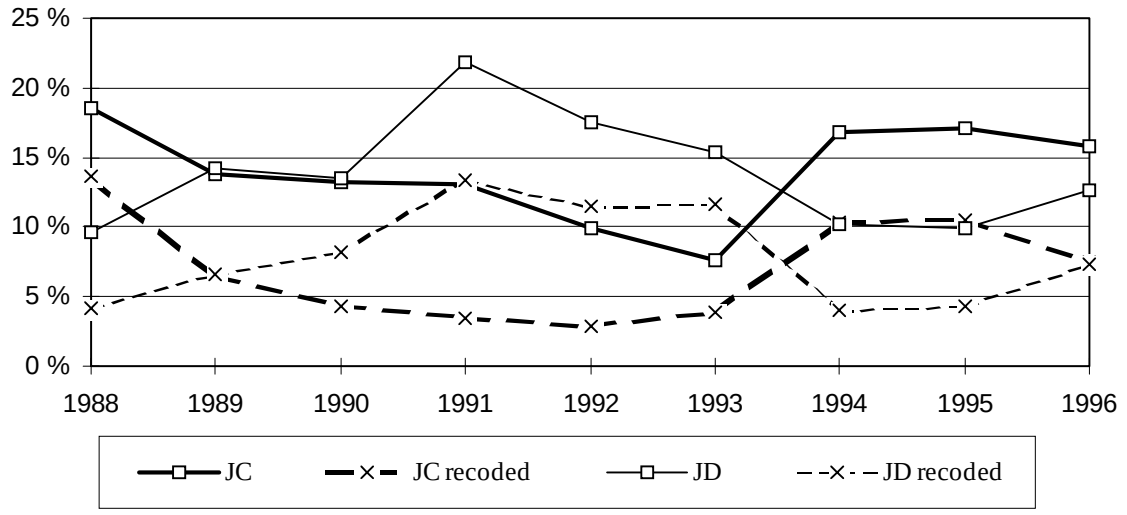
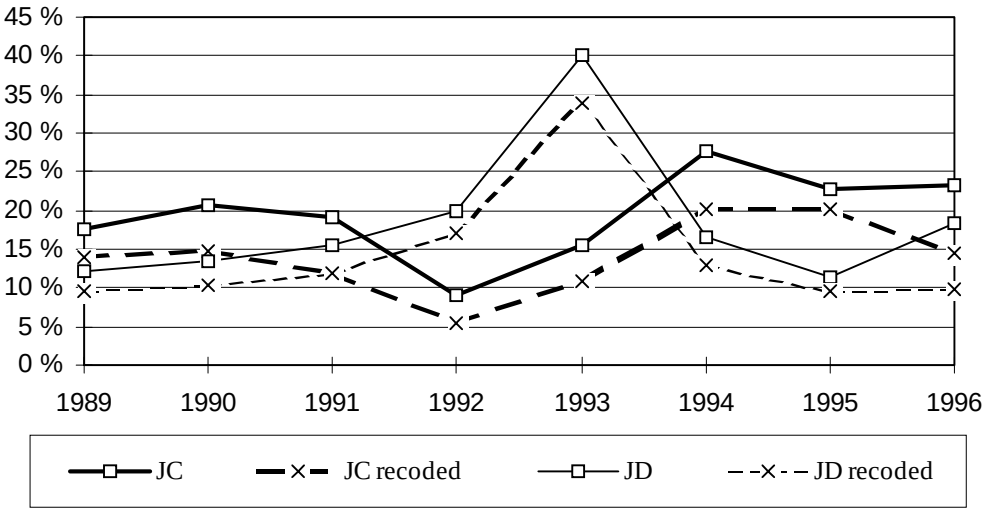


Figure 6.7 Job creation rate and job destruction rate in worker panel companies counted by original codes and after recoding in 1989-1996, business services



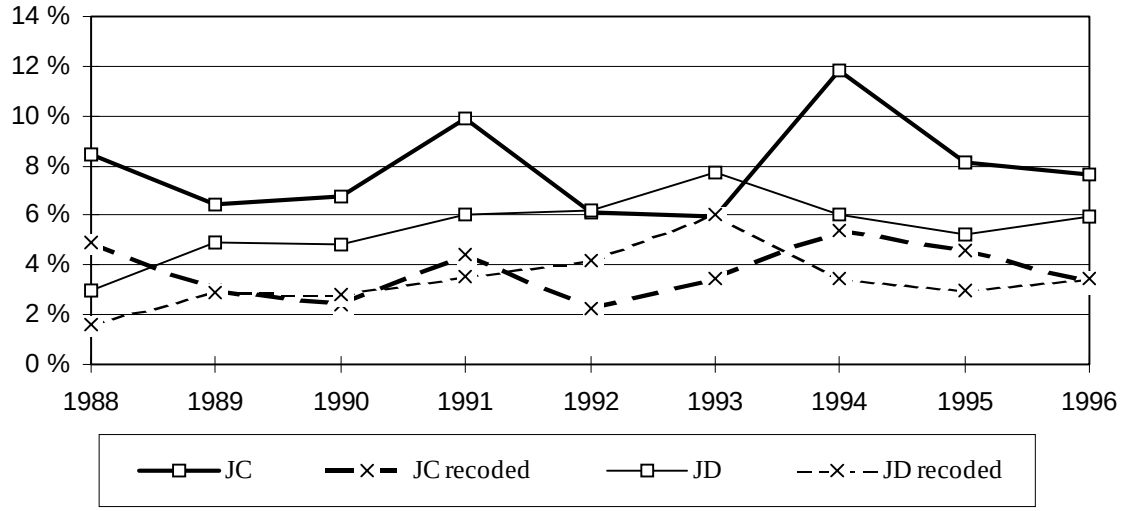
When the births and closures of companies were considered, we found large differences across industries. The same applies with respect to the flow measures. As an example, the job flows in manufacturing and business services are presented in Figures 6.6 and 6.7. Contrary to the business services, the difference between the raw and demography-corrected flow rates is substantial in manufacturing. Although the companies that enter and exit the market are typically small in all industries, the average company size as well as the average size of merging and dispersing companies are relatively large in the

manufacturing sector compared to other industries. This explains why eliminating artificial company births and closures has a stronger effect on the job flow figures of the manufacturing sector. Note also that job flows are generally slightly higher in business services.

Persson (1998) has studied the impact of establishment births and closures on employment in Sweden with correcting for spurious changes in establishment identifiers. It was found that real births contribute only one half of job creation associated with administrative establishment births in Sweden over the period 1987-1994. Similarly the real closures of Swedish establishments account only for 58 per cent of total job destruction associated with administrative closures. Since the majority of Finnish companies consists of a single establishment, one might expect to get similar results with the Finnish data using the company identifiers instead of establishment identifiers.

Figure 6.8 depicts the job creation and job destruction rates associated with company births and closures as measured by using both the original and demography-corrected company identifiers. It turns out that only 47 per cent of total job creation associated with administrative company births can be attributed to real births, whilst real company closures account for 61 per cent of total job destruction associated with the administrative births.[28]

Figure 6.8 Job creation rate and job destruction rate that is caused by births and closures in worker panel companies counted by original codes and after recoding in 1988-1996, all industries



[28] These figures are averages over the observation period.

Figure 6.9 Job creation rate and job destruction rate that is caused by births and closures in worker panel companies counted by original codes and after recoding in 1988-1996, manufacturing

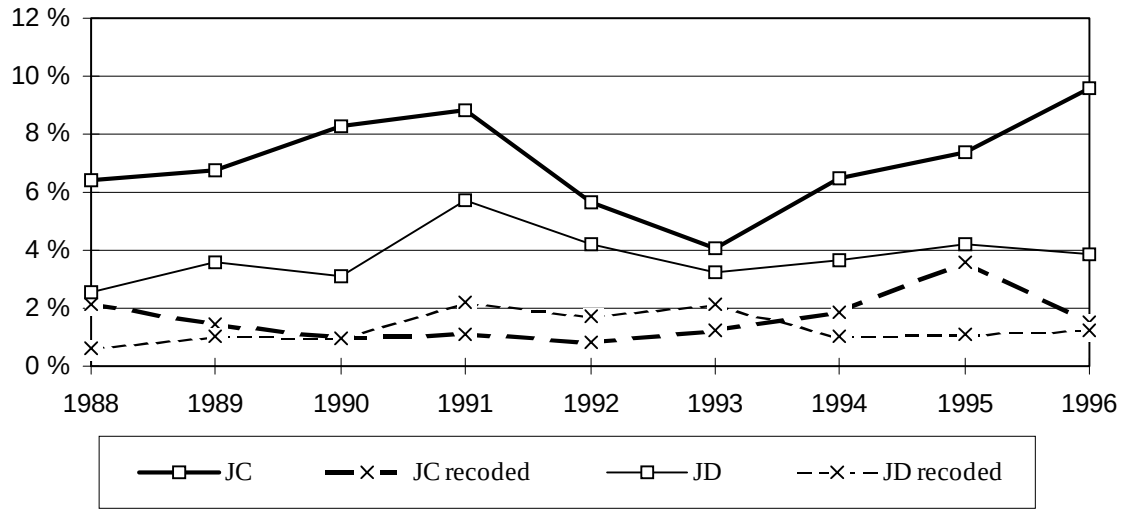
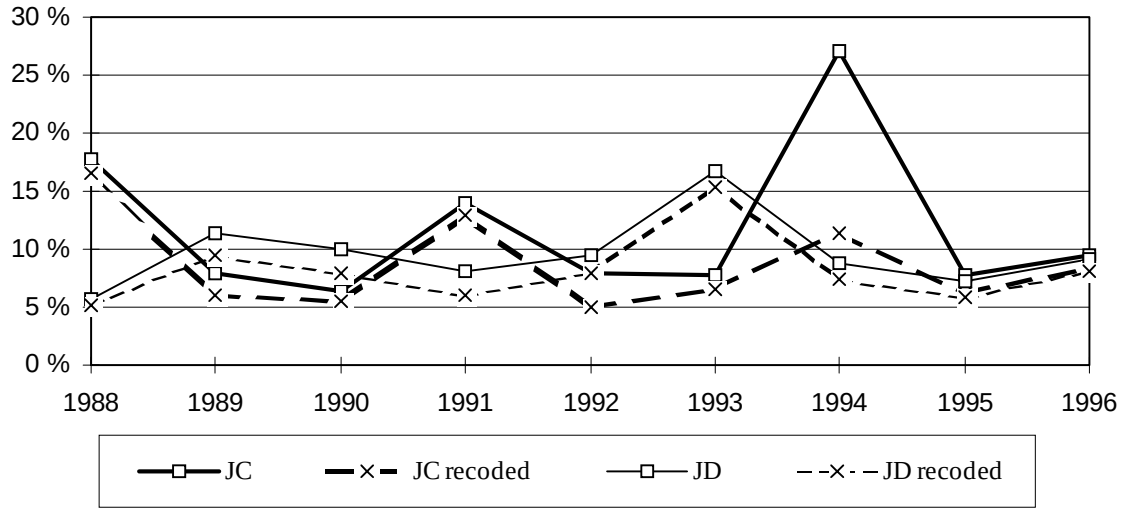


Figure 6.10 Job creation rate and job destruction rate that is caused by births and closures in worker panel companies counted by original codes and after recoding in 1988-1996, non-industry-classified companies



As in the previous cases, inter-industry differences abound. Once again, manufacturing serves as an example of the industry exhibiting relatively low and stable job creation and destruction rates with a sharp difference between the rates counted by original and demography-corrected company identifiers (see Figure 6.9). A good counterpoint in this case is the 'non-industry-classified' category which in our case contains the synthetic companies that we were unable to classify in any of the industry groups as well as companies in the industries not covered by the company panel. The bulk of the companies in this category are

small, employing only five persons on average. In this category job creation and destruction due to births and closures is of larger quantity and more volatile over time, but the decoding difference is quite small (Figure 6.10).

Recall that correcting for artificial events had only a minor effect on the aggregate numbers of company births and closures. On the basis of our job and worker flow measures it is however of great importance to control for spurious changes in the company identifiers if one wants to evaluate the employment impact of companies that enter and exit the market.

6.2 Comparing the employment measures in the worker panel and the company panel

To test the link between two sources of data, the aggregate number of workers obtained from the company panel is compared with the corresponding figures counted from the worker panel. We obtained aggregate employment figures from the company and worker panel in the following ways:

1. Aggregating the labour forces of synthetic companies formed from the worker panel;

2. Counting the aggregate number of workers by industry code in the worker panel;

3. Counting the aggregate number of workers in the synthetic companies who have been employed at least 6 months in the considered year;[29]

4. Aggregating the labour forces of companies in the company panel as reported in the financial statements;[30]

5. Counting the aggregate number of workers in the worker panel who can be matched to the companies covered by the company panel.[31]

Of course, some discrepancies between different figures are expected because of differences in the definitions and time of measurement. No major differences should occur however, if both data sets are considered to be consist with each other.

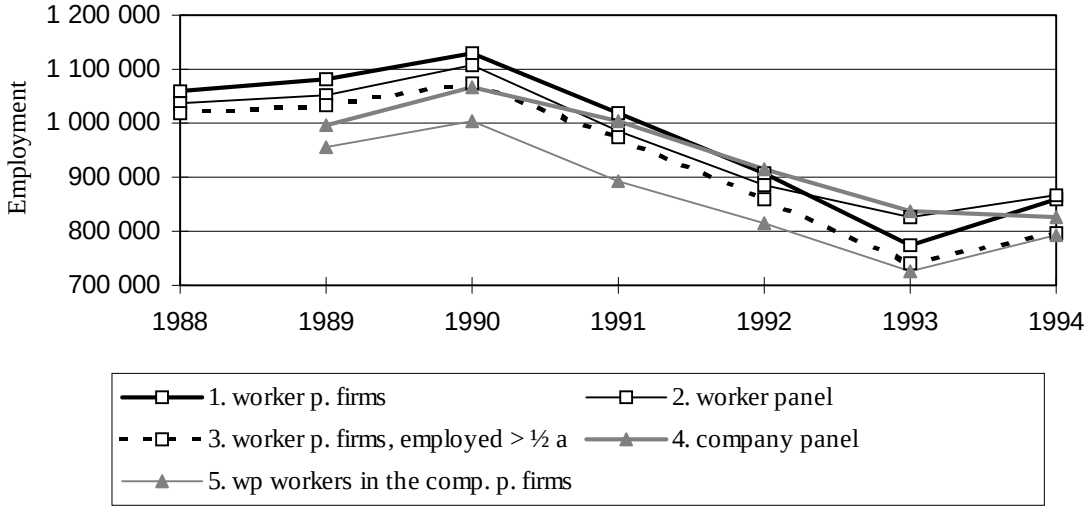
[29] This restriction was done in order to reduce the impact of the difference in the time of measurement between the company and worker panel.

[30] The number of employees of each company in the company panel is weighted by its sampling probability.

[31] Workers are weighted by using the sampling probabilities of their employers in the company panel.

Differences can be expected to arise especially between the figures obtained from the company panel and worker panel. Companies surveyed for the FSS report the average labour force during the accounting period as measured in person years (i.e. two half-time workers are reported as one). In contrast, workers in the worker panel are attached to their employers at the last week of each year irrespective of their employment history before that week and independently of whether they are employed on a half or full-time basis.

Figure 6.11 Comparing different employment measures counted from the worker panel and the company panel, industries covered by company panel



In Figure 6.11 are plotted the aggregate employment figures over the industries covered by the company panel. The time period is restricted to cover the period 1988-1994 since only half of the company panel covers 1995. We also lack observations on companies in business services for 1989.

All the employment measures counted from the worker panel move in unison and the levels are close to each others. Only exception is the year of 1993 when the aggregate number of employees in the covered industries (measure 2) exceeds the figure obtained by aggregating the labour forces of the synthetic companies (measure 1). It is hard to explain this phenomenon as both figures were obtained by counting more or less the same workers.

Recall that when forming the synthetic companies we first collected all workers associated with the same company identifier together, after which the industry code of the emerging company was defined as the mode of the industry codes of all its employees. For some of the employees of a given synthetic company, the industry code can be missing or differ from that of the company. This explains why the aggregate employment figure based on the synthetic companies may differ from the figure obtained by counting individual workers grouped by the

industry code. But it does not explain why the ranking of these two measures changes unexpectedly for one period. We suspect that this observation must be attributed somehow to the revision of Standard Industrial Classification (SIC) of Statistics Finland which took place in 1993.

It turns out that aggregate employment counted from the company panel does not react as severely to the recession at the beginning of the 1990s and the recovery seems to begin more slowly. The drop in employment from 1990 to 1993 is approximately 30 per cent when counted from the worker panel but only 21 per cent on the basis of the financial statement figures. This discrepancy reflects partly the difference in the time of measurement as well as the increasing number of temporary and fixed-term employment contracts over the recession. It is also possible that the sampling of the FSS did not entirely capture the full weight of company closures and births on total employment in the company panel.

The difference between the employment measure counted from the company panel (measure 4) and that based on the number of the workers matched to their employers in the company panel (measure 5) varies from 4 to 13 per cent over time, the average difference being 8 per cent. Note that the difference increases for the recession period 1991-1993. We believe that this is a result of an increase in the relative number of temporary and fixed-term employment contracts during the recession. Overall, the quantitative discrepancy between the measures is relatively small and towards the expected direction. Thus it seems that there are no particular problems in matching worker and company records.

We also computed the similar employment figures industry by industry, though the 1993 revision of SIC and the fact that the industrial classification underlying the company panel differs somewhat from that used in the worker panel caused some troubles for that effort. The results of this exercise are not reported here but can be obtained from the authors on request.

Additionally, we tested to which extent the company panel represents the private sector as a whole by counting the aggregate employment of the *whole* private sector from the worker panel to be compared with the total employment obtained from the company panel. It was found that the industries represented by the company panel account approximately for 70 per cent of total private-sector employment.

6.3 Wages in the worker panel and the company panel

Some of the most important variables in the worker panel and the company panel are the different wage measures. To confirm that one can obtain reliable wage measures from the worker panel to be attached to the companies, we counted wage sums over workers in the worker panel to be compared with the figures

based on the financial statements of companies. In particular the following measures were counted:

1. The wage sum obtained from the company panel based on the financial statements;[32]

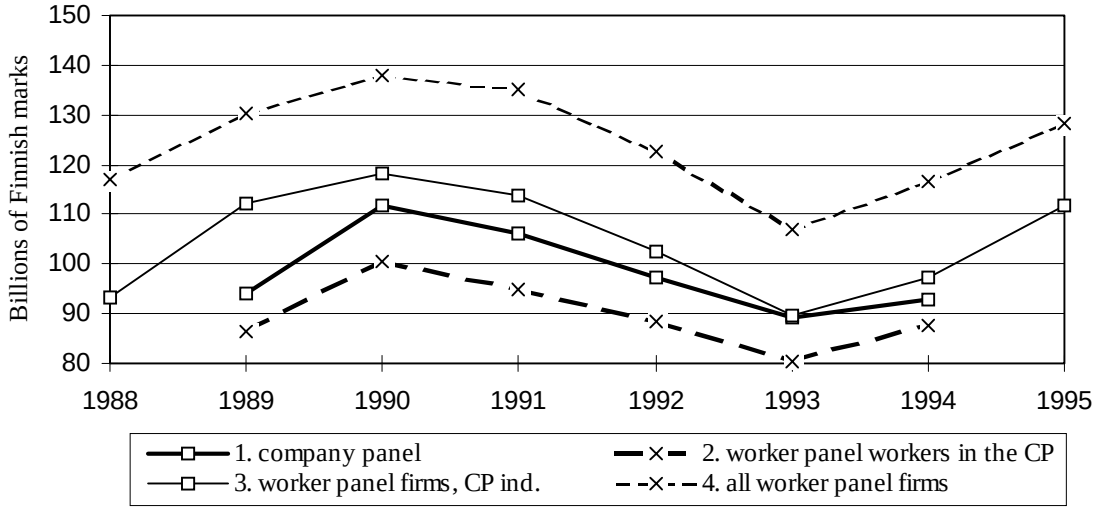
2. The wage sum over workers in the worker panel who can be matched to companies in the company panel;[33]

3. The wage sum over workers in the synthetic companies in the industries covered by the company panel;

4. The wage sum over workers in *all* private-sector synthetic companies.

These wage measures are plotted in Figure 6.12. As expected, the patterns of time series in Figure 6.12 resemble closely those reported in Figure 6.11. Wage sums move in unison and the three lowest curves are close to each others. The pattern of the discrepancy between the wage sum of synthetic companies in the industries covered by the company panel (measure 3) and the two wage sums related to the company panel (measures 1 and 2) is consistent with the pattern of the corresponding discrepancy in the employment figures in Figure 6.11. Overall, there should be no major problems in measuring the wages either from the company panel or from the worker panel.

Figure 6.12 Wage sums counted from the company panel and worker panel



[32] The wage sum of each company in the company panel is weighted by its sampling probability.

[33] The wage of each worker is weighted by using the sampling probability of his or her employer in the company panel.

It appears that the wages paid in the industries covered by the company panel (measure 3) account for almost 85 per cent of the total wage sum of the private sector (measure 4). Recall that the industries covered by the company panel represent only for 70 per cent of the total employment in the private sector, indicating that the excluded industries are dominated by low-wage companies.

We also counted the wage sums industry by industry, and the largest differences between the financial statements and corresponding worker panel wage sums emerged in real estate, cleaning and rental services, being 26 per cent, and in construction, being 21 per cent. The difference was found to be smallest in wholesale trade, 1 per cent, and quite small in energy and water supply, 4 per cent. These differences across industries are mainly due to seasonal differences. The seasonal fluctuations are greatest in the former industries, leading employment to fall during the winter periods. Consequently, there are relatively less workers to be attached to the companies in the last week of the year, i.e. in the time interval captured by the worker panel figures, compared to the accounting-period averages as reported in the financial statements.

7. Conclusion

In this paper we have described the Integrated Panel of Finnish Companies and Workers which contains longitudinal information on both sides of the labour market. The company records of the IP data form a dynamically representative panel of companies from different sectors of the economy for the period 1989-1995. The worker records are in turn available basically for all private-sector workers over the period 1988-1996. The identifiers of the employing companies attached to employed workers in the worker panel provide a link for workers who have been employed with the companies appearing in the company panel.

The only major drawback of the IP data is perhaps a relatively small number of cross-sectional observations on the most of small and medium-sized companies. Although all large companies are observed over the whole sample period, smaller companies may vanish from the sample too quickly for some purposes of dynamic panel data estimation and skew the size distribution of the estimation sample. Moreover, our application of enterprise demography in studying job and worker flows suggests that one should take a good care of separating the real company births and closures from the artificial ones when counting labour market flows or detecting changes in the company population.

Since the IP data cover a considerable portion of the economic activity in Finland over an interesting time period, it provides ample opportunities to tackle questions that earlier were unanswerable due to the lack of suitable data. We believe that this novel data set will find its use in many future research projects and will assist especially in probing the dynamics of the early 1990s recession in Finland.

It is finally worth emphasising that the construction of the IP data was based purely on combining the existing databases, maintained by Statistics Finland, without any effort on collecting new data. With this respect it serves as a good example of how it is possible to use the existing data sources in creating new data sets that open completely new possibilities for research. Since the collection and maintenance of large administrative databases is costly, exploiting them in a way we have done in the case of the IP data does not only extend the utilisation of the existing databases but also increase the effectiveness of the Finnish statistical system. We hope that the kind of collaboration between researchers and statistical authorities which was needed to create this unique data can take place also in the future.

References

Abowd, J. M. – Kramarz (1999): "The Analysis of Labor Markets Using Matched Employer-Employee Data" in Ashenfelter O., and D. Card eds., *Handbook of Labor Economics,* Vol. 3B, Elsevier.

Baldwin, J – Dupuy, R. – Penner, W. (1992): "Development of Longitudinal Panel Data from Business Register: Canadian Experience," *Statistical Journal of the United Nations*, ECE 9, 289-303.

Davis, S. J. – Haltiwanger, J. – Schuh, S. (1996): *Job Creation and Destruction*, The MIT Press, Cambridge.

Ilmakunnas, P. – Maliranta, M. (2000): "Työpaikkojen syntyminen ja häviäminen ja työvoiman vaihtuvuus Suomessa 1988-1996," Ministry of Labour Studies No. 209.

Mustaniemi, T. (1997): "Enterprise Demography as a Method of Studying Real Enterprise Births" in Laaksonen, S. ed., *The Evolution of Companies and Industries,* Statistics Finland.

Persson, H. (1998): "Job and Worker Flows in Sweden 1986-1995: Who Gets the New Jobs and Who Loses the Old Jobs?" Swedish Institute for Social Research, Working Papers 7 / 98.

Piekkola, H – Böckerman, P. (2000): "On Whom Falls the Burden of Restructuring? – Evidence from Finland," The Research Institute of the Finnish Economy, Discussion Paper No. 714.

Tegsjö, B. (1995): "Demography of Enterprises and Establishments in Sweden. An Employment Approach to Measure the Dynamics Among the Units," Statistics Sweden, paper presented in the 9th Roundtable of Business Registers.

Appendix: Information sources

Register of Enterprises and Establishments

Statistics Finland's Register of Enterprises and Establishments (REE) is a basic data which covers practically all Finnish companies and their establishments with a reasonable level of activity. Until 1993, all registered employers and companies subject to turnover tax and satisfying some minor preconditions were included in the data. Since the introduction of value added taxation in 1994, the criterion for inclusion has been basically that the company is subject to value added taxation.

The REE covers some 95 per cent of the activity of Finnish non-agricultural companies when measured by personnel. The main purpose of the register is to serve as a sampling frame for survey studies, and hence its information content is quite limited. The data on companies and establishments are derived mainly from two sources. The records of the National Board of Taxation provide information on company openings and closures, home municipality, preliminary branch of economic activity, turnover and wages. In addition questionnaires are addressed to all new registered employers and companies subject to valued added taxation (or previously to turnover tax), to all multi-establishment companies and to all single-establishment companies with more than 20 employees. The questionnaires ask only for basic information on the size of personnel, turnover, industry, municipality and type of ownership.

Financial Statement Statistics

Financial Statement Statistics (FSS) is a survey that Statistics Finland compiled annually on the basis of corporate income statement and balance sheet data. It is used to estimate profitability, financial position and liquidity of different branches of the economy.

The statistical unit in the FSS is an independent company. Groups of companies are not covered, nor corporate establishments. Also excluded are central and local government companies, voluntary associations, foundations and, for the most part, own-account workers. The company identifiers used in the FSS are identical to those used in the Register of Enterprises and Establishment. The FSS is organised into five broad groups of sectors on the basis of the company's principal activity, which is defined as the activity for which the aggregate value added in establishments is the largest. These sector groups are manufacturing, construction, trade, transport and business services.

The FSS is based on a sample survey and information is gathered from companies using official Statistics Finland questionnaires. The information content varies slightly between the sector groups. The data for period *t* in the FSS are from the accounting periods ending between the 1st April in year *t* and the last day of March in year *t* +1. The

accounting period of most companies is the calendar year.[34] The survey is statutory and mandatory, and the collected data are subject to strict privacy protection. In addition there are several rounds of sending reminding letters to non-respondents with suggestions to answer to the inquire. In spite of these the survey suffers from the response bias and incomplete answers that are common for all surveys. However, the share of missing and rejected questionnaires is relatively small being around 15 per cent annually.

In 1989 the sampling method, industry classification and information content of the FSS changed so that the data for previous periods are not consistent with the later periods. Moreover, the sampling method changed again in 1995 in the manufacturing and construction sectors and in 1996 in the other sector groups. Since we have used the FSS data only from the accounting periods 1989-1995 in compiling the company panel of the IP data, we describe below the sampling method used by Statistics Finland over that period.

The target population for sampling is derived from the Register of Enterprises and Establishment for the previous year, and it comprises basically all companies in operation. The annual sample is selected by stratified sampling using the company's activity and size group of personnel as stratification variables. A total of 64 sector groups and 6 size groups are used in the stratification. All large companies are surveyed, whereas a sample is drawn from smaller companies. The threshold size for large companies is 100 employees in manufacturing and trade, and 50 employees in construction and trucking. There are so few companies in the other branches of transport that they are all included in the annual survey. In the business service sector, which is dominated by small companies, the threshold size is lower than in other sectors being 20 employees.

Instead of drawing a totally fresh sample of companies in each year, the old sample of small and medium-size companies is rotated annually by replacing a portion of the oldest companies in each stratum with new ones. The rotation sampling is applied in order to keep the sample representative over time and to reduce the respond burden amongst small and medium-sized companies.

Employment Statistics

Employment Statistics (ES) is a major database of Statistics Finland which main objective is to produce regional figures about the economic activity of the Finnish population. Since 1987, Statistics Finland has updated the ES database annually by combining information from over 20 administrative registers. The statistical unit being an individual person identified by unique social security number, the ES covers effectively all people with a permanent residence in Finland.

[34] In cases where the length of the accounting period deviates significantly from 12 months, the data are adjusted to correspond to a normal-length accounting period.

The most important registers and administrative sources used in compiling the ES database are: Population information system of the Population Register Centre, registers maintained by the tax authorities, employment registers maintained by the Central Pension Security Institute, the State Treasury and municipal pension insurance programmes, Statistics Finland business register and register on the non-corporate public sector, the National Pensions Institute's pensioner register, student registers, Ministry of Labour registers on job applicants, Statistics Finland register on degrees and examinations, and the conscript register.

The list of variables available in the ES database is far to extensive to be described here. Amongst others, for each person the ES database includes information on education, occupation, annual incomes, employment and unemployment histories and family background. In addition the ES database includes the company and establishment identifiers of the employing company for persons who are employed at the end of the year.

Appendix: Definitions of flow measures

Definition 1. (Gross) Job creation (C) at time t equals employment gains summed over all companies f that expand or start up between $t-1$ and t,

$$C = \sum_{f \in E^+} \Delta E_f \ ,$$

where E^+ is the set of companies where employment is E_f and $\Delta E_f > 0$.

Definition 2. Job destruction (D) at time t equals employment losses summed over all companies that contract or shut down between $t-1$ and t,

$$D = \sum_{f \in E^-} \left| \Delta E_f \right| ,$$

where E^- is the set of companies where employment is E_f and $\Delta E_f < 0$.

Definition 3. Job creation rate (JC) at time t equals job creation divided by total employment averaged over $t-1$ and t,

$$JC = \frac{C}{\frac{1}{2}\left(\sum_f E_f^{t-1} + \sum_f E_f^t \right)} .$$

Definition 4. Job destruction rate (JD) at time t equals job destruction divided by total employment averaged over $t-1$ and t,

$$JD = \frac{D}{\frac{1}{2}\left(\sum_f E_f^{t-1} + \sum_f E_f^t \right)} .$$

Definition 5. Hires (HIR_f) for company f at time t equals the number of workers whose company identifier $\neq f$ at time $t-1$ and company identifier $= f$ at time t. Total hires (HIR) are then

$$HIR = \sum_f HIR_f \ .$$

Worker inflow rate is defined as total hires divided by total employment averaged over $t-1$ and t:

$$WIF = \frac{HIR}{\frac{1}{2}\left(\sum_f E_f^{t-1} + \sum_f E_f^t\right)}.$$

Definition 6. Separations (SEP_f) for company f at time t equals the number of workers whose company identifier $= f$ at time $t-1$ and company identifier $\neq f$ at time t. Total separations (SEP) are then

$$SEP = \sum_f SEP_f \ ,$$

and worker outflow rate is defined as total hires divided by total employment averaged over $t-1$ and t:

$$WOF = \frac{SEP}{\frac{1}{2}\left(\sum_f E_f^{t-1} + \sum_f E_f^t\right)}.$$

Appendix: List of variables for the company panel

This is a somewhat shortened variable list for the company panel. Here are listed the most important variables (in alphabetical order) that are common to all industries. In addition to a host of industry specific variables, many sub-categories of the financial statements variables are not reported here.

1. Accruals and deferred income
2. Advance payments and construction in progress
3. Advance payments paid
4. Advances received
5. Appropriations
6. Beginning and closing date of the accounting period
7. Bills of exchange payable
8. Buildings
9. Capital and reserves, total
10. Cash in hand and at banks
11. Change in depreciation reserve
12. Change in untaxed reserves
13. Company identifier
14. Convertible debentures
15. Country of origin for the largest foreign investor
16. Creditors, total
17. Current creditors, total
18. Debentures
19. Debtors
20. Debtors, total
21. Deductions (delivery price)
22. Depreciation
23. Depreciation of investments held as non-current assets
24. Depreciation reserve
25. Development expenses
26. Dividend distribution
27. Dividend income
28. Dividend payments
29. Exchange rate differences
30. Exports included in turnover
31. Extraordinary expenses
32. Extraordinary income
33. Financial assets
34. Financial result
35. Fixed assets and other non-current investments
36. Gross margin
37. Group contributions, paid
38. Group contributions, received

39. Increase in share capital, liable to charge
40. Increases (purchase price)
41. Industrial classification
42. Institutional sector classification
43. Intangible assets
44. Intangible rights
45. Interest expenses
46. Interest income
47. Investments
48. Investments held as current assets
49. Investments held as current assets, total
50. Land and water
51. Legal form
52. Legal reserve
53. Length of the accounting period
54. Liabilities subject to interest
55. Loan receivables
56. Loans from credit institutions
57. Machinery and equipment
58. Materials and consumables
59. Net investments, intangible assets
60. Net investments, tangible assets
61. Net result
62. Non-current creditors, total
63. Number of establishments
64. Of which: profit from sales / loss
65. Of which: revaluation at the end of the financial year
66. Operating profit (loss)
67. Other capitalised long-term expenses
68. Other current creditors
69. Other debtors
70. Other direct taxes
71. Other equity
72. Other financial expenses
73. Other financial income
74. Other investments
75. Other non-current creditors
76. Other operating charges
77. Other operating income
78. Other tangible assets
79. Other untaxed reserves
80. Pension loans
81. Personnel, total
82. Prepayments and accrued income
83. Profit / loss for the financial year
84. Profit / loss from operations before depreciation
85. Provisions
86. Purchases during the financial year

87. Reduction in value
88. Reduction in value of current assets
89. Research expenses
90. Result for the financial year
91. Retained earnings (loss)
92. Revaluation reserve
93. Revaluation during the financial year:
    - Tangibles
    - Intangibles
94. Share issue (unregistered share capital)
95. Share of foreign ownership
96. Shares and holdings
97. Staff expenses, total
98. Stocks
99. Subordinated loans
100. Subscribed capital
101. Tangible assets
102. Total assets
103. Total liabilities
104. Total result
105. Trade creditors
106. Trade debtors
107. Turnover
108. Untaxed reserves, total
109. Valuation items
110. Variation in stocks of finished goods
111. Variation in stocks, (+ for increase, - for decrease)
112. Work performed by the undertaking for its own purpose and capitalised

Appendix: List of variables for the worker panel

This is a cursory list of variables which gives an overview of the information content of the worker panel of the IP data without details on the classification codes underlying certain variables. Except for the person identifier and sex, all listed variables are defined on an annual basis.

1. Person identifier
2. Age
3. Annual earned income
4. Annual entrepreneur income
5. Annual income subject to state taxation
6. ATV job spell:[35]
   - starting date
   - ending date
7. Education (level and field)
8. Employing company:[36]
   - legal status
   - owner type
   - company identifier
   - establishment identifier
   - industry classification code
9. Home municipality code
10. Indicator for entrepreneur
11. Labour market state at the end of year
12. Last terminated job spell:
    - starting date
    - ending date
    - reason for termination
13. Last terminated spell of job replacement:
    - starting date
    - ending date
    - reason for termination
14. Last terminated unemployment spell:
    - starting date
    - ending date
    - reason for termination
15. Marital status
16. Months unemployed
17. Months worked
18. Number of terminated job spells
19. Number of terminated spells of job replacement
20. Number of terminated unemployment spells
21. Pension type
22. Sex
23. Starting date of end-of-year labour market state spell

[35] ATV job spell is the longest job spell during the year.

[36] Refers to the job spell in force at the end of year, i.e. to the TVM job spell.

24. TVM job spell:[37]

- starting date

25. *X*th terminated unemployment spell, *X*=1,2,3,4:

- starting date
- ending date
- reason for termination

[37] TVM job spell is the job spell in force at the end of year.