STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 19
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iii): Editing of administrative data

USE OF ADMINISTRATIVE DATA IN POPULATION CENSUSES – DEFINITION OF MAIN TYPE OF ACTIVITY AS AN EXAMPLE

Invited paper

Submitted by Statistics Finland[1]

Abstract: Since 1987, data on the population's economic activity have been produced annually in Finland from administrative register data. The most important of the registers exploited are the Central Population Register, the Register of Buildings and Dwellings, and the Register of Enterprises and Establishments. The additional registers used include registers of employment pensions, taxation, unemployed, pensioners and students. Population's main type of activity is based on information on the person's activity during the last week of the year. The population is divided into persons belonging to the labour force and persons outside the labour force. The information about the person's main type of activity is based on about 30 different registers. It is determined so that the classified population diminishes after each category, since the persons already classified are no longer included.

I. INTRODUCTION

1. In Finland the use of administrative data and registers already has a history of over 30 years. The decisive step towards a register-based population system was taken at the end of the 1960s when the Central Population Register was established. By this system, an identifying personal code was issued to each resident of Finland. The same personal identity code was used in other administrative registers, such as in taxation and in the employment pension insurance system. The information in the register was employed for the first time for the 1970 Population Census: personal data (personal identity code, name, address and some demographic data) were obtained from the Central Population Register and income data from the taxation register (Myrskylä 1991).

2. The use of registers increased in the 1970s and 1980s so that in the 1985 Census only data regarding persons' economic situation were inquired with a questionnaire. For the 1990 Census questionnaire inquiries were no longer sent directly to the population since almost all data were derived straight from registers. However, not all data are available from registers. Data on part-time or full-time jobs or the way of travelling to work cannot be obtained from any registers. In addition, it is not possible to obtain the exact location of persons' workplaces from registers for all employed persons. For this reason it is necessary to ask multi-establishment enterprises for the location data of their personnel's workplace in the present register-based census system.

II. ADMINISTRATIVE DATA AND REGISTERS AS SOURCES OF STATISTICS

A. The basic structure of statistics production based on registers

3. Since 1987, all census data have been produced annually in Finland from data in administrative registers. Each year, Statistics Finland produces demographic and employment statistics, building, dwelling, household and family statistics and statistics on housing conditions. The most important of the exploited registers are the Central Population Register (CPR; total number of residents, demographics, families), the Register of Buildings and Dwellings (RBD; buildings and dwellings), and the Register of

[1] Prepared by Kaija Ruotsalainen, kaija.ruotsalainen@stat.fi

Enterprises and Establishments (REE; all private sector enterprises and public sector establishments). Additional registers used include registers of employment pensions, taxation, unemployed, pensioners, and students. (Myrskylä and Ruotsalainen 1997)

4. The RBD contains data of importance in defining regional statistics. The information is linked via identification data to other statistical units. Units are identified through their personal identity code, building number and enterprise number.

5. The REE contains all private sector enterprises and their establishments, as well as government establishments and municipal establishments. It provides information on branch of industry, type of ownership, legal form and institutional sector.

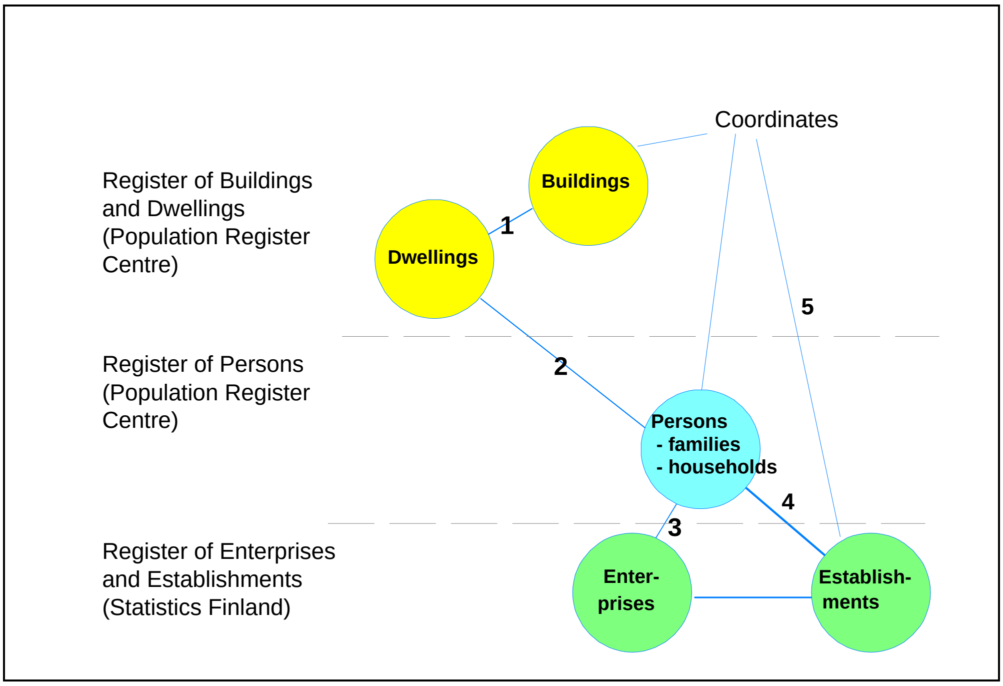
Coordinates
Register of Buildings and Dwellings (Population Register Centre)
Buildings
Dwellings
1
5
2
Register of Persons (Population Register Centre)
Persons
- families
- households
4
3
Register of Enterprises and Establishments (Statistics Finland)
Enter-prises
Establish-ments

Figure 1. Units belonging to the register-based statistical system and links between them

6. Statistical units such as persons, buildings, dwellings and establishments are linked together via different codes. All dwellings (Links 1) are linked to a building via a building code maintained by the CPR. The building code provides the co-ordinates for the respective unit. Persons and dwellings (buildings and map co-ordinates) are linked (2) via domicile codes.

7. Employed persons are annually linked (3,4) with employers and their establishments. With the help of the address it is possible to link (5) the establishment with the property number and co-ordinates in the RBD. In the case of some entrepreneurs, such as farmers, the industry is inferred from pension insurance data and from type of income. The location of the establishment is the home address of the entrepreneur.

B. Conditions for register-based statistical production

8. The registers and administrative data must meet certain conditions in order to be of use in statistics production. All the conditions listed below need not necessarily be fulfilled for an individual register but rather for a group of registers describing some phenomenon. The following conditions are based on the publication “Tilastoista tiedoksi korkea-asteelle” (Statistics into data for high level) edited by Niemi and Tourunen.

9. The utilisation conditions for registers can be divided into six main categories: content and conceptual conditions, reliability, up-to-dateness, technical conditions, administrative conditions and costs.

10. Meeting the content and conceptual conditions can be examined from three perspectives. The first used register or group of registers has to cover its whole target group: the whole population resident in the country, unemployed or employed persons. For example, the CPR contains all permanent residents of Finland and the Ministry of Labour's register on job applicants covers almost all unemployed persons. Only those who seek work outside the official job application system lie outside the register. The number of such persons is very small, however, because the condition for obtaining the unemployment benefit is to register to the official labour exchange. An example of using several registers for one phenomenon could be statistics on employment: no single register covers the employees of all sectors but about ten employment relationship materials from different sectors or employers are in use.

11. Secondly, the target population and the data of the register must describe units: persons, buildings, enterprises, etc. If the register does not contain basic data about units, it cannot be used. Thirdly, the register data must be relevant for statistical use and they should correspond to statistical concepts, such as employed person, unemployed person, establishment, and so on.

12. From the viewpoint of reliability, the registers used should also be sufficiently comprehensive. Individual data must also be as reliable as possible. This concerns both identifying data and attribute data of the units.

13. In order to utilise registers as well as possible in statistics production, they must be as up-to-date as possible. In Finland the fastest registers can be used for statistical purposes already a few weeks or a month from the reference period (Population Register, Register on job applicants), but the slowest registers or administrative data are available for statistical use only after one year or so (Taxation Register, part of employment relationship data).

14. By their technical attributes, registers must meet at least the following conditions. They must have good and detailed descriptions, on the basis of which the user will gain a picture of the function, structure and data content of the register and can assess the usability of the register for statistical purposes. The register must be in electronic format so that it can be used for compiling statistics. It is important that the units of the register have generally used identifiers. Otherwise it can be difficult to use the register. In some cases absence of identifiers can be replaced by name and address data. The attribute data of the register units should be in coded or numerical form so that they can be classified. If the data is in plain language, as is often the case for occupational titles, for example, the data can usually be coded automatically almost completely.

15. Fulfilment of administrative conditions refers to the right of statistical authorities to obtain register data in the possession of different authorities for statistical use, unhindered by secrecy regulations. In Finland this is specified in the Statistics Act.

16. The costs of utilising registers cannot be higher than those of some alternative collection method are. This is not generally the case: the costs of the Finnish register-based population census are just a fraction of the costs of questionnaire collection.

C. Advantages and disadvantages of register-based statistics production

17. One of the main advantages of a population census system that is based on administrative registers is that total data can be produced annually. In the past, population censuses were taken at five or ten year intervals in Finland, as is still done in most countries of the world even today. Although data at the whole country or provincial levels were obtainable from the sample-based Labour Force Survey, for example, small area statistics and statistics made with very detailed industrial and occupational classifications could be produced from population censuses. Data produced at intervals of five years were quickly out-of-date as both regional and social structures change.

18. As the 1980s arrived a clear demand had arisen for annual data. Changing over to annual production became viable thanks to the considerably lower cost of compiling statistics from register data instead of from data collected by questionnaires. In today's money, the questionnaire-based 1980 Finnish Census cost approximately EUR 34 million, while estimates put the cost of the 2000 Census at around EUR 840 000. The transition to annual production was also supported by the fact that it was easier to

maintain the processes and routines of continuous production than to build up a new system every few years.

19. The fact that individuals and buildings can also be assigned exact area coordinates also adds to the possibilities of producing different small area statistics. It also enables the compilation of flow statistics. Flow statistics refer to the temporal monitoring of a certain population, e.g. studying how unemployed jobseekers become employed or those having completed their education are placed on the labour market, etc. Flow statistics can also embrace the regional aspect: what happens to a population in a certain area over a monitoring period of, say, five years.

20. In the register-based statistical system the data on a person's branch of industry and workplace location, for example, are obtained from the REE. Thus, all persons working at the same establishment are assigned precisely the same branch of industry and workplace location. In a questionnaire-based census the data on a person's branch of industry and workplace location were generally based on a person's own reporting. This meant that the data could contain discrepancies even when they concerned persons who worked at the same establishment. Data that are based on registers are thus more reliable and coherent than those collected with questionnaires are. (Ruotsalainen 2001)

21. The use of registers for statistical purposes is by no means free from problems. Dependence on data suppliers can be mentioned as the biggest disadvantage. Statistics compilers and statistics exist as it were completely on the terms of the register holder. An amendment in legislation or some other administrative change can cause changes to the data content of the register as well. Data previously available from a register may become meaningless for the register holder, in which case it is not updated in the register any more. An example of this is the abolition of occupational data from the taxation register. Since the tax proposal procedure was adopted in Finland in the late 1990s, the majority of employees need not any more fill in the tax return and occupational data are thus not updated in the information system of the taxation authorities. In such cases some other source must be found for recording this information in the statistics. For the 2000 Census it was decided that occupational data would be partly collected from employers.

22. Changes in data suppliers' information systems can also bring about problems for the statistical authorities, at least delays from the normal production rate. A register-based statistical system is dependent on all of the different registers. If the completion of some register used as source material is delayed, this generally means that the completion of statistics will also get postponed. The schedule of the statistics is thus dependent on the completion schedule of the last source material.

23. The coverage of registers may in some cases be defective for some data, although the register itself would contain all the units to be described. For example, the Finnish Population Information System contains fairly reliable information on persons, births, deaths, marriages and divorces, but notices of removal, for example, are neglected. It has been estimated that domicile data are erroneous for about three per cent of the population. In addition, occupational data are updated in the Population Information System only in connection with notices of removal, which means that the data are obtained reliably from it only on those who have moved. (Tilastoista tiedoksi korkea-asteelle 1996)

24. Consistency problems may arise when linking information from different registers. For example, when linking personal data on workplace, occupation and income, it is not always certain that they describe the same employment relationship. (Tilastoista tiedoksi korkea-asteelle 1996)

III. USE OF REGISTER DATA IN STATISTICS COMPILATION

A. Ways of utilising registers

25. Finland’s population census system is hence almost completely based on registers. Register data are used primarily in two different ways. The easiest and simplest way is to use directly the information contained in the register. Such data are a person’s demographic data, such as age, gender, marital status, citizenship or income data. These data are obtained directly from one register and they need not generally be much edited or corrected.

26. Another way is to form new variables with the so-called register estimation method. The aim is to estimate for each statistical unit the value of the target variable as close to the statistical concept and definition as possible. This is done by using all the existing data available and a set of decision rules to estimate the value of the statistical variable. The sources can be with partial coverage, maybe overlapping and with different degrees of quality.

27. An example of this method is the forming of families for which several items of data from one register are employed. Another example of using multiple registers for forming one item of data is the definition of a person’s main type of activity. Data from about 30 different registers or administrative materials are used for that. This estimation method is given as an example in this paper.

B. Sources of Finland’s population census system

28. The source materials used for the census are mainly administrative registers and other register-based data materials. Direct data collection is made only for determining establishment data for those working for multi-establishment enterprises and municipal operating units. In all, data from about 40 different registers or data materials are usually used for completing the statistical data for the census. Central of these are:

- Population Information System (persons, buildings, dwellings, free-time residences);
- Various taxation materials;
- Different private sector employment relationship registers;
- Central and local government employment relationship registers;
- Ministry of Labour’s register on job applicants;
- The Social Security Institution’s and the Central Pension Security Institute’s pension registers;
- Different student registers;
- The conscript register of the General Staff of the Armed Forces;
- Statistics Finland’s Business Register and the register on the non-corporate public sector;
- Statistics Finland’s Register of Completed Education and Degrees.

29. In addition to registers, some questionnaire inquiries are also made:

- An inquiry to multi-establishment enterprises requesting information on personnel by establishment in connection with the Business Register inquiry;
- An inquiry concerning the establishment of those working in multi-establishment municipal operating units.

C. Inference of main type of activity and selection of employment relationship on the basis of register data

30. The main type of activity is determined for a person as cross-sectional data. The reference period is usually the last week of the year. The validity of the data in the last week of the year is not totally certain for all data.

31. The main type of activity is inferred on the basis of the register-based parallel statistics made in connection with the 1985 Population Census (Korpi 1989). In 1985 the census was carried out as a questionnaire-based census as concerns data on the economic activity of the population. At the same time the registers were used for inferring the main type of activity, occupational status and employer sector of

the population to examine the comparability and reliability of the planned register-based employment statistics in comparison to the questionnaire-based census. In addition, the data inferred from the register were compared to Statistics Finland’s Labour Force Survey, and this comparison is still made yearly.

32. The inference order is formed in principle in accordance with international recommendations. According to the ILO and UN Population Census recommendations, employment should thus be given priority when inferring a person’s activity in the reference period. Therefore, the person should also be classified as employed on the basis of a very short employment relationship or even a small amount of work (for example employment came before being a student or pensioner). In some cases the nature and quality of the register had an effect on the inference order. For instance, unemployed persons were inferred before employed persons in order not to produce a third unemployment figure in addition to Statistics Finland’s Labour Force Survey and the Ministry of Labour’s Register on job applicants. The inference rules were formed by testing the inferences in different ways and in different order. The aim was to come up with such inference rules by which the numbers of persons in different groups would be as close as possible to the data in the questionnaire-based census and the proportion of those belonging to the same group would be as high as possible.

33. In addition, persons in military and non-military service are inferred before employed persons because the registers on them are considered to be of very high quality and it is assumed that the person cannot be working while doing military/non-military service. Moreover, the person’s employment relationship is preserved when he is in military/non-military service, in which case according to the employment register the person may have a valid employment relationship simultaneously with military/non-military service. If the inference order of employed persons and conscripts were opposite, many persons in military service would be inferred to be employed persons.

34. A similar, but much wider, comparative study was also carried out in connection with the 1990 Population Census (Evaluating study of the 1990 Census, 1994). The 1990 Census was the first Population and Housing Census based completely on registers. In the reliability study, data were collected on about two per cent of buildings, dwellings and persons. The coherence of the data collected from registers and with questionnaires was compared in the reliability survey.

35. The population is classified according to main type of activity, which is based on information on the person's activity during the last week of the year. The population is divided into persons belonging to the labour force and persons outside the labour force. Both principal categories are divided into subgroups in the following manner:

Labour force

- employed
- unemployed.

Persons outside the labour force

- 0-14-year-olds
- students
- conscripts, conscientious objectors
- pensioners
- other persons outside the labour force.

36. The information about a person's main type of activity is based on about 30 different registers. The person's main type of activity is determined in the manner and order described below. The classified population diminishes after each category, since the persons already classified are no longer included. Appendix 1 presents the inference of the main type of activity in 1999 by rule and the proportions of persons determined according to each rule.

1. All persons under fifteen years of age are included in the group "0-14- year-olds" and all those over 75 years of age are classified as "pensioners".
2. All those who according the Ministry of Labour's register on job applicants were unemployed the last weekday of the year are considered unemployed.

3. All those who were doing their military service or non-military service during the last week of the year are classified as conscripts.
4. Persons who have only a valid self-employment pension’s insurance in the last week of the year are classified as self-employed persons.
5. Persons who have both a valid self-employment pension’s insurance and employment pension insurance (employee relationship) and whose entrepreneurial income is larger than earned income are classified as self-employed persons.
6. Persons who according to registers do not have a valid employment relationship but whose entrepreneurial income is higher than EUR 810 and larger than earned income and according to the taxation data the person is an employer are classified as self-employed persons.
7. Persons who according to registers do not have a valid employment relationship but whose entrepreneurial income is higher than EUR 810 and larger than earned income and the person is not retired is classified as self-employed persons.
8. Persons who have a valid employment pension insurance and who have earned income are classified as employees.
9. Persons who do not have employment relationship data but who are employed with labour policy measures are classified as employees.
10. Persons who according to the student register study full-time in the autumn term are classified as students.
11. Those 15-year-olds who were not classified into any other group at earlier stages are classified as students.
12. Persons in labour market training in the last week of the year are classified as students.
13. Persons who have received study aid either in the autumn term or both in the spring and autumn terms are classified as students.
14. Persons who at the turn of the year receive old-age, disability, unemployment or special farmers’ pension are classified as pensioners. In addition, all those whose pension is higher than their income from work and above EUR 580 and the pension is not a survivor's pension or part-time pension are considered pensioners.
15. For the rest of the population employees are defined based on the size of their income. If it exceeds the income as a self-employed person, it being over EUR 4 860, the person is classified as an employee.
16. Persons who do not meet any of the above criteria are included in the category "Other persons outside the labour force".

37. These rules have been nearly unchanged since 1987. There have been some small disparities between different years due to diverse materials. However, the basic principle has been the same in all these years. The income limits used in the inference were originally derived by testing. For self-employed persons and pensioners the original income limits have been raised yearly in accordance with the Cost-of-Living Index. For employees the income limits are defined annually on the basis of the number of employees in the Labour Force Survey in the following way. Persons whose earned income is larger than their entrepreneurial income and/or pension income are sorted in the descending order by earned income. The limit for earned income is determined by the point where the number of employees is the same as in the Labour Force Survey starting from the highest earners.

38. A person may have several valid employment relationships at the same time. If a person has at the end of the year two or more valid employment relationship, the one from which the calculated monthly earnings are higher is, as a rule, chosen as the valid employment relationship. For this reason the monthly earnings by each employment relationship are important. Part of employment relationship data already include the earnings paid for the employment relationship. Then the monthly earnings are calculated directly from that material by dividing the earnings indicated by the duration of the employment relationship. Earnings data are not included in the data on private sector employment relationships. Earnings data are also missing from part of the central and local government employment relationship data. Earnings data are derived for these employment relationships from employers’ annual reporting data.

39. The majority of employed persons have just one employment relationship and then there is no selection problem. However, about 200,000 persons have at least two simultaneously valid employment relationships and then it is necessary to select for the person both primary and second employment relationships, referred to as the person’s secondary employment relationship.

IV. CONCLUSION

40. In Finland administrative data and registers have been used successfully for population censuses for as long as 30 years. The use of registers started in the 1970 Census, when part of personal data were obtained from the Central Population Register and income data were linked to persons from the taxation register. The use of registers increased gradually until all data for the 1990 Census were produced on the basis of registers.

41. Register data are used primarily in two different ways. The easiest way is to use directly the information contained in the register, which need not generally be much edited or corrected. Another way is to form new variables with the so-called register estimation method, in which data of several registers are generally employed. An example of this method is the forming of main type of activity, one of the key variables of employment statistics. Inference rules were formed by utilising the parallel statistics of the 1985 Census for testing the inferences in different ways and in different order. The aim was to come up with such inference rules by which the numbers of persons in different groups would be as close as possible to the data in the questionnaire-based census and the proportion of those belonging to the same group would be as high as possible. The basic principle in the definition of a person’s main type of activity is that first the whole population is included and then the number of persons keeps falling until all have been classified into some group.

42. The reliability of register-based data was considered sufficient on the basis of reliability studies on the 1990 Census and on register-based parallel statistics made in connection with the 1985 Census. In addition, personal data are compared on individual level to the Labour Force Survey as concerns main type of activity, occupational status and industry.

43. A register-based census can produce almost all the data produced earlier in questionnaire-based censuses, that is, all the data in compliance with the EU and UN Population Census recommendations. In addition to conventional census data, the use of registers enables more varied linking and utilisation of data. Moreover, almost all data can be produced annually as total data and the exploitation of existing data sources in statistics is very advantageous from the perspective of costs.

References

Evaluating study of the 1990 Census (1994). Population Census, Vol. 9B. Statistics Finland, Helsinki.

Korpi, Helena (1989). *Main Type of Activity and Occupational Status in the 1985 Census: Register-based Parallel Data*. Studies 152. Central Statistical Office of Finland, Helsinki.

Myrskylä, Pekka (1991). Census by Questionnaire - Census by Registers and Administrative Records: The Experience of Finland. *Journal of Official Statistics Vol. 7.* No 4, pp.457-474. Statistics Sweden.

Myrskylä, Pekka and Kaija Ruotsalainen (1997). Annual System of Small Area Statistics Based on Administrative Records and Registers. *Contributed Papers for the 51st Session of the ISI, Istanbul*. Book 1, 511-512.

Ruotsalainen, Kaija (2001). *Annual System of Small Area Statistics Based on Administrative Records and Registers - the Possibilities and the Problems.* A paper prepared for the forty-ninth plenary session of the Conference of European Statisticians, Geneva, 11-13 June 2001.

Virallisten tilastojen tiedonkeruu (1996) (Data collection for official statistics; in Finnish only). In *Tilastoista tiedoksi korkea-asteelle* (From statistics into data for high level), pp. 111-119. Ed. Niemi, Hannu and Kalevi Tourunen. Statistics Finland, Helsinki.

Appendix 1. Breakdown of the population into different main type of activity groups in the 1999 employment statistics. The reference period is the last week of the year (25-31 December) unless otherwise stated. The classified population diminishes after each category, since the persons already classified are no longer included.

	Rule	Main type of activity group	Proportion of whole population		Cumulative accumulation	
			Persons	%	Persons	%
	Whole population	**Whole population**	**5 171 302**			
1	Persons aged under 15	0-14-year-olds	943 001	18.2	943 001	18.2
2	Persons aged over 74	Pensioners	331 385	6.4	1 274 386	24.6
3	Persons who are unemployed in the register on job applicants on the last weekday of the year	Unemployed	353 586	6.8	1 627 972	31.5
4	Persons in military or non-military service	Conscripts	20 025	0.4	1 647 997	31.9
5	Persons with a valid self-employed person’s insurance only	Self-employed	222 016	4.3	1 870 013	36.2
6	Persons with a valid self-employed person’s insurance and employment pension insurance but higher entrepreneurial income than earned income	Self-employed	10 086	0.2	1 880 099	36.4
7	Persons without a valid employment pension insurance. Entrepreneurial income is higher than EUR 810 and larger than earned income and the person is an employer according to taxation data	Self-employed	777	0.0	1 880 876	36.4
8	Persons without a valid employment pension insurance. Entrepreneurial income is higher than EUR 810 and larger than earned income and the person is not an employer or retired	Self-employed	14 234	0.3	1 895 110	36.6
9	Persons with a valid employment pension insurance and have income from work	Employees	1 872 285	36.2	3 767 395	72.9
10	The person is employed by labour policy measures	Employees	5 056	0.1	3 772 451	72.9
11	Persons who study full-time in the autumn term according to the student register	Students	298 627	5.8	4 071 078	78.7
12	15-year-olds	Students	61 109	1.2	4 132 187	79.9
13	Persons in labour market training	Students	21 975	0.4	4 154 162	80.3
14	Persons who have received study aid either in the autumn term or in both autumn and spring terms	Students	13 445	0.3	4 167 607	80.6
15	Persons receive old-age, disability or special farmers’ pension at the end of the year	Pensioners	725 956	14.0	4 893 563	94.6
16	Persons receive unemployment pension at the end of the year	Pensioners	50 500	1.0	4 944 063	95.6
17	Persons whose pension income is over EUR 580 per year and the person does not receive a survivor’s pension or part-time pension at the end of the year	Pensioners	6 523	0.1	4 950 586	95.7
18	Persons whose earned income is higher than entrepreneurial income and over EUR 4860 per year	Employees	49 431	1.0	5 000 017	96.7
19	Rest of the population	Other persons outside the labour force	171 285	3.3	5 171 302	100.0