UNITED NATIONS

STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

INTERNATIONAL LABOUR ORGANIZATION

Joint ECE-Eurostat-ILO Seminar
on Measurement of the Quality of Employment
(Geneva, 3-5 May 2000)

Topic 3

Wages by level of education and occupation from the Dutch Structure of Earnings Survey

Invited paper submitted by Statistics Netherlands[1]

I.   Introduction

1.   Based on the Structure of Earnings Survey (SES), data are compiled every year on the earned wages of employees in relation to their level of education and occupation. Statistics Netherlands uses a method for the SES which does not require additional surveys. Information of three distinct sources is combined. The three sources used in the matching procedure are:

- the Annual Survey on Employment and Earnings, a business survey, which collects mainly payroll data from the full range of establishments and in which the public sector is well represented;
- the Insured Persons Register, which contains an even larger number of records and in which the private sector is very well represented, but the number of variables is smaller;
- the Labour Force Survey (LFS), a household survey, which collects data on the employment situation, but also on education and occupation.

2.   When the records are lined up, the matching process is started. Payroll and Insured Persons Register micro data are matched with LFS micro data, using linking variables: address, postal code, city, date of birth and sex. Only exact matches are allowed since the aim of the survey is a structural analysis of earning patterns.

3.   This method has been applied for the first time to the 1995 SES. For 1996 the method used has been improved. The 1997 SES has been constructed the same way as for 1996. Therefore, the results of 1996 and 1997 are well comparable.

---

1 Prepared by Eric Schulte Nordholt.

4. The possibilities of compiling statistics on earning patterns in the Dutch economy have been explored on different ways in the past. Information on the structure of earnings used to be collected from companies by Statistics Netherlands every few years. However, payroll administration data on educational levels and other background characteristics of employees are fairly inexact. The quality of these variables in the LFS is much higher. Moreover, the response burden on the companies was large. To obtain more reliable data on earning patterns, reduce the response burden and be able to increase the frequency of these statistics, Statistics Netherlands decided to explore the possibilities of obtaining the information needed by matching the records of three main source statistics mentioned above. This decision would not have been possible without the enlargement of the number of records in the Annual Survey on Employment and Earnings and the availability of the Insured Persons Register on the Statistics Netherlands' premises.

5. The methodology of the 1995 SES in general and the imputation procedure used in particular are described in Schulte Nordholt (1998a). This paper describes the 1997 SES and compares some results with those of the 1996 SES. In section 2 the matching procedure is discussed in more detail. The problem of missing values can either be solved by imputing or by weighting methods. In matching records their number proved to be too large to use the random hot deck method. Therefore a sequential hot deck method was opted for that is discussed in section 3. In that section the weighting procedure is also discussed briefly. Section 4 contains the key results and in section 5 the conclusions are drawn.

II. Matching procedure

6. As the purpose of the SES is to give a description of the earnings structure, only some selected variables from the three sources are considered relevant, although more variables are added for matching, imputation and weighting procedures. Data on 1995, 1996 and 1997 were available from the Annual Survey on Employment and Earnings and the Insured Persons Register for the 1995, 1996 and 1997 SES. To obtain reliable data from the Labour Force Survey, for every SES data on three consecutive years were combined: the year before the year of the survey, the year of the survey and the year following that year. This may have affected some variable scores, but working with cumulative data on three consecutive years was seen as an acceptable compromise between merely relying on data of the survey year or opting for data for more than three years. For reasons of expediency, the few survey records with a missing score on one of the relevant variables were dropped. The remaining records were raised to the population totals in the weighting process.

7. Only exact matches are allowed. However, this requirement does not prevent mismatches: i.e. records, which refer to the same statistical unit according to the sources, but not in real life. Missed matches may also occur: typing errors in postal codes often cause matches, which should turn up to be missed. Analysis showed that mismatches and missed matches notwithstanding, an extensive set of matched records could be obtained from the three sources. This set was reliable enough that no recourse had to be taken to synthetic matching procedures. The payroll data from the business survey were taken as the general framework. Also included were some of the records from the Insured Persons Register that belong to the target population of the Annual Survey on Employment and Earnings, but were not available from that source due to the sample design or because of non-response. Such Insured Persons Register records that matched survey data were included in the data set.

8. The resulting micro data set consisted of a compilation of five subsets that contained the records of matched sources. Table 1 shows these five subsets and the number of records they contain for 1997, and also which groups of variables were chosen from which source. For 1995 and 1996 tables with somewhat smaller numbers of records resulted.

Table 1. Outline of the record sets in the 1997 SES.

Subset	Source			Variables from payroll and register files	Variables from payroll	Variables from the Labour Force Survey	Number of records
	pay rolls	reg-istra-tions	sur-vey				
1	yes	yes	yes	payroll data	payroll data	hh survey data	38,450
2	yes	yes	no	payroll data	payroll data		1,432,118
3	yes	no	yes	payroll data	payroll data	hh survey data	30,123
4	yes	no	no	payroll data	payroll data		1,102,027
5	no	yes	yes	register data		hh survey data	80,892

III. Imputing and weighting

9. In subset 5 of Table 1 some variables were missing. This problem of missing values was be solved by imputation. Auxiliary variables in this imputation process were the variables available in both the payroll and the registration files.

10. The missing scores on the payroll variables were imputed using some auxiliary variables that were available from both the Annual Survey on Employment and Earnings and the Insured Persons Register. Examples of the variables that have to be imputed are gross wages per month, gross wages for overtime per month and the number of holidays. The auxiliary variables for the imputation are sex, type of employment contract, age, gross wages per day, economic sector and (for 1996 and 1997) firm size. All together, 26 payroll variables were imputed using these six (five for 1995) auxiliary variables. The classification of the auxiliary variables was chosen in such a way that it resulted in homogeneous groups containing approximately the same number of records. As there is a big difference in the scores on the variables that have to be imputed between different economic sectors, this auxiliary variable was categorised in homogeneous groups that do not all contain approximately the same number of records. The auxiliary variables appeared to be of good quality and did not contain missing values themselves.

11. As deterministic imputations distort the distribution of the imputed variable and as distributions of variables are of major concern in the statistics we are aiming at, stochastic imputations are necessary in the imputation process. The question was which stochastic imputation method was best suited. An easy choice would have been a stochastic regression imputation, but this does not always lead to feasible imputed values. Therefore, a hot deck method proved a better alternative. As the number of records was too large to use the random hot deck method, the sequential hot deck method was selected as imputation method. A random element is introduced in this method by sorting the non-imputed data set randomly before the imputation process starts. Help arrays with so-called potential donor values were created for 4,480 (for 1995: 2,240) different combinations of scores on the categorised auxiliary variables.

12. The first time a missing value from a record is found with that combination of scores on the categorised auxiliary variables, the first score of the relevant help array is copied. The second time, the second score of the help array is copied, and so on. If all records of an array have been used once, a second iteration through the help array starts. The record from which the imputed value is copied is called the donor record. Although we encountered 2,602,718 potential donor records for 1997 (subsets 1-4) that could be included in the help arrays, we needed to be careful with the number of categories of the auxiliary variables. If too many of these categories were created, the risk would arise of a record having to be imputed with an empty help array, which is obviously not feasible. If the help array contains a few values but many records

have to be imputed using this array, there is the problem of multiple donors' use that will often lead to underestimation of the variance of the imputed variable. Therefore empty or almost empty help arrays have to be combined with other help arrays. Methodologically this corresponds with the introduction of a priority ordering of the help variables in the random hot deck method. Also, in that case we cannot impute all records using all auxiliary variables categorised in the finest categorisation we have available.

13.   To avoid inconsistencies between related variables as a result of the imputations, record matching was used for the imputation. This means that related variables were imputed simultaneously using the same imputation model. In this way covariances between imputed variables were better preserved, which is important for the analysis. More information about applying imputation methods in official statistics can be found in Schulte Nordholt (1998b).

14.   Having finished the imputation, the problem was obviously to find out how accurate the imputations were. As the real values were not known, this is a difficult problem. Although the performance record of an imputation method in simulation experiments may inspire confidence, it is never certain how accurate it will turn out to be in practice. An important criterion in judging a hot deck imputation is to see whether we encounter the problem of the multiple use of donors. Table 2 presents the distribution of the used donor records to impute subset 5 of Table 1 for 1995, 1996 and 1997 by the number of times that these records were used.

Table 2. Percentage of used donor records by the number of times that these donor records were used for the 1995, 1996 and 1997 SES[a].

Number of times used	Percentage of used donor records for 1995	Percentage of used donor records for 1996	Percentage of used donor records for 1997
1	99,22 %	99,45 %	98,91 %
2	0,59 %	0,46 %	0,99 %
3	0,05 %	0,06 %	0,09 %
4	0,05 %	0,02 %	0,01 %
5	0,03 %	0,00 %	0,00 %
6	0,01 %	0,00 %	-
>=7	0,04 %	0,00 %	-
Total	100,00 %	100,00 %	100,00 %

[a]:   because of rounding effects sums of rounded values are not necessarily equal to the total.

15.   From Table 2 we see that some donors are used more than once, but that most used donor records are only used once. The maximum number of times a donor record was used in the imputation process is 16 for 1995, 7 for 1996 and 5 for 1997. Therefore we conclude that the multiple donor problem does not play a major role in the current imputation process and this will increase our confidence in the applied strategy.

16.   Subsets 2 and 4 of Table 1 do not have values for LFS variables. This problem was tackled in a weighting procedure, as imputation is undesirable here for a number of reasons. In the first place not much auxiliary information is available to base a large-scale imputation upon. Secondly, a mass imputation would give dramatic results, if not only the imputed variable is analysed but also the crossing of the imputed variable by other variables not taken into account in the imputation process. Analysing the structure of the data set is the main aim of the SES and therefore it is recommendable to limit ourselves to the weighted and imputed data of the subsets 1, 3 and 5. The imputed data set is raised to the population totals of the Annual Survey on Employment and Earnings in the weighting process.

17.   Two sets of weights have been produced: one set for all employees in the data set and one set for the employee data from the Annual Survey on Employment and Earnings only (subsets 1 and 3 of Table 1). This last set of weights was produced especially for users that did not want to use imputed data. Based on both sets of weights population estimates can be derived that are to some extent consistent. The resulting estimates are the same for those tables that are used in the weighting procedure. If tables are produced that make use of other variables some differences between the tables based on the two different sets of weights may arise. Also if variables are published in more detail such differences may arise. Moreover, users of the small data set do not only have to take into account that the mean weight is smaller, but also that the variance of the weights is much larger than the variance of the weights of the set for all employees. This is a result of the selectivity of the different subsets in Table 1.

IV.   Key results

18.   In this section the results can only be described briefly. In Boerdam, Loeve and Ruijs (1998) the results of the 1995 SES are described extensively and the 1997 results are presented in Schulte Nordholt and Ruijs (2000). In December 1997 employees in the Netherlands had in all 6.1 million jobs. The gross mean hourly wage was 31 Dutch guilders (14 Euros). This gross mean hourly wage is calculated as follows. The gross monthly wage (of December) excluding overtime is multiplied by 12 and divided by the yearly hours of work. The yearly hours of work is the contractual weekly hours of work multiplied by the number of weeks per year, minus not-worked hours because of reduction in working hours, (public) holidays and extra leisure for older employees. The yearly hours of work is calculated using information of December. The gross hourly wage of an employee depends on individual characteristics as e.g. level of education and occupation.

IV.1   Wage by level of education

19.   The survey results of 1997 show, among other things, that hourly wages increase as the level of education increases. Employees with primary education earn 24 Dutch guilders (11 Euros) an hour and employees with university education earn 49 Dutch guilders (22 Euros) an hour. In Figure 1 the hourly wage by level of education is shown graphically. Table 3 shows the distribution of the mean hourly wage by level of education and age. In this table the wages are rounded to tens.

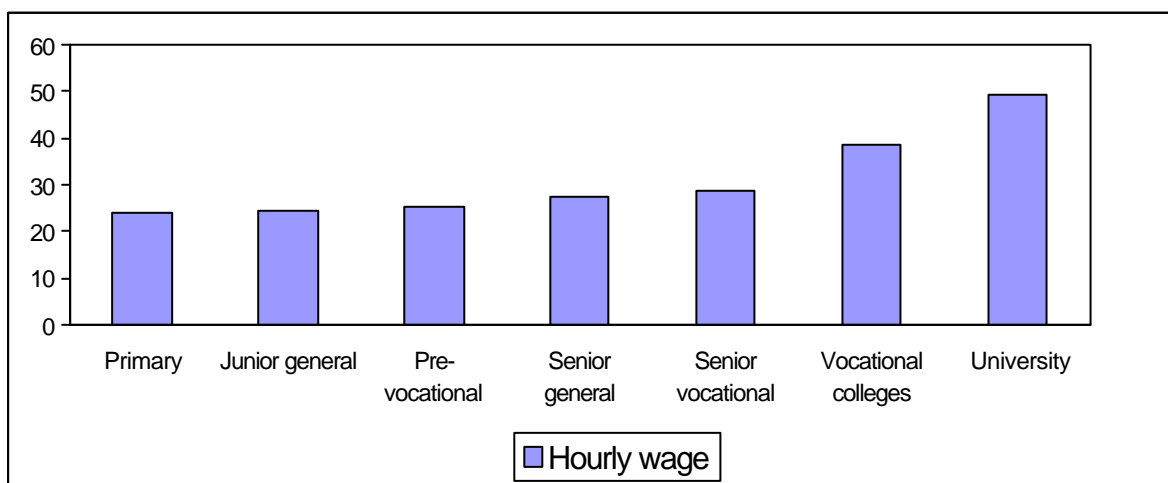Figure 1. Hourly wage of employees by level of education, 1997 SES

Table 3. Hourly wage of employees by level of education and age, 1997 SES

Age	Primary education	Junior general secondary education	Pre-vocational secondary education	Senior general secondary education	Senior vocational secondary education	Vocational colleges	University education	Total
24 years	14.30	13.60	15.80	14.60	17.90	20.10	21.20	16.50
25 – 34 years	23.10	23.90	24.60	26.00	26.00	30.30	35.10	27.10
35 – 44 years	25.00	28.50	27.50	34.40	32.00	40.70	50.80	34.10
45 – 54 years	27.20	31.60	28.80	43.60	35.10	47.50	65.10	38.10
55 – 64 years	27.30	31.60	29.80	48.00	36.00	53.60	72.70	40.10
Total	24.20	24.30	25.10	27.50	28.90	38.50	49.40	30.80

20.   Also age influences the wage level of employees as we can see from Table 3. In the age category 55 - 64 years they earn on average 2.4 times as much as employees under 25 years. Wages of employees with primary education vary less among age categories than wages of employees with university education. For employees with primary education those in the age category 55 - 64 earn 1.9 times as much as those under 25 years. For employees with university education those in the oldest age category earn 3.4 times as much as those in the youngest age category.

21.   Men have a higher mean hourly wage than women. Furthermore, the wage differences between men and women increase with age and level of education. In Table 4 the mean hourly wage of female employees by level of education is shown. The data are expressed in percentages of the equivalent hourly wage for men. Remarkable is that for employees with vocational colleges or university education the wage difference between males and females is larger than for those with lower levels of education.

Table 4. Hourly wage of female employees in percentage of equivalent male hourly wage by level of education, 1997 SES

Primary education	Junior general secondary education	Pre-vocational secondary education	Senior general secondary education	Senior vocational secondary education	Vocational colleges	University education	Total
77.2	78.2	76.1	73.9	77.4	72.3	73.4	76.4

IV.2   Wage by level of occupation

22.   In Figure 2 we see that hourly wages increase also as the level of occupation increases. Wages of employees in elementary occupations vary less among age categories than wages of employees in academic occupations. From Table 5 we can learn the following. For the elementary occupations employees in the age category 55 - 64 years earn on average 1.7 times as much as employees under 25 years, whereas for the academic occupations employees in the age category 55 - 64 years earn on average 3.3 times as much as employees under 25 years.

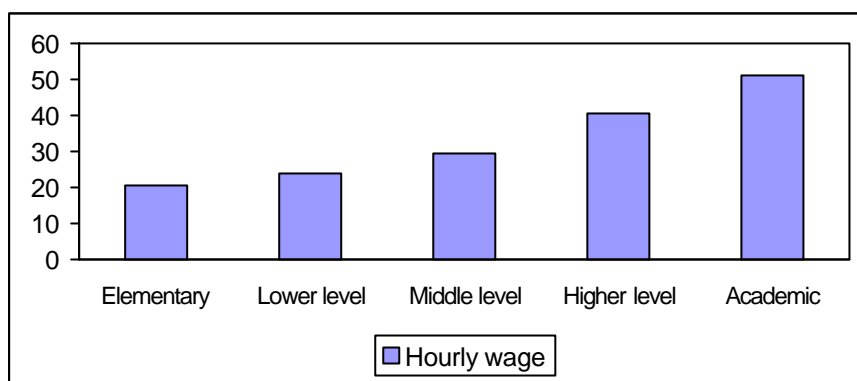Figure 2. Hourly wage of employees by level of occupation, 1997 SES

Table 5. Hourly wage of employees by occupation and age, 1997 SES

Age	Elementary occupation	Low level occupations	Middle level occupations	High level occupations	Academic occupation	Total
24 years	13.70	15.70	18.00	20.80	20.80	16.50
25 – 34 years	22.10	24.20	26.60	31.50	35.60	27.10
35 – 44 years	23.10	26.70	32.40	41.80	50.20	34.10
45 – 54 years	24.00	27.50	34.90	47.60	62.60	38.10
55 – 64 years	23.70	27.30	36.20	53.10	68.40	40.10
Total	20.50	23.80	29.40	40.20	51.20	30.80

23.   Occupations that require more knowledge and experience are better paid than occupations that do not require specific skills. In Table 6 we see that employees working in low level occupations in home economics and service trades have the lowest hourly wage and employees working in academic level occupations in management positions the highest hourly wage. Employees in transport, traffic and communications have a higher mean than the grand mean for low, middle and high level occupations. For middle level occupations the lowest wage is earned by teachers and staff in education and the highest in juridical, public administration, law enforcement and security. The lowest wages for high level occupations are earned in medical and paramedical occupations and for academic level occupations in mathematics and natural sciences. Combinations of occupation level and major skill specialisation that do not contain any occupations are denoted by '-' and combinations that are taken together with others as a consequence of insufficient cell numbers are denoted by 'O'.

Table 6. Hourly wage of employees by occupation and major skill specialisation, 1997 SES

Major skill specialisation	Elementary occupation	Low level occupations	Middle level occupations	High level occupations	Academic occupation
General	-	27.20		-	-
Teachers and staff in education	-	25.80	25.10	36.90	48.90
Agricultural	-	21.90	27.10	37.80	48.00
Mathematics and natural sciences	-	25.80	31.20	34.50	44.10
Technical	-	25.70	30.00	41.30	46.80
Transport, traffic and communications	-	26.40	33.20	54.30	
Medical and paramedical	-	20.70	27.50	32.10	51.80
Economics, clerical and commercial	-	21.40	29.50	41.50	48.80
Juridical, public administration, law enforcement and security	-	26.90	35.40	46.30	47.30
Language and culture	-	-	28.50	34.20	
Social behaviour and society	-	-	28.00	37.80	46.30
Home economics and service trades	-	19.90	25.50	36.20	
Management	-	-		61.20	63.60
Total[a]	20.50	23.80	29.40	40.20	51.20

[a]: including occupations without further major skill differentiation.

IV.3   Changes from 1996 to 1997

24.   The gross hourly wage rose 4.1 percent from December 1996 to December 1997. Figure 3 depicts the large differences in the changes from 1996 to 1997 by level of occupation. In this period employees at the academic level of occupation had the highest increases in wages: 6.2 percent. The mean wage increase at the academic occupation level of those in the age category 25 - 35 years was even 6.8 percent. In Table 7 all changes between 1996 and 1997 are shown by level of occupation and age category. In every age category the mean rise is lower than the grand mean of 4.1 percent. This is due to the ageing effect: the percentage of older employees is rising and this effects the mean rise positively. Finally, we see in Table 7 that young and old employees get a similar wage increase. If these changes are stable over time the differences in wages by age will remain in the future.

Figure 3. Percentage change of the hourly wage of employees by level of occupation, 1996-1997 SES
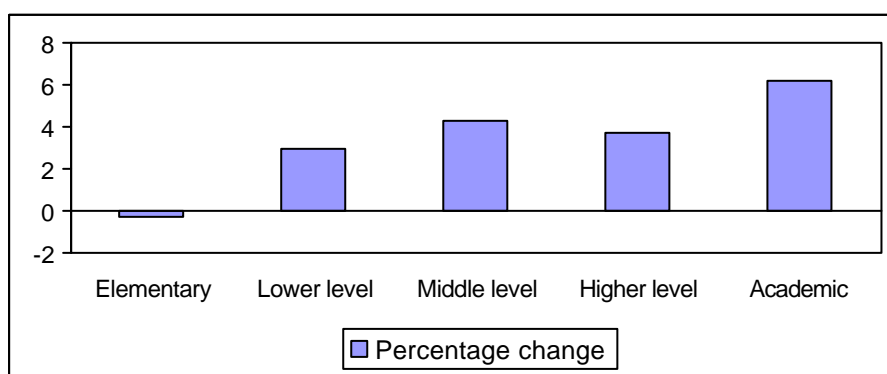
Table 7. Percentage change of the hourly wage of employees by occupation and age, 1996-1997 SES

Age	Elementary occupation	Low level occupations	Middle level occupations	High level occupations	Academic occupation	Total
24 years	1.6	1.6	5.2	5.5	5.6	3.1
25 – 34 years	1.8	2.7	3.6	3.5	6.8	3.8
35 – 44 years	-1.0	2.5	3.7	4.5	4.6	3.4
45 – 54 years	-1.1	3.1	3.0	2.9	5.1	3.3
55 – 64 years	-4.2	4.5	3.2	4.6	2.8	3.8
Total	-0.3	3.0	4.3	3.7	6.2	4.1

V.   Conclusions

25.   Applying state of the art statistical methods enables Statistics Netherlands to get census like information on the structure of earnings on a yearly basis without having to set up a separate and elaborate survey. Level of education and occupation pays: better educated employees and employees in higher levels of occupation have on average higher wages. At the same time remarkable differences between different jobs at the same occupation level can be found. Also age influences hourly wages: older employees get better paid than their younger colleagues. We found that from 1996 to 1997 employees in academic occupations in the age category 25 - 34 years had the largest increase in their mean hourly wage.

References

Boerdam, A.A., J.A. Loeve and G.P.C.M. Ruijs, 1998. Wages by level of education and occupation: the Structure of Earnings Survey 1995. In: Sociaal-economische maandstatistiek, Volume 15, March 1998, pp. 31-57. [in Dutch]

Schulte Nordholt, E., 1998a. Imputation, the alternative for surveying earning patterns. In: Netherlands Official Statistics, Volume 13, spring 1998, pp. 14-15.

Schulte Nordholt, E., 1998b. Imputation: methods, simulation experiments and practical examples. In: International Statistical Review, Volume 66, Nr. 2, pp. 157-180.

Schulte Nordholt, E. and G.P.C.M. Ruijs, 2000. Wages by level of education and occupation: the Structure of Earnings Survey 1997. To appear in: Sociaal-economische maandstatistiek, Volume 17, April 2000, pp…-... [in Dutch]