Combining information from different sources to estimate the annual working hours of part-time workers

Sandra Barragán, National Statistics Institute (Spain), sandra.barragan.andres@ine.es

Elisa Martín, National Statistics Institute (Spain), elisa.martin.hernandez@ine.es

Abstract

Structural statistics on earnings provide comparable information on relationships between the level of earnings, individual characteristics of employees (sex, age, occupation, length of service) and their employer (economic activity, size of the enterprise). In Spain there is a main survey carried on 4-year periodicity. Additionally, due to administrative records, most of the information can be given annually as well. However, some problems arise when different administrative records are used together and joined with survey results to obtain the final dataset. Administrative data were originally collected for a definite non-statistical purpose that might affect the quality of the data. In addition, there is a coherence issue when more than one source of information are integrated to obtain data for the same variable.

Coherence is an essential part of official statistics so it is receiving such a growing attention to ensure quality in the data. Statistical offices focus on the importance of obtaining coherent results specially when they come from a process of data integration. Along this process, we deal with the inconsistency of the individual values by taking into consideration the additional uncertainty due to the difference between the administrative concept and the statistical variable. Therefore, we present the methodology used in the last publication of the Annual Structure of Earnings Survey to obtain the annual working hours of part-time workers. We have combined three sources of information in order to obtain a trustworthy value of the working hours for each individual. We have developed an algorithm with decision rules built by using expertise on the field. This algorithm is an initial solution to maintain coherence at microdata level in the combination of different sources for the same variable.

Keywords: Data integration, Administrative records, Combining information, labour market, working hours

1. Introduction

1.1. Quality issues

The Annual Structure of Earnings Survey (ASES) arose for the purpose of filling the information gap on earnings between each 4-year Structure of Earnings Survey (SES). Most of information required by the users is obtained by merging administrative records with a small survey. Thus, the respondent burden is reduced significatively and frequency and timeliness are improved.

Nonetheless, administrative records have their own issues in terms of reliability and coherence. These sources are collected for a definite non-statistical purpose (administering taxes, benefits or services) that might affect the treatment of the source unit. Sometimes the administrative definition does not match with the statistical definition. In fact, the same information can be collected in different data sets with their own specific definitions.

Coherence is an essential part to ensure the quality of the data so that it is receiving such a growing attention over the last decade in statistical offices, in particular when data come from the integration of different data sources. To guarantee coherence is the motivation of this work. There is a need of measuring the statistical variable of the annual working hours but there is neither register nor survey information where it can be obtained at individual level for part-time workers, from now on *employees* as self-employed workers are not taken into consideration.

The ASES is the result of various sources: administrative records from different institutions and a survey. The annual working hours are essential to obtain the salary per hour that is used to compare employees with different workday. Although there is no availability of the statistical variable of interest, some other variables have strongly related information. There are three sources: agreed hours, contractual hours and contributed hours. The first is obtained from another survey, the Quarterly Labour Cost Survey (QLCS). The rest of the sources are registers of Social Security records. Once the available information is integrated, there are substantial differences among the values of the three variables for the same individual. Moreover, there is no pattern which leads to a direct decision about which source could be the most reliable. It depends on the individual.

Summarizing, there are two main problems of coherence: 1) conceptual differences between the available variables and the statistical variable and 2) differences among the available values for the same individual. Both problems can be faced in the microintegration process of the different sources. Then, the context of microintegration is shown in the following subsection.

1.2. Microintegration background

"Micro-integration is the method that aims at improving the data quality in combined sources by searching and correcting for the errors on unit level", (Bakker, 2011). The general idea is to combine different processes such as data editing, derivation of statistical variables and imputation in order to correct data on the unit level to produce consistent files.

Similar questions have appeared in the field of data fusion, where various data sets with different units are merged to generate synthetic data sets (Wald, 1999, Castanedo, 2013). Data fusion searches for a consensus which summarize all the information under the assumption that all the sources are trustworthy. In the case of this work, reliability cannot be confirmed for all the sources and individuals. For this reason, it is not possible to use directly the methods developed for data fusion. However, some ideas were the origin to develop the methods presented in this paper. In the engineering field some solutions have been developed regarding information management (Hall and Llinas, 1997, Bloch, 1996, Esteban et al., 2005). There are also procedures to find a consensus among candidates or experts (Macdonald and Ounis, 2006). In cognitive psychology the combination of data fusion with the lack of reliability in some sources has leads to the emergence of information integration theory (Anderson, 1971, 2016, Birnbaum et al., 1976, Mullet and Wolf, 2016).

In official statistics, this problem appears in particular when administrative registers are integrated with information from survey questionares (Laitila et al., 2011, Zhang, 2012, Schnell, 2013). The particular case of *multi-valued variables* has been studied (Wallgren and Wallgren, 2007) with the assumption that a single individual can have acceptable different values for the same variable, for example the occupation. Apart from that, there is a project of the European Statistical System about data integration (ESSnet, 2011). Regarding integration at micro level, there are some clear guidelines for record linkage and statistical matching problems, when the problem is to find the individuals in different sources. However, the problem of this work is about harmonisation on a conceptual level, which needs data reconciliation and there is no general algorithms for this purpose (Eurostat, 2018). In terms of consistency, the

consistent repeated weighting procedure obtains estimations by assigning weights to records in a (combined) data file for each table (Bakker, 2011). Nevertheless, neither the conceptual differences nor the lack of agreement among sources would be solved with those procedures.

In this work, a new method is presented to deal with having different information for the same concept, see Section 2. This method has been applied in the ASES of the year 2015 (published in June 2017), the first time that all the different sources for the working hours were available. The results obtained are compared with previous years in Section 3. Finally, in Section 4 the final conclusion of this work is shown, there is an improvement in quality but there is a need of developing general procedures of microintegration.

2. Distance-based method

2.1. General description

Let H_i $\forall i = 1, \ldots, n$ be the statistical variable of interest for n individuals in the survey. The variable H has no possibility of being measured directly. Let x_{i1}, x_{i2}, x_{i3}, $\forall i = 1, \ldots, n$ the measurements of n individuals from three different variables strongly related with H, and ω_{i1}, ω_{i2}, ω_{i3}, $i = 1, \ldots, n$; the corresponding weights which are proportional to the level of confidence in each source for each individual.

The method described below was developed with the aim to solve the coherence issue due to the different values for the same individual. It has been observed that when some individual data is not reliable the value for that source is further from the other two sources than those sources between themselves. Then, the first step is to find and reject the value at a greater distance from the others.

First of all, the two nearest values are chosen,

$$(x_{ir}, x_{is}) = \arg \min_{(x_{ij}, x_{il}) \in \{x_{i1}, x_{i2}, x_{i3}\}} d(x_{ij}, x_{il}) \tag{1}$$

where $d(x_{ij}, x_{il}) = \sqrt{(x_{ij} - x_{il})^2}$ is the Euclidean distance between the values of j and l for the i-th individual. Then, the final estimation of H is calculated for each individual

with those two values as a weighted mean as follows,

$$\hat{H}_i = \frac{\omega_{ir} * x_{ir} + \omega_{is} * x_{is}}{\omega_{ir} + \omega_{is}}. \quad (2)$$

2.2. Application to working hours

The motivation of this work is the estimation of the statistical variable H: the annual working hours. In the ASES the information about the working hours comes from the following three sources.

Agreed hours: x_1. The first variable is obtained from the QLCS. As this survey obtains the information of the agreed hours in each workplace (CCC) for the full-time and part-time employees separately. Then, the first variable contains the average annual working hours of part-time employees.

Contractual hours: x_2. The second variable is obtained by using the part-time coefficient, name as $coef$. This coefficient is set by the employer at the moment of the beginning of the contract and it is registered in the Social Security record. The meaning of $coef$ is the expected proportion of time that a part-time employee is going to do in comparison with a comparable full-time employee. Then, as we have the agreed hours for the full-time employees of the same CCC, name as fth, we can obtaine easily the contractual annual hours for each employee, $x_2 = coef * fth/100$.

Contributed hours: x_3. The third variable is the total number of annual working hours registered in the Social Security record and it is the base for social contributions. They could not correspond exactly to working hours.

Auxiliary information from the ASES is used in order to assign properly the weights, (ω_1, ω_2, ω_3), that represent the level of confidence in each source for each individual. That auxiliary information comes from variables directly related to the salary that are registered in the Social Security records. These variables have information about the workplace: Region (Nuts 2), economic activity (2 digits of NACE Rev.2), size of the workplace; about the employee: contribution group, type of contract, the number of days registered as employee, named as annual contributed days. Four cases (A, B, C D) are considered depending on the availability of x_1, x_2 and x_3.

Case A

Specific decision rules have been created to assign those weights in the case of all the three variables are available: x_{i1}, x_{i2}, x_{i3}.

Initially, we set all weights to $\omega_{ij}^{(0)} = 1$, $\forall i = 1, \ldots, n$, $\forall j = 1, 2, 3$. Then, these values are updated per individual depending on the interpretation of each available variable and the auxiliary information. The general idea is to evaluate how reliable is the value for the hours in each individual compare with other similar individuals. This similarity is given by the group they belong, such as their CCC or the same values of some auxiliary variables. We have conducted different analyses to obtain the set of variables that yields more homogeneous groups. The final decision has been to use the workplace (CCC) for x_1 and the intersection of contribution group (CG) and size of workplace (SC), CGxSC for x_2 and x_3.

Under the circumstances of x_1, there is a relation with the contribution bases to social security, named as cb_i, so that the decision rule is: $\omega_{i1}^{(1)} = \omega_{i1}^{(0)} - 0.5$ if $cb_i \notin [Q1_{1g}, Q1_{3g}]$, where $Q1_{1g}$ and $Q1_{3g}$ are the first and the third quartile of contribution bases (cb) in the g-th CCC to which the i-th employee belongs.

Under the circumstances of x_2, cb_i is also used but in combination with x_2. Let $bh2_i = cb_i/x_{i2}$ be the contribution bases per contractual hour and the decision rule to reassign the weights is: $\omega_{i2}^{(1)} = \omega_{i2}^{(0)} - 0.5$ if $bh2_i \notin [Q2_{1k}, Q2_{3k}]$, where $Q2_{1k}$ and $Q2_{3k}$ are the first and the third quartile of $bh2$ in the k-th group made by CGxSC to which the i-th employee belongs.

Under the circumstances of x_3, cb_i is used again, $bh3_i = cb_i/x_{i3}$ is the contribution bases per contributed hour and the decision rule to reassign the weights is: $\omega_{i3}^{(1)} = \omega_{i3}^{(0)} - 0.5$ if $bh3_i \notin [Q3_{1k}, Q3_{3k}]$, where $Q3_{1k}$ and $Q3_{3k}$ are the first and the third quartile of $bh3$ in the k-th group made by CGxSC to which the i-th employee belongs.

The variable x_3 is specially problematic due to its vague meaning for the employer who registers the values. It has been observed that sometimes they set x_3 as the extra hours in addition to the contractual hours. The annual contributed days, named as d_i $\forall i = 1, \ldots, n$, give information about the contributed hours. Then, the following decision rules are added: $\omega_{i3}^{(2)} = \omega_{i3}^{(1)} - 0.15$ if $d_i > 0$ or $\omega_{i3}^{(2)} = \omega_{i3}^{(1)} + 0.15$ if $d_i = 0$. The last rule could cause a final weight greater than 1, however this is not a problem to obtain the estimation of the working hours with the definition in (2).

Case B

In the cases where not all the three variables are available, instead of using the distance-based method, the decision rules give directly the estimation of the annual working hours.

Under the assumption of having x_1 and x_2 but not x_3, we have three possibilities.

1) $\hat{H}_i = (x_{i1} + x_{i2})/2$ in case of: 1.a) $bh2_i \notin [Q2_{1k}, Q2_{3k}]$ and $cb_i \notin [Q1_{1g}, Q1_{3g}]$; or 1.b) $bh2_i \in [Q2_{1k}, Q2_{3k}]$ and $cb_i \in [Q1_{1g}, Q1_{3g}]$.

2) $\hat{H}_i = x_{i2}$ if $bh2_i \in [Q2_{1k}, Q2_{3k}]$ but $cb_i \notin [Q1_{1g}, Q1_{3g}]$.

3) $\hat{H}_i = (x_{i1} * G_i)/\overline{G}_{CCC}$ if $cb_i \in [Q1_{1g}, Q1_{3g}]$ and $bh2_i \notin [Q2_{1k}, Q2_{3k}]$, where G_i is the annual salary of the i-th employee and $\overline{G}_{CCC}$ is the mean salary in the corresponding CCC. Since x_1 is an average of hours from different employees, it is adjusted by using the individual salary relative to the mean salary in that CCC.

Case C

Under the assumption of having x_1 and x_3 but not x_2, we have three possibilities.

1) $\hat{H}_i = (x_{i1} + x_{i3})/2$ in case of: 1.a) $bh3_i \notin [Q3_{1k}, Q3_{3k}]$ and $cb_i \notin [Q1_{1g}, Q1_{3g}]$; or 1.b) $bh3_i \in [Q3_{1k}, Q3_{3k}]$ and $cb_i \in [Q1_{1g}, Q1_{3g}]$.

2) $\hat{H}_i = x_{i3}$ if $bh3_i \in [Q3_{1k}, Q3_{3k}]$ but $cb_i \notin [Q1_{1g}, Q1_{3g}]$.

3) $\hat{H}_i = (x_{i1} * G_i)/\overline{G}_{CCC}$ where G_i is the annual salary of the i-th employee and $\overline{G}_{CCC}$ is the mean salary in the corresponding CCC. This is used if $cb_i \in [Q1_{1g}, Q1_{3g}]$ and $bh3_i \notin [Q3_{1k}, Q3_{3k}]$.

Case D

There are some individuals with just the value of x_1, then $\hat{H}_i = (x_{i1} * G_i)/\overline{G}_{CCC}$ is the final estimation where G_i is the annual salary of i-th employee and $\overline{G}_{CCC}$ is the mean salary in the corresponding CCC.

2.3. Computational issues

The implementation of the distance-based method and its corresponding decision rules are put together in an algorithm developed in `R` language. Athough this algorithm requires to look over all individuals and there are some loops, it is efficient for medium size data.

3. Results and comparisons with previous years

All this administrative information related to the working hours was available, for the first time in Spain, for the year 2015. In that year, the total amount of employees in the sample was 228.560 with n=39.942 part-time employees of which 87% had values in the three variables x_1, x_2 and x_3 (Case A). In this section, some results of the estimation of the annual working hours are shown. The source for all the figures and tables of this section is the Structure of Earnings Surveys: ASES and SES (INE, Spain).

First, a brief analysis of the available variables is shown in the boxplots of the Figure 1 and Figure 2 for all part-time employees in the samples of 2015 and 2016.

Figure 1. Boxplots for the data of 2015

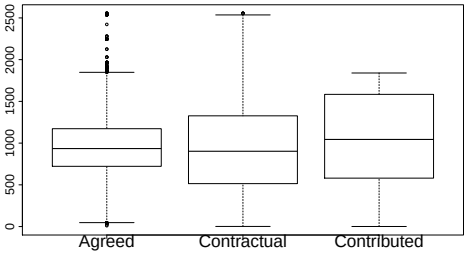
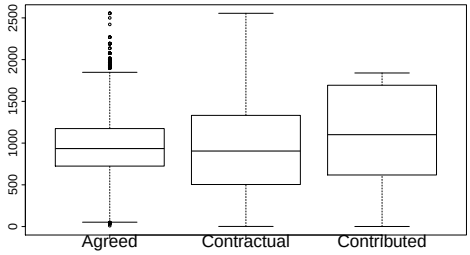


Figure 2. Boxplots for the data of 2016



Second, the estimated annual working hours are compared to the values of previous years. In 2013, there was information only about the agreed hours on average in the CCC, x_1. In 2014, reference period of the 4-year survey, the working hours were obtained directly from the questionnaire. For 2015 and 2016, the annual working hours are estimated by using the distance-based method presented in Section 2.

A brief numeric comparison is shown in Table 1 where it can be observed that the deciles of 2015 and 2016 are much more similar to 2014 than the deciles of 2013.

Table1. Deciles of the annual working hours

	0%	1%	10%	25%	50%	75%	90%	99%	100%
2013	21.00	189.00	431.00	701.00	917.00	1154.00	1353.00	1698.36	1840.00
2014	9.00	89.00	276.00	598.00	920.00	1338.00	1618.00	1752.00	1840.00
2015	2.29	91.03	289.22	592.56	950.26	1287.24	1537.25	1840.00	1840.00
2016	4.00	91.19	297.00	605.00	954.59	1305.06	1561.00	1840.00	1840.00

In the graphical comparison among years and methods, it is observed how the distribution of hours in 2013 (Figure 3) is similar to a Normal distribution. Note that x_1 is an average of the hours in each CCC. In the cases of 2015 (Figure 5) and 2016 (Figure 6), it is clear how some peaks appear in similar values for both years which, in principle, makes them more similar to the profile of the observations in 2014 (Figure 4). Therefore, it can be said that the new estimation is an improvement in quality since the values have a distribution more similar to the working hours from the survey.

Figure 3. Annual working hours for part-time employees in 2013 (n=37933)

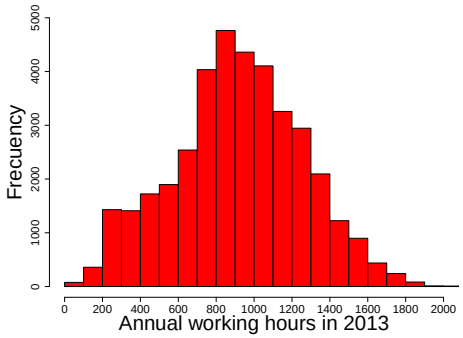


Figure 4. Annual working hours for part-time employees in 2014 (n=36919)

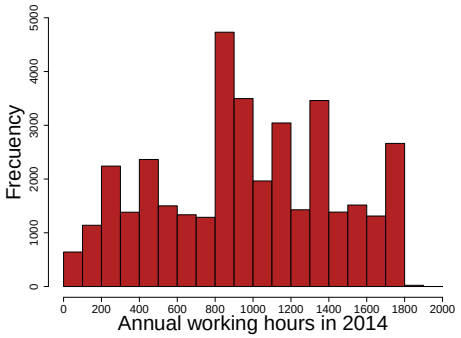


Figure 5. Annual working hours for part-time employees in 2015 (n=39942)

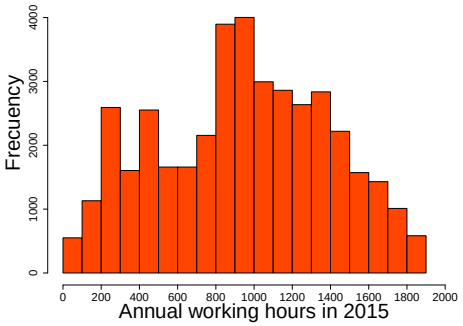
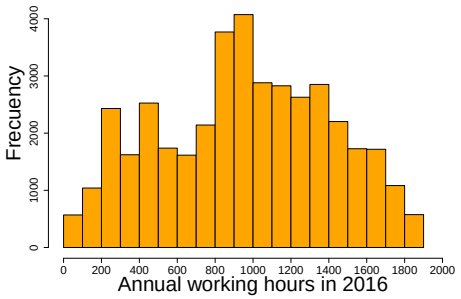


Figure 6. Annual working hours for part-time employees in 2016 (n=40020)



4. Conclusion and final remarks

+ There is a considerable improvement in the quality of the values for the annual working hours although the weights are assigned with ad-hoc decision rules completely thought to this specific data.

+ In the comparison with previous years, a good approximation to the values obtained directly from the 4-year survey is observed.

+ The respondent burden is reduced while the quality of the data is kept.

- Notice that this method is neither easily generalized for more than three sources of information, nor to solve problems of similar form but in a different field and meaning.

5. References

Anderson, N. H. (1971). Integration theory and attitude change, *Psychological review* **78**(3): 171.

Anderson, N. H. (2016). Information integration theory unified psychology based on three mathematical laws, *Universitas Psychologica* **15**(3).

Bakker, B. F. (2011). Micro-integration: State of the art, *Report on WP1 State of the art on statistical methodologies for data integration* . ESSnet on Data Integration.

Birnbaum, M. H., Wong, R. and Wong, L. K. (1976). Combining information from sources that vary in credibility, *Memory & Cognition* **4**(3): 330–336.

Bloch, I. (1996). Information combination operators for data fusion: A comparative review with classification, *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* **26**(1): 52–67.

Castanedo, F. (2013). A review of data fusion techniques, *The Scientific World Journal*

ESSnet (2011). *Data integration*, European Statistical System Network.

Esteban, J., Starr, A., Willetts, R., Hannah, P. and Bryanston-Cross, P. (2005). A review of data fusion models and architectures: towards engineering guidelines, *Neural Computing & Applications* **14**(4): 273–281.

Eurostat (2018). *Handbook on Methodology of Modern Business Statistics*, Memobust.

Hall, D. L. and Llinas, J. (1997). An introduction to multisensor data fusion, *Proceedings of the IEEE* **85**(1): 6–23.

Laitila, T., Wallgren, A. and Wallgren, B. (2011). Quality assessment of administrative data, *Statistiska centralbyrån* .

Macdonald, C. and Ounis, I. (2006). Voting for candidates: adapting data fusion techniques for an expert search task, *Proceedings of the 15th ACM international conference on Information and knowledge management*, ACM, pp. 387–396.

Mullet, E. and Wolf, Y. (2016). New frontiers in information integration theory, *Universitas Psychologica* **15**(3).

Schnell, R. (2013). Linking surveys and administrative data, *Improving Survey Methods: Lessons from Recent Research* pp. 273–287.

Wald, L. (1999). Some terms of reference in data fusion, *IEEE Transactions on geoscience and remote sensing* **37**(3): 1190–1193.

Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*, Vol. 553, John Wiley & Sons.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration, *Statistica Neerlandica* **66**(1): 41–63.