DACSEIS
IST-2000-26057

Workpackage 7

Questionnaire on the use of register data for Labour Force Surveys

Deliverable 7.1

2002

List of contributors:

Paul Knottnerus, Statistics Netherlands;
Rolf Wiegert, University of Tübingen.

Main responsibility:

Paul Knottnerus, Statistics Netherlands;
Rolf Wiegert, University of Tübingen.

IST–2000–26057–DACSEIS

The DACSEIS research project is financially supported within the IST programme of the European Commission. Research activities take place in close collaboration with Eurostat.

`http://europa.eu.int/comm/eurostat/research/`

`http://www.cordis.lu/ist/`

`http://www.dacseis.de/`

Preface

In order to get insight into the use of registers by the NSI's in Europe, one of the activities in working package WP7 of the DACSEIS project was to make a questionnaire on the use of auxiliary information for Labour Force Surveys. Based on the questionnaires and some additional sources this report gives a summary of the various manners in which the NSI's are using the auxiliary information from the available registers.

Paul Knottnerus — Statistics Netherlands
Rolf Wiegert — University of Tübingen

Contents

List of Tables

Chapter 1

Introduction

In helping to solve the conflicting requirements of more socio-economic information on the one hand and a reduction of the response burden on the other hand, population registers and other registers on persons and households are indispensable. Registers permit complex sampling techniques and more accurate socio-economic information on the survey population at lower costs while producing satisfactory quality. Additionally, registers may help to realise a higher degree of harmonisation between the various national statistical systems, provided that the registers are based on the same international standards – until now not yet realised. In order to get more insight into the state of the art with respect to the use of registers, DACSEIS has sent a questionnaire to the distinct NSI's.

After we had made a first draft of the questionnaire on the use of registers for Labour Force Surveys (LFS) in Europe, our attention was drawn to a publication of EUROSTAT (2001), we were not aware of at the time. The publication was entitled "Labour Force Survey: Results 2000". One of the consequences of the additional information in this publication was that we could drop some questions in the first draft, in particular on the weighting procedures used by the various countries. For the final draft of the questionnaire see the Annex. For further details on some of the Labour Force surveys of persons and households see the deliverables of WP1 of DACSEIS.

Chapter 2 discusses a number of methods for using auxiliary information from registers. Based on the questionnaire and the publication mentioned before, Chapter 3 summarises the different ways in which registers are used in the European countries.

Chapter 2

The use of registers

This chapter briefly describes some estimators based on auxiliary information from registers. For the sake of convenience we only consider simple random samples without replacement (SRS). For an extensive treatment of unequal probability sampling see Särndal *et al.* (1992). Since the DACSEIS project is mainly concerned with variance estimation, we will pay only little attention to imputations based on registers.

2.1 (Post)stratification

A classical method for using population totals or means is the stratified estimator. This estimator, denoted by $\widehat{\tau}_{Y,ST}$, is written in the classical way as

$$\widehat{\tau}_{Y,\mathrm{ST}} = \sum_{h=1}^{H} N_h \overline{y}_{h,s} \qquad \left(\text{note: } \overline{y}_{h,s} \equiv \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \quad \text{and} \quad \overline{y}_h \equiv \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi} \equiv \frac{\tau_h}{N_h} \right)$$

$$\mathrm{var}(\widehat{\tau}_{Y,\mathrm{ST}}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h} \qquad \left(\sigma_h^2 \equiv \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \overline{y}_h)^2 \right) .$$

From the variance formula it emerges that compared to the SRS estimator a substantial variance reduction can be realised when the strata are homogeneous. In the extreme case of perfect homogeneity with $\sigma_h^2 = 0$ the stratified estimator has zero variance. When the n_h are random one calls this estimator the poststratification estimator. This occurs in practice after one has drawn, for example, an SRS sample from the whole population. Usually in such a situation one neglects the random character of the n_h leading to the so-called conditional variance; see Holt and Smith (1979). Applying prior stratification one needs a sampling frame for each stratum separately rather than for the population. From the formula it follows straightforwardly that the weights for each sample observation of the hth stratum are equal to

$$w_{hi} = \frac{N_h}{n_h} \qquad (h = 1, ..., H;\ i = 1, ..., n_h).$$

2.2 The regression estimator

Another way to use prior knowledge of population totals or means is the regression estimator, denoted by $\widehat{\tau}_{Y,REG}$. It is defined by

$$\widehat{\tau}_{Y,REG} = N\overline{y}_s + \widehat{\beta}^{\,T}(\tau_x - N\overline{x}_s) \qquad \left(\widehat{\beta} = \left[\sum_{i=1}^{n} x_i x_i^T\right]^{-1} \sum_{i=1}^{n} x_i y_i\right)$$

$$\mathrm{var}(\widehat{\tau}_{Y,REG}) \approx N^2\left(1 - \tfrac{n}{N}\right)\tfrac{\sigma_y^2}{n}(1-\rho^2)$$

$$\rho^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \qquad \left(e_i = y_i - \beta^T x_i;\ \overline{y} = \tau_Y/N\right);$$

see COCHRAN (1977), p. 194. The vector x_i is a k-vector of auxiliary variables, including the intercept ($i = 1, \ldots, n$). The sample mean of the x_i is denoted by $\overline{x}_s$. The quantity ρ^2 is the so-called squared multiple correlation coefficient from a regression of y on the auxiliary variables stacked in the x-vector. By substituting the formula for $\widehat{\beta}$, we can write the regression estimator in an alternative manner as a weighted mean

$$\widehat{\tau}_{Y,REG} = \sum_{i=1}^{n} w_i y_i \qquad \left(w_i = \frac{N}{n} + [\tau_x - N\overline{x}_s]^T \left[\sum_{i=1}^{n} x_i x_i^T\right]^{-1} x_i\right).$$

For further details on the regression estimator see, e.g., SÄRNDAL *et al.* (1992). Although it can be shown that the poststratification estimator and the regression estimator are equal when the auxiliary variables can be seen as binary stratum dummies, their variance formulas lead to somewhat different outcomes; see, e.g., KNOTTNERUS (2003), p. 132. From the formulas on the regression estimator it emerges directly that the use of known population totals leads to a substantial reduction of the variance when ρ^2 is high or, equivalently, when the target variable y is highly correlated with the auxiliary variables or the variables used for dividing the population into strata.

2.3 Calibration

DEVILLE and SÄRNDAL (1992) show in their article how a general class of calibration estimators can be written as

$$\widehat{\tau}_{Y,\,CAL} = \sum_{i=1}^{n} w_i y_i.$$

The weights w_i are the solution of the optimisation problem

$$\min \sum_{i=1}^{n} G\left(\frac{n w_i}{N}\right) \times \frac{N}{n} \qquad \text{subject to} \qquad \sum_{i=1}^{n} w_i x_i = \tau_x,$$

where $G(.)$ is a strictly convex, differentiable, and nonnegative function with $G(1) = G'(1) = 0$ and $G''(1) = 1$ while τ_x is a given vector of population totals. By choosing, for instance, $G(u) = (u-1)^2/2$ the calibration method amounts to the regression estimator. When estimating a contingency table with given margins, the choice of $G(u) = u\ ln(u) - u + 1$ leads to the weights of the well-known *iterative proportional fitting* (IPF) method, also called *raking*. Deville and Särndal (1992) give a thorough proof that under mild regularity conditions the variance of the calibration estimator is asymptotically equal to the variance of the regression estimator for all $G(.)$. Additionally, a more intuitive explanation of this maybe somewhat surprising asymptotic result is as follows. First, similarly to $(\tau_x/N - \overline{x}_s)$, the quantity $(w_i n/N - 1)$ is of order $1/\sqrt{n}$ under mild regularity conditions; note that $w_i n/N = 1$ when $\overline{x}_s = \tau_x/N$. Second, under the above regularity conditions a second order Taylor series expansion of an arbitrary distance function $G(u)$ at $u = 1$ gives $G(u) \approx (u-1)^2/2$, i.e., the distance function corresponding to the regression estimator. Therefore, both the calibration estimator and the regression estimator have the same asymptotic variance.

2.4 Intermezzo

It emerges from the formulas mentioned above that registers with detailed information on all sample or population elements are not strictly necessary in order to apply these formulas. Only the population totals are needed, provided that the x-variables observed in the sample are identical with the x-variables in the registers; this should be checked carefully in practice. However when less unambiguous variables are involved in the analysis, matters become less straightforward. Consider, for instance, the labour classification variable. In practice it is difficult to observe this variable in a sample because, in general, persons in the sample hardly know what their labour classification (LC) is according to the NACE classification. When one feels that the actual employment variable (E) is highly correlated with LC, it is worthwhile to combine the register of LC data with the data from the sample for each individual record observed in the sample, provided that it is possible to match the elements in the register and the sample. Assuming that E and LC are highly correlated, the accuracy of the resulting regression estimators may increase substantially. The same remarks apply when registers with lagged data on the employment variable or job-seeking are available. Needless to say that variance reduction is equivalent to cost reduction when one keeps the required precision of the estimates from the sample on the same level as before. Often one will try to realise both variance reduction and cost reduction. In addition, the accuracy and harmonisation of the surveys can be improved when registers are based on the same international definitions of difficult variables such as labour classification.

2.5 Reduction of nonresponse bias

In order to correct for nonresponse bias, one often divides the population into homogeneous strata or groups based on, for instance, sex, age, and region. Assuming that within

each stratum the nonresponse is not selective, one can reduce the nonresponse bias. Matters become more complicated when nonresponse is highly correlated with, for instance, the unobserved education variable while education is highly correlated with the target variable employment. In such cases one can only reduce the nonresponse bias by means of stratification when the education variable is highly correlated with the available variables sex, age, and region.

2.6 Differences between register and sample

It may occur in practice that the variable in the register is more or less different from the variable observed in the survey. This may be due to a difference of the underlying definitions as well as to a poor quality of the data in the register or vice versa, the variable from the survey is not well defined or misunderstood by the sampled person, household, or business. Additionally, the reference periods of the register data and the present survey might be different. All these kinds of differences are to be seen as a warning against a direct use of the population total from the register. Nevertheless, when differences are small or, equivalently, the target variable in the sample and the register variable are highly correlated, the available register variable can still be used for the regression estimator and, subsequently, for weighting the sample observations, provided that the sample and the register can be matched appropriately in order to calculate the sample mean of the corresponding register variable.

However, before using a register in combination with a survey, one has to consider a number of characteristics of the register data:

- the register data are available in a form that a modern statistical use is possible, e.g., an electronic register;
- the definitions of units should be equivalent or at least comparable with respect to temporal, areal and factual criteria;
- levels of classification and aggregation should be mutually consistent;
- there is no limitation in the use of the data with respect to disclosure control, and it is possible to correctly combine the data and their units by means of unique matching keys.

A potential user who wants to combine survey and register data should be prepared that in practice often at least one of these criteria is violated which has to be taken into account with the evaluation of the results. For further details see MÜNNICH and WIEGERT (2002) or MÜNNICH and WIEGERT (2004).

Chapter 3

Results of questionnaire

This chapter summarises the various ways in which the registers are employed by the national statistical offices in Europe. As stated above, the results given in this chapter are based on the Eurostat publication "Labour Force Survey: Results 2000" and the questionnaire, entitled "Questionnaire on the use of auxiliary information (register data) for Labour Force Surveys (LFS) or other surveys containing Labour Force Data"; for the questions see the Annex. Table 3.1 shows the countries to which we have sent the questionnaire.

Table 3.1: Countries and their abbreviations

1. Austria (AT)	7. Greece (GR)	13. Norway (NO)
2. Belgium (BE)	8. Ireland (IE)	14. Portugal (PT)
3. Denmark (DK)	9. Israel (IL)	15. Spain (ES)
4. Finland (FI)	10. Italy (IT)	16. Sweden (SE)
5. France (FR)	11. Luxembourg (LU)	17. Switzerland (CH)
6. Germany (DE)	12. The Netherlands (NL)	18. United Kingdom (UK)

Between parentheses the abbreviations used in the remainder of this report are given.

After 1 rappel there were two nonrespondents which, by assumption, are treated in the same manner as the countries not using registers. We have also sent the questionnaire to the candidate Member States Bulgaria, Czech Republic, Estonia, Lithuania, and Slovenia. However, we didn't receive a reaction.

3.1 Questions 1 and 2

The first two questions of the questionnaire were general questions of which other voluntary or mandatory surveys are related to LFS. Four countries mentioned other surveys:

1. BE has the Household Budget Survey;
2. CH has the Household Income & Consumption Survey;

3. IL has the Income Survey, Family Expenditure Survey and Social Survey;

4. UK has the General Household Survey, the Family Resource Survey, and the Expenditure & Food Survey.

BE, IL and FR answered that all questions of their LFS are mandatory. In DE most questions are mandatory, in AT only some questions are mandatory, whereas in the other countries the LFS is voluntary.

3.2 Countries not using registers

Globally spoken, the countries can be divided into two groups: (i) countries using registers and (ii) countries not using registers. The majority of the countries don't use registers for several reasons. See the following table.

Table 3.2: Reasons for not using registers

Countries	Reason
AT, DE, and LU	law/privacy
CH and LU	matching key problems
ES	no quality improvement
CH, IE, and UK	no (suitable) population register
IL	bias registers and frame errors
IT	no possibilities
LU	no (lagged) LFS variables in register

AT, CH, IL, and LU have plans to use registers in the future.

3.3 Countries using registers

There are several reasons for using registers in combination with the LFS and, correspondingly, several methods for employing the registers in an estimation procedure; see questions 4 and 14. The group of countries that use registers consists of seven countries: BE, DK, FR, FI, NL, NO, and SE. Table 3.3 summarises the main "why's and how's" of the use of registers.

Some comments are to be made. First, Table 3.3 gives an overview of the answers given in the questionnaire. That is, a country is mentioned in the second column of the table only when it said to use the corresponding method. It does not necessarily mean that other countries do not use that method/approach. It is also noteworthy that nobody called the method of bounding weights in order to avoid negative weights and/or too large weights. Although some techniques are more or less identical, we mention all of them because these differences illustrate the different points of view as well as the different software packages used by the NSI's.

Table 3.3: "Why's and how's" of the use of registers

NL and SE	stratification
BE, DK, FI, and NO	poststratification
BE, FI, NO	calibration
BE, FI, SE	regression estimator
BE, FR	raking method
FI, FR, NL	nonresponse correction, bias reduction, increase of accuracy
BE, FR, DK	imputation of social security, labour class and education, respectively
DK	sample selection
BE, FR, NL	equal weights for members of the same household

The following table summarises the registers that the NSI's employ in their weighting and/or imputation procedures.

Table 3.4: Registers used by the NSI's

BE	population (daily), social security (4)*
DK	population (3), unemployment (2), education (12), labour class (12)
FI	population (1), job-seekers (1)
FR	business register (for labour classification) (daily)
NL	population (12), employment (1)
NO	population (daily), employment (3), tax (12)
SE	population (weekly), employment (12), job-seekers (daily)

Between parentheses the update period is given (in months).

Consistency and readiness (questions 7 and 8)

All countries report that there is an almost-perfect consistency with respect to the variables. Only BE and NL mention an inconsistency problem with the employment variable, partly due to a time lag. Most register data are ready for immediate use by the NSI's, only FI and NL need some additional calculations with respect to the registers of job-seekers and employment, respectively.

Matching (questions 9-13)

With the exception of SIRENE in France, all registers have a unique matching key. FR uses name and address as additional key, while NL mentions birth date, address, and sex as additional key information. The number of missed matches in DK, FI, NO, and SE is zero, while the number of missed matches in BE and NL is less than 1%. The number of missed matches in FR (SIRENE) is not available or unknown. Only BE, FI, and NL mentioned other institutes that provide the statistical offices with population totals required for the estimation procedure.

Weighting

Table 3.5 shows for each country the variables used in the weighting scheme. The main source used for Table 3.5 is Eurostat's publication "Labour Force Surveys: Results 2000".

Table 3.5: Variables included in the weighting schemes

AT	sex, age, region, and nationality
BE	sex, age, and region
CH	sex, age, region, nationality, and marital status
DE	sex, age, region, nationality, and household size
DK	age, income, labour class, education, and employment status
ES	age and region
FI	sex, age, region, and job-seeking
FR	sex and age
GR	region
IL	sex, age, and region
IT	sex, age, and region
IE	sex, age, and region
LU	sex, age, nationality, and household size
NL	sex, age, region, ethnicity, and marital status
NO	sex, age, region, income, and employment status
PT	sex, age, and region
SE	sex, age, region, labour class, and job-seeking
UK	sex, age, and region

Appendix A

Questionnaire on the use of auxiliary information (register-data) for Labour Force Surveys (LFS) or other surveys containing Labour Force Data

Name: .....

Organisation: .....

Address: .....

Country: .....

Phone: .....

e-mail: .....

Fax: .....

1. Which regular sample surveys on labour force or multipurpose surveys with labour force information do you/does your institution apply in your country? Are those surveys voluntary or mandatory?

   .....

   .....

2. Do you use data of registers. If not, go to question 15.

   .....

   .....

3. Please indicate the type of register data, e.g. municipal, regional, national, other! Are these registers part of official administration or provided by non-official institutions? What type of variables are used in the registers.

   .....

   .....

4. Please, indicate the purpose of the usage of the register data (for nonresponse error correction, auxiliary information for estimators or unequal probability designs, poststratification, marginal distributions).

   .....

   .....

5. What is the timeliness of the registers employed by your statistics office? Please, indicate (approximately) the time lag between the survey observation period and the renewal (date) of the registers.

   .....

   .....

6. Are the registers updated every year/quarter/currently or otherwise?

   .....

   .....

7. Are the units and the variables in the samples and registers mutually consistent regarding their definitions and/or classifications?

   .....

   .....

8. Are the register data ready for immediate use in the sampling procedure or do they prior to their use need some kind of numerical adjustment or transformation?

   .....

   .....

9. Do the registers you use have unique identification numbers for the units? Can they be used for matching purposes?

   .....

   .....

10. Which other matching keys (including variables) do you use for combining registers and the survey?

    .....

    .....

11. Please, describe briefly your methods for matching records of surveys and registers!

    .....

    .....

12. Is there any indication for the percentage of survey records that are wrongly matched?

    .....

    .....

13. In case you use population totals, averages or data on a higher level of aggregation, do you obtain these totals from registers or from external sources/publications or from previous surveys ?

    .....

    .....

14. Please, could you briefly describe the estimation procedure in case you use auxiliary information. Do you use regression/ratio estimators, raking methods, special weighting procedures, stratification? Do you use imputation for auxiliary variables in the sample? Do you impose additional restrictions on the final weights, e.g. bounding of weights, equal weights for members of the same household?

    .....

    .....

15. Please could you, if there are, indicate specific reasons or obstacles for not using registers.

    .....

    .....

16. Do you have plans to use registers in future?

    .....

    .....

Please return this questionnaire to:

Statistics Netherlands

Methods and Informatics Department

Dr. Paul Knottnerus

Prinses Beatrixlan 428

2273 XZ Voorburg

The Netherlands

e-mail: pkts@cbs.nl

References

Cochran, W. G. (1977): *Sampling Techniques*. Third edition. New York: Wiley.

Deville, J. C. and Särndal, C. E. (1992): Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Eurostat (2001): *Labour Force Survey: Results 2000*. Luxembourg: Eurostat.

Holt, D. and Smith, T. M. F. (1979): Poststratification. *Journal of the Royal Statistical Society, A* **142**, 33–46.

Knottnerus, P. (2003): *Sample Survey Theory: Some Pythagorean Perspectives*. New York: Springer-Verlag.

Münnich, R. and Wiegert, R. (2002): On the influence of limited assignment of survey and register units on data quality. In *ICIS 2002 conference proceedings*. `http://www.icis.dk`.

Münnich, R. and Wiegert, R. (2004): German register data for regression estimation in survey sampling - A study on the German Microcensus respecting for data protection. *Jahrbücher für Nationalökonomie und Statistik* **224/1-2**, 247–259.

Särndal, C. E., Swensson, B. and Wretman, J. (1992): *Model Assisted Survey Sampling*. New York: Springer-Verlag.