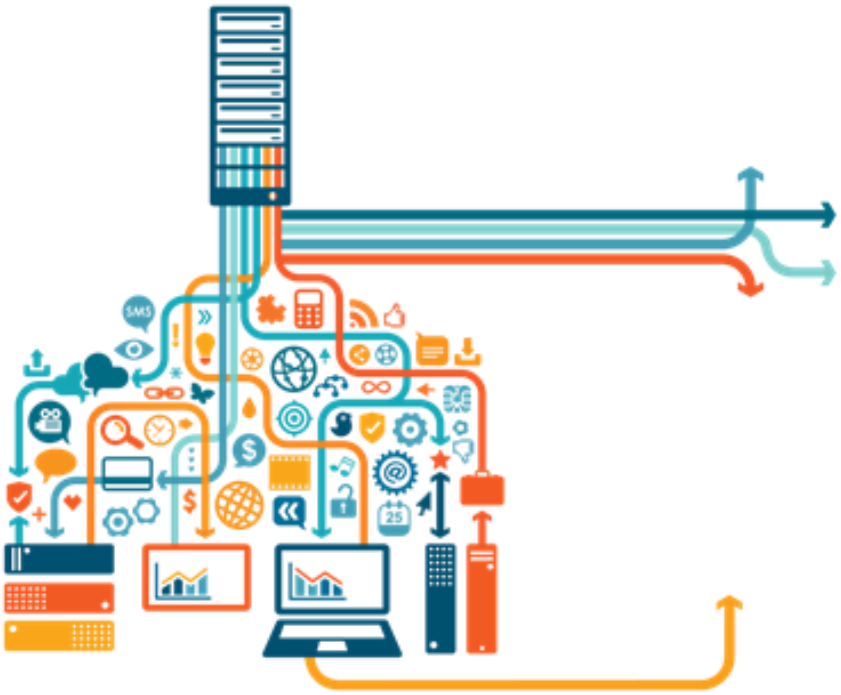
Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-1: Modelling measurement error in admin and survey variables on turnover

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-1: MODELLING MEASUREMENT ERROR IN ADMIN AND SURVEY VARIABLES ON TURNOVER

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252

Luxembourg, February, 2017

Title of the case study: **Modelling measurement error in admin and survey variables on turnover**	
Executive summary	
Usages (Deliverable 1)	Data validation
Statistical Tasks (Deliverable 2)	Alignment of measurements
Data configurations (Deliverable 1)	4
Methods (Deliverable 5)	Variable harmonisation based on latent variable models
Presentation of the Case study	
Agency – country	Statistics Netherlands
Topic	Measurement errors in admin and survey variables on turnover

Description of the problem	Since 2011, the quarterly short-term statistics (STS) on turnover at Statistics Netherlands are based on a combination of administrative data on value-added tax (VAT) turnover from the Dutch tax authorities for small to medium-sized enterprises and survey data for all large and/or complex enterprises. The main STS output consists of estimated growth rates of turnover for different business domains (sectors of the economy classified by NACE code). In addition, the total turnover level in each domain is estimated as input for other statistics, in particular the national accounts. It was found in previous studies that the values of VAT turnover and survey turnover for the same enterprise are often not equal (Van Delden et al., 2016). Part of these differences are due to random measurement errors or linkage problems between fiscal units and statistical units. For some NACE codes also systematic differences between the two turnover measurements occur. This may be explained by conceptual differences between the definitions of turnover according to the tax authorities and according to the Eurostat regulations for STS. For instance, certain types of economic activity may be partly exempt from taxes (and thus not included in VAT turnover) but relevant for statistical purposes (and thus included in survey turnover). As another example, for certain types of trade in second-hand goods the tax authorities require only the profit margin made by businesses, whereas the statistical definition requires the total turnover. For each (group of) NACE code(s) that constitutes a publication domain, Statistics Netherlands had to decide whether the VAT turnover could be used for small to medium-sized enterprises, possibly after applying a correction formula to account for conceptual differences. For domains where it was not possible to use VAT turnover – because the conceptual differences where too large and/or too erratic to be corrected reliably – an existing sample survey of enterprises remained in place.
Input data	Administrative data on VAT turnover, survey data on turnover, both for the year 2012
Expected output	Microdata with harmonised administrative turnover values, to be used for further processing

Technical summary

Methods	During the development of the new production process for STS at Statistics Netherlands, the relation between VAT and survey turnover was analysed for each domain using a robust weighted linear regression model, with survey turnover as the dependent variable and VAT turnover as the independent variable (Van Delden et al., 2016). For domains where a systematic difference between the two turnover variables was found, the estimated regression line was used directly as a correction formula for VAT turnover. In this way, the two turnover variables were harmonised. Two drawbacks of this approach are that it does not explicitly account for (a) random measurement errors in the VAT data and (b) systematic measurement errors in the survey data. An alternative method was investigated based on a structural equation model (SEM).

	In the SEM, both the administrative and survey variables are treated as error-prone measurements of an unobserved (latent) true variable. In the present application, the latent variable represents the true value of turnover for each enterprise (according to the statistical definition). The model includes regression equations for the relation between each observed variable and the latent variable. If the model includes several latent variables, the relations between these concepts can also be modelled using regression equations. By explicitly modelling the errors in each observed variable, the SEM approach can avoid the two above-mentioned drawbacks of a direct linear regression model. A more detailed introduction to SEMs is given in the module "Variable harmonisation based on latent variable models" in Deliverable 5. Several interesting quality indicators for the administrative and survey variables can be derived from the estimated parameters of the SEM. The validity of each observed variable is defined as the absolute value of its correlation to the latent (true) variable that it is supposed to measure; this is a value between 0 and 1, with values close to 1 indicating good measurement. In addition, the intercept and slope parameters of the regression equation that links an observed variable to the true variable can be used to quantify bias due to systematic measurement errors. In the absence of bias, we would expect the intercept and slope to be equal to 0 and 1, respectively. From these estimated parameters, again a correction formula can be derived if necessary to harmonise the observed variables. It should be noted that such a harmonisation step is useful only if the validity of the observed variables is close to 1, because the observed variables should be highly correlated to their latent ones. To obtain an identified SEM, some assumptions need to be made about the scale of the latent variables. In the application below, model identification was achieved by including additional error-free measurements for a random subsample of the original data set (an audit sample). These audit data were obtained by letting subject-matter experts re-edit part of the original observations, with the aim of obtaining the true values according to the statistical definition of turnover.

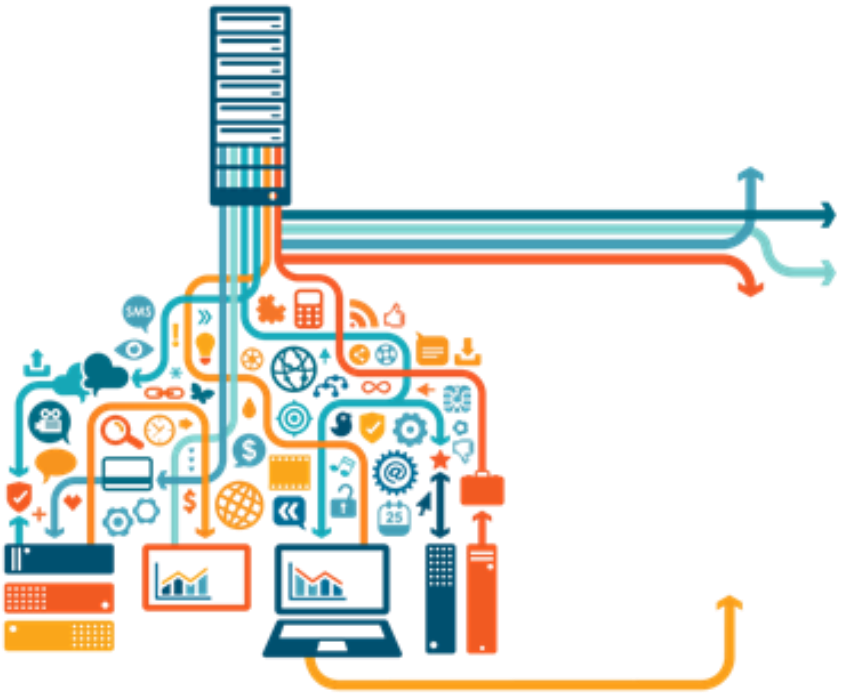
Application	Scholtus et al. (2015) describe an application of the SEM methodology to estimate the validity and bias in VAT turnover for STS at Statistics Netherlands. The authors used annual turnover values for 2012 from three sources: the VAT data, the structural business survey (SBS) and an additional administrative source, the so-called Profit Declaration Register (PDR). SBS data were used instead of STS data because the SBS sample is larger and the turnover definitions for these surveys are identical for nearly all NACE codes. An SEM was set up which included, besides turnover, three other concepts: number of employees, costs of purchases and total operation costs. By including more concepts, it was easier to obtain an identified model. Observed values for the other concepts were taken mainly from the PDR data. The model was estimated for eight different publication domains, four from the sector "Trade" and four from the sector "Transport". Model fit was evaluated using several common fit measures for SEMs. For most domains, a model with a good or acceptable fit could be found. For the smallest domains (i.e., with relatively few units), finding a well-fitting model was more problematic. From the estimated models, it was found that the observed variables usually had a validity close to 1. In particular, the estimated validity of VAT turnover was larger than 0.95 in six of the eight domains, including all "Trade" domains. From the estimated intercept and bias parameters, it was found that VAT turnover was systematically biased downwards in several domains. Given the good validity values, it was possible to derive a formula to correct for this bias from the model. For instance, in the domain "Sale and repair of passenger cars and light motor vehicles", the estimated SEM indicated that the validity of VAT turnover was 0.98 and that a harmonised turnover value could be obtained as Estimated true turnover = 0.11 + 1.22 × VAT turnover, with turnover measured in millions of Euros. Although the main interest here was with VAT turnover, the same model also provided estimates of validity and bias for the other observed variables from the SBS survey and the PDR data. See Scholtus et al. (2015) for detailed results.
Main findings (lessons learnt)	
Advantages	• The SEM can be used to correct for systematic measurement errors in observed variables, by deriving a harmonised variable. • The SEM can account for systematic and random measurement errors in all observed variables (both administrative and survey variables). It is not necessary to take one source as the 'gold standard' a priori.

Disadvantages	• In its current form, the method does require 'gold standard' audit data for a random subsample of the original data. Collecting such additional data may be expensive or difficult in practice. Moreover, the assumption that the audit data are truly error-free cannot be tested within the modelling framework. (The need for audit data can be avoided if one is willing to assume that one of the sources, e.g. the survey, is at least free of systematic bias. Again, this assumption cannot be tested with the available data.) • The standard SEM estimation methodology is not robust to outliers in the original data. • Estimated SEM parameters may suffer from small-sample bias. This may be a problem when a domain contains relatively few (sampled) enterprises.
Gap analysis	• To make the method more suitable for use in practice, the above disadvantages should be addressed. • In the application, only annual turnover values were analysed, but the STS is published on a quarterly basis. There may also be systematic differences between VAT turnover and survey turnover that affect the quarterly values but not the annual values, for instance due to seasonal reporting differences by enterprises.
Other remarks	N/A
References	A. van Delden, R. Banning, A. de Boer and J. Pannekoek (2016), Analysing Correspondence between Administrative and Survey Data. Statistical Journal of the IAOS 32, 569–584. S. Scholtus, B.F.M. Bakker and A. van Delden (2015), Modelling Measurement Error to Estimate Bias in Administrative and Survey Variables. Discussion Paper 2015-17, Statistics Netherlands, The Hague. Available at www.cbs.nl.

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-2: Estimating classification errors under edit-restrictions in combined register-survey data

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-2: ESTIMATING CLASSIFICATION ERRORS UNDER EDIT-RESTRICTIONS IN COMBINED REGISTER-SURVEY DATA

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252

Luxembourg, February, 2017

Title of the case study:
Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages (Deliverable 1)	Data validation
Statistical Tasks (Deliverable 2)	Alignment of measurements
Data configurations (Deliverable 1)	4
Methods (Deliverable 5)	Variable harmonisation based on latent variable models

Presentation of the Case study

Agency – country	Statistics Netherlands
Topic	Classification errors in admin and survey variables on home ownership
Description of the problem	Statistics Netherlands has access to a registration of addresses and buildings (BAG). This registration can be linked to the Dutch population register and the information in the BAG can be used to derive whether a person owns or rents a home. The variable home ownership is useful for many analyses in social research. These analyses may be affected by classification errors in the BAG, i.e., cases where a person who in reality rents a home is classified as owning a home in the BAG, or vice versa. To investigate the size of the classification error problem in the BAG, Statistics Netherlands has linked the BAG data to survey data from the so-called LISS panel (Longitudinal Internet Studies for the Social sciences). The LISS panel is a web panel, administered by CentERdata, based on a random sample from the Dutch population register. Respondents in the LISS panel were asked whether they own or rent a home. Thus, by using the overlapping units in the BAG and LISS data, two observed variables that measure home ownership were obtained for a sample of the population. The data were analysed using a latent class model to account for classification errors. In addition, the LISS survey contained questions on marital status and whether the respondent received rent benefit. These were included as covariates in the model, to aid the estimation of classification errors. The fact that only people who rent a home can receive rent benefit implies an edit rule for the data. It is useful to incorporate this edit rule into the model, as it helps to identify some of the classification errors.
Input data	Administrative data (BAG) on home ownership, survey data (LISS panel) on home ownership, marital status and rent benefits. All data refer to the year 2013.

Expected output	Estimated classification error probabilities for home ownership in both data sources, as well as a multiply-imputed data set with estimated true home ownership status values.

Technical summary

Methods

Boeschoten et al. (2016) used a latent class model to analyse the combined register-survey data. In this model, the true classification of each unit with respect to the variable of interest is denoted by a latent (unobserved) variable. The model describes, for each unit, the probability that it is classified in a particular category according to each observed variable, given its true category according to the latent variable. Thus, the model describes the probability of correct classification as well as the probability of each possible classification error. A more detailed introduction to latent class models is given in the module "Variable harmonisation based on latent variable models" in Deliverable 5.

-Let X denote the true home ownership status of a person, with two possible categories: "own" and "rent". This variable is unobserved (latent). The observed home ownership status in the BAG and LISS panel is denoted by Y_1 and Y_2, respectively. Finally, the covariates are denoted by Q (which may be a vector).

In the latent class model, the marginal probability of observing a particular response pattern $P(Y_1 = y_1, Y_2 = y_2)$ is modelled as the sum of the joint probabilities $P(Y_1 = y_1, Y_2 = y_2, X = x) = P(X = x)P(Y_1 = y_1, Y_2 = y_2 | X = x)$ over all latent classes $x \in \{\text{own}, \text{rent}\}$. Furthermore, two assumptions are made:

- The classification errors in different observed variables are independent, given the score on the true variable (latent class).
- The classification errors in the observed variables are independent of the covariates, given the score on the true variable.

This leads to the following model:

$$P(Y_1 = y_1, Y_2 = y_2 | Q = q) = \sum_{x \in \{\text{own,rent}\}} P(X = x | Q = q) \prod_{l=1}^{2} P(Y_l = y_l | X = x).$$

Here, $P(X = x|Q = q)$ denotes the probability that an individual has true home ownership status x, given his/her scores on the covariates; also, $P(Y_l = y_l|X = x)$ denotes the probability of observing a home ownership status y_l on the l-th observed variable, given that the true home ownership status is x. Thus, the classification error probabilities for the BAG under this model are given by $P(Y_1 = \text{rent}|X = \text{own})$ and $P(Y_1 = \text{own}|X = \text{rent})$, and similarly for LISS.

As mentioned above, people who own their home cannot receive rent benefit. Therefore, if the rent benefit indicator is included as one of the covariates in the model, say Q_1, then it must hold that

$$P(X = \text{own}|Q_1 = \text{receives rent benefit}) = 0.$$

This can be added as a restriction on the parameters of the latent class

model.

From the estimated model, posterior membership probabilities for each latent class $x \in \{\text{own}, \text{rent}\}$, given the observed values of Y_1, Y_2 and Q, can be obtained by applying Bayes' rule. Boeschoten et al. (2016) propose to use these posterior probabilities to impute an estimated true home ownership status for each person.

To reflect the uncertainty in these estimated true values, they apply a multiple imputation on the basis of the estimated posterior probabilities. This leads to a data set with, in addition to the original observed variables, M imputed variables with predicted true home ownership status values. These imputed variables can be used in further statistical analyses to correct for classification errors; see Boeschoten et al. (2016) for more details.

Application

After combining the BAG registration with the LISS panel data, Boeschoten et al. (2016) were left with 3011 individuals for which both Y_1 and Y_2 were available. The model was fitted to this combined data set. Values of marital status were also available from the LISS panel for all these persons. The indicator whether a person receives rent benefit was available from the LISS panel for only 779 individuals, due to the routing in the questionnaire. Missing values on this covariate where taken into account by applying Full Information Maximum Likelihood to estimate the model. Estimation of the model was done using the Latent Gold software (Vermunt and Magidson, 2013).

We discuss the results for two different latent class models. Both models included marital status and rent benefit as covariates. In the first model, the above restriction that people who receive rent benefit cannot own a home was not explicitly included in the model (unrestricted model). In the second model, this restriction was included (restricted model).

The following table shows the estimated classification error probabilities in the two sources, for both models. It is seen that, according to the model, the two sources are not error-free. However, the proportion of misclassified persons appears to be relatively small in both sources. In other words, the quality of measurement is good. In the LISS panel, errors where persons who rent a home classify themselves as home owners appear to be more frequent than misclassifications in the reverse direction. In the BAG, both types of misclassification occur about equally often according to the restricted model.

Data source	Probability	Unrestricted	Restricted
BAG	$P(Y_1 = \text{rent} \mid X = \text{own})$	0.0251	0.0475
	$P(Y_1 = \text{own} \mid X = \text{rent})$	0.0500	0.0504
LISS	$P(Y_2 = \text{rent} \mid X = \text{own})$	0.0003	0.0008
	$P(Y_2 = \text{own} \mid X = \text{rent})$	0.1062	0.0666

The next table shows the estimated proportion of home owners in the population P(X=own) according to both sources separately and according to

both latent class models. The model-based estimates were obtained by applying multiple imputation as described above and pooling the results.

	Point estimate	95% confidence interval
Observed in BAG	0.6450	[0.6448; 0.6451]
Observed in LISS	0.6830	[0.6829; 0.6832]
Unrestricted model	0.6505	[0.6503; 0.6506]
Restricted model	0.6591	[0.6590; 0.6593]

The original estimated proportions based on a single data source differ significantly. With both models, the estimated proportion of home owners after correction for misclassification lies closer to the BAG estimate than to the LISS estimate. See Boeschoten et al. (2016) for more results, including an illustration of the use of the imputed data set in a logistic regression of home ownership on marital status.

Main findings (lessons learnt)

Advantages

- The latent class model can be used to correct for classification errors in observed variables, by estimating and imputing a harmonised variable. The uncertainty in this harmonised variable can be taken into account through multiple imputation.
- The latent class model can account for classification errors in all observed variables (both administrative and survey variables). It is not necessary to take one source as the 'gold standard' a priori.
- The method can also take edit rules into account in the form of restrictions on the model parameters.

Disadvantages

- The model is based on two strong assumptions: that classification errors are independent across observed variables and do not depend on covariates, conditional on the latent (true) class. In practice, these assumptions may not always hold and then the model-based results may be biased.
- Classification errors in the covariates are not explicitly accounted for.

Gap analysis

- To make the method more suitable for use in practice, the above disadvantages could be addressed.
- Estimated relations of the latent class variable to covariates that are not included in the model may be biased. It is not always possible or desirable to include all potentially relevant covariates in the model. It would be good to have a method that can correct estimated relations between the latent variable and covariates not included in the latent class model.

Other remarks

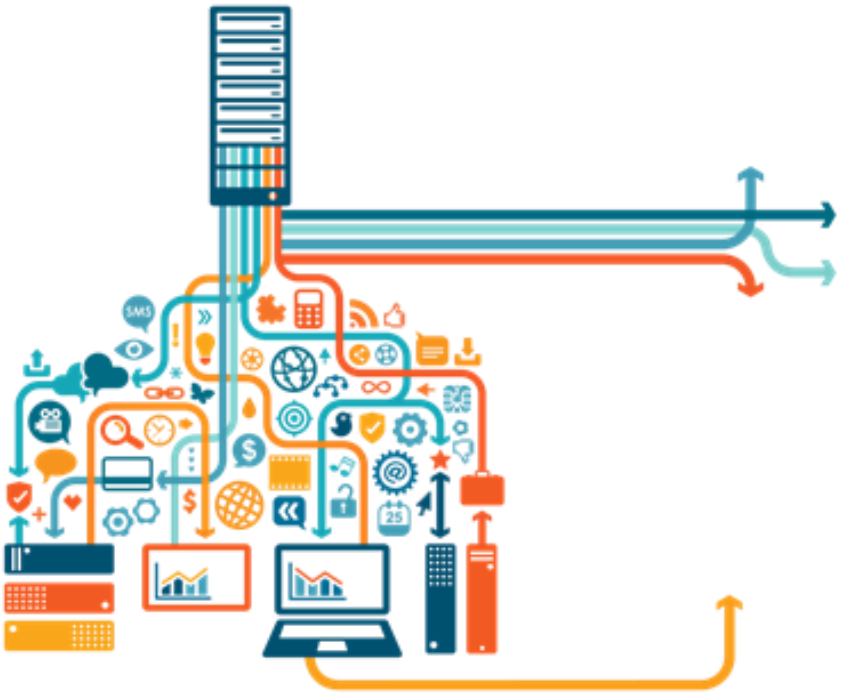
/

References	L. Boeschoten, D. Oberski and T. de Waal (2016), Estimating Classification Error under Edit Restrictions in Combined Survey-Register Data. Discussion paper 2016-12, Statistics Netherlands, The Hague. Available at www.cbs.nl. J.K. Vermunt and J. Magidson (2013), Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax. Statistical Innovations Inc., Belmont, MA.

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-3: The creation of a Social Policy Simulation Database - SPSD/M (Stat Canada)

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-3: THE CREATION OF A SOCIAL POLICY SIMULATION DATABASE - SPSD/M (STAT CANADA)

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252

Luxembourg, February, 2017

Title of the case study:
Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages	Database for Microsimulation
Statistical Tasks (Deliverable 2)	Creation of joint statistical micro data
Data configurations (Deliverable 1)	3
Methods (Deliverable 5)	Statistical matching

Presentation of the Case study

Agency – country	Istat (Italy) on a case developed by Statistics Canada
Topic	The creation of a Social Policy Simulation Database - SPSD/M
Description of the problem	Statistics Canada publishes a Social Policy Simulation Database and Model (SPSD/M), which is a tool designed to assist those interested in analysing the financial interactions of governments and individuals in Canada. In other words it can help one to assess the cost implications or income redistributive effects of changes in the personal taxation and cash transfer system. Apart the microsimulation model, this tool needs a non-confidential, statistically representative database (SPSD) of individuals in their family context, with enough information on each individual to compute taxes paid to and cash transfers received from government. No one dataset from a given survey or administrative source provides a sufficiently detailed and integrated picture of Canadian households to support the analysis of costs and distributional impacts of the entire tax/transfer system as it moderates the flows of money between governments and individuals. The SPSD was constructed by combining individual administrative data from personal income tax returns and unemployment claimant histories with survey data on family incomes, employment and expenditure pattern. As far as administrative data are concerned, only samples of anonymized records have been used available.
Input data	**Survey of Labour and Income Dynamics (SLID):** the main source of data on the distribution of income amongst individuals and families and served as the host (recipient) dataset. **Personal income tax return data:** a sample of over 400,000 personal income tax (T1) returns used as the basis of Canada Revenue Agency's annual Income Statistics publication. **Employment Insurance claim histories:** a 10% sample of histories from

	Human Resources Development administrative system. **Survey of Household Spending:** Statistics Canada's periodic survey of very detailed data on Canadian income and expenditure patterns at the household level including information on net changes in assets and liabilities (annual savings). Adjustments have been made to data in order to ensure consistency with known control totals, as the number of people by age and sex as reported in the Census and the number of high income Canadians as reported by Canada Revenue Agency.
Expected output	The output is a synthetic data set of more than 200000 individuals in over 80000 households in ten Canadian provinces (Yukon and Northwest Territories were excluded) that preserves the confidentiality of individual information without compromising statistical validity, consisting of almost 600 variables covering detailed socio-economic and demographic data as well as information on weekly employment histories, expenditure patterns and itemized tax deductions.
Technical summary	
Methods	The overall method applied to create SPSD is rather complex (fully described in detail in Statistics Canada (2014)) and involves many steps, including imputation and population completion. Here we focus only on how the data sets have been matched. The idea is to use statistical matching methods for the creation of synthetic data files based on hot-deck imputation techniques. These techniques use the data sets to match as either a recipient data set (always the SLID file) or a donor, i.e. a file that provides only the additional variables to attach to the records of the recipient. For each record of the recipient file, a record in the donor is found so that it can be considered as the most "similar" to the recipient according to some characteristics. This is achieved by creating "bins" of records from SLID and the file to match, where each bin is composed by a fixed pattern of categories of the matching variables (e.g. all records with the same gender, age class, region of residence). For the pair of bins created according to the same values of the common variables, records are duplicated or weights adjusted so that the bins consist of the same number of records. Finally, income as available in SLID and the file to match is used in order to rake the records, from the lowest to the highest income. Individuals with the same rake are matched (the creation of bins corresponding to applying a "categorically constrained" statistical matching; duplication and adaptation of the weights is described in Kovacevic and Liu (1991) as the "weight-split algorithm").

Application	Hennessy et al (2015) take data from SPSD 2009 and supplemented the model's database with health data to enable analysis of alternative health care financing options in Canada. Data on health status, disability, disease status, health service, medication use and out-of-pocket spending on health care were drawn together from population-based surveys (including the CCHS, the Canadian Health Measures Survey (CHMS), Survey of Household Spending (SHS)), health administrative data (the Discharge Abstract Database (DAD)) and estimates of health service use and cost available in the literature. Costs of health services and drugs were assigned to individuals in the CCHS 2009/2010 (host health dataset) through imputation. The resulting health dataset was merged with the SPSD. Models using the enhanced SPSD -Health will allow health policy makers and academics to "try-out" alternative health care financing options, and consider their monetary impacts on individuals and families in Canada The method used in this application is still a categorically constrained hot-deck method based on ranks in bins through the weight-split algorithm.

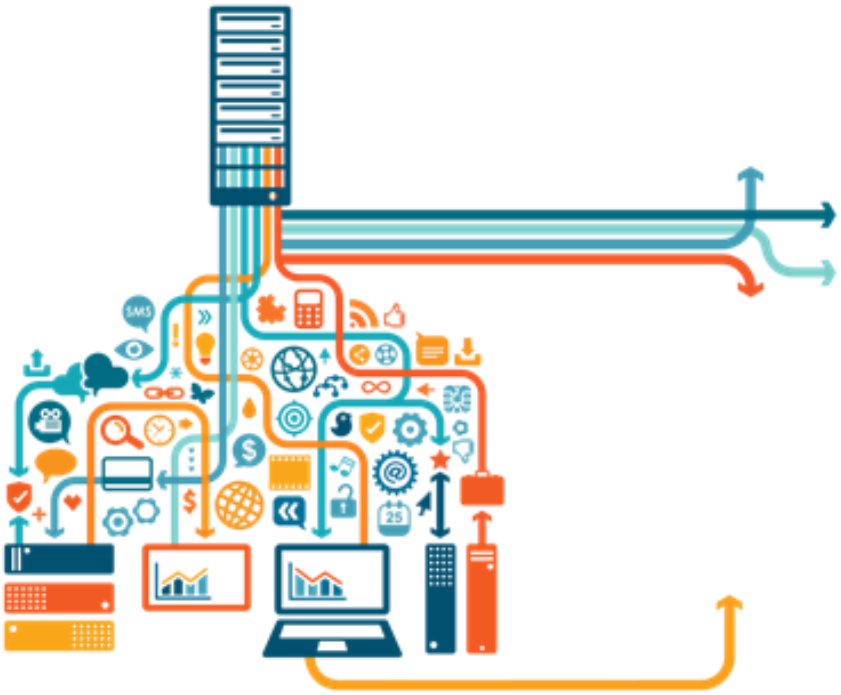
Main findings (lessons learnt)	
Advantages	The synthetic files are a very good source for microsimulation purposes, it is a product successfully produced by Statistics Canada in many years.
Disadvantages	The question is if this file could also be used for any statistical analysis on the 600 variables gathered together imputing successively the SLID with the specific content of the other three files. Under this point of view, the main disadvantage seems to be the fact that the data set could be affected by the so called conditional independence assumption, i.e. the data set imputes a specific variable from one source independently of specific variables in SLID given the matching variables. This assumption seems to be very attenuated for the joint analyses involving income, given the use of the rake procedure for matching records in the homologues bins (e.g. Donatiello et al, 2016). Anyway it is not sufficient in order to say that the conditional independence assumption has been completely avoided.
Gap analysis	/
Other remarks	The method could possibly make use of the notion of uncertainty in statistical matching. This can be achieved by means of micro data sets using also statistical matching methods based on multiple imputation.
References	Donatiello D., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M. (2016) The role of the conditional independence assumption in statistically matching income and consumption. Statistical Journal of the IAOS, 32, 667-675, DOI 10.3233/SJI-161000. Hennessy, Sanmartin, Eftekhary, Plager, Jones, Onate, Mc Evoy, Hicks, Deber (2015) Creating a synthetic database for use in microsimulation models to investigate alternative health care financing strategies in Canada.

	International Journal Of Microsimulation, 8(3) 41-74. Statistics Canada (2014) SPSD/M database creation guide', Social Analysis and Modelling Division, Statistics Canada, Ottawa, Ontario. (http://www.statcan.gc.ca/eng/microsimulation/spsdm/spsdm).

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-4: Dutch Population and Housing Census

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-4: DUTCH POPULATION AND HOUSING CENSUS

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Luxembourg, February, 2017

Title of the case study:
Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages (Deliverable 1)	Indirect estimation
Statistical Tasks (Deliverable 2)	Integrate microdata with macrodata; integrate macrodata with macrodata
Data configurations (Deliverable 1)	6, 7
Methods (Deliverable 5)	Macro-integration

Presentation of the Case study

Agency – country	Statistics Netherlands
Topic	Population and Housing Census
Description of the problem	Description of the problem to solve, characteristics of the input data, characteristics of the expected output, constraints if any, main findings Statistics Netherlands pursues a 'one-figure' policy for the Dutch Census. With this 'one-figure' policy Statistics Netherlands aims to publish only one estimate for the same phenomenon occurring in different tables of the Dutch Census. The Dutch Census is a virtual census in the sense that is based on a number of administrative data sources and surveys rather than on a complete enumeration of the population. To enforce the 'one-figure' policy population totals, either known from an administrative data source or previously estimated, are imposed as benchmarks, provided they overlap with an additional survey dataset that is needed to produce new output statistics.
Input data	Various administrative data sources, and the Labour Force Survey
Expected output	Frequency tables

Technical summary

Methods	Macro-integration is the process of reconciling statistical figures on an aggregate level. These figures are usually in the form of tables, obtained from different sources. When macro-integration is applied, only estimated figures on an aggregated level are adjusted. The underlying microdata are not adjusted or even considered in this adjustment process. Several methods for macro-integration have been developed, see, De Waal (2016) for an overview. Traditionally, macro-integration has mainly been applied in the area of macro-economics. At Statistics Netherlands macro-integration is applied to

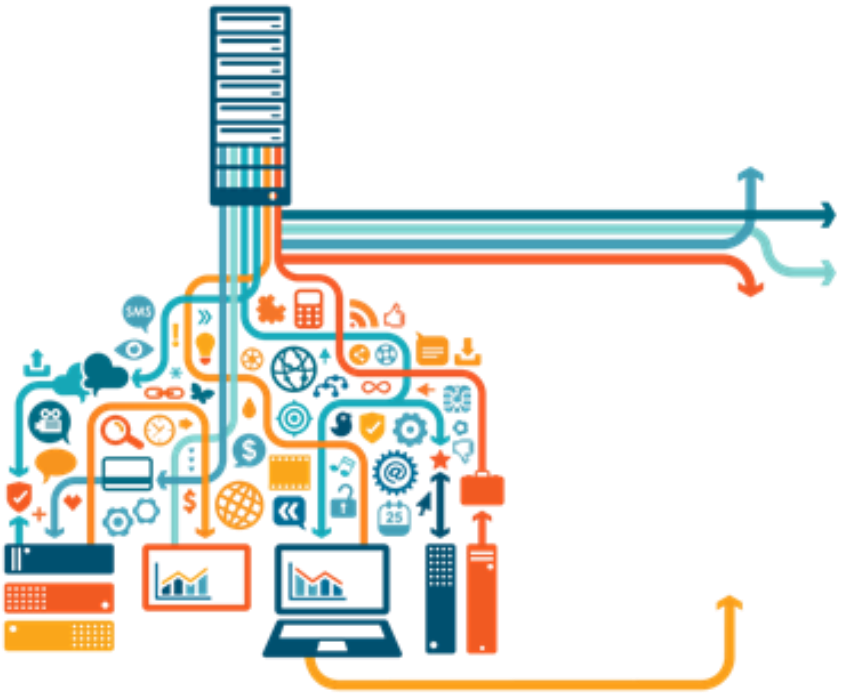
	reconcile estimates for the National Accounts. Also, applications in other areas have been studied at Statistics Netherlands, namely for the reconciliation of tables of Transport and Trade Statistics, for combining estimates of labour market variables, and for the Dutch Census 2011 (see Mushkudiani, Daalmans and Pannekoek 2012). The starting point of macro-integration is a set of estimates in tabular form. These can be quantitative tables or frequency tables. If the estimated figures in these tables are based on different sources and (some of) the tables have cells in common, these cell values are often conflicting. In order to apply a macro-integration method later on, it is important that (an approximation or indication of) the variance of each entry in the tables to be reconciled is computed. During the reconciliation process the entries of the tables are adjusted by means of a macro-integration technique so all differences between tables are reconciled and the entries with the highest variance are adjusted the most. In a macro-integration approach often a constrained optimization problem is constructed. This is, for instance, the case for the so-called Denton method (see, e.g., Mushkudiani, Daalmans and Pannekoek 2012). A target function, for instance a quadratic form of differences between the original and the adjusted values, is minimized, subject to the constraints that the adjusted common figures in different tables are equal to each other and internal cell values of the adjusted tables sum up to the corresponding marginal totals. Inequality constraints can be imposed in these quadratic optimization problems. In the literature also Bayesian macro-integration methods have been proposed.
Application	The backbone of the Dutch 2011 census is the central population register (PR), which combines all the municipal population registers. In the 2011 census, PR data for 1 January 2011 were used as the basis for the set of tables that needs to be published. The tables focus on frequency counts, not on quantitative information. Data not available or derivable from the PR were taken from other registers. All register variables are now available from Statistics Netherlands' System of social statistical datasets (SSD), and their quality has been improved by applying micro-integration techniques. Micro-integration entails checking the data and adjusting those that are incorrect. It is widely assumed that micro-integrated data provide more reliable results, as they are based on a maximum amount of information. They also provide better coverage of subpopulations: if data are missing in one source, another source can be used. In the 2011 Census, only two variables were not taken from a register: 'occupation' and 'educational attainment'. Records from the Labour Force Survey (LFS) in a three year period around the enumeration date (1 January 2011) were used to estimate values for these two variables, which are included in 23 of the 60 tables. For the Dutch 2011 Census, table consistency was guaranteed by using the

	so-called repeated weighting approach for these 23 tables (see Statistics Netherlands 2014). However, the use of macro-integration instead of repeated weighting was studied in an attempt to overcome some of the disadvantages of repeated weighting (see Mushkudiani, Daalmans and Pannekoek 2012). The conclusion of that study was that macro-integration leads to good results and can be applied efficiently in practice.
Main findings (lessons learnt)	
Advantages	• Estimates obey the 'one-figure' policy. • Logical relationship between figures in tables can be taken into account (e.g. that the estimated number of persons with a driver's license in a certain region is at most equal to the estimated number of persons that are eligible to have a driver's license in that region) In principle, all tables can be reconciled simultaneously rather than sequentially as is the case for repeated weighting.
Disadvantages	• In some cases exceedingly large optimization problems have to be solved. • The direct relation between microdata and final estimates may be lost.
Gap analysis	Solving macro-integration problems for very large cases.
Other remarks	/
References	De Waal, T. (2016), Obtaining Numerically Consistent Estimates from a Mix of Administrative Data and Surveys. Statistical Journal of the IAOS 32, pp. 231–243. Mushkudiani, N., J. Daalmans and J. Pannekoek (2012), Macro-Integration Techniques with Applications to Census Tables and Labour Market Statistics. Discussion paper, Statistics Netherlands. (https://www.cbs.nl/nl-nl/achtergrond/2012/06/macro-integration-techniques-with-applications-to-census-tables-and-labour-market-statistics) Statistics Netherlands (2014), Dutch Census 2011: Analysis and Methodology. Report, Statistics Netherlands. (https://www.cbs.nl/en-gb/publication/2014/47/dutch-census-2011)

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-5: Validation of combined sources

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-5: VALIDATION OF COMBINED SOURCES

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Luxembourg, February, 2017

Title of the case study:
Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages (Deliverable 1)	Data validation
Statistical Tasks (Deliverable 2)	Editing and imputation
Data configurations (Deliverable 1)	1,2
Methods (Deliverable 5)	Macro editing

Presentation of the Case study

Agency – country	Statistics Netherlands
Topic	Effect of classification errors on turnover estimates
Description of the problem	When statistical estimates based on survey data are edited, often a selective editing procedure is used. In this procedure a score function is computed where the largest scores concern records with a large potential error and a large influence on the outcomes. The records are edited starting with the one with the highest scores, and the editing process is ended when the remaining total error of the estimate at hand is expected to be smaller than the sampling error. In the current case study we consider a somewhat different situation of data editing, namely where estimates are based on a combination of data sources, such as the integration of administrative with survey data. In addition, we have data on all units in the population; this corresponds to the data configurations 1 and 2 (deliverable 1). In the specific case study, turnover derived from VAT is available for the small and medium sized units and turnover derived from a census survey is available for the large and complex units. The statistical unit on which the output is based (after harmonisation of the units in the data sources) is the enterprise. The turnover levels of the integrated data are published for different economic industries, that are based on NACE codes. The NACE codes are obtained by linking the data to a general business register. We know from a previous study (Burger et al., 2015) that the estimated turnover levels are affected by NACE code errors and we would like to correct the output for those errors. Throughout this case study we assume that classification errors are the only errors that occur. In this situation of data editing, we wish to apply a form of selective editing: thus the most important records should be consulted first. Also, we need some criterion that helps us to decide when the editing process can be ended. In our situation we cannot use the sampling error as a criterion, since we have data for all units in the population.

Input data	Value added tax data used to derive turnover; census sample survey data on turnover; a general business register; audit sample.
Expected output	Quality evaluation of estimated turnover levels

Technical summary

Methods	In this case study we need two methods: (1) a method to estimate the effect of the classification errors on the turnover level estimates of the integrated data and (2) an editing method to improve the quality of the estimates (before publication). *Effect of the classification errors.* Assume that we know the probability for each unit i of the target population that it is observed (in the business register) in industry code h given that its true, but unknown, industry code is g. We construct a transition matrix for each unit i where in rows we put the true industry codes, in the columns the observed industry codes and in the cells are the transition probabilities. Hopefully, the diagonal probabilities in this matrix are the largest, which implies that most units are classified correctly. For each unit, we now repeatedly draw a new industry code according to the transition probabilities using a bootstrap procedure. Each bootstrap run draws a sample with industry codes for all units in the population and the result is used to compute a turnover estimate classified by industry. The full set of bootstrap runs can be used to compute the accuracy of the turnover estimates. This accuracy is expressed in terms of the root of the mean squared error, relative to the estimated turnover level itself, abbreviated by RRMSE. In practice we do not know the transition matrix for each unit i, but we estimated the transition probabilities using an audit sample. It was assumed that these probabilities varied across units only as a function of several background variables, such as size class and complexity of the enterprise. We applied the bootstrap procedure to the observed industry codes per enterprise and we used their turnover values. *Editing of classification errors.* Experts in statistical production estimated that they know approximately the 25 largest units in the population per industry that they are responsible for, implying that those units are free of classification errors. For these units the diagonal transition probabilities were set to 1. This current situation is referred to as editing effort 1. We studied the effect of editing on the RRMSE of the estimates. To that end, we compared four levels of editing effort, referred to as 0, 1, 2 and 3. Editing effort 2 corresponds with manually editing twice the amount of editing effort 1 and editing effort 3 refers to editing three times the amount of editing effort 1. Editing effort 0 corresponds to no manual editing. In the editing procedure always the largest enterprises are selected first. Furthermore, we tried out the effect of two different ways of allocating the enterprises to be edited within the population over the underlying industries: - fixed: each industry is allocated the same number of enterprises to be

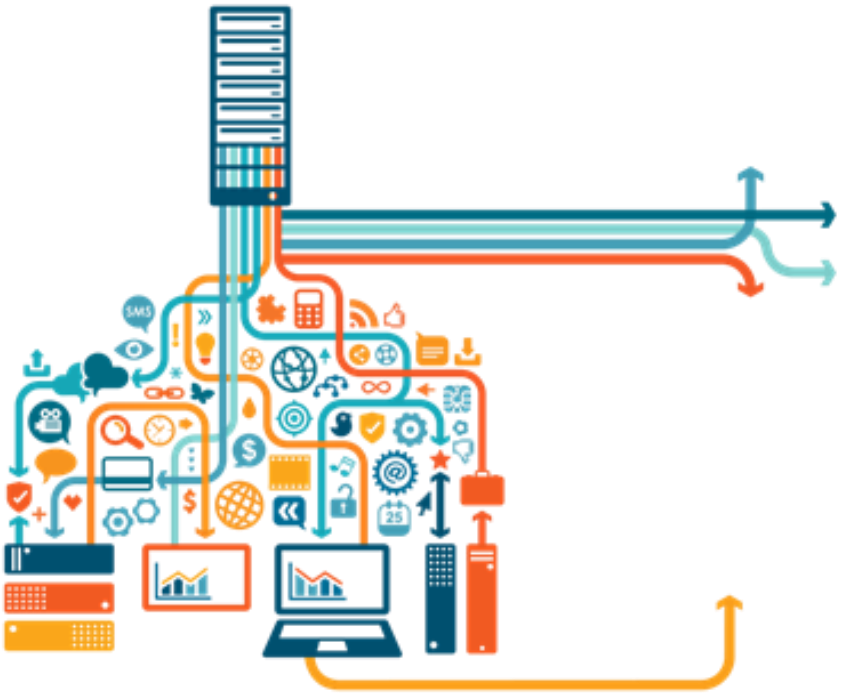
	edited; - pro rata: the number of enterprises to be edited per industry is relative to the product of the $RMSE(\hat{Y}_h)$ x population size per industry": $$n_h^E = \frac{RMSE(\hat{Y}_h)N_h}{\sum_{h=1}^{H} RMSE(\hat{Y}_h)N_h} n^E,$$ where n_h^E is the number of units to be edited per industry h, $RMSE(\hat{Y}_h)$ is the root mean squared error of the estimated turnover per industry $(\hat{Y}_h)$, N_h is the population size per industry, $h = 1, \ldots, H$ are the indices of the target industries and n^E is the total number of enterprises to be edited, dependent on the level of editing effort. The pro-rata allocation is similar to the so-called Neyman allocation. The Neyman allocation describes the allocation of sample units over a population that is divided into strata, and it yields the smallest total variance due to sampling errors for stratified sampling. We therefore expected the pro-rata allocation to yield the smallest RRMSE for a given number of additionally edited enterprises.
Application	The case study is applied to the quarterly turnover growth rates that are computed for the short-term statistics. We applied it to the Dutch car trade population, which consists of nine underlying industries. The industry with the largest total quarterly turnover is industry code 45112 (sale and repair cars and light motor vehicles) and the one with the smallest quarterly turnover is code 45194 (sale and repair of caravans). The outcomes of the RRMSE were computed from the first quarter of 2012 till the second quarter of 2014. Since the outcomes of the RRMSE did not vary much over that period, we computed the two editing scenarios only for one quarter, namely the first quarter of 2013. The relative editing efforts 0, 1, 2 and 3 correspond with manual editing of 0, 225, 450 and 675 enterprises in the car trade population as a whole (the sum over all nine industries). For relative editing effort 1 and the fixed allocation, which corresponds to the current actual editing effort at SN, the RRMSE was about 0.33 per cent which was judged to be acceptable by the statisticians that are responsible for the daily production. After tripling the current editing effort, thus going from relative editing effort 1 to 3 for the fixed allocation, the RRMSE was reduced to about 0.24. Surprisingly, the pro-rata allocation resulted in lower accuracies than the fixed allocation. Editing effort 3 for the pro-rata allocation for instance yielded a RRMSE of 0.37, which is considerably higher than for the fixed allocation. To understand the difference between the two allocations better, we analysed the two components underlying the RRMSE, namely the relative standard deviation (CV: coefficient of variation) and the relative bias (RB). We found that the better accuracy of the fixed allocation was mainly because the RB was smaller in the fixed than in the pro-rata allocation. Also

	in the nine underlying car trade industries we found that an increased editing effort did not always lead to a reduced RRMSE; this was the case for both allocations. Especially, the RB in case of the pro-rata allocation tended to increase in some of the industries (e.g. 45111, 45402 and 45401) with increased editing effort. By a further analysis of the data, we could explain the results as follows. The bias of a turnover estimate for an industry is the net result of the effects of the turnover inflow from units that are wrongly classified into that industry (but in fact belong to another industry) and the outflow from units that are wrongly classified into another industry (but in fact belong to the industry under consideration). Thus, when the transition probabilities with editing effort 1 are such that inflow and outflow happen to be in balance, then there is no bias. When the editing effort in a particular industry is increased, the inflow component is decreased. However, there may still be an outflow component: units that are erroneously classified into another industry. Due to this effect the balance of in- and outflow may become disturbed when increasing the editing effort, because the editing effort is not done over all industries in the economic domain but only for a limited number of target industries. In the fixed allocation, when units are divided evenly over the nine car trade industries the effect of an imbalance of in- and outflow is limited. In the pro-rata allocation however, the number of edited units varies by industry, which leads, in some of the industries, to the mentioned imbalance. Details of the results can be found in Van Delden et al. (2015).
Main findings (lessons learnt)	
Advantages	Accuracy (RMSE) as affected by classification errors, can be improved with increased editing effort using a simple scenario: a fixed number of enterprises per industry.
Disadvantages	Manual editing is a very costly activity and the fixed allocation is not very effective: the RMSE decreased only gradually with increased editing effort.
Gap analysis	We expect that there exists an editing scenario for classification errors that is more effective than the fixed allocation. This more effective scenario still needs to be developed.
Other remarks	The method can be extended to study the effect of other types of (non-sampling) errors and also to other configurations.
References	Burger, J., A. van Delden, and S. Scholtus (2015). Sensitivity of Mixed-Source Statistics to Classification Errors. Journal of Official Statistics 31: 489–506 Delden, A. van, Scholtus, S. and J. Burger (2016). Accuracy of mixed-source statistics as affected by classification errors. Journal of Official Statistics 32, 619-642.

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-6: Dutch Population and Housing Census

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-6: DUTCH POPULATION AND HOUSING CENSUS

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Luxembourg, February, 2017

Title of the case study:

Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages (Deliverable 1)	Indirect estimation
Statistical Tasks (Deliverable 2)	Integrate microdata with macrodata
Data configurations (Deliverable 1)	6
Methods (Deliverable 5)	Repeated weighting

Presentation of the Case study

Agency – country	Statistics Netherlands
Topic	Population and Housing Census
Description of the problem	Statistics Netherlands pursues a 'one-figure' policy for the Dutch Census. With this 'one-figure' policy Statistics Netherlands aims to publish only one estimate for the same phenomenon occurring in different tables of the Dutch Census. The Dutch Census is a virtual census in the sense that is based on a number of administrative data sources and surveys rather than on a complete enumeration of the population. To enforce the 'one-figure' policy population totals, either known from an administrative data source or previously estimated, are imposed as benchmarks, provided they overlap with an additional survey dataset that is needed to produce new output statistics.
Input data	Various administrative data sources, and the Labour Force Survey
Expected output	Frequency tables

Technical summary

Methods	Repeated weighting has been developed at Statistics Netherlands in the late 1990s in order to enforce the 'one-figure' policy for the Dutch Census. In this approach a separate set of weights is assigned to sample units for each table of population totals to be estimated. The tables to be estimated for the Dutch Census are estimated sequentially and each table is estimated using as many sample units as possible in order to keep the sample variance as low as possible. The combined data from administrative data sources and surveys are divided into rectangular blocks. Such a block consists of a maximal set of variables for which data on the same units has been collected. The data blocks are chosen such that each table to be estimated is covered by at least one data block.

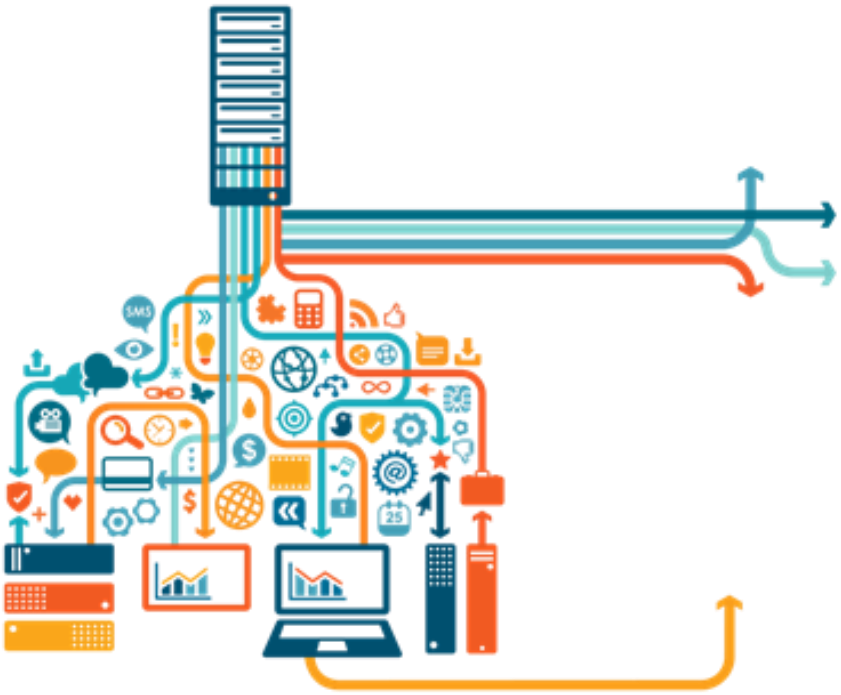
	Data from an available administrative data source covering the entire population can simply be counted. Data only available from surveys are weighted by means of regression weighting (see Särndal, Swensson and Wretman 1992). In that case weights must be assigned to all units in the block. For a survey one usually starts with the inverse inclusion probabilities of the sample units, corrected for response selectivity. These weights are then further adjusted by calibrating them to previously estimated totals. For a data block containing the overlap of two surveys, one usually begins with the product of the standard survey weights from each of the surveys as starting weight for each observed unit, and then corrects these starting weights by calibrating to totals known from administrative data sources and previously estimated totals. When estimating a new table, all margins of this table that are known or have already been estimated for previous tables are kept fixed to their known or previously estimated values, i.e. the regression weighting is calibrated on these known or previously estimated values. This ensures that the margins of the new table are consistent with previous estimates. For more on repeated weighting, see Houbiers (2004).
Application	The backbone of the Dutch census is the central population register (PR), which combines all the municipal population registers. In the 2011 census, PR data for 1 January 2011 were used as the basis for the set of tables that needs to be published. The tables focus on frequency counts, not on quantitative information. Data not available or derivable from the PR were taken from other registers. All register variables are now available from Statistics Netherlands' System of social statistical datasets (SSD), and their quality has been improved by applying micro-integration techniques. Micro-integration entails checking the data and adjusting those that are incorrect. It is widely assumed that micro-integrated data provide more reliable results, as they are based on a maximum amount of information. They also provide better coverage of subpopulations: if data are missing in one source, another source can be used. In the 2011 Census, only two variables were not taken from a register: 'occupation' and 'educational attainment'. Records from the Labour Force Survey (LFS) in a three year period around the enumeration date (1 January 2011) were used to estimate values for these two variables, which are included in 23 of the 60 tables. Table consistency was guaranteed by using repeated weighting for these 23 tables.
Main findings (lessons learnt)	
Advantages	• Estimates obey the 'one-figure' policy • Based on well-known weighting techniques
Disadvantages	• When repeated weighting is used, tables are reconciled sequentially, which may lead to a suboptimal solution. • Several technical problems can occur, for instance when a survey does not

	contain any observation of a situation that is known to occur in the population. Another example of a technical problem is that certain logical relations may be violated, for instance that the estimated number of persons with a driver's license in a certain region is higher than the estimated number of persons that are eligible to have a driver's license in that region. See also Chapter 7 of Statistics Netherlands (2014) for more on the disadvantages of repeated weighting.
Gap analysis	Overcoming the technical problems of repeated weighting.
Other remarks	The method can be extended to study the effect of other types of (non-sampling) errors and also to other configurations.
References	Houbiers, M. (2004), Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics 20, pp. 55-75. Särndal, C.E., B. Swensson and J. Wretman (1992), Model Assisted Survey Sampling. New York: Springer-Verlag, 1992. Statistics Netherlands (2014), Dutch Census 2011: Analysis and Methodology. Report, Statistics Netherlands. (https://www.cbs.nl/en-gb/publication/2014/47/dutch-census-2011)

Specific contract n°000052 ESTAT
N°11111.2013.001-2016.038
under framework contract Lot 1
n°11111.2013.001-2013.252

eurostat

ESTIMATION METHODS FOR THE INTEGRATION OF ADMINISTRATIVE SOURCES – DELIVERABLE D4



3/20/2017

Template T4-7: Statistics on road accidents

Responsible person at Commission: **Fabrice Gras,** Eurostat – Unit B1

Written by: **Nicoletta Cibella, Ton de Waal, Marco Di Zio, Mauro Scanu, Sander Scholtus, Arnout van Delden, Tiziana Tuoto, Li-Chun Zhang**

Estimation methods for the integration of administrative sources – deliverable D4

TEMPLATE T4-7: STATISTICS ON ROAD ACCIDENTS

Project manager: Laurent Jacquet
Sogeti Luxembourg: Sanja Vujackov

FOREWORD

Task 4: Literature review presenting one actual example in the NSIs for each of the type of use of administrative sources and for the steps that have been previously identified

This activity aims at providing examples of actual usages of statistical estimation methods when using administrative sources based on NSIs experiences. The examples are referred to the usages and steps identified in Tasks 1 and 2.

For each identified example is provided an executive summary presenting the method, the contextual framework and the main results.

ISTAT experts: Marco Di Zio, Nicoletta Cibella, Mauro Scanu, Tiziana Tuoto
CBS experts: Ton de Waal, Arnout van Delden and Sander Scholtus
with help of Li-Chun Zhang

Deliverable 4. An in-depth inventory of the data needs of EUROMOD for each country. Specific contract n°000052 ESTAT N°11111.2013.001-2016.038 under framework contract Lot 1 n°11111.2013.001-2013.252
Luxembourg, February, 2017

Title of the case study:
Estimating classification errors under edit-restrictions in combined register-survey data

Executive summary

Usages (Deliverable 1)	Direct tabulation
Statistical Tasks (Deliverable 2)	Creation of joint statistical micro data
Data configurations (Deliverable 1)	1, 4, 5
Methods (Deliverable 5)	Probabilistic record linkage

Presentation of the Case study

Agency – country	Istat, Italian National Institute of Statistics - Italy
Topic	Statistics on victims and injured people of road accidents
Description of the problem	Istat carries out the "road accidents survey", which collects all road accidents resulting in deaths (within the 30th day) or injuries, involving at least a vehicle circulating on the national road net and documented by a Police authority or military corps. The survey is an exhaustive and monthly based data collection, achieved by Istat, with the cooperation of several different other public national institutions. The data collection system is adapted to the local level organisation and needs, so the collection of information is usually variable at local level. The unit of analysis is the single accident and information collected, such as circumstances and data on involved vehicles and dead or injured persons, are referred to the moment when the accident occurred. <u>Characteristics of the task:</u> 1. Duplications of events: that is, a. to collect twice data referred to events occurring at the end of each month, which should be registered the month later. b. the several, independent data providers (local police authorities and/or military corps) could duplicate the same event. 2. Lack of information: the event identification task is complicated by the serious circumstances in which data are collected, that is the presence of victims and injured people. Some data on personal information (names, ages) are not correctly provided or not provided at all. 3. Different time and unit references: monthly data are compared yearly with the data on mortality by cause, from the Italian National Vital Statistics Death Registry on causes of death managed by Istat, in order to achieve a complete and enriched list of victims occurred on the national roads. Obviously, the two data sources are only partially

	overlapped. The comparison is further complicated by the different reference units in the two sources: "accident" in the road accidents survey and "death" in the latter source. Furthermore, the possible different reference time of the accident and death could make problematic the detection of the same events , together with the lack of personal information, as described above. Characteristics of the expected output: The expected output of the integration of data sources is a complete list of deaths and injured people in road accidents, involving at least a vehicle circulating on the national road net and documented by police authorities, provided with some characteristics of both the roads where the accident happened and the involved people. This list allows direct tabulation of counts of mortality by this specific cause and further studies on dangerousness of both roads and individual behaviours in relation with some characteristics.
Input data	Administrative data on road accidents coming from police authorities; administrative data on deaths by cause of death Registry
Expected output	Counts, micro-data for studying relationship between events and their determinants.

Technical summary

Methods	Due to the characteristic of the data, that is missing and uncorrected values, the data integration is realised by the probabilistic record linkage, according to the classical theory of Fellegi and Sunter (1969). Shortly, given two data sets A and B of size N_A and N_B respectively, let us consider $\Omega=\{(a,b), a\in A \text{ and } b\in B\}$ of size $N=N_A x N_B$. The linkage between A and B can be defined as a classification problem: the pairs that belong to Ω should be assigned to two subsets M and U independent and mutually exclusive, such that: M is the set of matches $(a=b)$ and U is the set of non-matches $(a\neq b)$. In order to classify the pairs, K common identifiers (matching variables) have to be chosen so that, for each pairs, a comparison function is applied and a comparison vector γ is obtained. The ratio r between the probabilities of γ given the pair (a,b) membership either to the subset M or U, say respectively m and u, is used to classify the pair. The probabilities m and u are often estimated by means of the EM algorithm. Therefore all the pairs can be ranked according to their ratio r and a classification criterion based on two thresholds T_m and T_u $(T_m > T_u)$ is applied. More precisely, those pairs for which r is greater than T_m can be considered as linked; those pairs for which r is smaller than T_u can be considered as not-linked, if r falls in the range (T_m , T_u) no-decision is made and the pair is held out for the clerical review so to be solved. The thresholds are chosen so to minimize false match rate and false non-match rate.

Application	As example, we consider the probabilistic record linkage aiming at identifying death persons, between road accidents survey and cause of death register, in the Toscana region and the 2008 year. Road accident survey (RAS) consists of 291 records, while causes of death register (CDR) corresponds to 321 records, according to ICD-10 codes for motor vehicle traffic accidents on public roads. The variables useful for the linkage purpose in RAS are: name, surname, gender and age of the victim, day, month, municipality, province of the accident. The variables selected for the linkage purpose in CDR are: name, surname, gender and age of the dead person, day, month, municipality, province of the death. The difference in the reference units, that is the accident for the former set and the single person for the latter one, is mainly influential when an accident involves more than one person, because from the road accidents data is not possible to associate at individual level variables "name and surname" with variables "age and gender", due to the structure of the form for data collection. So, while the whole set of variables has a high identification power, the use of only a subset causes a serious loss. The variables "name and surname" have been preferred when an accident involves more than one dead person. In order to successfully apply probabilistic linkage methods, new linking variables are generated starting from the former ones. Note that variables "year" and "region" are equal for all records due to the starting selection; in fact, they can be considered as blocking variables in a standard search space reduction procedure. The size of the selected data sources do not requires further reduction procedures, and the cross product of all records can be considered. For the comparison between the values of "name and surname" a distance function, based on the Jaro string comparator, has been considered. Prior error rates are set in order to accept as matches those pairs with posterior linking probability greater or equal to 0.95 and to refuse as non-matches those pairs with posterior linking probability smaller or equal to 0.50. The described linkage process identifies 189 pairs as Matches and 14 pairs as Possible-Matches. A clerical review of the Possible-Matches suggests to accept 13 of them as Matches and to reject one of them. So, the whole linkage result proposes 202 matches. The associate errors, estimated from the model, are 0.054 and 0.002 for the false non-match rate and false match rate, respectively.
Main findings (lessons learnt)	
Advantages	Identifying the same units in different sources even in absence of unique identifiers. Treatments of missing and incorrect values in identifying variables.
Disadvantages	Some manual check are required to be sure to avoid false matches and missing matches. Without the availability of generalized tools for probabilistic record linkage, the knowledge of probabilistic record linkage

	methodologies should be considered a disadvantage. Fortunately, several tools are available for performing such integration procedure, e.g. Relais is particularly set to official statistics tasks (Relais, 2015).
Gap analysis	In order to evaluate cases missed by both sources, a capture-recapture approach should be applied, thanks to the availability of two independent sources.
Other remarks	The analysis can be extended to several years as well as to other administrative data sources, that is the hospital admission forms, so to create reliable and stable maps of road accident and injured people surveillances.
References	Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, 1183-1210 RELAIS, (2015). User's guide version 3.0, available at http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais Tuoto T., Valentino L., Baldassarre G., Bruzzone S., Cibella N. and Pappagallo M. "Towards an integrated surveillance system of road accidents in Italy" in Proceedings of XLVI Scientific Meeting of Italian Statistical Society, "Sapienza" University, June 2012