Swiss Structural Business Statistics: Data Harmonization for the Construction of Full-time Equivalents

Desislava Nedyalkova, Daniel Assoulin [1]

Abstract

From 2011 onwards, the Swiss Business Census is replaced by the new Swiss structural business statistics (STATENT). The construction of full-time equivalents (FTE) for STATENT is based on the integration of register and survey data. Register data comes from the OASI[2] social security register (SR) and from the business register (BR). Full-time equivalents (FTE) of employment by gender is an important target variable in STATENT which measures the work capacity of an enterprise. In the social security register FTE are not directly available and must be constructed by means of a linear prediction model applied on matched data (survey-registers) and with explanatory variables issued from the registers. However, we observe inconsistency in employment variables coming from the different sources. These differences are treated in order to make FTE coherent with SR data. A first method which treats the differences by a simple ratio adjustment was applied. Our analyses has shown that this is not the optimal solution. Another approach was thus developed. This method permits to treat the differences by taking into account information about employment.

Key Words: data harmonization, administrative data, survey data, model prediction

1. Introduction

The Swiss Business Census (BC), held for the last time in 2008, played an important role for producing various statistics on the structure of the Swiss economy. For the reference year 2011 it was replaced by an integrated system called STATENT (Swiss Structural Business Statistics). STATENT is mainly based on the business register (BR), the social security register (SR) and complementary surveys like the Quarterly Survey of Employment (JobStat).

The transition from the BC to STATENT induces several changes in definition and methodology. The principal differences concern the covered units, the definition of employment [3] and the periodicity. For instance, the BC was conducted every 3-4 years whereas STATENT appears each year. The BC referred to an exact date, whereas STATENT refers to the last month of the reference year.

Another major difference between BC and STATENT is the way FTE are calculated. In the past, FTE were derived from the information in the BC (occupational levels) which is not available in STATENT. For this reason the construction of FTE is an important challenge for STATENT.

For enterprises not included in a complementary survey, FTE for STATENT are constructed using a linear prediction model based on explanatory variables coming from the

[1] Desislava Nedyalkova, Daniel Assoulin, Swiss Federal Statistical Office, Espace de l'Europe 10, 2010 Neuchtel, email: desislava.nedyalkova@bfs.admin.ch, daniel.assoulin@bfs.admin.ch.

[2] "The old-age and survivors' insurance, better known as OASI, is the main pillar of the Swiss social security system. Its aim is to compensate, at least partly, the reduction or loss of income from employment due to old-age and death." Source: http://www.zas.admin.ch/org/00723/00725/index.html?lang=en

[3] The BC counted employees that worked at least 6 hours per week in an enterprise or an establishment. In STATENT all persons working in an enterprise (as wage-earner or as independent) and paying their mandatory contributions to the OASI for a minimum annual wage of CHF 2 300 are counted for (criteria for reference year 2011)

register. This model is fitted on matched data coming from the register and some complementary surveys. In case that an enterprise has FTE collected from a complementary survey, these FTE will be in principle used in STATENT. The integration of data coming from different sources reveals the existence of inconsistencies regarding the number of employees. In such cases, FTE coming from the survey need to be adapted in order to reflect the employment data from the register.

We begin by describing the FTE model, based on matched data available in survey and registers. Next, we present two different methods for treating inconsistencies between the different sources. We describe a first approach based on a ratio adjustment and show how this approach is employed in the FTE model. Then, we present a second approach in which differences are treated by taking into account information about employment. Finally we show how FTE for STATENT are constructed.

2. FTE model

For the construction of the model, we used survey data (fourth quarter of 2011) matched with register data on the enterprise level. These are mainly single-establishment enterprises (EUNT) for which we know FTE, annual standardized wages and some other variables like NUTS2 region and NACE. The model is estimated separately for the subpopulation of men (m) and women (w) in each of the two economic activity sectors (s2 and s3). For the sake of simplicity, the same notation is used to describe the estimated models.

On the basis of the information contained in the survey about occupation levels and on the salary distribution in the register, we construct, for each NACE section, four salary classes. These classes form the basis for the construction of the explanatory variables of the model. The variable of interest is the number of FTE.

The model we want to estimate is the following:

$$y_i = \alpha_1 . V_{i1} + \sum_{j=2}^{4} \alpha_{jkl} V_{ij} + \epsilon_i, \tag{1}$$

where:

- y_i, is the number of FTE of an enterprise i,
- V_{ij}, the number of employees of enterprise i in the salary class j $(j = 1, ..., 4)$ ($\sum_j V_{ij} =$ EMPTOT_R$_i$, total number of employees in enterprise i according to the register)
- α_1, regression coefficient for V_{i1},
- $\alpha_{jk\ell}$, regression coefficient for V_{ij} in NUTS 2 k $(k = 1, ..., 7)$ and NACE section ℓ,
- ϵ_i, residual with $E(\epsilon_i) = 0$ and Var$(\epsilon_i) = \sigma^2$EMPTOT_R$_i$.

3. Harmonization of employment variables

Matched data on which FTE are estimated present certain differences between the number of employees from the survey (EMPTOT_S) and those from the register (EMPTOT_R). There exist different methods for variable adjustment which can be used to treat these differences, e.g. *prorating* and *generalized ratio adjustment* (Panekoek, 2011; Panekoek, 2014). For instance, the *prorating* method represents a simple multiplicative adjustment which is applied on variables employed in control rules. In what follows we present the different treatments that we have done in order to overcome inconsistency.

3.1 First approach

Based on the methods described above, for each enterprise i, we define a new variable FTE_R (by gender) as follows:

$$\text{FTE\_R}_i = \eta_i \text{FTE\_S}_i, \tag{2}$$

where $\eta_i = \text{EMPTOT\_R}_i / \text{EMPTOT\_S}_i$ and FTE_S_i is the survey FTE.

This new variable, *harmonized with the register*, will be used for modelling. In this way predicted FTE will be consistent with the values of EMPTOT_R. If the inconsistencies are treated according to Equation (2), we can rewrite the equation of Model (1) as follows:

$$\eta_i \text{FTE\_S}_i = \alpha_1 . V_{i1} + \sum_{j=2}^{4} \alpha_{jkl} V_{ij} + \epsilon_i, \tag{3}$$

or

$$\text{FTE\_S}_i = \alpha_1 . \frac{V_{i1}}{\eta_i} + \sum_{j=2}^{4} \alpha_{jkl} \frac{V_{ij}}{\eta_i} + \frac{\epsilon_i}{\eta_i}, \tag{4}$$

In the case $\eta_i > 1$, we can interpret Equation (4) as follows: The adjustment between EMPTOT_R and EMPTOT_S is done uniformly in the four salary classes by reducing the number of employees by η_i. This procedure can be justified in the case where the inconsistences are independent of the salary classes.

3.2 New approach

We present an alternative of Model (2) in which inconsistences in the variables number of employees are not treated uniformly in the different salary classes. We will examine the problem for the following cases:

- Case 1: EMPTOT_R $>$ EMPTOT_S.
- Case 2: EMPTOT_S $>$ EMPTOT_R.
- Case 3: EMPTOT_S $=$ EMPTOT_R.

Let diff_ab $=$ EMPTOT_R $-$ EMPTOT_S (by gender) and diff_ba $=$ EMPTOT_S $-$ EMPTOT_R (by gender).

3.2.1 Treatment of Case 1

We suppose that EMPTOT_R $>$ EMPTOT_S. Knowing the number of employees in each salary class, we estimate the coefficients of the following model (by gender and economic activity sector):

$$\text{diff\_ab}_i = \sum_{j=1}^{4} \beta_j V_{ij} + \epsilon_i,$$

under the hypothesis $\text{Var}(\epsilon_i) = \sigma^2 \text{EMPTOT\_R}_i$. This is not done with the aim to predict the difference between EMPTOT_R and EMPTOT_S. We are rather interested in the estimated coefficients, $\widehat{\beta}_j$, which can be interpreted as an estimation of the proportion of persons in the salary class j which are in the register but not in the survey.

Table 1: Values of $\widehat{\beta}_j$ and the estimated standard errors

	$\widehat{\beta}_1$		$\widehat{\beta}_2$		$\widehat{\beta}_3$		$\widehat{\beta}_4$	
	Estimator	StdErr	Estimator	StdErr	Estimator	StdErr	Estimator	StdErr
m/s2	0.659	0.063	0.562	0.016	0.153	0.019	0.039	0.004
m/s3	0.808	0.025	0.624	0.012	0.329	0.012	0.088	0.003
w/s2	0.724	0.023	0.495	0.015	0.241	0.013	0.051	0.006
w/s3	0.693	0.012	0.390	0.009	0.125	0.006	0.129	0.004

Table 1 contains the estimated coefficients, $\widehat{\beta}_j$, as well as their standard errors, obtained with the ROBUSTREG procedure in SAS with weights proportional to $1/\text{EMPTOT\_R}_i$. It indicates, for example, that the coefficients for the class of employees having the smallest wages (salary class 1) are larger than the coefficients for the salary class 2. We would conclude that the proportion of employees not counted in the survey is higher for the salary class with the small wages. Thus, at least in this subpopulation (Case 1), a uniform treatment using Equation (2) seems not to be appropriate.

Initial idea - probability proportional to size (PPS) sampling of fixed size

For each enterprise i, we suppose that the persons which are not in the survey come from a sample s_i of fixed size $n_i = \text{diff\_ab}_i$ of probabilities proportional to $\widehat{\beta}_j$ (see Table 1). The probability that a person d belonging to the set of wage-earners in the salary class j is not counted in the survey is therefore given by:

$$P(d \in s_i) \quad = \quad n_i \frac{\text{mos}_d}{\text{mos}_i} = \frac{n_i \widehat{\beta}_j}{\sum_{j=1}^4 \widehat{\beta}_j V_{ij}} = \widehat{\beta}_j \frac{n_i}{\widehat{n}_i} = \widehat{\beta}^*_{ij},$$

where $\text{mos}_d = \widehat{\beta}_j$ and $\text{mos}_i = \sum_{d \in i} \text{mos}_d = \sum_{j=1}^4 \widehat{\beta}_j V_{ij} = \widehat{n}_i$.

If $\widehat{\beta}^*_{ij} \geq 1$, then the person will be automatically removed (Srndal et al. (1992, p.89)).

We should note that $\widehat{\beta}^*_{ij}$ can be seen as a $\widehat{\beta}_j$ adjusted so that $\sum_{j=1}^4 \widehat{\beta}^*_{ij} V_{ij} = n_i$.

Calculation of the average number of persons which should be removed in each salary class

Some inconveniences of using PPS sampling of persons are its random aspect with a potential impact on comparability over years and the difficulty of implementation in production (matched data is on enterprise and not on person level). This has led us to develop a general procedure to replace the random PPS sample. Instead of drawing a sample we calculate the expected number of persons in class j to be selected in the sample s_i. This number is given by:

$$E(\sum_{d \in \mathcal{V}_{ij}} 1(d \in s_i)) \quad = \quad V_{ij} P(d \in s_i) = V_{ij} \widehat{\beta}^*_j. \tag{5}$$

where $\mathcal{V}_{ij}$ denotes the set of employees of enterprise i in the salary class j.

As in the case of PPS sampling, our procedure first removes all persons for which $\widehat{\beta}^*_{ij} >= 1$. Next, we calculate the mean number of persons which have to be eliminated according to Equation (5). At the end of this procedure we obtain new variables:

$$\widetilde{V}_{ij} = V_{ij} - V_{ij}\widehat{\beta}^*_{ij},$$

such that $\sum_{j=1}^{4} \widetilde{V}_{ij} = \text{EMPTOT\_S}_i$. These new variables will replace the variables V_{ij} in the estimation of Model (1) where y_i will be given by FTE_S_i.

3.2.2 Treatment of Case 2

We suppose that EMPTOT_S > EMPTOT_R. Knowing the number of employees working at part time III (T_{i1}), part time II (T_{i2}), part time I (T_{i3}) and full time (T_{i4}) from the survey data, we estimate the following model (by gender and economic activity sector):

$$\text{diff\_ba}_i = \sum_{j=1}^{4} \gamma_j T_{ij} + \epsilon_i,$$

The used procedure is PROC ROBUSTREG in SAS with weights proportional to $1/\text{EMPTOT\_S}_i$. The estimated coefficients and their standard errors are given in Table 2. These estimated coefficients can be seen as estimation of the proportion of persons working at part time III, for example, that are in the survey but not in the register. It can be seen from the table that the coefficients for the persons working at part time III are larger than the coefficients for the persons working full time.

Table 2: Values of $\widehat{\gamma}_j$ and the estimated standard errors

	$\widehat{\gamma}_1$		$\widehat{\gamma}_2$		$\widehat{\gamma}_3$		$\widehat{\gamma}_4$	
	Estimator	StdErr	Estimator	StdErr	Estimator	StdErr	Estimator	StdErr
m/s2	0.469	0.039	0.657	0.041	0.256	0.032	0.071	0.002
m/s3	0.448	0.014	0.237	0.011	0.203	0.014	0.123	0.003
w/s2	0.548	0.025	0.329	0.019	0.063	0.018	0.065	0.007
w/s3	0.442	0.010	0.251	0.008	0.002	0.007	0.140	0.005

Adaptation of the harmonization procedure to Case 2

We know that in order to eliminate the differences between the employment variables EMPTOT_S and EMPTOT_R and make them consistent we have to eliminate, for each enterprise, a fix number of persons, $n^*_i = \text{diff\_ba}_i$. Using the coefficients $\widehat{\gamma}_j$, we apply the same procedure as for the Case 1, with the required modifications. In this way we obtain the new variables $\widetilde{T}_{ij}$ such that $\sum_{j=1}^{4} \widetilde{T}_{ij} = \text{EMPTOT\_R}_i$.

Let suppose that FTE_S can be modeled as follows:

$$\text{FTE\_S}_i = \sum_{j=1}^{4} \delta_j T_{ij} + \epsilon_i, \qquad (6)$$

Table 3: Values of $\widehat{\delta}_j$ and the estimated standard errors

	$\widehat{\delta}_1$		$\widehat{\delta}_2$		$\widehat{\delta}_3$		$\widehat{\delta}_4$		
	Est.	StdErr	Est.	StdErr	Est.	StdErr	Est.	StdErr	R^2
m/s2	0.110	0.006	0.300	0.007	0.632	0.005	0.998	0.000	0.999
m/s3	0.078	0.004	0.278	0.003	0.645	0.004	0.998	0.001	0.997
w/s2	0.095	0.008	0.293	0.006	0.654	0.006	0.994	0.002	0.997
w/s3	0.091	0.004	0.281	0.003	0.660	0.003	0.988	0.002	0.992

where $\text{Var}(\epsilon_i) = \sigma^2\text{EMPTOT\_S}_i$. This model, by gender and economic activity sector, is estimated using PROC GLM from SAS with weights proportional to $1/\text{EMPTOT\_S}$. The estimated coefficients and their standard errors are given in Table 3.

Using the estimated coefficients of Model (6), we calculate a new adjusted variable FTE_S, denoted by FTE_R, which is coherent with the variable EMPTOT_R:

$$\text{FTE\_R}_i = \min\left(\text{FTE\_S}_i \frac{\sum \widehat{\delta}_j \widetilde{T}_{ij}}{\sum \widehat{\delta}_j T_{ij}}, \text{EMPTOT\_R}_i\right). \tag{7}$$

We can explain the minimum in Equation (7) by the fact that mean occupational level (MOL) of harmonized data, $\text{FTE\_R}/\text{EMPTOT\_R}$, should be bounded by 1.

3.3 Variables used to estimate the FTE model after data harmonization

Table 4 presents the variables used in the estimation of the model of Equation (1) in cases 1, 2 and 3, respectively. From the table, it can be seen that in case 1 it is the explanatory variables that are adjusted in order to correspond to FTE recorded in the survey. In case 2, where the total number of employees is larger for the survey than for the register, the explanatory variables remain unchanged but the dependent variable for the model is adjusted. The three data sets were set together and unique dependent and independent variables were created.

Table 4: Variables used in model 1

Case	Variable of interest	Explanatory variables
1	FTE_S	$\widetilde{V}_{ij}$
2	FTE_R	V_{ij}
3	FTE_S	V_{ij}

3.4 Consequences of data harmonization for constructing FTE in STATENT

In this section, we show how we calculate FTE for enterprises for which survey data is available. For all the other enterprises, the prediction model 1 and register data are used.

- Case 1: Let us denote

$$\text{FTE\_R}_{i,model} = \sum_{j=1}^{4} \widehat{\alpha}_j V_{ij}$$

the FTE calculated using the estimated FTE model parameters and the variables V_{ij} and

$$\text{FTE\_S}_{i,model} = \sum_{j=1}^{4} \widehat{\alpha}_j \widetilde{V}_{ij}$$

the FTE calculated using the FTE model parameters and the variables $\widetilde{V}_{ij}$. Then, the ratio of these variables is applied to the FTE_S in order to produce the harmonized FTE, denoted by FTE_R. Thus,

$$\text{FTE\_R}_i = \min\left(\text{FTE\_S}_i \frac{\sum_{j=1}^{4} \widehat{\alpha}_j V_{ij}}{\sum_{j=1}^{4} \widehat{\alpha}_j \widetilde{V}_{ij}}, \text{EMPTOT\_R}_i\right). \tag{8}$$

The minimum is explained by the fact that mean occupation level of harmonized data should be bounded by 1.

- Case 2: the harmonized FTE_R is calculated by applying a multiplicative or enhanced ratio adjustment as defined in Equation (7).
- Case 3: in this case we have that FTE_R $=$ FTE_S.

  To summarize: the simple ratio adjustment based on the number of employees is replaced by an enhanced ratio adjustment using full-time equivalents.

Note that an enterprise for which the variables EMPTOT present extreme differences either in the subpopulation of men or in the subpopulation of women will be treated as if survey data is missing.

3.5 Effects of the harmonization

We illustrate the effects of harmonization on the data by a few examples taken from real data. Figure 1 shows a comparison between the mean occupation levels of survey and OASI after harmonization for Case 1. From the graph it can be seen that in general the mean occupational levels based on the enhanced ratio adjustment (mol_harm) are smaller than the mean occupational levels based on the simple ratio adjustment (mol_ratio) defined in Section 3.1. This can be explained by the fact that the surplus of employment in the OASI corresponds rather to small FTE while the simple ratio adjustment is based on the hypothesis that mean occupational levels in OASI are the same as in the survey (the values on the diagonal of the graph).

Next, Table 5 shows the values of $\widetilde{V}_{ij}$ for a given enterprise. For the second economic sector, model for women, the estimated FTE model coefficients are respectively, $\widehat{\alpha}_1 = 0.223$, $\widehat{\alpha}_2 = 0.364$, $\widehat{\alpha}_3 = 0.601$ and $\widehat{\alpha}_4 = 0.940$. For this particular example it can be seen that the enterprise has 9 employees according to the register (variable EMPTOT_R) and 5 employees according to the survey (variable EMPTOT_S). Value of FTE according to the survey (FTE_S) is equal to 4. This enterprise falls in the Case 1 for data treatment. Most of the employees are in the first salary class. Proportionally we will eliminate most of the employees in the first salary class ($\widehat{\beta}_1 = 0.7239$). The values of the harmonized FTE for this enterprise are given in Table 6. For instance, FTE_R$_{i,model} = 3.96$ and FTE_S$_{i,model} = 4.915$. The ratio of these two values applied to the survey FTE gives the value of the harmonized FTE using the enhanced adjustment as given in Equation (8). We can see from Table 6 that the harmonized FTE (enhanced adjustment) is much more in accordance to the value of FTE_S than the one using the simple ratio adjustment (according to Equation (2) and that the mean occupational level is smaller after harmonization. Thus, this result is consistent with the results of Figure 1.

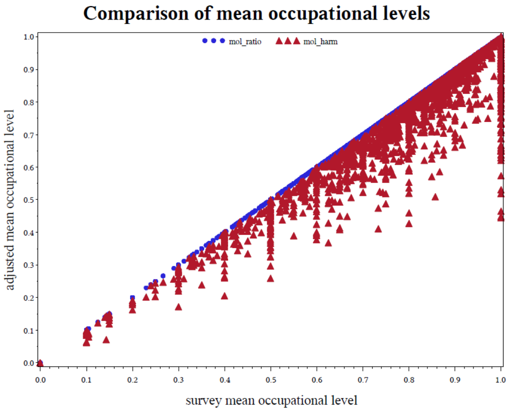

Figure 1: Mean occupational levels, women /sector 2

Table 5: Number of employees per salary class

V_1	V_2	V_3	V_4	Sum
4	0	2	3	9
$\widetilde{V}_1$	$\widetilde{V}_2$	$\widetilde{V}_3$	$\widetilde{V}_4$	Sum
0.718	0	1.455	2.827	5

Table 6: Effects of harmonization on FTE

	Source		Adjustment	
	Survey	SR	simple	enhanced
EMPTOT	5	9		
FTE	4		**7.2**=$4 \times \frac{9}{5}$	**5.325**=$4 \times \frac{4.915}{3.692}$
FTE model	3.692	4.915		
MOL			**0.8**	**0.59**

4. Conclusion

This paper presents the methodology used to treat the inconsistences between the different data sources used for the construction of FTE for the Swiss Structural Business Statistics. The analyses put into question the application of a simple ratio adjustment according to Definition (2). The divergences seem rather to be due to low wages or small occupational levels. The presented harmonization based on PPS sampling takes into account information on the employment type (salary class, occupation level). However, this procedure has the inconvenience to be random and difficult to apply in practice. So, we use expected sample sizes instead of random sampling for adjusting the number of employees in the different salary classes, in order to overcome this inconvenience. This new approach was tested and used for STATENT 2011.

REFERENCES

Pannekoek, J. (2011), “Models and algorithms for micro-integration,” in *Memobust Handbook on Methodology of Modern Business Statistics*, http://www.cros-portal.eu/content/wp2-development-methods.

Pannekoek, J. (2014a), “Method: Reconciling Conflicting Microdata,” in *Memobust Handbook on Methodology of Modern Business Statistics*, http://www.cros-portal.eu/content/reconciling-conflicting-microdata-method.

Pannekoek, J. (2014b), “Method: Prorating,” in *Memobust Handbook on Methodology of Modern Business Statistics*, http://www.cros-portal.eu/content/prorating-method.

Pannekoek, J. (2014c), “Method: Generalised Ratio Adjustments,” in *Memobust Handbook on Methodology of Modern Business Statistics*, http://www.cros-portal.eu/content/generalised-ratio-adjustments-method.

Srndal, C.E., Swensson, B. et Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer.