Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Home Affairs FDHA
Federal Statistical Office FSO
Statistical Methods

Construction of Full Time Equivalent for the Swiss Business Frame

Monique Graf and Jann Potterat,
Statistical Methods Unit, Swiss Federal Statistical Office, Neuchâtel

1 Introduction

Business statistics in Switzerland face a paradigm shift. The business census was held for the last time in 2008. It will now be replaced by the use of registers and complementary surveys. The two main sources are the business register (BR) that provides information at the enterprise and local unit levels, and the social security register (SR) that provides information at the enterprise level only. The record linkage (on the basis of the enterprise name and address) between BR and SR is ongoing and should be finished by the end of 2011. The business register will record the new businesses and update the economic activity. The social security register will provide information about gender, employment and wages at the level of the enterprise. A survey called “profiling” will (among others) allocate employment to local units for enterprises with more than one local unit.
The employment data recorded in SR are the months worked per employee. From this information, it is possible to deduce the number of employees per month and gender, but not the corresponding full-time equivalents (FTE). Full-time equivalents per gender (FTE) will be reconstructed using a model based on the combined register and results of the Quarterly Survey of Employment (JobStat). This survey gives the total employment and FTE by gender for approximately 36’000 enterprises. The model will be first applied for production in 2013 on the 2011 data.
The purpose of the presentation is to describe the foreseen model, and methods in place for its validation.

2 Method

2.1 Principle

Remark: The method described below will not be used for the very large companies. In their case, a direct contact will furnish the necessary data.

Let us call “register” the result of the record linkage of the business and the social security registers.

1. Information from the register: Predictive variables will be characteristics of the enterprise’s wage distribution, complemented by its economic activity and regional information. This gives rise to a set of monthly explanatory variables for each enterprise in the register. Notice that only one out of 3 months will be used (in phase with the Quarterly Survey of Employment).

Federal Statistical Office FSO
Monique Graf
Espace de l'Europe 10, CH-2010 Neuchâtel
Phone +41 32 71 36615, Fax +41 32 71 36093
Monique.Graf@bfs.admin.ch
www.statistics.admin.ch

2. Information from the Quarterly Survey of Employment (Jobstat): For the enterprises recorded in the survey, we couple the employment data (full time equivalents and total employment) with the information from the register on the month containing the JobStat reference date.
3. The total monthly employment per gender is recorded both in the register and in the survey. We take advantage of the information: instead of predicting the full-time equivalents directly, we use the model to predict the ratio of the FTE to the total employment, i.e. the average occupation level per company. Then we obtain the predicted full-time equivalent by multiplication with the total employment recorded in the register.

This program is detailed below.

2.2 Estimation of company's monthly wage distribution from the register

Information from the register:

Company level: company's economic activity and region,

Employee level: working months, annual salary and gender.

From there, we can estimate an average monthly wage for each employee (Table 1).

Table 1: Example of the wage computations at the employee level

Employee λ	J	F	M	A	M	J	J	A	S	O	N	D	Annual wage bill (thousand CHF)	Monthly wage (thousand CHF)
Working months	1	1	1	1	1	1	0	0	0	0	0	0	30	5
Estimated wage	5	5	5	5	5	5	0	0	0	0	0	0		

Working months : 1= yes ; 0= no. We suppose (for lack of anything better) that the monthly salary is the same each month of the year. This employee contributes for 1 unit to the total employment in the months Jan to June, and 0 unit in Jul to Dec.

In this way, we construct the total monthly employment and the monthly wage distribution by gender at the company level. Predictive variables will be characteristics of this wage distribution, complemented by the company's economic activity and regional information. In the prototype (see below), the wage distribution characteristics considered are the mean and standard deviation of wages at the logarithmic scale. The mean occupation level should be positively correlated with the mean log wage, so this last variable is a rather natural predictive variable. Contrary to the occupation level, wages are not bounded from above. To take this fact into account, we use the standard deviation of the log wages as a predictive variable. As expected its regression parameter is negative, so if there is a large discrepancy between salaries within the company, the large standard deviation will apply a downwards correction to the prediction given by the mean log salary.

2.3 Coupling with JobStat

The quarterly JobStat sample, giving the full time equivalent (FTE) and total employment (TOT) by gender, is coupled with the register data on the basis of the business identification number. This enhanced sample is the input for the model. A quality check on the comparability of the information provided by both sources will be provided by the comparison between the total employment by gender observed in JobStat (TOT, see Table 2) with the one recorded in the register (TOT').

2.4 Estimation (mass imputation) in the register

The result of the model will be a predicted value of the mean occupation level (MOL) by gender for each company in the register. The FTE estimate is the MOL multiplied by the known total employment (TOT' in Table 2).

Table 2: Model variables (company level) by quarter and by gender

	JobStat		Register		Common Variables	
Input Variables	FTE	TOT	TOT'	Wage distribution	NACE2	NUTS2 or NUTS3
	Dependent variable		Independent variables			
Model Variables	Mean Occupation Level (MOL) MOL = FTE / TOT		Mean, Standard Deviation and skewness of log wages		NACE2	NUTS2 or NUTS3

The question whether the predicted or the JobStat observed FTE value should be used for those companies answering the employment survey, raises an issue. From a purely modeling perspective, it is not desirable to mix observed and predicted values, because they do not have the same variability. On the other hand, from an official statistics point of view, it is hardly acceptable to replace an observed quantity by a synthetic one, when there is no reason to question the validity of the observation.

3 Feasibility study and validation

The Swiss Earnings Structure Survey (SESS) is a business survey that provides data on the occupancy rates and wages. The stratification is a combination of economic activities, NUTS2 and business size classes. In each stratum a two-stage cluster sample is selected: first a random sample of enterprises and within each enterprise a random sample of wages. The SESS makes it possible to test the feasibility of a FTE model with all the information on the same source. However, it cannot be the solution for updating the FTE because it is only a biennial statistics. Several models have been tested on the SESS data, using for the while the 2002 definition of the NACE. Below we present two of them. The SESS should also in the near future undergo a revision that (among others) will cover the social security definition of wages and extend the type of workers to include all those who are liable to pay the social security contributions. For the while, for instance, young workers on apprenticeship are not included. The advantage of the SESS over the foreseen solution using the register and JobStat is that all variables are provided by the same source. The tests with the SESS represent a somewhat simpler situation. If the results were not satisfying in the SESS context, then we should reconsider the whole process.

3.1 SAS procedure TRANSREG

In our model, the explanatory variables are a mix of categorical data (NACE, region, gender) and quantitative data (characteristics of the wage distribution). The dependent variable is the logarithm of mean occupancy level in the company. For estimation, our plans are to use the SAS procedure TRANSREG that allows dealing with categorical, ordinal and continuous data. The method behind is the generalized additive model, see e.g. Hastie and Tibshirani (1990). An extensive documentation on this procedure can be found in the SAS manuals (Kuhfeld, SAS Online Documentation) and details on the algorithms in Kuhfeld (1990). A similar method of estimation should be available in R, see Wood (2006). Among the many possibilities offered by TRANSREG, we used Fisher's optimal scoring for the categorical variables. This is implemented with the keyword OPSCORE. The chosen characteristics of the company's wage distribution (mean, standard deviation and skewness of log wages) are untransformed (keyword IDENTITY). Below, an example of our pretests using the SESS is presented.

3.2 Some results with the private sector

For the while, only the private sector was used in our tests. The following models are only pretests, because the gender is not introduced. Two models are considered here. Using the notations in Table 2, the models (either 1 or 2) are specified in Equation (1):

Table 3: Comparison of the benchmark FTE with 2 prototypes of the model

NACE2 (2002)	Net sample size (# businesses)	Extrapolated total (# businesses)	Extr. FTE benchmark	Group for Model 1	Extr. FTE Model 1	rel. diff. with benchmark %	Extr. FTE Model 2	rel. diff. with benchmark %
Total	42'315	157'458	2'217'796.7		2'239'798.2	0.992%	2'231'444.0	0.615%
10-14	187	253	3'879.9	2	3'851.0	-0.744%	3'862.4	-0.450%
15	926	1'634	49'230.7	1	49'515.9	0.579%	49'822.1	1.201%
16	9	9	2'668.5	2	2'660.6	-0.293%	2'655.3	-0.494%
17	306	391	9'425.9	2	9'445.4	0.206%	9'550.8	1.324%
18	206	296	2'752.3	1	2'669.3	-3.016%	2'716.4	-1.305%
19	75	104	1'449.5	1	1'362.6	-5.997%	1'383.4	-4.565%
20	701	3'240	24'385.2	1	24'408.8	0.097%	24'388.9	0.015%
21	157	191	11'374.0	2	11'335.9	-0.335%	11'439.9	0.580%
22	1'090	2'272	34'884.6	1	36'414.9	4.387%	36'571.5	4.836%
23,24	570	685	59'185.9	2	60'035.5	1.436%	60'308.8	1.897%
25	521	630	22'226.1	2	22'080.2	-0.656%	22'195.0	-0.140%
26	531	685	15'574.6	2	15'429.2	-0.933%	15'664.8	0.579%
27,28	1'401	4'770	83'312.5	2	82'703.8	-0.731%	83'181.8	-0.157%
29,34,35	1'590	2'615	108'042.5	2	108'526.9	0.448%	108'982.4	0.870%
30-32	938	1'185	55'299.4	2	54'715.1	-1.057%	54'951.3	-0.629%
33	1'383	1'897	75'796.0	2	75'565.9	-0.304%	75'901.8	0.140%
36,37	900	1'521	20'881.7	1	20'759.9	-0.583%	20'871.2	-0.050%
40,41	212	240	14'917.1	2	15'048.5	0.881%	15'129.1	1.421%
45	2'484	18'774	211'399.4	1	210'291.3	-0.524%	210'699.0	-0.331%
50	1'211	7'601	54'608.1	4	54'332.2	-0.505%	54'401.2	-0.379%
51	3'407	9'862	144'930.2	4	144'058.9	-0.601%	144'580.3	-0.241%
52	2'348	18'170	238'929.4	3	244'441.0	2.307%	238'064.5	-0.362%
55	2'242	19'797	137'819.1	3	138'642.6	0.597%	136'447.5	-0.995%
60	1'019	3'119	45'538.8	4	45'310.8	-0.501%	45'600.0	0.134%
61	58	80	1'597.3	4	1'581.3	-1.000%	1'580.5	-1.050%
62	97	107	6'737.9	4	6'584.8	-2.272%	6'528.0	-3.116%
63	1'215	1'636	40'644.2	4	40'807.1	0.401%	41'008.4	0.896%
64	223	319	29'085.9	4	30'307.3	4.199%	30'311.5	4.214%
65	981	1'117	105'086.8	4	109'090.2	3.810%	109'534.3	4.232%
66	194	237	46'224.7	4	46'123.9	-0.218%	46'118.1	-0.230%
67	1'422	1'753	16'781.9	4	16'625.6	-0.931%	16'650.8	-0.781%
70,71	1'378	2'200	19'956.9	4	19'672.3	-1.426%	19'869.2	-0.439%
72,74	4'185	25'013	233'216.6	4	230'277.9	-1.260%	230'742.5	-1.061%
73	227	287	8'867.7	4	8'894.0	0.297%	9'000.7	1.500%
80	1'415	1'918	29'620.8	4	29'750.2	0.437%	30'792.7	3.956%
85	2'896	14'359	186'733.5	3	200'976.5	7.627%	193'973.4	3.877%
90	239	316	3'601.5	4	3'546.4	-1.529%	3'550.9	-1.403%
91	1'204	1'617	19'491.5	4	19'425.8	-0.337%	19'661.6	0.873%
92	1'383	2'287	26'358.0	4	27'039.5	2.586%	27'440.6	4.107%
93	784	4'271	15'280.2	3	15'489.1	1.367%	15'311.5	0.205%
						min -5.997%		min -4.565%
						max 7.627%		max 4.836%
						median(abs) 0.744%		median(abs) 0.870%

$$\log(MOL) = \text{OPSCORE(NUTS2)} + \text{OSPCORE(NACE2)} + \text{IDENTITY(mean\_log\_wage stdev\_log\_wage skew\_log\_wage)} \quad (1)$$

The difference between the models is that in Model 1, the estimation is done separately in 4 groups (sectors x 2 NACE groups with high or low median wage), whereas in Model 2, the model specification takes the whole economy altogether. We used the sampling weights in the estimation. For Model 1, the four R^2 vary between 0.7168 and 0.8441 (better prediction for Industry than for Services); in Model 2, R^2=0.8006.

As said before, the predicted FTE is obtained by multiplication of the predicted MOL by the total employment in the company. In Table 3 the resulting predictions are extrapolated at the NACE2 level (with some groupings). The benchmark extrapolated FTE is defined as the extrapolated total of the actual occupancy levels corresponding to the surveyed wages. The relative difference (in %) between the extrapolated FTE and the benchmark is given for each model. We can see that the range of the relative errors when separate estimation in four groups is processed (Model 1) is slightly lower than in the simpler Model 2.However, the average R^2 of Model 1 is similar to the R^2 in Model 2.

4 Provisional conclusions

A FTE predictive model seems to be a sensible approach to the problem of mass imputation of full time equivalents at the micro-level, when wage information is provided. The tests with the SESS show a good predictive power (R^2 around 0.8). Note that there is no size effect (that would imply a large R^2), because the dependent variable is related to the mean occupancy level in the company, and no wage bill is used in the model. As a consequence, we can use the total employment recorded in the register: when multiplied by the predicted mean occupancy level, it furnishes the desired FTE prediction.

Both of our models contain a categorical explanatory variable with NUTS2 levels. We observed that distinct models applied to 4 groupings do not bring any significant improvement over a unique model. Even if it may be due to a poor choice of groupings, we think that it is a sign that separate models do not bring much improvement. Investigations using truncated wages for the computation of the wage distribution characteristics show similar results as those presented here. Maybe we stay one the safe side when using some kind of truncation.

Definitive conclusions cannot be drawn for the while, because the data that will be used in the actual imputation of the FTE do not yet exist. Nevertheless we are confident that the result will be satisfying.

References

Hastie, T.J. and R.J. Tibshirani (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability)

Kuhfeld, W.F. SAS OnlineDoc 9.1.3. The TRANSREG Procedure. http://support.sas.com/onlinedoc/913/docMainpage.jsp.

Kuhfeld, W.F (1990). SAS Technical Report R-108: Algorithms for PRINQUAL and TRANSREG Procedures. http://support.sas.com/kb/23/addl/fusion_23806_1_r108_59040.pdf

Wood, S.N. (2006). *Generalized Additive Models. An Introduction with R.* Chapman & Hall/CRC Texts in Statistical Science Series.