Federal Statistical Office of Germany April 2015
Roland GÜNTHER
roland.guenther@destatis.de

Eurostat's workshop on labour costs in Rome, May 2015

Item 1.4: Sampling (including precision requirements)

1 The precision requirements of the EU regulations on the Labour Cost Survey

Council regulation (EC) No 530/1999 of 9 March 1999 concerning structural statistics on earnings and on labour costs fixes in Article 10 "Quality" the – rather general – precision requirements for the Labour Cost Survey (LCS):

> "1. The national authorities shall ensure that the results reflect the true situation of the total population of units with a sufficient degree of representativity."

In Commission regulation (EC) 1737/2005 amending Regulation (EC) No 1726/1999 as regards the definition and transmission of information on labour costs the details of the transmission are laid down, including the populations to be covered by LCS in terms of region and economic activity:

> "Annex III – Transmission of data including breakdowns by economic activity, enterprise size class and country or region
> Three files are to be provided, corresponding to Tables A, B and C:
> — Table A contains national data (one record for each economic activity at the section and division levels of NACE Rev. 2)
> — Table B contains national data by size class (one record for each economic activity at the section and division levels of NACE Rev. 2, for each of the size classes)
> — Table C contains regional data at the NUTS 1 level (one record for each economic activity at the section and division levels of NACE Rev. 2, for each of the regions)."

For Germany, the most detailed breakdown for transmission is table C, with its cross-tabulation of the 82 divisions of NACE Rev. 2 sections B to S by the 16 NUTS1 regions of Germany. In Germany, we understand "the true situation of the total population of units" to apply to this detailed breakdown. Consequently, to fulfill the regulation the "sufficient degree of representativity" should be guaranteed for this detailed breakdown. By doing so, we fulfill the other, less demanding breakdowns of tables A and B automatically. Since “sufficient degree of representativity” is not further defined by the regulations, we apply our usual national standard regarding the accuracy of results for statistics of earnings and labour cost. The standard requires a coefficient of variation of less than 10% for the key estimates.

Further, national legislation on LCS defines a maximum national sample size of 34 000 enterprises. This is about 11% of the population of German enterprises with ten or more employees of NACE Rev. 2 sections B to S. This sample size had been shown empirically to be sufficient and a balance between data needs and burden to the respondents. For the LCS 2012 there was need to further reduce burden. So it was decided to not use the maximum number, but 32 000 enterprises only.

Finally, the statistician’s task is to distribute the sample size of 32 000 over the strata to meet the requirements of the regulations. This allocation should be optimal in the sense of

- sampling efficiency (maximum accuracy for the given sample size) and
- balance between the dimensions of data needs (the cells of tables A, B, C).

2 The principle of graded precision for sample allocation

In most cases of German business statistics the allocation of a national sample over sampling strata is done using the so-called principle of graded precision (*Prinzip der abgestuften Genauigkeit*). This principle is based on the idea, that the statistician gains control over the accuracy of the statistical results for subpopulations and places a certain scheme of accuracy as a function of the subpopulations’ size or importance. The subpopulations’ accuracy shall be measured by the coefficient of variation of a given key estimate. The coefficient of variation shall be a function of a given total of the subpopulation. The total shall be a measure of size or user need for the subpopulation, like number of employees or sum of turnover. Consequently, “important” subpopulations will be measured with greater accuracy than “unimportant” subpopulations. But also small or “unimportant” subpopulations will be measured with sufficient precision. Mathematically:

$$\varepsilon_h = \frac{Const}{Y_h^b} \qquad (1)$$

ε_h = coefficient of variation of the key variable of subpopulation *h*

Y_h = Total of size variable *Y* of subpopulation *h*

$Const$ = a positive, real constant

b = the grading parameter

The parameter *b* is key to the results of the allocation. It is to be fixed between 0 and 0.5. A value of zero yields an allocation of equal precision for all subpopulations. A value of 0.5 (and all means and variances equal) yields the known case of the optimal *Neymann* allocation. In practice *b* is fixed to a value in the interval [0.1, 0.3].

For the LCS the sample allocation is done in two steps. In both steps the principle of graded precision is used.

3 Step 1: sample allocation over NUTS1

The first step of sample allocation is done over the 16 regions of NUTS1. In the federal system of German official statistics the field work is done by the offices of the 16 regions (*Länder*). Therefore the allocation over the 16 regions means the allocation of the workload to the *Länder* offices too.

For step 1 we use the principle as follows:

Y_h = number of employees of the enterprises in region *h*

b = 0.3

ε_h = coefficient of variation of the total of *Y*

Therefore we can use for ε_h a simple function of regional sample size n_h:

(2) $$\varepsilon_h = \frac{\sqrt{\left(\frac{N_h}{n_h} - 1\right) \cdot N_h}}{N_h} \cdot CV_{Y,h}$$

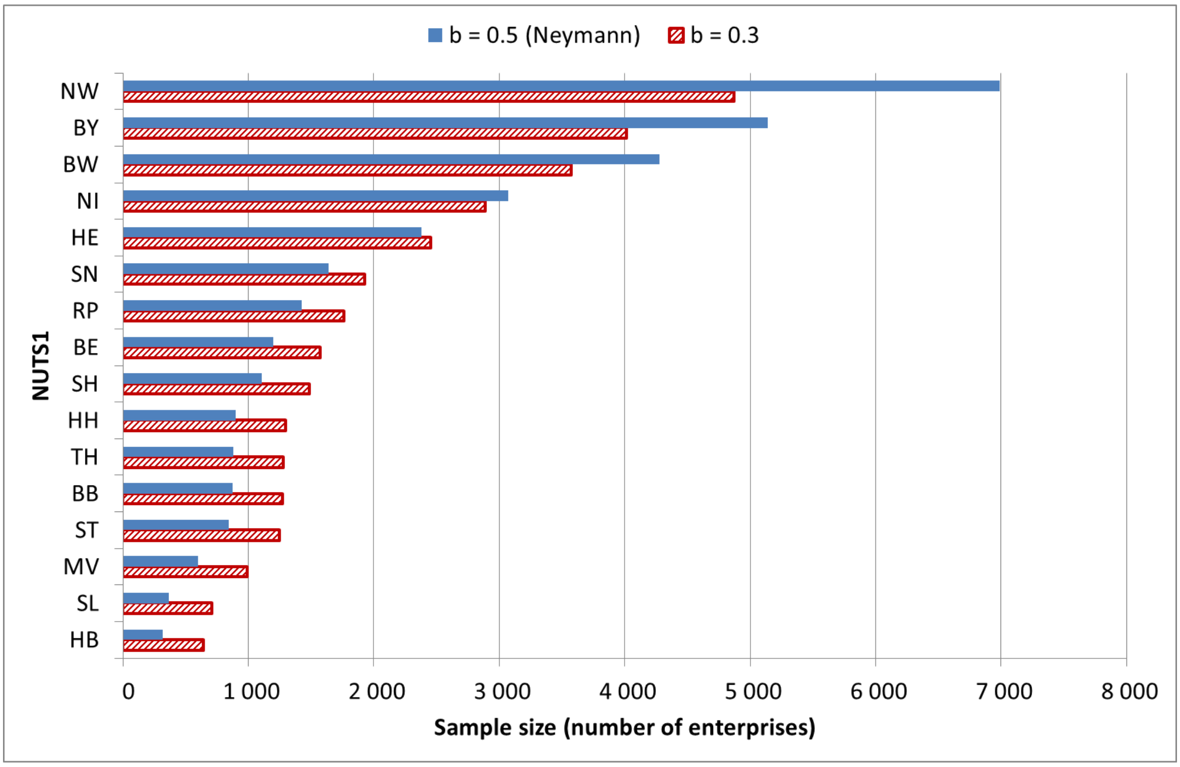
N_h = Number of sampling units (enterprises) in the population of region *h*

$CV_{Y,h}$ = coefficient of variation of *Y* in region *h.*

For simplicity we further assume equal $CV_{Y,h}$ for all regions *h*. The resulting mathematical problem is solved for n_h in a simple Microsoft Excel sheet by using Excel's solver tool.

Figure 1 shows the result of step 1 for the actually used *b* equal to 0.3 in comparison to the situation of a *b* equal to 0.5. The situation of a *b* equal to 0.5 is the well-known and widely used *Neymann* allocation. The sample allocation for LCS (and for many other business surveys in Germany) deviates from *Neymann* allocation in the sense of allocating sample size from the largest regions in Germany to the small and medium sized. This also means that we give regional statistics some more precision and we pay for it with somewhat less precise or less optimal federal results.

Figure 1: Step 1 of the sample allocation – sample size by NUTS1



4 Step 2: sample allocation over divisions and size classes

In step 2 each of the 16 regional samples is allocated to the final sampling strata built by 81 divisions of economic activities (the 82nd division "O84 public administration and defense; compulsory social security" is not covered by the LCS sample but another source) and 5 size classes.

For step 2 we use the principle in the same way as for step 1 but with another value of *b*:[1]

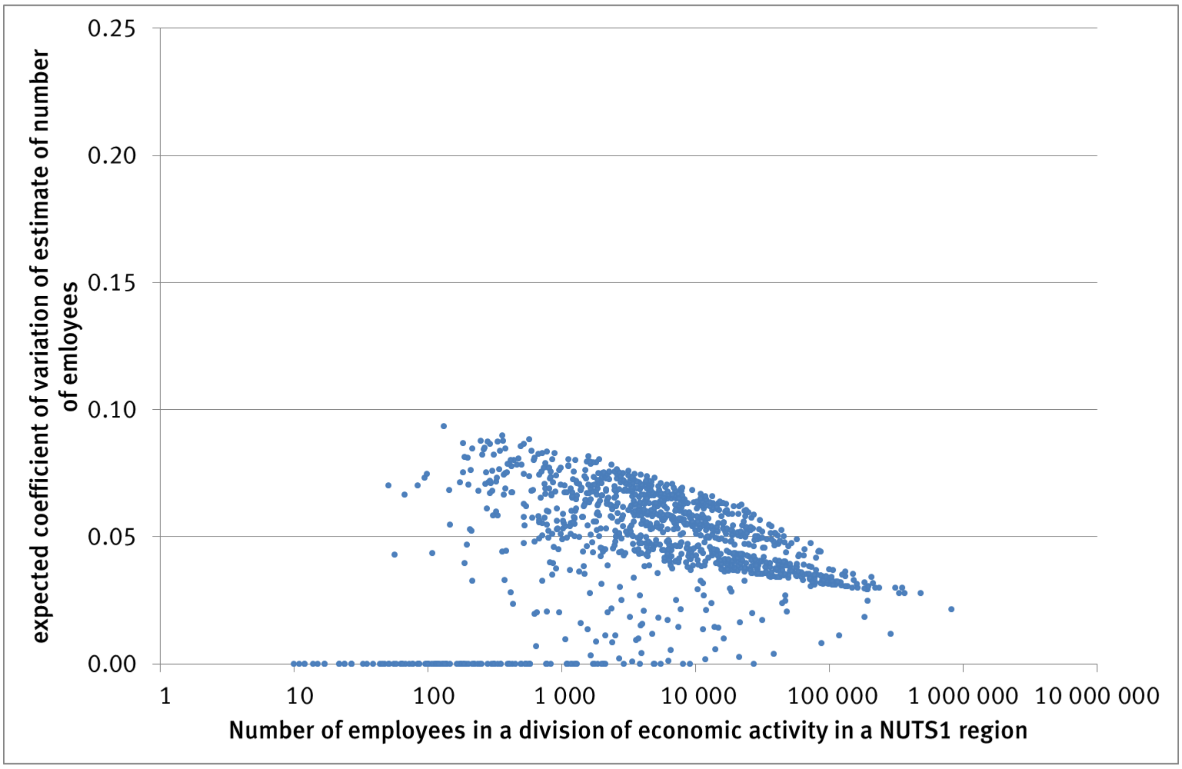
b = 0.1

The rather low value of *b* forces the algorithm to provide small variances also for small economic activities.

Figure 2 shows the results of step 2 for all 16 regions. Here the results are displayed not as sample size as in figure 1 but as the expected coefficients of variation for the key estimate - total number of employees. For all divisions' results of table C the sample of 32 000 enterprises lets expect coefficients of variation of less than 10% and hence below the national requirement of maximum sampling error.

[1] Here was done a simplification for the sake of simplicity and understanding of the procedure. In fact we used a special SAS macro %OPTALLOC by our office which does the allocation in two steps. The 1st step is an allocation over divisions of economic activity according to parameter *b*. The 2nd step is an allocation within the division of economic activity over the 5 strata of size classes using *Neymann* allocation. Step 1 and 2 are redone iteratively until an optimal solution is reached.

Figure 2: Expected coefficients of variation for NACE Rev. 2 divisions of NUTS1 regions, LCS 2012



5 Actual results of sampling error including non-response and weighting

So far we talked about the design of the sample and the *expected* outcome, another thing is the true outcome of the survey. Deviations from our plan are to be expected due to:

- non-response of sampling units,
- the true, unknown variation in the population,
- the difference in concepts between sampling enterprises but analyzing local units and
- effects due to the weighting procedure.

Non-response is not a big problem in the German LCS. Only few enterprises, about 1.2% do not respect the legal obligation and do not take part in the survey. But we have errors in the sample frame leading to a loss of about 6% of the sample size. We try to capture that in the weighting process. For LCS 2012 for the first time we applied the Generalized regression method (GREG) using the SAS macro CLAN by Statistics Sweden.

Table 1 and figure 3 show the results of the estimated coefficients of variation of the key indicator hourly labour cost for tables A (divisions) and C (divisions by NUTS1) of the LCS in relation to the number of employees in the division. For only 3.9% of the divisions of table C the national objective of a coefficient of variation of less than 10% could not be reached.

Figure 3: Estimated coefficients of variation for NACE Rev. 2 divisions, LCS 2012

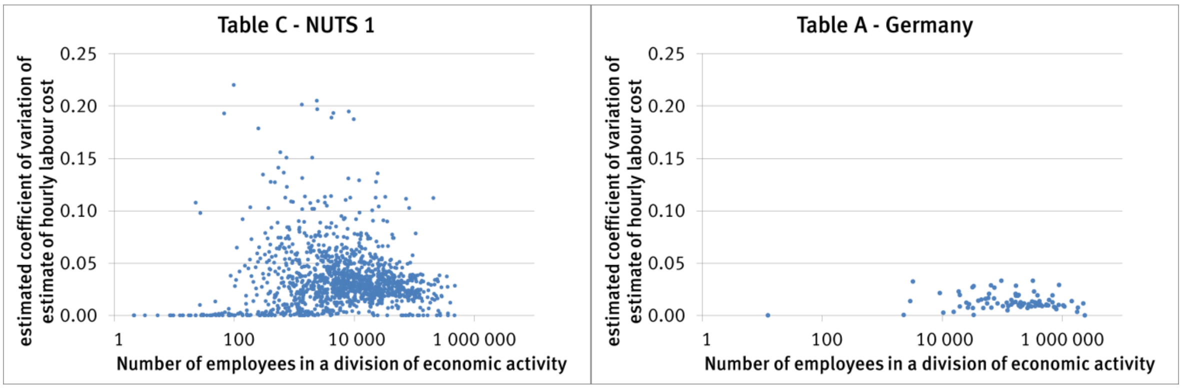


Table 1: Number of NACE Rev. 2 divisions of economic activities by size of estimated coefficients of variation of hourly labour cost, LCS 2012

Coefficient of variation from … up to less than … %	Table C – NUTS 1		Table A – Germany	
	Count	%	Count	%
0 – 1	196	*15.6*	34	*41.5*
1 – 5	817	*64.9*	48	*58.5*
5 – 10	196	*15.6*	-	-
10+	49	*3.9*	-	-
Total	1258	*100*	82	*100*

6 Conclusion

EU regulation on the LCS requires representative results not only at national level but via table C for regional level too. In Germany the national standard method of sample allocation – the principle of graded precision – helps to fulfill the requirements. Although the desired control over the accuracy could not perfectly be reached for LCS 2012, the empirical results give satisfaction.

Annex

NUTS 1 regions in Germany

ISO 3166-2	NUTS	
BB	DE4	Brandenburg
BE	DE3	Berlin
BW	DE1	Baden-Württemberg
BY	DE2	Bayern
HB	DE5	Bremen
HE	DE7	Hessen
HH	DE6	Hamburg
MV	DE8	Mecklenburg-Vorpommern
NI	DE9	Niedersachsen
NW	DEA	Nordrhein-Westfalen
RP	DEB	Rheinland-Pfalz
SH	DEF	Schleswig-Holstein
SL	DEC	Saarland
SN	DED	Sachsen
ST	DEE	Sachsen-Anhalt
TH	DEG	Thüringen