WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

B. Data processing: Coding of economic activity, occupation and educational attainment

B1– Classification of occupations. Trends in self-reported occupational titles, and the dependence of auxiliary variables for classification

Ole Villund- Norway

Statistisk sentralbyrå
Statistics Norway

Paper presented to the
7th Workshop on LFS Methodology in Madrid, May 10–11, 2012
Ole Villund, Statistics Norway

More words, less meaning?

Trends in self-reported occupational information, and the consequences for classification of occupation based on text input data.

Background

In the Norwegian Labour Force Survey, employed people are asked about their occupation and main tasks. The answers are recorded as two text variables, which serve as the basis for coding (classifying occupations). Coding is solely done in-house by dedicated personnel; there is no field coding by the interviewers. To aid the coding process, supplementary information is available, such as education and industry.

The standard classification for occupations (ISCO-88) has recently been revised (ISCO-08) after being used for many years. Over this period in time, the labour market has seen major changes. The structure of skills and tasks has undergone changes, such as increased level of education and growth in the use of information technology. Furthermore, we have suspected a trend towards "flashy" titles – wordier and less useful answers about occupation.

This study set out to answer two questions: Has the coding work become more difficult over time, and did the revision of the standard classification make the work easier?

Challenges

We can point to several challenges of using text data in general, and some specific regarding classification of occupations:

- The relationship between the multitude of actual tasks and the limited number of official occupation categories is not simple. The coding process means that we allocate each job to one category in the standard classification. Each occupation is defined by a more or less well-defined collection of tasks that "naturally" go together. Ideally, to prove that an occupation exists, you should demonstrate that the groups of tasks are found together in actual jobs. On the other hand, a classification more based on traditional assumptions is less useful for describing the actual situation. International standard classifications contain compromises, as they are made to serve different purposes.
- The relationship between the actual tasks and the recorded responses depends on both the interviewer and the respondent. Each job involves a more or less well-defined collection of

tasks. The occupational title may be more or less meaningful in describing such a collection of tasks. Thus, it will vary how well the response maps the actual tasks. Memory effects, stereotyping and other cognitive matters may also affect the response. In addition, it is likely that the interviewers do some formatting and editing before recording their perception of the response.

- In the coding process, each job is manually allocated to exactly one occupation class. This allocation is based on the recorded text about title and tasks. In many cases, it relies on auxiliary information such as education and industry. In some cases even the company name, type and size is used in order to find the right occupation.
- The frequency distribution of words in natural texts is often found to follow *Zipf's law*. It states that the frequency of a word is inversely proportional to its rank in the frequency table. In short, some words are used very often; most words are used very seldom. In addition there seem to be a negative association between frequency and meaningful content. Together, these properties lead to important consequences for our text-based coding: Most frequent words are not useful, and most useful words are not frequent.
- Both text responses and official classification names are nominal data, values that falls into unordered categories. In principle the variables are discrete and finite, but the great variation in responses can make the text input data seem “somewhat continuous” and “seemingly infinite”. Usually to construct a categorical model, you have one or a few classes and a nice set of independent variables. Due to the “wild” nature of the texts on one hand and the rather detailed classification on the other, such models have not been practical in this study.
- Responses vary in how precisely they reflect the categories in a standard classification. A coding index acts as a link between responses and the official classification, thus facilitates the coding process. A good coding index can take care of most cases of synonyms, slang, etc. However, we have found that orthographic variation (misspelling, dialects, etc.) poses significant and not easily solved problems.
- Cases where the response is the same as a class name seem trivial and suggest a more or less automatic process. However, we have observed cases where titles assumed to be unambiguous were used for completely different occupations. Manual inspection and auxiliary information was necessary to discover and correctly classify these cases.
- As mentioned above, useful words often have a low frequency. Thus, the sample variance can make it hard to generalize results from text data, even from large surveys.

Data

The analysis uses data from 1st Quarter 1996 up to and including 4th Quarter 2011, in all 64 quarterly samples. Each sampled person is interviewed every 13 week for 8 consecutive quarters, and dependent interviewing is used for most job variables. For people who report no change in position, previous occupational information is retrieved instead of asking again. Including every record for each person would mean data with many repetitions, since people don’t change occupation very often. In order to construct a sample with independent units, we select only the first interview for each person.

In the Norwegian LFS, self-reported occupation data is written down as text in two variables: title and tasks, each with a maximum length of 40 characters. The occupation class is a 4 digit variable, according to the national version of ISCO. The text variables are not verified or restricted in any other way than the limitation in length. The code variable is checked for validity.

When comparing the standard classifications, we use two smaller data sets: combined sample from 2010 (coded according to ISCO 88) and 2011 (ISCO 08).

Method

Given the complex nature of text data, we try to simplify the situation by defining only three possibilities for classifying any given response:

- Identity: the text is exactly equal to a class name.
- Synonymity: the text represents the same as a class.
- Homonymity: the text can represent two or more classes.

There are cases where the text is too vague or does not seem to represent any class, but we will regard this as a special case of homonymity.

We define a quantitative measurement of how "occupation-specific" a text is. By this we want to evaluate how useful the text is for classifying occupations. Indirectly, this would measure the (inverse) importance of auxiliary variables. For now, we assume that responses and coding is correct, and that the text or text plus auxiliary variables were sufficient. In this limited study we do not intend to evaluate response errors or measurement errors.

We define:

$p_{t,c} = \frac{n_{t,c}}{n_t}$ Proportion classified from text *t* to occupation class *c*.

$M_t{}^* = \sum_{c=1}^{C}(p_{t,c}^2)$ Specificity (monopoly function)

For a singular text ($n_t = 1$) this measurement ($p_{t,o} = 100\%$) isn't very helpful, because we cannot decide if the text was good enough or if auxiliary data were required.

Therefore, we also construct a discontinuous function which kicks out uncertain cases:

$$M_t = \begin{cases} \sum_{c=1}^{C}(p_{t,c}^2) & |n_t \geq n_0 \\ 0 & |n_t < n_0 \end{cases}$$

for some limit n_0. How low this limit should be, depends on how much sample variance you can stomach.

Furthermore, we assume that the data contains some random errors. That means we should allow specificity slightly lower than 100% to be called "specific".

Following these premises, we can group all texts into the following categories:

Specific:	$M_t \geq m_0$
Unspecific:	$0 < M_t < m_0$
Uncertain:	$M_t = 0$

for some limit m_0. How low this limit should be, depends on how many errors you believe there are.

Before comparing the standard classifications, we try to establish a trend or baseline for specificity. When comparing the 2010 and 2011 data, we assume that this trend continues i.e. that there are no abrupt changes in the use of occupation titles.

Results

Have the titles become longer?

Yes, there is a slow but significant trend in the period 1996–2011.

The average was around 11.5 in the beginning of the period, rising to around 12.5 in the end.

The length of "tasks" text varies more during the same period.

Have the titles become less specific?

Not really, our data and method indicates the opposite – if any trend at all.

The proportion of specific and not too rare titles hovers around 25–28 percent, possibly increasing slowly during the period.

Is the new standard classification easier to work with?

Conclusion so far: no.

Macro level comparison shows that the proportion of specific titles have decreased (from 42 to 38 percent).

Micro level analysis shows more cases of text that was specific under the old standard classification became unspecific under the new, than vice versa (18 vs 11 percent).

Comments

We have confirmed a trend in longer occupation titles, but cannot confirm that titles are becoming less meaningful.

Both the coding workers experience and our data point to harder work. This means, for instance, that the coding process is more reliant on extra information.

The new standard classification may not be easier for the coding work, but hopefully the resulting statistics is more useful.

One fallout from our analysis is that up to ⅓ of the jobs can be coded automatically from the titles alone. This requires an empirical, high-quality coding index. Combined with adequate auxiliary information, the potential for automatic coding could be higher.