Administrative data as input and auxiliary variables to estimate background data on enterprises in the CVT survey 2011

Eva-Maria Asamer
Statistics Austria, Guglgasse 13, Vienna, eva-maria.asamer@statistik.gv.at

Abstract: In this paper an example for adding information from administrative data is introduced. A step-by-step method to transfer information from one to another survey with some administrative data in common is described in detail. For the example of total hours worked, it is shown that this method, using days worked as auxiliary variable, leads to good results on enterprise level.

Keywords: administrative data, combining surveys, linear regression

1. Introduction

Our approach of combining different, already existing administrative sources as well as integrating other survey data to reduce the response burden will be presented for the 4th Continuing Vocational Training Survey (CVTS4).
For the CVTS4 in Austria a number of items are not collected directly, but added using administrative sources. This includes the total number of persons employed, total labour costs and the total number of initial vocational training participants. These variables can be determined using mainly administrative sources, like the social security and tax registers. Those data are linked on enterprise level. For data with missing links, other sources are investigated.
In contrast, the total number of hours worked cannot be determined by administrative data directly. For CVTS, a procedure to estimate this variable using other surveys, with administrative data (primarily the number of days worked full time and part time per enterprise) as auxiliary variables, was developed. Here, different reference periods and unequal definitions of the population have to be taken into account.
This year the total number of hours worked is being collected directly as well as being estimated, providing the opportunity to evaluate the estimation procedure. Furthermore the estimated values are used for control and imputation.
The matching and estimation procedures along with first results will be discussed.

2. Data Sources

The Survey on continuing vocational training (CVTS) is an EU survey performed every five years, asking a variety of questions on initial and continuing vocational training in enterprises. Apart from these variables a number of structural variables are asked such as the NACE (Nomenclature générale des activités économiques dans les Communautés européennes) category or the number of people employed. Further, the total labour costs

and the total hours worked are asked as reference values as well as to determine indirect costs of vocational training due to the non-productive time while attending a training course. To ease the response burden for enterprises, as many variables as possible are added using existing administrative data. This is especially important as it is an optional survey and the more questions have to be completed the less likely the form will be completed at all.

For business data, a variety of administrative data is available in Austria. The main sources, social security and tax information, as well as data of the chamber of commerce are included in the statistical business register (BR) maintained by Statistics Austria. Every enterprise has a unique key in this register, and foreign keys from external data sources are matched to this key.

Preparing the first register based census in Austria 2011, an Austrian Activity Register (EVA (German: Erwerbstätigen-Versicherten-Arbeitslosen-Datenbank)) was build. Here, data from social security, tax register, and unemployment register is available on personal basis containing an anonymous key. In EVA, the business register key is available as a foreign key. Also, the key for the local units of work is used according to the business register. The links are not available for all combinations, but there is steady work going on to improve the linking within EVA as well as to the BR.

Some information, like some parts of labour costs, or the hours worked by person or by enterprise, is not included in administrative data. But there are some business surveys containing this information. The labour cost survey (LCS) is held every four years in Austria, the last time in 2008. In this survey, enterprises are asked about the average number of employees, subdivided by full-time and part-time workers, with apprentices counted separately. The main focus of the survey is labour costs, which are asked in detail. Furthermore the total amount of hours actually worked is asked for the subgroups separately.

Every month, data is collected for a short term statistics survey in industry and construction (KJE (German: Konjunkturerhebung)), which is published every year. Small companies do not have to answer the survey, but their values are estimated. For all bigger companies some information is included from the BR, other data is asked directly. The economic sectors of KJE are industry and construction, which is only a part of the enterprises of interest for CVTS. But for those enterprises within the NACE and size categories, data for the same period of time as CVTS is available.

3. From administrative to statistical data on employment

A key variable of an enterprise is the number of persons employed. Employed persons contain employees and self-employed persons, with persons in training counted separately. Further, male and female should be counted separately as well. The number of persons is asked on two reference days, 31.12.2010 and 31.12.2009, as well as an average for the year 2010.

In EVA, all periods of employment since 2002 are stored. But whereas the periods concerning employees usually hold a connection to the enterprise they work in, working proprietors hold no connection to their enterprise in EVA. Thus in a first step all variables are calculated for employees only. For all enterprises in the CVTS sample these preliminary variables could be determined. For working proprietors and family members a matching procedure is being developed at the moment for the census 2011, using data from 2009. First results from these matching procedures are used to add self-employed persons to the number of persons working.

To determine the total labour costs for the year 2010, administrative data was used too. In EVA, tax and social security information is available per person (with an anonymous key) and per enterprise. The social security information is used to determine all persons working in the companies asked in the given period of time (the whole year 2010). Furthermore, there is some income information in the social security data, but only above the marginal income and up to a maximum income, so this income information is only used if no link to tax information exists. For the majority of people the salary from the tax register is processed. A simple model is applied to determine the statuary social security contributions for the enterprises per person employed.
This linking of tax information to persons and enterprises is already standardized in EVA, using different data sources to improve missing links. This linking information is available in so-called "linking-tables" containing an additional quality attribute.

4. Estimation from other surveys

EVA contains no exact information about the hours worked for a company. In Austria, there exists no administrative data source (apart from some very small subgroups) where hours worked are registered, neither on personal nor on enterprise level. There exists a variable full time or part time, a variable whether a person is holding a marginal job, and some information about the yearly income and the period of time a person worked in the past years. Total hours worked are included in a variety of partly compulsory business surveys, though, which can be used as basis for estimation.

In 2006, for CVTS3 a procedure to estimate hours worked from another survey was developed, but as the true values were not known, the operating department was not sure about the quality of these estimates. This time the item was asked directly, but a high item-non-response rate is expected. So the estimation process was performed for CVTS4 too.

First challenges to be faced are different basic populations (e.g. different NACE sections, different minimum employees) and different definitions of hours worked (e.g. breaks or waiting time either counted or not). So as preparatory work, a variety of surveys performed by Statistics Austria were analyzed and the population range as well as the definitions and subdivisions were compared. This has to be done for every new project, and it is recommended to monitor changes in questions and definitions of the surveys analyzed as those can change over time as well.

For CVTS the labour cost survey (LCS) proved to be appropriate regarding the definitions of hours actually worked, as well as the subgroups available. The KJE survey is available only for half of the NACE categories, and only data with similar definitions and for wider groups are available. On the other hand, data for the same period of time as CVTS will be available soon. This information can be used to verify the assumption that there are no changes in the relation of administrative information (e.g. number of days worked) and hours worked.

To build a model to transfer worked hours from one survey (LCS) to another (CVTS), a stepwise process was performed.

First, administrative data for the period of time and the enterprises in the LCS was extracted. From this data a reference value, in this case the average number of employees in the reference year 2008 was derived. This value is asked in the survey and can be determined from the administrative source as well. We found that such a reference value is very useful to eliminate data sets which are not plausible and would therefore worsen the quality of the model. These data sets can appear for a number of reasons, like missing links in administrative sources, different definitions of units, different classification of employees, or wrong data in the survey. As benchmark the proportion of the administrative value (X_{admin}) to the survey value (X_{survey}) is used:

$$|1-(X_{admin} / X_{survey}) | \leq \alpha \qquad (1)$$

Only data which lay within this threshold (α) is used in the next steps.
In our sample, there were 6585 enterprises in the NACE categories of interest, 6105 data sets were lying within the threshold $\alpha = 0.25$ and were therefore used for estimation.

In a second step the amount of part time and full time workers was derived from administrative data. If those values differ strongly from the amount in the survey, but the overall number of employees is similar, a model without distinction by these categories is used, as here the definitions of full and part time are apparently not the same in administrative data and the survey.

Next, auxiliary variables are derived from administrative data. In our case, the number of days worked in the reference year, for full time employees, part time employees, and apprentices separately, as well as for all employees together, are calculated. First for enterprises in LCS, then these variables are calculated analogously for the reference period and the enterprises of the CVTS, so they can be used as independent variables in our model.

Before the model is estimated, the scope of the samples has to be considered. For CVTS, the NACE sections B to N and R, S are part of the survey. For LCS, the sections P and Q are surveyed too. These NACE sections do not need to be modeled, as they won't be transmitted to the CVT-Survey. A different scope of size of the enterprise could be considered as well, for CVTS and LCS they are nearly the same, so no further cut is made.

The following formulae are assumed:

$$
\begin{aligned}
d_{ft} * w_{ft} &= h_{ft} \\
d_{pt} * w_{pt} &= h_{pt} \\
(d_{ft} + d_{pt}) * w_{tot} &= h_{em}
\end{aligned}
\qquad (2)
$$

d_{ft} … days worked full time
d_{pt} … days worked part time

w_{ft} ... average hours actually worked per day by full time employees
w_{pt} ... average hours actually worked per day by part time employees
w_{tot} ... average hours actually worked per day by employees (without apprentices)

h_{ft}... total hours actually worked by full time employees
h_{pt}... total hours actually worked by part time employees
h_{tot}... total hours actually worked by employees (without apprentices)

Theoretically, there should be no intercept, as zero days of work imply also zero hours of work, but this constraint was not set in advance.

It is assumed that the total hours worked per day is not equal for all NACE categories, so regressions are estimated by groups of categories, which we found by analysing the data. Data is also divided in different classes of size, to ensure the dependent variable h_{tot} is normally distributed.
First models for full and part time employees together were estimated. The models show a good quality with an adjusted R^2 between 0.75 and 0.99, and corroborate the hypothesis that w_{tot} is significantly different for groups of NACE categories. As we found some outliers with a high leverage in our data, we decided to use robust regression to minimize these effects.

In a second step full and part time employees were examined separately, but these models did not lead to a significant improvement. Apparently, dividing part and full time employees according to administrative information is not similar to the separation performed by the enterprises for answering the survey.

These models were now used to estimate the hours worked in the enterprises of the CVT survey. First raw data has been just available and is used to determine the quality of the estimation models outside the data used for building the models. Also for CVT data, a benchmark variable is used to determine whether the questionnaire is answered for the same unit as identified in the administrative data. For CVTS, the number of persons employed on 31th December 2010 is asked as well as calculated from administrative data. Again, formula (1) is used to filter those subjects with a similar value in this variable. For these data sets the estimation model is applied.

The estimated values are then compared with the values answered in the survey. From 1230 enterprises for which the survey has been answered so far, 986 responded on the item total hours worked, 890 of them are within the filter of formula (1).
For 80% of them, the ratio of estimated and survey value is between 0.75 and 1.25.

For the smaller enterprises the results are plotted in Figure 1. Here the squares belong to enterprises which did not pass the filter of formula (1), the circles belong to the remaining.
The dotted line corresponds to estimated values = survey values, the solid line is the mean line of our data, whereas the dash-dotted line would be the mean line of all data, including the squares.
The dashed lines is the (0.75, 1.25) interval. For those further apart, a plausibility check is performed by the operating department, especially for those with a survey close to zero.

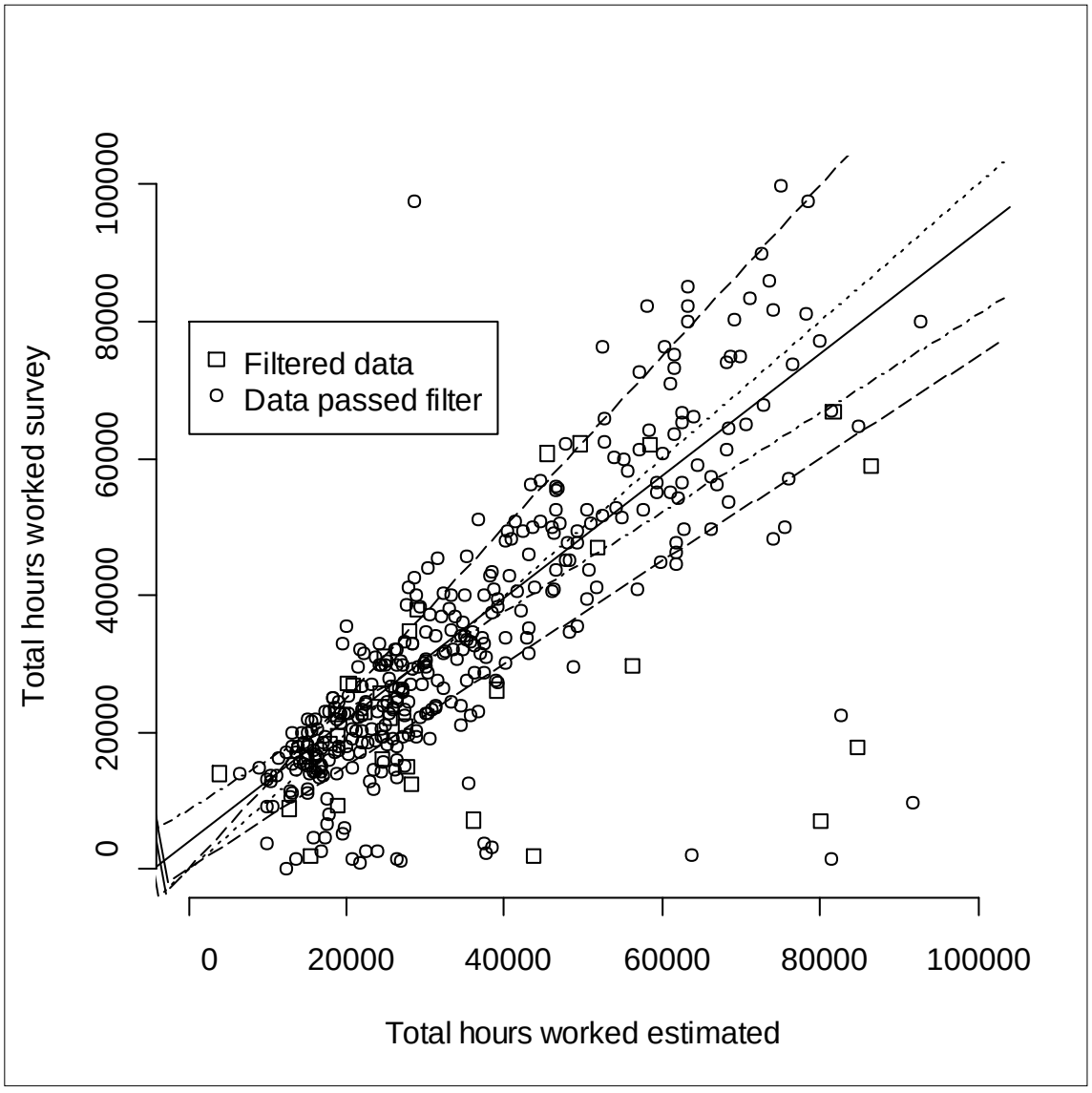

Figure 1: *Estimated vs. Collected Data on hours of work*

5. Conclusion and further work

In our case study, estimating total hours worked using total days worked derived from administrative data as auxiliary variables led to reasonable results. Considering that the survey value is not always exact for this item, a use of models to lower the response burden or improve the quality of imputation seems advisable. Filtering data with a benchmark variable avoided applying the model where administrative and survey data did not correlate. For these units other methods of imputation should be used.

When KJE data for 2010 is available, there will be a model estimated using this data and an analysis about possible changes in the weights w_i will be done.

The whole procedure of processing the administrative data will be automated.
In a next step, also the selection of the usable data files from the "donor" survey will be standardized, as well as the selection of the data files from the receiving survey.
Calculation of labour costs will be done using already processed data from EVA.

A further problem to be tackled will be the determination of the quality of these variables in the new survey.

References

Čiginas A., Kavaliauskienė D., Overview of use Administrative Data in STS, *ESSnet Seminar,* Rome, March 2010.

Bundesstatistikgesetz 2000, BGBl. I Nr. 163/1999, Austria.

Fox J. (2002) Robust Regression in: *An R and S Plus Companion to Applied Regression, Appendix,* Sage.

Salfinger B., Sommer-Binder G., Erhebung über betriebliche Bildung (CVTS3) in: *Statistische Nachrichten 12/2007*, p. 1106 – 1119.

Silva D.B.N. and Clarke P., Some Initiatives on Combining Data to Support Small Area Statistics and Analytical Requirements at ONS-UK, *IAOS Conference*, Shanghai, 2008.