Testing the Importance of Cleansing Procedures for Overlaps in German Administrative Data.

Patrycja Scioch
Institute for Employment Research, Patrycja.Scioch@iab.de

Abstract

Process-generated and administrative datasets have become increasingly important for labour market research over the past ten years. Major advantages of this data are large sample sizes, absence of retrospective gaps and unit nonresponses. Nevertheless, the quality and validity of the information remain unclear. This paper contributes to this subject, focusing on the variation of research results due to alternative data cleansing procedures. In particular, the paper replicates an evaluation study presented by Wunsch/Lechner (2008) and investigates the robustness of the results. Variance is induced by using different cleansing methods for overlapping and parallel observations. The descriptive results show that the differences between the data sets (based on the different methods) are statistically not significant but show different magnitudes on some of the attributes. The same result comes up for the subsequent analysis of the treatment effects. This clearly indicates that the 'real' effect is also significant and only the extent depends on the underlying cleansing method.

Keywords: administrative data, cleansing procedures, data quality

1. Introduction

Process-generated and administrative datasets have become increasingly important in research over the past ten years. Kluve (2006), for example, reports that almost 80% of all microeconomic evaluation studies in Europe are based on that type of data. So far there are only a few studies that focus on the quality of German administrative data. Jaenichen et al. (2005) or Bernhard et al. (2006) refer to the requirement of data preparation and data cleansing. Recent work also focuses on the connection of research results and data cleansing procedures (e.g. Kruppe et al. 2008; Waller 2007). Based on a German data set (Integrated employment Biographies – IEB) we investigate the impact of different cleansing procedures on data overlaps and inconsistencies between different data sources. In order to analyse the effect of different data cleansing procedures we use an evaluation study and replicate the data cleansing approach suggested by Wunsch/Lechner (2008). In a second step we develop variations of the data cleansing procedures and analyse their effects on point estimators.

This paper is structured as follows. In the next section we describe the database and discuss problems that may occur when using the data. Section 3 explains the replication and identification of variance before Section 4 summarises the descriptive results. The first part of section 2 values the quality of replication and the second the

---

1 We thank Conny Wunsch and Michael Lechner for supporting the work related to this paper and for giving us access to their programme codes and regional data.
variance in the evaluation samples. Section 5 contains the estimation results for the replication and the investigation of variance and section 6 concludes.

2. Database

The database used in this study is the Integrated Employment Biographies (IEB) of the Institute for Employment Research (IAB), which is a longitudinal data set merged from four distinct process generated data sources in Germany. Each of these sources offers a brought set of attributes and covers different periods of observation.

- The first data source is the Employment Histories containing employment periods captured by the social insurance register back until 1990.
- The second data source contains data on spells of unemployment from the Benefit-Recipient-History.
- Most of the individual characteristics in the IEB data arise from the Applicants-Pool data, which contains information on job-searching spells since 1999.
- Finally the data set on Active Labour Market Programme Participation provides information on periods spent in promoted schemes.2

The IEB data are organised on a daily basis and allow to control for time varying covariates. It is important to note that the sources are not cross-validated, which may cause the existence of parallel observations (overlaps). Individuals can have several jobs at the same time or they might be employed and searching for a new job or receiving benefits while on job search or participating in labour market programmes. These spells can be completely parallel, one may embed the other or they are overlapping.

The existence of parallel observations is twofold: It may offer additional information, like periods of promoted employment. However, it may also cause problems when information is contradictory. In the latter case one must decide which data source to believe - which is the subject of data cleansing procedures.

3. Replication and identification of variance

3.1 The general framework

The general framework for the data cleansing procedure used in this study follows a study by Wunsch/Lechner (2008) in which they perform matched pair comparisons. This procedure uses statistical twins with respect to the likelihood to participate in a certain promotion scheme (Rosenbaum/Rubin, 1985; Heckman et al, 1998) to identify counterfactual observations. They use a 2% random sample of the IEB supplemented with additional characteristics taken from the different data sources as well as characteristics from regional statistics.3

The reference study identifies potential comparisons between participants and non-participants taken out of the total of inflows into unemployment between January 1992

---

2 For a detailed description of the data generating process of the participation in measure data see Engelhardt et al. (2008)
3 See Wunsch/Lechner (2008) for a detailed description of the data.
Participants are limited to individuals who have started a programme during the next 18 months after becoming unemployed and have received unemployment benefits directly before the programme start. A reference date (the non-observed begin date) for the non-participants is imputed by using regression methods.

In order to cope with parallel observations per individual Wunsch/Lechner (2008) define time frames and rules of priority for possible parallel states to clearly identify one state at each point in time and transform the data into a panel data set. Labour market programmes are treated with the highest priority (followed by periods of benefit receipt and times of employment). The lowest priority gets information out of applicants’ pool data.

### 3.2 Variation based on data cleansing

Referring to reference analysis we use the same set up for the analysis based on a more recent draw of the IEB. We produce multiple subsamples based on different rules of priorities for individuals with overlapping observations. The resulting different final states for the individuals cause variation in the composition of the subsamples used for the matched comparisons analysis.

Similar to Wunsch/Lechner (2008) we split the spell data into frames of two weeks. Within these time frames it is now possible to isolate one state. In a first step the parallel observations are sorted by length (sorting rule one). If two or more parallel observations have the same length the respective data source is used as a proxy of the validity to order the observations (sorting rule two).

An illustration of this approach is given in figure one, where the left part displays overlapping observations and the table on the right hand side shows the related data matrix. The time-period is divided into (seven) time windows. In the whole period six different observations occur, each coming from a distinct source. For example: There is a benefit-spell going from time-frame one until the end of time-frame three and an employment-spell beginning in time-frame one and ending in the mid of time-frame four. The aim of the data cleansing is the identification of one valid observation per time-frame. The right hand side shows the transformation of the observations into a data matrix, with one row representing one time-frame and each state in one column (e.g. time-frame one - see t₁ - covers two states and period five contains four states- see t₅ - ).

**Figure one: Identification of the final state**

![Figure one: Identification of the final state](image-url)
The most important step in this approach refers to the sorting routines, therefore the order of state across the column displayed in figure one is crucial. The first row displays two observations from different sources, one from the receipt of benefit source and the other from the employment histories. The first column contains the observation with the longest period in window one. If we observe multiple observations with the same length we need to sort the observations by heuristic routines (see time-frame five). All observations with participation in a training scheme are classified with the highest priority, since we are interested in the evaluation of them. For not being associated with any type of payment observations out of the job search register (two possible states: searching and unemployed) are less valid and therefore classified with a lower priority.

For any further data cleansing only the first two states – independent of the number of overlaps are accounted. Both are now sorted only based on the sorting rule two (source priority). To demonstrate the choice of the final state the example continues in the illustration. In time window two observations from the unemployment benefit register and the employment history occur. Following the priority we define the first as the final state. Likewise, in period five the final state (further vocational training) arises because unemployment assistance has a lower priority than the participation in a labour market programme. The results are displayed in column five.

Changing the priority alters the definitions of the final states and leads to different data samples. Finally, three different methods of data cleansing procedures are developed:

- Method V0 follows the approach of Wunsch/Lechner (2008). Evaluating labour market programmes the participation in a programme gets the highest priority. Sources associated with payments (benefit-recipient-history, employment history) follow on second and third priority. The job search register with lots of optional informations is considered to be less valid. This method is used as reference method.
- The first variation occurs in method V1, where the priority of the two sources with money payments is reversed. As mentioned above both are regarded as valid without a clear indication which one to prefer. This may lead to a significant change of the number and duration of employment spells in the analysis sample.
- Method V2 assumes the participants-in-measure database as not fully valid. Thus, method two degrades the priority of this data. However, since some participations come along with benefits and the interest of any evaluation focuses on the effects of participation we do not degrade them completely, but order participation as priority two. Assigning them behind benefit receipt may lead to a dramatic reduction of the participations used for subsequent evaluation studies.

4. Descriptive results

The first part of this section presents the quality of the replication by comparing means statistics of the evaluation samples of the study of Wunsch/Lechner (2008) and the replication V0. The second part focuses on the variance in the different evaluation samples (V0 vs. V1; V0 vs. V2).
4.1 Quality of the replication

Replicating a study based on administrative data is not as easy as it sounds. In contrast to survey data they change over the time due to innovations and modifications in data structure or just by an updating. This means not only adding new data but it also affects previous information by the completion of missing data and the correction of wrong information. Therefore a replication is more promising using the identical data as the original study than working with the same but more recent data.

The goodness of the replication can be assessed by comparing the means and shares of the variables presented in the Wunsch/Lechner (2008) evaluation study. Unfortunately, in their paper the means are listed without decimal places and standard deviation. For simplicity we assume that both samples are drawn out of the same population. With the sample variance of V0 and the number of observations of V0 and the given ones in the paper we calculate the sample variance for the means listed in their study. Given this variance we calculate the confidence interval of the original data which can be used for a rough comparison of both samples.

Due to the more recent data the number of observations is higher in sample V0 compared to the original data used in Wunsch/Lechner (2008). Most variables are equal and not significantly different. The majority of the significant differences occur in the employment history variables and the regional information. The regional characteristics stem from a source which is not one of the four data sources described above, but has the same drawbacks concerning the possibilities of replication. Note that some of the variables relate to each other, e.g. the occupational sector of the desired job is parted into six subcategories. This inter-correlation leads to an overestimation of the sample differences when focusing on the number of variables with significant differences.

Overall a wide range of variables from personal characteristics, information about the desired job, profession, remaining unemployment benefit claim, characteristics of the employment history over the last 10 years and a range of regional information were tested. The number of significant differences varies between the distinct types of programmes and indicates differences in the samples, but due to the fact of using more recent data and the low magnitude of differences method V0 can be interpreted as a satisfying approximation to the sample used in Wunsch/Lechner (2008).

4.2 Variance in the evaluation samples

To gauge the influence of the different cleansing procedures the evaluation samples with the different underlying order of priority are compared. The comparison focuses on testing differences of the sample means (V1, V2) in reference to sample V0 and conditional to the participant-state.

We find no significant differences between the methods in the group of the participants, what indicates that simple decision rules are sufficient. Even in cases with significant differences, which occur in the group of non-participants, the differences range between one and four percentage points and are not of any substantial relevance. An example: The duration of last unemployment in method V0 with 5.39 months decreases to 5.26 months in method V1 and is statistically significant on the 96%-level. However, substantially this is only a few days, which seems to be of no practical importance.

---

4 Short training (ST), short combined measures (SCM), jobseeker-assessment (JSA), job-related training (JRT), degree course (DC), general further training ≤6 months (GT6), > 6 months (GT6+)
5. Estimation Results

5.1 Replication of Point Estimators

Figure two presents the matching estimation effects\(^5\) of programme participation compared to non-participation for unsubsidised employment based on method V0 which replicates the analysis of Wunsch/Lechner (2008). Each type of programme with respective population of participants is displayed by one line over a 30 months period after programme start. The dots indicate a significant effect on the 5% level. As found before all types of programme show the typical lock-in effects and recover at different rates. The developing and the amount of the values is consistent with the original findings too, but in contrast the programme types ST, SCM, GT6 and GT6+ show even significant positive effects. The absence of significant positive effects is a major point in the study of Wunsch/Lechner (2008). Because the whole approach of data cleansing has been replicated it is necessary to discuss reasons for this difference and to check the robustness of the results.

Figure two: Effects of programme participation compared to non-participation

One possible reason are certainly the significant differences of the replication sample V0 compared with the sample used by Wunsch/Lechner (2008). As mentioned above the main differences occur in the regional characteristics, because version V0 uses updated regional information.

Figure three illustrates the difference that is induced by updating the regional data. Exemplarily, we focus on the effects of programme participation for unsubsidised employment for participation in general further training with duration over 6 months.

\(^5\) Mahalanobis matching with caliper
(GT6+). The abscissa displays the months after programme start and the ordinate the effects in percentage-points. The dots again indicate a significant effect (5% level). The pink line represents the effects estimated with the updated regional characteristics (replication) and the green line displays the effects estimated with the regional data used in the reference study (original). Comparing the evaluation results based on the updated and non-updated information shows only small differences – which are not statistically significant at any time.

Figure three: Influence of differences in regional information on the effects

5.2. Variance in the estimation results

When comparing the effects of the estimations based on the different methods (comparing V1, V2 in reference to V0), one would not expect great differences due to the absence of significant differences in the descriptive results. However, differences may occur in multiple ways, e.g. all effects significant but with different amount; one effect is significant and the other is not.

Figure four shows the effects of programme participation compared to non-participation over a 30 months period after programme start exemplarily for general training with duration longer than 6 months (GT6+). These effects are displayed for each of the three methods (V0: pink; V1: green; V2: blue). Moreover the figure adds the confidence interval of reference method V0 (black lines).

During the lock-in effect the three different methods of data cleansing lead to differences up to 10 percentage points. Smaller differences occur in the end of the observation period (months 22-24) where method V0 leads to significant positive effects of approximately 10 percentage points and method V2 to insignificant effects of nearly 3 percentage points. However, the results based on the methods V1 and V2 lay within the confidence bands of the results based on method V0. Not displayed:
Testing the significance of the differences between the methods V1 and V2 also shows insignificant results.

Figure four: Variation in the estimation results

Although the differences between the methods are not significant questions raise concerning the great difference between Method V0 and V2 during the lock-in effect. A possible explanation is the priority of the data source. Method V0 grants programme participation the highest priority and employment is ordered on third position. In method V2 being employed is preferred to programme participation in cases of parallel information. When the programme starts the effects are nearly equal but the lock-in effect is 10 percentage points lower for method V2. This could be due to participants who dropout earlier and start to work. The dropout is not always (seldom) registered and two parallel observations occur in the data. Method V0 continues counting this as participation whereas in method V2 employment is the final state and therefore the lock-in effect decreases.

6. Summary and Discussion

This paper investigates the influence of variations in cleansing procedures on overlaps in a merged administrative data set. The study presents the cleansing methods and the effects of data cleansing that yield to distinctive analysis samples. First we replicate the preparation procedures applied in Wunsch/Lechner (2008), which yields a sufficient approximation of the data compared to the one used in the reference analysis. However, some differences occur which concentrate on the regional characteristics. A part of this difference may be explained by using updated information in the replication study. In a second step we develop and apply two
variations of the cleansing procedures: by changing the order of priority in cases of overlapping observations the decision rule changes and thus also the final states. The differences between the data sets are statistically not significant but show different magnitudes on some of the attributes. The same result comes up for the analysis of the treatment effects. This clearly indicates that the ‘real’ effect is also significant and only the extent depends on the underlying cleansing method, which is consistent with the findings of previous studies (e.g. Waller 2007).

Even though lots of researchers work with administrative data there are only few studies about the data themselves and their quality, especially in Germany. This study is just a start and more have to follow to assess the effects. Further research focuses on an application of a ‘ naïve’ procedure which prefers observations of one distinct data source without considering aspects of possible and allowed combinations of states. Or even more interesting could be the implementation of the developed procedures in others than the labour market or a comparison with other countries, who use administrative data more intensive.

References

Jaenichen, U. et al. (2005) You can split it if you really want: Korrekturvorschläge für ausgewählte Inkonsistenzen in IEB und MTG, FDZ Datenreport, 04/2005