

Chapter 5

A rough guide to the database

"Forty-two!" yelled Loonquawl. "Is that all you've got to show for seven and a half million years' work?"

"I checked it very thoroughly," said the computer, "and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is."

Adams (1979)

5.1 Introduction

The aim of this paper is to provide a rough guide for the data cleaning of an administrative database which has been used for several research projects in empirical labour economics at the University of Louvain-la-Neuve, Belgium.¹

The raw database contains information about 140.075 people who are 15 to 26 years old and who received in 1998 for at least one day either waiting-benefits or transition-benefits.² The database follows these individuals for the period from 1998-2002. In addition a reduced set of

¹See Cockx et al. (2004a) and Cockx and Göbel (2005) for research projects which use the data.

²Waiting-benefits (allocation d'attente) and transition-benefits (allocation de transition) are a special form of unemployment benefits which are provided to workers either after the end of schooling or during the period of part-time schooling (for 15-18 year old individuals).

variables is available for the year 1997.

The following guide serves several purposes. First of all it should enable a researcher to understand the main issues and the underlying structure of the data. Second we document the process of data cleaning and preparation of the database for empirical research. It is, however, not intended to provide a complete discussion of all decisions which have been made during the data cleaning process. Instead this paper aims to be a starting point for a researcher who wants to use the database for his/her own research. To reach this goal we explain also the used infrastructure: the structure of the data on the SAS-server and the organisation of the programs which have been used to fulfil the tasks which have been necessary during the data-cleaning process.

Moreover this paper completes the existing online-documentation which can be found in the Internet under <http://www.ires.ucl.ac.be/Axe4/doc/docstart.html>. The online documentation focuses on the contents and the nomenclature of the variables. It contains numerous details and links to various resources. Finally the online-documentation provides access to the original documents which have been used during the data-cleaning process.

The paper is structured as follows. Section 5.2 describes the origins of the raw data. In section 5.3 we give an overview of the variables in our raw data. The discussion of the main issues when organising our data is outlined in section 5.4. Section 5.5 provides an overview of the data cleaning process and its successive steps. Section 5.6 gives some indication on where to start with a new project. Finally section 5.7 points to additional resources which may be helpful and section 5.8 concludes.

5.2 Where does the administrative raw data come from?

The raw data has been delivered by the Belgian *Crossroads Bank for Social Security*,³ an administrative organisation which is in charge of infor-

³The French name is “Banque Carrefour de la Sécurité Sociale”.

mation exchange and process optimisation for the different institutions involved in the social insurance system of Belgium.⁴

One of the activities of the Crossroads Bank of Social Security (CBSS) is the provision of administrative data. The so-called “*Datawarehouse labour market and social protection*”⁵ is a service which is collecting data regarding the labour market. The data of the CBSS comes from different administrative institutions of the social insurance system.

In Belgium every citizen and every company has a unique administrative identification number. This makes it possible to match the data from different institutions to the individuals as well as individuals to the employers. Since the CBSS has access to the administrative data of the different institutions the Datawarehouse can provide rich information concerning the labour market outcomes of citizens and companies. The information in the database is made anonymous and provided in a form which can be used for statistical analysis and research.

The raw data, for this paper contains information from the following institutions: The national unemployment office (ONEM), the national office for social security (ONSS), which is in charge of the collection of social insurance contributions from the employer and employees of the private and public sector (with exception of the individuals which are covered by the ONSSAPL). The national office for social security of the local and provincial administration (ONSSAPL). The national institute for the Social Security of the self-employed (INASTI). The national office for family allowances for employees (ONAFTS). Finally there is some information from the national institute of statistics (INS).⁶ Since the names of the original data sets use the Flemish acronyms for these institutions we refer to Table 5.1 in the appendix for an overview of the institutions, their Flemish and French names and their acronyms.

⁴See Deprest and Robben (2003) for a detailed description of the Belgian Crossroads Bank for Social Security.

⁵This is translated literally from French “*Datawarehouse marché du travail et protection sociale*”.

⁶There are additional institutions which are in the Datawarehouse labour market and social protection. We concentrate on the institutions from which we have information in our database. For more information we refer to Deprest and Robben (2003).

Our database contains information from all these institutions for a sample of young workers. Without going into details here, it is worthwhile to mention that the data is organised in different ways in each of these institutions and the raw-databases reflect this heterogeneity.

5.3 What kind of information is in the different databases?

To understand the structure of the database as well as the different steps of the data cleaning process it is indispensable to outline the genesis of database. At the origin of the database there are two data requests which has been part of a research project. This project has been done on behalf of the Belgium Services Fédéraux des Affaires Scientifiques, Techniques et Culturelles (SSTC)⁷.

In this section we first summarise the two data-demands.⁸ Then we present the additional data which has been integrated into the database. In subsection 5.3.4 we are going to provide an overview of the variables in the different databases and we link the information to the different administrative institutions. Finally we describe where the raw databases are stored physically.

5.3.1 1st data request

The first data request has been formulated as a part of a research project which has been supported financially by the SSTC.⁹ The aim of the project has been the evaluation of active labour market policies for young unemployed people in Belgium. The idea has been to exploit the detailed information and the longitudinal character of the “*Datawarehouse labour market and social protection*” to overcome the selection-problems which typically appears in non-experimental microeconomic

⁷As a result of a reform of the federal administration in Belgium the SSTC has changed the name and is now called “Politique scientifique fédérale”.

⁸The original data-requests are available via the online documentation on <http://www.i-res.ucl.ac.be/Axe4/doc/docstart.html>.

⁹The original French version of the data-request is available in the online-documentation.

evaluation studies.¹⁰

The sample population includes individuals who have between 15 and 26 years of age in 1998¹¹ and who received for at least one day in 1998 either waiting-benefits or transition-benefits. In the first data request we asked for all information up to the year 2000, which has been available in the Datawarehouse for the sample population at that time. In order to be able to control for initial condition problems we also requested the available information for 1997. However for the year 1997 not all administrative institutions are available in the Datawarehouse, notably the information from the national unemployment office (ONEM) and from the national institute for social security of the self-employed (INASTI) is missing.

To ensure that the data has a certain degree of consistency with respect to the timing of the information the data is generated at the end of each quarter, whenever possible (end of March, end of June, end of September and end of December) for the time period 1997-2000.

5.3.2 2nd data request

This data request is basically an extension of the first request. The information for the original selection of individuals has been completed.

Most important is the inclusion of the year 2001 and 2002 into the database. In addition we received information for the education of the individuals and corrected information on the start of the first unemployment spell. The information concerning the *highest school degree* of the individuals is only available at the end of the year 2002.

Table 5.2 in the appendix contains an overview of the variables which have been requested in 1st and 2nd data request.

¹⁰We refer to Heckman et al. (1999) or Costa-Dias and Blundell (2002b) for a recent survey on microeconomic evaluation techniques.

¹¹More exactly the following rule for the birthday has been applied: 03/04/1972 < birthday < 01/01/1983.

5.3.3 Introduction of additional information

In addition to the data which has been provided by the CBSS, following the two data-requests we have also manually integrated the gender-specific local unemployment rates.

The data on the *local unemployment rates* are from the “Bulletin Mensuel” a monthly publication of the Belgian national unemployment office. The information has been filled into Excel-tables by hand and has been transformed into a SAS-database afterwards. The Bulletin Mensuel provides a gender-specific administrative measure of the unemployment rate at the level of 42 administrative districts.¹² The unemployment rate is measured by the number of unemployed who receive full unemployment benefits¹³ at the end of the month as a percentage of the population of individuals who are covered by the unemployment insurance.

Since we use the unemployment rate for an analysis based on quarterly data we only use the unemployment rate at the end of each quarter. To avoid endogeneity problems, which could result from the fact that the unemployment rate could also be the outcome of the underlying economic process, we map the unemployment rate (taken at the end of a quarter) to the *subsequent* quarter. For example the unemployment rate for the first quarter of 1998 is actually the unemployment rate at the end of the preceding quarter (i.e. the last quarter of 1997 or equivalently at the end of December 1997).

In order to be able to merge the data on the local unemployment rates with the data of the workers, we have to identify the administrative district to which the workers belong. However this information is not directly available in the raw data. To be able to merge the data on the local unemployment rate with the workers data we use the information on the local unemployment office to which a worker is associated.¹⁴

¹²Information is available at the level of “Provinces et Arrondissements” (ONEM (1998)). The information for 1997 is only available at the level of the “Direction Subrégionale” (ONEM (1997)).

¹³In French: “Chômage complet indemnisé”.

¹⁴Belgian has 10 administrative provinces plus the Brussels Region. Each of these provinces has up to five unemployment offices. The territories associated to the un-

5.3.4 Overview: databases and variables of the raw data

Table 5.3 in the appendix contains an overview for the 356 raw databases, their origin, their names and the time for which they contain information. Like explained above, the databases from the national employment office (ONEM) and the national institute for the social security of the self-employed (INASTI) are not available before 1998. Most of the databases have been delivered in quarterly units. An exception are the databases which are based on the information from the national institute for statistic (INS). These databases are delivered in yearly units. Finally there is a database from the ONEM which contains corrections and educational information and a database which contains the gender specific local unemployment rates.

The second row of Table 5.3 contains the letters which are used as a symbol for a column of databases in the following. For example, if we refer to database “A” we refer to all raw databases in the second column of Table 5.3, which come from the ONSS and which have a name of the form `bestand_RSZ_YYQ_gecodeerd.sas7bdat` where YY stand for the last two digits of the year and Q refers to the quarter.¹⁵

Now we are going to summarise briefly the contents of the different types of raw-databases. Table 5.4 in the appendix provides a complete list of the variables in the different databases together with a short description of each variable.

Database A contains information from the national office for social security (ONSS). The ONSS provides mainly information on employment contracts.¹⁶ This administrative database contains information on the contracts, the status of the employee, reductions of social insurance contributions and information on the employer.

employment offices correspond in many cases to an administrative district. When the territories of unemployment office and administrative districts are different then we either merge the individual data with the unemployment rate of the province or with the unemployment rate of the closest administrative district.

¹⁵For example `_981_` refers to the first quarter of 1998 and `_024_` refers to the fourth quarter of the year 2002.

¹⁶In Belgium the employers have to declare employment for the working contracts at the end of each quarter at the ONSS.

Database B contains information from the national office for social security for the local and provincial administration (ONSSAPL). Roughly speaking this administrative institution has the same role as the ONSS but is in charge if the employer is a local or provincial administration. The variables in database B are therefore similar to the variables of database A.

Database C contains information concerning unemployment, from the national employment office (ONEM). The database contains also information on the status just before unemployment, the duration in unemployment the type of benefits and the source (reason) for the right to receive unemployment benefits.

Database D contains additional information from the ONEM concerning the participation in active labour market programmes: The type of the programmes, start and end of the programmes and information on the importance of the programme (full/part-time). Most of the information is available at a monthly basis.

Database E contains information on child benefits from the national office for family allocation for employees.

Database F contains information from the national institute for the social security of the self-employed. The delivered variables provide information on the start and the end of self-employment activities and the profession of the self-employed.

Database G contains information about the socio-economic status at the end of each quarter and takes information from different institutions into account.

Databases H, I and J contain information from the national institute for statistics. The year of birth, the nationality, gender, sub-region of living and the family relation with respect to the head of the household are in database H. Database I contains information about the quarter of

death and database J contains information about the number of individuals in different age classes of the household.

Databases K - S are databases which contain additional computed variables. These variables are specified in the appendix of the first data-request, which is available in the online-documentation.

Database T contains information from the ONEM that has been requested in the second data-request. This includes corrected information about the start of the first unemployment and information about education.

Database U contains information which has *not* been delivered by the Crossroads Bank for Social Security. The database contains information about the gender specific local unemployment rates and has been collected manually.

The available raw databases are fundamentally different with respect to the type of contents. Consequently also the underlying structure of the different databases varies. We will see in section 5.4 and 5.5 that this has important consequences for the organisation of the databases during the process of data cleaning.

5.3.5 Where are the raw databases stored?

The raw data is stored on the SAS-server of the statistical institute at the University of Louvain-la-Neuve. The SAS-server can be reached under the address: `sas.stat.ucl.ac.be`. Access to the SAS-server is reserved to authorised persons. A login and a password are required. For an access to the database we recommend the use of a FTP (File-Transfer-Protocol) client.

All raw databases for the period 1997-2002 can be found on the SAS-server in the following folder: `/bbbc/carrefour/eerste`. To save disk space on the server, the data files are stored in compressed form in the

file eerste.7z.¹⁷

5.4 Several issues when organising data

In this section we are going to outline some of the fundamental issues that we had to overcome to be able to exploit the databases for economic research.

Most statistical packages require the availability of a single database for computation.¹⁸ However, as we can see from Table 5.3, we have received 356 raw databases. One of the first tasks is therefore a sensible integration of these databases. An additional problem comes from the fact that the databases have been delivered in separate pieces over a total period of 3 years. To be able to work on the data it was necessary to build a flexible framework which enables us to integrate newly received databases into our analysis.

To handle the large number of databases we make extensive use of the macro capabilities of the SAS-programming language. The use of macros has advantages whenever similar tasks have to be repeated several times. At the downside, the use of macros implies a higher degree of abstraction and leads consequently to a reduction of the readability of SAS-programs.

Due to a lack of computational resources and stocking capacities it was not feasible to merge all databases into one big database. In addition the fundamentally different structure of the databases could lead to an important waste of storage space. In order to avoid these problems we regroup the databases with a similar structure, in a first step.

In the description of the contents of the databases in subsection 5.3.4 it turned out that the type of information varies with the source of the data. Now we are going to look into the differences of the structure

¹⁷See <http://www.7-zip.org> for a free program which is able to decompress the file. The file eerste01.zip contains the data for the period 1997-2001.

¹⁸Even in cases where the statistical package can cope with several databases at the same time it is often convenient to merge the data before starting with a statistical analysis.

and explain how we regroup the raw databases in order to reduce the number of databases.

First note that the information in the databases is associated to different entities: an individual, a working contract, an employer or a region. Second the information is available for different time spans e.g. available from 1997-2002 or available from 1998-2002. Finally the information is aggregated at different time unit: singular events, monthly, quarterly or yearly information.

Since our analysis is on the labour market outcome of the workers (in the following we will simply use the term “individual”) and not the employer we are going to present the structure of the different databases on the level of an *individual*.

Table 5.5 provides an overview of the databases and the underlying structure. The databases A, B, M, N, Q, R, S, P contain quarterly information for the period 1997-2002, which is related to the employment contract of the individuals. Note that individuals have working contracts with up to 5 different employers within one quarter and up to 4 working contracts with the same employer. The databases C,D,O,L contain quarterly and monthly information on unemployment for the period 1998-2002. In addition these databases can contain information about the employers. Database E contains quarterly information for the period 1997-2002. Databases F and G contain quarterly information for 1998-2002. The information from the national institute for statistics (database H, I and J) are available yearly. Database I contains the information on the quarter of death of the individuals. The databases K and T contain information on the characteristics of the individuals. Finally database U contains the quarterly gender specific unemployment rates at the level of sub-regions for the period 1998-2001.

We regroup the databases with respect to the underlying structure and we get 7 distinct subgroups of databases. In the next section we will see that one of the first task when cleaning the data is to merge the databases with the same underlying structure.

5.5 Overview of the data cleaning process

In this section we summarise the steps for the data cleaning process. To get a complete picture it is indispensable to take a close look to the SAS-programs which have been used during the data cleaning process. As for the sections above, the online documentation provides complementary information to this section.¹⁹

5.5.1 1st step: cleaning and reorganising of raw data

The first step of data cleaning process provides only basic data-cleaning and the reorganisation of the databases. In order to be able to effectively use the macro capabilities of SAS, all variables have been renamed in this step. In addition the databases are regrouped according to the underlying data structure and information which is available several times is eliminated.

In the following we are going to list the names of the SAS-programs in the correct order. For each program we provide a short description of what the program is doing. The SAS-programs are stored on the SAS-server in the folder `/bbbc/_final`.

`/bbbc/_final/_nodup_x.sas`: Eliminates duplication of datalines and copies the files into the folder `/bbbc/nodupxc`.

`/bbbc/_final/_chi_x.sas`: Renames the variables, creates new variables and merges the available databases within a quarter, per type (A...S). Write the resulting databases into the folder `/bbbc/nodupxc`. For the renaming of the files this SAS-program uses files which are stored in the folder `/poem/nom`. The online documentation provides a complete list for the renaming.

`/bbbc/_final/_xc2_.sas`: Corrects abnormal values for the duration in unemployment (`n_xC2`) in the database C.

¹⁹The first two steps of the data-cleaning process are described in documents which are available in the online-documentation: "First note on the treatment of the data of the activation project" and "La Construction de l'Echantillon 2".

/bbbc/_final/stkflx.sas: Selects individuals who belong to our definition of the stock and flow-sample.

/bbbc/_final/elu_x_.sas: Eliminates all information for all individuals which have not been retained for our stock/flow sample.

/bbbc/_final/___qi_fus.sas: Merges databases horizontally according to the underlying structure. This SAS-program uses a set of different sub-programs which can be found in the folder /bbbc/_final: `_jj_x.sas`, `qi_an.sas`, `qi_as.sas`, `qi_bp_.sas`, `qi_co_-.sas`, `qi_dl_.sas`, `qi_eg.sas`, `qi_hj.sas`, `qi_ii_.sas` and `qi_kk.sas`.

/bbbc/_final/fusver_.sas: Merges the resulting databases vertically (over time).

/bbbc/_final/eluPP_.sas: Eliminates the individuals which have a professional experience at the start of their first unemployment spell.

/bbbc/_final/_i_i_i_X.sas: Creates a database which contains the singular individual information on the death of the individuals on a quarterly basis.

At the end of the first step the regrouped data-files are made available. The files which contain all available databases from 1997-2001 can be found in the folder /bbbc/_bases00 on the SAS-server. The names of these files have a specific structure. For example one name is `zzXX7891.sas7bdat` where XX is replaced by the shortcut from Table 5.5 and the numbers indicate the years for which the data has been included. For example `zzan7891.sas7bdat` contains the variables from the databases A and N for the time period 1997-2001 and `zzco_-891.sas7bdat` contains the variables from the databases C and O for the time period 1998-2001.²⁰

²⁰The merged raw data files, obtained at the end of the 1st step, for the time period

5.5.2 2nd step: creation of aggregate information

In this step of the data cleaning process we use the databases which have been constructed in the first step to generate information on the labour market states at the end of each quarter. To reach this goal we cleaned certain variables and finally aggregated some information.

We are going to list the names of the SAS-programs in the correct order. For each program we provide a short description of what the program is doing. All the files can be found on the SAS-server in the folder `/bbbc/_final`.

`/bbbc/_final/SFR_L_NL.sas`

`/bbbc/_final/SFR_qCBB.sas`

`/bbbc/_final/SFR_i_AB.sas`: Creates the variables TFR and TTP which contain information about the time in employment. This is done in three steps. Each of the three programs provides different level of aggregation (employer, quarter, individual).

`/bbbc/_final/RED_e_AB.sas`

`/bbbc/_final/RED_i_AB.sas`: Generates variables for the active labour market policies from the database AB. This is done in two steps which provide different levels of aggregation (employer, individual). The programs make use of formats as specified in `/bbbc/_final/-Xfor_red.sas`.

`/bbbc/_final/C1_et_C2.sas`: Generates duration variables from the database C. This program relies on a temporary structure which is provided by the SAS-command "summary".

`/bbbc/_final/OK_c_i_CO.sas`: Generates a variable about the labour market state at the end of the third month of each quarter from the database C.

1997-2000 are available in folder `/bbbc/X_bases`.

/bbbc/_final/POL_i_CO.sas: Generates variables which contain information about active labour market policies from database C.

/bbbc/_final/Pol_i_DL.sas: Generates variables which contain information about active labour market policies from database D.

/bbbc/_final/OKf_i_EG.sas: Generates variables for the periods of self-employment from database F.

/bbbc/_final/X_Emtm_X.sas: Merges all generated databases for all available quarters. This program calls the program `/bbbc/_final/dcl_emt.sas`.

/bbbc/_final/X_pol0_X.sas: Generates variables which define the labour market states for the individuals. This program calls the program `/bbbc/_final/dcl_emt.sas`.

/bbbc/_final/X_var0_x.sas: Generates the variables that describe the labour market states for the individuals. This program calls the program `/bbbc/_final/dcl_var.sas`.

/bbbc/_final/xh_i_aaq.sas: Cleans the individual data from the HJ database and provides the required data-structure.

Each stage described here generates files which contain the required variables in a similar form. There are two types of databases: Either for one individual there is one line for each quarter of the observation period or all information for one individual is in one line.

5.5.3 Integration of the year 2001 and 2002 into the database

The databases for the year 2001 and 2002 have been delivered after an extension of the first research project. Unfortunately, at the moment of their creation, the structure of the programs which have been used in

the first and the second step of the data-cleaning process have not been prepared for the integration of data beyond the year 2000.

In order to preserve the SAS-programs and databases for the period 1997-2000 we have created a parallel structure for the extension of the existing data, instead of reconstructing or modifying the existing SAS-programs.

Concretely, we have copied the existing programs from `/bbbc/_-final` into the folder: `/bbbc/v2001`. Then we have adapted the macros which are stored in the file `/bbbc/_x_.sas` and we have gone over the all the programs in the same order as described above. For each program we checked carefully that the new data has been generated correctly. Finally we created the SAS-programs `eins.sas`, `zwei.sas` and `drei.sas` which call all new programs in the required order.

Table 5.6 provides a comparison for the different SAS-programs which have been used for the period 1997-2000 and the year 2001-2002. In the left column we can see the programs of step one and two like described above. In the right column we see the programs `eins.sas`, `zwei.sas` and `drei.sas` and all subprograms. Programs on the same row are equivalent. Note that it is not necessary to re-execute the programs which are responsible for the selection of the sampling (stock/flow) since the individuals have been chosen for their characteristics in the years 1997 and 1998 and we simply add the data for the new years to the original sample.

The data for the year 2002 is not yet integrated into the final database since the data has not been completely available, when we prepared the data for the research project.

In order to integrate the data for the year 2002 on should execute the following tasks:

1. Adapt the macro variables in a way that the year 2002 is considered when executing the program `eins.sas`, `zwei.sas` and `drei.sas`.²¹

²¹It is useful to follow the include command at the start of program `eins.sas`. For example the command `%inc "/bbbc/v2001/_x_.sas";` indicates that the file `_x_.sas` is included at the start of the file `eins.sas`.

Another files which contains macro-variables and which should be checked is the file `/bbbc/poemes/forinoux.sas`.

2. Executing the program eins.sas and check carefully the corresponding log-file eins.log for errors and warning messages. If warning or error messages are printed then one has to adapt the program to fix the problems and re-execute the file eins.sas.
3. Redo the last point with the files zwei.sas and drei.sas (in this order). When only basic data cleaning (as described in subsection 5.5.1) is required, then this point can be skipped.
4. The generated databases can then be downloaded from the SAS-server and processed with statistical packages.

The files which contain all available databases from 1997-2001 can be found in the folder /bbbc/_bases00.

5.5.4 3rd step: Preparation of the data set which has been used for the duration analysis for school-leavers in Belgium

This step can be separated from the first two steps of the data cleaning process. Practically speaking at the start of this third step we take up the necessary databases which have been generated in the first two steps and we work towards a single database. The final database at the end of the third step contains one line per individual. The variables are aggregated on a quarterly level and contain one value for each quarters which is covered by the observation period.

Before the execution of the SAS-programs all used raw databases have been copied manually into a separate folder. The following SAS-programs can be found on the SAS-server in the folder /bbbc/christian the used databases have been copied into the folder /bbbc/-christian/data.²²

We are going to list the names of the SAS-program in the correct order. For each program we provide a short description of what the program is doing and the used databases.

²²The programs can be processed either on the SAS-server or, after minimal customisation, on a desktop computer.

data_x_empt0_x.sas: Prepares variables about the start of unemployment and entry into paid-unemployment.

input: x_empt0_x.sas7bdat, _chom_DN.sas7bdat

output: C_i_empt0.sas7bdat

data_xh_i_aaq.sas: Prepares variables which are available on a yearly basis. Apply several correction to the variables and bring the data into the required form for the merge with the final individual database. Generate indicator variables for qualitative variables. Merge the unemployment rates. The variables which are treated in this program cover the position of the individual in the household, nationality, unemployment office, gender, year of birth, age profile, the number of persons in the household and the local unemployment rates.

input: zzhj7891.sas7bdat, bureaux_apres4.sas7bdat

output: h_i_empt0.sas7bdat

data_VAR_i_CO.sas: Aggregates monthly data to quarterly data.

input: zzco_891.sas7bdat

output: VAR_m_CO.sas7bdat, VAR_q_CO.sas7bdat

data_CO_i_a.sas: Reshapes the quarterly data, which has been generated by the program `data_VAR_i_CO.sas` into the required form.

input: VAR_q_CO.sas7bdat

output: m_i_empt0.sas7bdat

data_constr_policy.sas: Constructs variables that indicate the participation in active labour market policies for each quarter. This step requires the use of several data-sources.

input: zzco_891.sas7bdat, zzbp7891.sas7bdat, zzan7891.sas7bdat

output: pol_q.sas7bdat

data_policy_i_q.sas: Brings the data on the participation in active labour market policies into the required form.

input: pol_q.sas7bdat

output: p_i_pol.sas7bdat

data_verify_selection_i_.sas: Introduces the corrected variables which have been delivered after the second data request (raw-database T). Prepare the data for a merge with the existing databases. Integrate the information on the education of the individuals into the data.

input: bestand_gegenvens_rva_gecodeerd.sas7bdat

output: d_i_verify.sas7bdat

data_i_merge.sas: Merges the files which have been generated until now and apply some verifications. Generate separate sub-samples, which contain a flow or a stock sample for men and women.

input: h_i_empt0.sas7bdat, v_i_empt0.sas7bdat, c_i_empt0.sas7bdat, m_i_empt0.sas7bdat, p_i_pol.sas7bdat, d_i_verify.sas7bdat

output: d_i_empt0.sas7bdat, pol_homme_flow_B1.sas7bdat, pol_femme_flow_B1.sas7bdat, pol_homme_stock_B1.sas7bdat, pol_femme_stock_B1.sas7bdat.

data_employer_AN_BP_q_i_merge.sas: Prepares the variables concerning the characteristic of the employer. Keeps only the information for the “most important employer” (per quarter).²³ Apply corrections and bring the data into the required form.

input: zzan7891.sas7bdat, zzbp7891.sas7bdat, ab_q_sfr.sas7bdat

output: zAB_i_7891.sas7bdat

data_i_empl_merge.sas: Merges the flow data sets with the information about the most important employer per quarter.

input: pol_homme_flow_B1.sas7bdat, pol_femme_flow_B1.sas7bdat, zAB_i_7891.sas7bdat

output: pol_homme_flow_B.sas7bdat, pol_femme_flow_B.sas7bdat

²³For the definition of the “most important employer” see page 5 of the document “La Construction de l’Echantillon 2” in the online documentation.

Fspell111_fin.sas: Prepares the database for the application of duration models. Generates all necessary variables and indicators which allow to model the labour market trajectories of the individuals.

input: pol_homme_flow_B.sas7bdat, pol_femme_flow_B.sas7bdat
output: FS1_h.sas7bdat, FS1_f.sas7bdat

5.6 Where to start a new project?

Now - where to start with a new project? The universal answer to this question is obviously that this depends on the project. If the new project is relatively close to the existing ones it might only be necessary to modify a few programs to generate new data. However if the new project has fundamental differences it may be necessary to step in much earlier. The following use-cases might be helpful for a first orientation.

Pull new variables into the existing projects: Here we have in mind a use-case where one tries to extend the existing final databases by variables which have not yet been included. The first step would be the identification of the original name of the required variable. Table 5.2 and Table 5.4 provide an overview of the existing variables. In the second step one has to identify the name of this variable at the end of the first step of data-cleaning. The online documentation provides searchable lists for the translation of the original names. The Table 5.5 shows in which database the variables can be found. The best strategy for the data cleaning and introduction is to extend the existing programs of the third step of data cleaning process. Sub-section 5.5.4 gives an overview of the programs and the databases which have been used. It might be useful to consult the online documentation for the nomenclature of the variables. Finally one has to adapt the programs which prepare the data for the econometric analysis.

Analyse a different active labour market policy with duration models: Now we assume that the new project aims to analyse the effect of

an active labour market policy on the unemployment or employment duration of participating long-term unemployed workers. The essential step here would be to adapt the programs in the third step of data cleaning in which the information on the active labour market program are generated.

Sub-Section 5.5.4 reveals that the programs `data_constr_policy.sas` and `data_policy_i_q.sas` aggregate the relevant information from the different databases. Again, it might be useful to consult the online documentation for the nomenclature of the involved variables. Finally all subsequent programs have to be updated and executed.

Prepare the data for a completely different type of analysis: If a new project requires a structure which is fundamentally different from the provided one (e.g. monthly or yearly aggregation, or a different sampling) then it is probably wise to start with the databases provided at the end of the first step. The programs in place might be instructive on how to handle the underlying data-structure.

Of course it is not possible to give fix rules for the data-preparation of a new project. This section provides only a first orientation and it might turn out that there exist different ways to generate the required data.

5.7 Additional resources

This chapter provides a short overview of additional resources which might be helpful when working on the data.

Programming resources: The data preparation requires an extensive use of the SAS-program. While SAS provides flexible tools for data handling it takes some effort to get familiar with the programming language. The home-page of the SAS-server at the statistical institute provides access to the complete documentation of SAS.²⁴

²⁴Throughout the data cleaning process we have used the version 8.2 of SAS.

<http://sas.stat.ucl.ac.be/>

Programming tools: Working on a remote server can be troublesome since it requires that all files have to be transferred into the right place before they can be executed. The following set of free tools can make the life of a researcher easier:

Jedit is a free and multi-platform programming editor. Most important it can be extended by plug-ins via the build in plug-in manager. The editor can highlight SAS-syntax and provides block-wise editing. The FTP plug-in makes it possible to edit and save the programs on the SAS-server almost as if they would be on the desktop computer. The Jdiff plug-in provides easy comparison between similar programs. The editor can be obtained via <http://www.jedit.org/>.²⁵

FileZilla is a free and powerful FTP-client. FileZilla is a tool which is useful to transfer programs or databases between a desktop computer and the remote SAS-server.²⁶

ssh: For the execution of SAS-programs on the server we recommend the use of a Secure Shell (SSH) which provides access to the command line on the server. Ssh-clients enables one to execute programs on the server. The ssh client which is provided by Cygwin is free and stable.

Additional documentation: The already mentioned online documentation provides links to all kinds of documents which have been used during the data cleaning process and the nomenclature of most of the variables:

<http://www.i-res.ucl.ac.be/Axe4/doc/docstart.html>.

The Belgian Crossroads Bank for Social provides own documentation on the homepage of the Datawarehouse:

http://ksz-bcss.fgov.be/fr/documentation/reg%20BCSS_1.htm.

²⁵Make sure you have installed the latest java JRE before installing Jedit (see for example: <http://java.sun.com/javase/downloads/index.jsp>)

²⁶The SAS-server requires you to use passive-mode for the FTP-program.

5.8 End

The federal public administration of different countries make administrative data available to researcher in the last years. These databases have the advantage of offering huge sample size combined with rich information, both of which is desired by economists with an empirical orientation.

However the complexity of administrative data and the huge diversity in administrative records of the individuals in the database have serious consequences for empirical research. First the preparation and data cleaning require a lot of resources in terms of knowledge, computational facilities and time. Second data aggregation and consequently the loss of information is often unavoidable to assure that the database is processable. Finally, the decisions which have to be made during the process of data-cleaning may have an impact on the results of the empirical analysis. Despite the importance of rigorous data-cleaning the work of data-cleaning is hardly recognised and does not play a significant role for scientific publications.

This paper tries to assure that the work which went into these databases is documented and preserved. We hope that this document is useful in future research projects.

Bibliography

- Adams, D. (1979). *The Hitchhiker's Guide To The Galaxy*. Pan Macmillan.
- Cockx, B. and Göbel, C. (2005). Évaluation microéconométrique du Plan Avantage à l'Embauche pour une population de jeunes défavorisés. In Cockx, B., Sneesens, H., and van der Linden, B., editors, *Evaluations micro et macroéconomiques des allègements de la (para)fiscalité en Belgique*, pages 29–96. Academia Press, Gent.
- Cockx, B., Göbel, C., and van der Linden, B. (2004a). *Politique d'activation pour des jeunes chômeurs de longue durée sans expérience de travail. Une évaluation*. Problèmes actuels concernant la cohésion sociale. Academia Press, Gent.
- Costa-Dias, M. and Blundell, R. (2002b). Alternative approaches to evaluation in empirical microeconomics. *Portuguese Economic Journal*, 1(2):91–115.
- Deprest, J. and Robben, F. (2003). E-government: the approach of the Belgian federal administration. *Paper, June 2003*. <http://www.ksz.fgov.be/En/CBSS.htm>.
- Heckman, J. J., LaLonde, R. J., and Smith, J. A. (1999). The economics and econometrics of active labor market programs. In Ashenfelter, O. and Card, D., editors, *Handbook of Labour Economics*, volume 3A, chapter 31, pages 1865–2097. Elsevier, Amsterdam.
- ONEM (1997). *Bulletin Mensuel Décembre 1997*. Office national de l'Emploi, Brussels.
- ONEM (1998). *Bulletin Mensuel Janvier 1998*. Office national de l'Emploi, Brussels.

Appendix

A Tables

Table 5.1: English/French/Flemisch names for Belgian institutions

National employment office	
Office nationale de l'emploi	ONEM
Rijksdienst voor Arbeidsvoorziening	RVA
National office for social security	
Office national Securite Sociale	ONSS
Rijksdienst voor Sociale Zekerheid	RSZ
National office for social security fo the local and provicial administration	
L'Office national de sécurité sociale des administrations provinciales et locales	ONSSAPL
Rijksdienst voor Sociale Zekerheid van de Provinciale en Plaatselijke Overheidsdiensten	RSZ/PO
National Institute for the Social Security of the Self-employed	
Institut National d'Assurances sociales pour travailleurs indépendants	INASTI
Rijksinstituut voor de Sociale Verzekeringen der Zelfstandigen	RSVZ
National office for family allowances for employees	
Office national d'allocations familiales pour travailleurs salarié	ONAFTS
Rijksdienst voor Kinderbijslag voor Werknemers	RKW
National institut for statistics	
Institut National des Statistiques	INS
Nationaal Instituut voor de Statistiek	NIS

Table 5.2: Variables in the data request

Data	Source	Data description	Possible Values
<i>Variables which allow to identify the individual labour market state at the end of each quarter</i>			
Persoon - Socio-economische positie	Di-verses (con-struit)	Labour market state at the end of the quarter	according to the number of states
Persoon - Overlijden	NIS	Date of death	quarter / year
Stage d'attente 6 mois	NIS (com-puted)	Indicator for the duration in the waiting period before the payment of unemployment benefits	0 = 9 months 1 = 6 months
<i>Variables which allow to identify the type of an active labour market policy</i>			
Activering - Arbeids-regime (BESLIS)	ONEm	Indicator for a full/part-time activation	2 values
Activering - Montant moyen_A	ONEm (com-puted)	Indicator for the mean of the activated benefits per quarter	in classes of 100 Euro
Activering - % Arbeidsduur (BREUKACT)	ONEm	% of time in employment compared to a full time employment	in percentage
Activering - Statuut RVA (Fiche 7)	ONEm	Type of activation programme	according to the number of types
Activering - Toeken-ningsvoorwaarden (TYPACT)	ONEm	Eligibility criteria (age, unemployment duration, etc.) for a programme	according to the number of criteria
Werkloze actief in PWA -Statuut (STATALE)	ONEm	Type of selection for an occupation in ALE	- office - spontaneous
Bijdragevermindering Montant moyen_B	ONSS (com-puted)	Mean value of reduction in the social insurance contribution per employer	in classes of 100 Euro
Bijdragevermindering Verminderingscode (Codred)	ONSS	Indicator for the type of reduction	according to the number of possible reductions
Arbeidsprestatie - Consolidatietype (Cntycd)	ONSS-APL	Indicator for the type of reduction in social insurance contributions	according to the number of types
Arbeidsprestatie - Consolidatiesubtype (Cnstcd)	ONSS-APL	Indicator for the type of reduction in social insurance contributions	according to the number of types
<i>Variables which allow to better identify the timing of transition between states</i>			
Persoon - Identificatienummer	NIS	Individual identification number (anonymous)	number

Werkgever - Stamnummer (Matric)	ONSS	Identification number for employer (anonymous)	number
Werkgever - Aansluitingsnummer	ONSS-APL	Identification number for employer (anonymous)	number
Activering - Begindatum (BEGINACT)	ONEm	Start of the activation period	quarter / year
Activering - Einddatum (EINDACT)	ONEm	End of the activation period	quarter / year
Activering - Vergoede dagen (DAGEN)	ONEm	Number of days in activation	number of days
Werkloosheid - (DUUR)	ONEm	Unemployment duration at the end of the quarter	number of months
Werkloosheid - Vergoede dagen (DAGEN)	ONEm	Number of days in paid-unemployment	number of days
Arbeidsprestatie - Begindatum	INASTI	Start of self-employment activity	quarter / year
Arbeidsprestatie - Schrappingsdatum	INASTI	End of self-employment activity	quarter / year
Arbeidsvolume - VTE incl gelijkgestelde dagen (...)	ONSS	Indicator per employer: the fraction of the quarter in employment ("jours assimilés" included)	continuous value
Arbeidsvolume - VTE incl gelijkgestelde dagen (Activte)	ONSS-APL	Indicator per employer: the fraction of the quarter in employment ("jours assimilés" included)	continuous value
<i>The (taxable) gross-wage</i>			
SALnet	ONSS/ONEm (computed)	Indicator for (taxable) gross wage	in classes of 100 Euro
<i>Socio-economic characteristics</i>			
Relation	NIS (derived)	Family relation with respect to the head of the household (1st January / 1st July)	
Nmembres	NIS (derived)	Number of members in the household per age-class (1st January / 1st July)	Number per age class : [0,1), [1,3), [3,6), [6,12), [12, 18), [18,30), [18,50), [50,65), [65,75), +75 ans.
Persoon - geboortedatum	NIS	Year of birth	year
Persoon - Geslacht	NIS	Gender	2 values

Persoon - Nationaliteit	NIS	Nationality (1st January /1st July)	Belgian European Union (not Belgian) Not EU
Kinderbijslag - Begin der betalingen (DEB_BET)	ON-AFTS	Start of payment per child	quarter /year
Kinderbijslag - Einde der betalingen (FIN_BET)	ON-AFTS	End of payment per child	quarter /year
Persoon - Woonplaats	NIS	Indicator for sub-region of living	according to the number of unemployment office
<i>Characteristics which are conditional on the labour market state</i>			
<i>- Unemployment</i>			
Werkloosheid - Bedrag dagvergoeding (DGNDMND)	ONEm (computed)	Indicates the daily average amount for unemployment benefits within a quarter (times 78 to normalise the amounts per quarter)	in classes of 100 Euro
Werkloosheid - Activiteit vooraf (S11END)	ONEm	Indicates the situation before unemployment	according to number of situations
Werkloosheid - Statuut (FICHE 7)	ONEm	Indicates the source (reason) for the right to receive unemployment benefits	according to number of sources
Werkloosheid - Vergoedingscategorie (FICHE 61)	ONEm	Indicates the type of received benefits	according to number of types
<i>- Employment</i>			
Arbeidsprestatie - Beroep (Beroepc)	INASTI	Indicates the profession for self-employment	1 position (when replaced by NACE-code: 2 positions)
Arbeidsprestatie - Consolidatietype (Cntycd)	ONSS-APL	Identifies the category to which an employee belongs to	according to the number of categories
Arbeidsprestatie - Consolidatiesubtype (Cnstcd)	ONSS-APL	Identifies the sub-category to which an employee belongs to	according to the number of categories
Arbeidsregime - Percentage deeltijds (Tauxpt)	ONSS-APL	Identifies per employer the % of part-time within a "homogeneous" period before the end of a quarter	%
Arbeidsvolume - VTE excl. Gelijkgestelde dagen (Vte)	ONSS-APL	Indicates per employer the equivalent to a full-time employment ("jours assimilés" excluded)	continuous value

SALbrut	ONSS- APL (com- puted)	Indicates per employer the quarterly gross salary expressed as equivalent of a full-time employment	in classes of 100 Euro
Werkgever - Dimensie (Codimp)	ONSS- APL	Code for the size of the employer	in classes
Werkgever - Nacecode (Nacecd)	ONSS- APL	Principal sector of the employer	first two numbers
Arbeidsprestatie - Werknemerscode (Codtra)	ONSS	Allows to identify the status of the employee (blue collar, white collar,...)	according to the number of statuses
Arbeidsprestatie - werknemersklasse (Clatra)	ONSS	Status of the employee	values from 1 to 4
Arbeidsprestatie - werknemersklasse speciaal (Clatr2)	ONSS	Delivers information on the status of the worker - this allows to understand the level of the salary	Two letters or numbers (16 different values)
Arbeidsregime - Percentage deeltijds (Tauxpt)	ONSS	Identifies per employer the % of part-time within a "homogeneous" period before the end of a quarter	%
Arbeidsvolume - VTE excl. Gelijkgestelde dagen (Eqtpsp)	ONSS	Indicates per employer the equivalent to a full-time employment ("jours assimilés" excluded)	continuous value
SALcoût	ONSS (com- puted)	Indicates per employer the labour-cost (salary)	in classes of 100 Euro
Loon - Verbrekingsvegoeding (Preavi)	ONSS (com- puted)	Indicates the amount for the "préavis rapporté à l'ONSS"	in classes of 100 Euro
Loon - Forfaitloon (Salfor)	ONSS (com- puted)	Indicates the salary (forfaitaire perçu) for certain categories of workers	in classes of 100 Euro
Werkgever - Dimensie (Codimp)	ONSS	Code for the size of the employer	in classes
Werkgever - Nacecode (Codnac)	ONSS	Principal sector of the employer	first two numbers
Werkgever - Regionalisatiecode (Codreg)	ONSS	Indicates if the enterprise has one or several subsidiaries	values 0 or 1
Werkgever - Paritair Comité (Nr-pc)	ONSS	Number of the "commission paritaire"	number
Werkgever - Sector (Secemp)	ONSS	Indicates if private or public sector	2 values
<i>Added after the second data-request</i>			

PREFIX	ONEm	Status of the unemployed at the moment of first payment of benefits	according to classes
SUFFIX	ONEm	Type of benefits at the moment of the first payment of benefits	according to classes
ETUDES	ONEm	Level of schooling (in 2002)	according to number of levels available
DATES04	ONEm	Date of first payment of benefits	date

Table 5.3: Overview: 356 databases

annual	U	.	.	.	X																	.	.	.	bureaux apres4...
ONEM	T	X																	.	.	.	..._GEGEVENS_RVA_...			
a8	S	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	ACHTSTE_AV_SALFOR_YYQ_...		
a7	R	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	ACHTSTE_AV_PEAUL_YYQ_...		
a6	Q	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	ZEVENDE_AV_YYQ_...		
a5	P	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	ZESDE_AV_YYQ_...		
a4	O	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	VJFDE_AV_YYQ_...		
a3	N	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	VIERDE_AV_YYQ_...		
a2	M	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	DERDE_AV_YYQ_...		
a1	L	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	TWEEDE_AV_YYQ_...		
INS family	K	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	EERSTE_AFGELIDE_VAR_...		
INS death	J	X																	.	.	.	...GEZINSLEDEN_YY_...			
INS	I	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	..._OVERLEDENEN_YY_...			
Socn-econ.	H	X																	.	.	.	..._PERS_GEG_YY_...			
INASTI	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	..._SOCECOPOS_YYQ_...		
ONAFIS	F	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	..._RSVZ_YYQ_...		
ONEM	E	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	..._RKW_YYQ_...		
ONEM	D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	..._RVA_YYQ_ACT_...		
ONEM	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	..._RVA_YYQ_OD_...		
ONSSAPL	B	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	..._PPO_YYQ_...		
ONSS	A	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	..._RSZ_YYQ_...		
YY/Q	97/1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	97/2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	97/3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	97/4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	98/1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	98/2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	98/3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	98/4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	99/1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	99/2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	99/3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	99/4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	00/1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	00/2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	00/3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	00/4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	01/1	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	01/2	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	01/3	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	01/4	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			
	02/1	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n			
	02/2	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n			
	02/3	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n			
	02/4	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n			

Note: “X” indicates that the data-cleaning process is complete for the database. “n” indicates that the database is not yet integrated (but is available as raw data). A “.” indicates that the database is either not available or not yet available as raw database.

YY refers to the last two digits of the year, Q refers to the quarter.

A typical example for the name of a raw database is: “bestand_PPO_981_gecodeerd.sas7bdat”. In the table we suppress the leading “bestand” and the following “gecodeerd.sas7bdat”.

Table 5.4: Overview of variables by database

	Variable-name	Description
A	CLATR2	Delivers information on the status of the worker - this allows to understand the level of the salary
A	CLATRA	Status of the employee
A	CODIMP	Code for the size of the employer
A	codnac	Principal sector of the employer
A	CODRED	Indicator for the type of reduction
A	CODREG	Indicates if the enterprise has one or several subsidiaries
A	CODTRA	Allows to identify the status of the employee (blue collar, while collar,...)
A	EQTPSP	Indicates per employer the equivalent to a full-time employment ("jours assimilés" excluded)
A	matric_gecodeerd	Identification number for employer (anonymous)
A	NR_CP	Number of the "commission paritaire"
A	SECEMP	Indicates if private or public sector
A	TAUXAA	Indicator per employer: the fraction of the quarter in employment ("jours assimilés" included)
A	TAUXPT	Identifies per employer the % of part-time within a "homogeneous" period before the end of a quarter
B	ACTIVE	Indicator per employer: the fraction of the quarter in employment ("jours assimilés" included)
B	CNSTCD	Indicator for the type of reduction in social insurance contributions
B	cnstcd_type	Identifies the sub-category to which an employee belongs to
B	CNTYCD	Indicator for the type of reduction in social insurance contributions
B	cntycd_type	Identifies the category to which an employee belongs to
B	CODIMP	Code for the size of the employer
B	codnac	Principal sector of the employer
B	gecodeerd_matric	Identification number for employer (anonymous)
B	TAUXPT	Identifies per employer the % of part-time within a "homogeneous" period before the end of a quarter
B	VTE	Indicates per employer the equivalent to a full-time employment ("jours assimilés" excluded)
C	DAGEN	Number of days in activation
C	DUUR	Unemployment duration at the end of the quarter
C	FICHE61	Indicates the type of received benefits
C	FICHE7	Indicates the source (reason) for the right to receive unemployment benefits
C	REFMAAND	Month
C	S11END	Indicates the situation before unemployment
D	beginact	Start of the activation period
D	BESLIS	Indicator for a full/part-time activation
D	BREUKACT	% of time in employment compared to a full time employment
D	eindact	End of the activation period
D	FICHE7	Type of activation programme

D	REFMAAND	Month
D	STATALE	Type of selection for an occupation in ALE
		Eligibility criteria (age, unemployment duration, etc.) for a programme
D	TYPEACT	
E	aantal_rechtgeevenden	Information on child benefits
F	begaansc	Start of self-employment activity
F	beroepc	Indicates the profession for self-employment
F	schaansc	End of self-employment activity
G	nomenc	Labour market state at the end of the quarter
H	geboortejaar	Year of birth
H	gecodeerd_insz	Individual identification number (anonymous)
H	geslacht	Gender
H	nationaliteit	Nationality (1st January/1st July)
		Family relation with respect to the head of the household (1st January/1st July)
H	relatie	
H	werkloosheidsbureau	Indicator for sub-region of living
I	overlijden	Date of death
		Number of members in the household per age-class (1st January/1st July)
J	lft	
J	Nmembres	Number of members in the household per age-class (1st January/1st July)
		Indicator for the duration in the waiting period before the payment of unemployment benefits
K	isstage	
L	FICHE7	Type of activation programme
		Indicator for the mean of the activated benefits per programme and quarter
L	montant_moyen_A	
M	CODRED	Indicator of the type of reduction for the employer
		Mean value of reduction in the social insurance contribution per employer
M	montant_moyen_B	
		Mean value of reduction in the social insurance contribution per employer
M	Smontant_moyen_B	
N	SALnet	Indicator for (taxable) gross wage
		Indicates the daily average amount for unemployment benefits within a quarter (times 78 to normalise the amounts per quarter)
O	montant_journalier	
O	REFMAAND	Month
		Indicates per employer the quarterly gross salary expressed as equivalent of a full-time employment
P	SALbrut	
Q	SALcout	Indicates per employer the labour-cost (salary)
R	Preavis	Indicates the amount for the "préavis rapporté à l'ONSS"
		Indicates the salary (forfaitaire perçu) for certain categories of workers
S	Salforf	
T	DATES04	Date of first payment of benefits
T	ETUDES	Level of schooling (in 2002)
		Status of the unemployed at the moment of first payment of benefits
T	PREFIX	
		Type of benefits at the moment of the first payment of benefits
T	SUFFIX	
U	percent	Local unemployment rate in %
U	percent_10	The variable "percent" times 10

Table 5.5: Overview: structure of the databases

raw databases	A,B,M,N, P,Q,R,S	C,D,O,L	E	E,G	H,I,J	K,T	U
entity	indiv.	indiv.	indiv.	indiv.	indiv.	indiv.	subregion
contracts per employer	max 4	max 4					
number of employers	max 5						
monthly data		X					
quarterly data	X	X	X	X	X		X
yearly data					X		
years available	6	5	6	5	6	singular	4
shortcut	AB,AN,BP	CO,DL		EG	HJ		

Note: indiv. means that the entity is an individual (a worker).
Individuals have working contracts with up to 5 different employers per quarter and up to 4 different contracts with the same employer.
Databases C,D,O,L are delivered quarterly but contain also monthly information. Database I is delivered yearly but contains also quarterly information. The available years for database U (the gender specific unemployment rate) could be extended manually.

Table 5.6: Comparison for SAS-programs: 1997-2000 vs 2001-2002

1997-2000	2001-2002
<code>_nodup_x.sas</code>	<code>eins.sas</code>
<code>_chi_x.sas</code>	<code>av02_dup.sas</code>
<code>_xc2_.sas</code>	<code>bv02_chi.sas</code>
	<code>cv02_xc2.sas</code>
<code>stkflx.sas</code>	<code>zwei.sas</code>
<code>elu_x_.sas</code>	<code>not necessary</code>
<code>_qi_fus.sas</code>	<code>not necessary</code>
<code> jjx_.sas</code>	<code>dv02_jj_.sas</code>
<code> qi_an_.sas</code>	<code>ev02_an_.sas</code>
<code> qi_as_.sas</code>	<code>skipped</code>
<code> qi_bp_.sas</code>	<code>fv02_bp_.sas</code>
<code> qi_co_.sas</code>	<code>gv02_co_.sas</code>
<code> qi_dl_.sas</code>	<code>hv02_dl_.sas</code>
<code> qi_eg_.sas</code>	<code>iv02_eg_.sas</code>
<code> qi_hj_.sas</code>	<code>jv02_hj_.sas</code>
<code> qi_ii_.sas</code>	<code>kv02_ii_.sas</code>
<code> qi_kk_.sas</code>	<code>skipped</code>
<code>fusver_.sas</code>	<code>lv02_fus.sas</code>
<code>eluPP_.sas</code>	<code>not necessary</code>
<code>_i_i_i_X.sas</code>	<code>mv02_iii.sas</code>
<code>SFR_L_NL.sas</code>	<code>drei.sas</code>
<code>SFR_qCBB.sas</code>	<code>nv02_FRa.sas</code>
<code>SFR_i_AB.sas</code>	<code>ov02_FRb.sas</code>
<code>RED_e_AB.sas</code>	<code>pv02_FRc.sas</code>
<code>RED_i_AB.sas</code>	<code>qv02_ABa.sas</code>
<code>C1_et_C2.sas</code>	<code>rv02_ABB.sas</code>
<code>OK_c_i_CO.sas</code>	<code>sv02_CC2.sas</code>
<code>POL_i_CO.sas</code>	<code>tv02_kCO.sas</code>
<code>Pol_i_DL.sas</code>	<code>uv02_pCO.sas</code>
<code>OKf_i_EG.sas</code>	<code>vv02_pDL.sas</code>
<code>X_Emtm_X.sas</code>	<code>wv02_kEG.sas</code>
<code>X_pol0_X.sas</code>	<code>xv02_Emt.sas</code>
<code>X_var0_x.sas</code>	<code>yv02_pol.sas</code>
<code>xh_i_aaq.sas</code>	<code>zv02_var.sas</code>
	<code>_v02_aaq.sas</code>

Note: Indented programs are called by the preceding non-indented programs. e.g. `av02_dup.sas` is a SAS-program which is called by the program `eins.sas`.

The SAS-programs for the period 1997-2000 are stored in the folder `/bbbc/_final/` on the SAS-server.

The SAS-programs for the period 2001-2002 are stored in the folder `/bbbc/v2001/` on the SAS-server.

Chapter 6

Conclusion

6.1 Ideas for future research

In a next step, we would like to test if the effect of participation depends on the duration spent in the subsidised employment. Some of our results suggest that it is likely that the maximum eligibility period could be shortened and at the same time the effect of participation could be preserved: In chapter two we find no effect on the direct transition rate into regular employment during participation. Instead, we find that the shortening of the duration until regular employment has two sources. Firstly a strong positive effect on individuals who leave subsidised employment before the maximum entitlement period. Secondly almost all of the workers who stay in subsidised employment period have a transition into regular employment at the end of the subsidised employment. These result makes us believe that the maximum entitlement period could be shortened and at the same time the desired effect of the participation could be maintained. However this hypothesis is based on the assumption that the effect during, after and at the end of participation stays roughly the same when shortening the entitlement period. This assumption could partly be tested by the integration of the duration spent in subsidised employment as an explanatory variable for the transition out of subsequent spells.

In the same line of thought, one could integrate the length of other preceding spells as explanatory variables for the transition rates in or-

der to check if the effect of participation depends on the duration of preceding spells. For example, one could verify whether the effect of participation changes with the duration in unemployment before the transition into the employment subsidy programme.

Another important extension would be to add employer information into the transitions out of employment spells, in the fourth chapter. This would allow us to check if our findings depend on characteristics of employers and therefore provide information which is necessary to generalise our findings, beyond the effect of the take-up. Moreover, one could use these estimates to simulate the effect of participation for the hypothetical case where the policy would be targeted to certain employer characteristics, e.g. sectors.

The following point is of a technical nature: It would be interesting to augment the number of points of support for the multivariate discrete distribution of the unobserved heterogeneity. Even though recent Monte-Carlo evidence suggests that even a few points of support can help to control for selection in unobserved characteristics, it would be valuable to test to which extent our findings are sensitive to the inclusion of additional points of support.

In chapter four we simulate the effects for different subgroups of the population. This decision has been driven by the desire to take the actual correlation of observed characteristics into account. For example, if policy makers would like to obtain a simple rule to target the employment subsidy programme, they could use the type of simulation provided in this thesis. An alternative would be to simulate how the effects depend on a change of the explanatory variables, one by one. This result would be informative on how the effect of participation changes if only on characteristic is changed.

Finally, as an extension to the empirical framework of chapter four, one could compare the effects of participation for school leavers versus young workers who have already had a employment experience in order to investigate if it is the *first* employment experience which makes a difference rather than a mere increase of the employment experience. This could for example be realised by including additional information on a second participation spell.

