Research Data Centre (FDZ)
of the German Federal
Employment Agency (BA)
at the Institute for
Employment Research (IAB)

FDZ·Methodenreport

03/2010
EN

Methodological aspects of labour market data

How to use data swapping to create useful dummy data for panel datasets

Peter Jacobebbinghaus
Dana Müller
Agnes Orban

Bundesagentur für Arbeit

How to use data swapping to create useful dummy data for panel datasets

Peter Jacobebbinghaus (Institute for Employment Research)
Dana Müller (Institute for Employment Research)
Agnes Orban (University of Mannheim)

FDZ-Methodenreporte (FDZ method reports) deal with methodical aspects of FDZ data and help users in the analysis of these data. In addition, users can publish their results in a citable manner and present them for public discussion.

Contents

Abstract 3

Zusammenfassung 3

1 Introduction 4

2 Purpose and characteristics of dummy data 5

3 Constraining the swaps 6

4 Summary 10

References 11

Appendix: Stata code of the important steps 12

Abstract

Many research data centres (RDCs) provide access to micro data by means of on-site use and remote execution of programs. An efficient usage of these modes of data access requires the researchers to have dummy data, which allows them to familiarize with the real data. These dummy data must be anonymous and look the same as the original data, but they do not have to render valid results. For complex datasets such as panel data or linked data, the creation of useful dummy data is not trivial. In this paper we suggest to use data swapping with constraints in order to keep some consistency and correlation between variables within cross-sections and over time. It is easy to be implemented even for datasets with many variables and many survey waves.

Zusammenfassung

Einige Forschungsdatenzentren (FDZ) bieten den Zugang zu Mikrodaten auch per Gastaufenhalt oder Datenfernverarbeitung an. Eine effiziente Nutzung dieser Zugangswege setzt voraus, dass sich die Nutzerinnen und Nutzer im Vorfeld der Analysen anhand von Testdaten mit der Struktur der Echtdaten vertraut machen. Diese Testdaten müssen absolut anonymisiert sein, in ihrer Struktur den Echtdaten entsprechen, aber keine validen Ergebnisse liefern. Für komplexe Datensätze wie Paneldaten oder verknüpfte Daten ist die Erstellung solcher Testdaten nicht trivial. In diesem Papier schlagen wir vor Data Swapping mit Restriktionen anzuwenden, so dass ein gewisser Grad an Konsistenz und Korrelation zwischen den Merkmalen sowohl Innerhalb eines Querschnittes als auch über die Zeit erhalten bleibt. Diese Methode ist auch für Datensätze mit vielen Variablen und Erhebungswellen einfach anzuwenden.

Keywords: Dummy data, data swapping, panel data, establishment data

1 Introduction

In recent years numerous research data centres (RDCs) were established to improve researchers' access to micro data. For micro data that are not available in anonymous versions, two ways of data access have gained importance: on-site use and remote data access. On-site use means that the researchers access the data on secure computers within RDCs. Remote data access either means that researchers log on an RDC computer from their workplaces or that they send programs to the RDCs and the RDC staff runs the programs and returns the results. The latter mode of data access is also referred to as remote execution. Efficient usage of on-site use and remote execution requires the researchers to have dummy data that allow them to familiarize with the real data. These dummy data must be anonymous and look the same as the original data, but they do not have to render valid results.[1]

For complex datasets such as panel data or linked data the creation of useful dummy data is not trivial because they have to meet certain requirements. Nevertheless, as far as we know, there is no literature on how to create useful dummy data, especially, how to do this efficiently. One problem in creating dummy data that occurs with panel data is to keep some consistency of the variables over time. In this paper we suggest to use data swapping with some constraints in order to keep some consistency and correlation between variables within cross-sections and over time. As it is easy to implement, even for datasets, many variables and many waves, we think this method is interesting for RDCs and other institutions that need to provide dummy data.

So far we applied this procedure to create dummy data for two of our RDC's datasets: the IAB Establishment Panel (IAB-EP) and the Establishment History Panel (BHP). The examples in this report relate to the implementation of our procedure for the IAB-EP. The IAB-EP is an establishment survey with about 340 variables in each of the 16 waves from 1993 to 2008 that are currently available. The participating firms are assured that the information they provide will not be published. As the identification of some of the firms based on the survey information can not completely be ruled out, data access is restricted to on-site use and remote execution. The dummy data we create by data swapping allows our users to prepare their programs at their workplaces to a well advanced stage.

In the following section we describe the purpose of dummy data and the requirements they have to meet. Section 3 describes how we constrain the data swapping to keep consistency. Section 4 summarises the report. The appendix provides code snippets of the important steps.

[1] Dummy data are also called test data or structural data. In Section 2 we state the characteristics of dummy data more precisely.

2 Purpose and characteristics of dummy data

The purpose of dummy data is that researchers can familiarize with data of restricted access before they get access to the real data. They are used to prepare programs before on-site use terms at RDCs in order to shorten on-site use terms. This saves time and money. Dummy data are also necessary for data access by remote execution, i.e. programs are sent to RDCs and results are returned. It is hard to write programs without bugs without data to check them.

The two basic requirements dummy data have to meet are:

1. Disclosure risk: Dummy data must be absolutely anonymous. This means that the risk of disclosure is zero. No information about single real persons or firms shall be inferable from the dummy data.

2. Utility: Programs run on the dummy data as far as they would with the original data and reveal in which direction final results might go.

Requirement 1 is fix, especially if the access to the dummy data is not restricted, e.g., if they are placed on the internet. Requirement 2, the degree of utility, depends on how complex the data is and how much effort can be spent on the design of the dummy data.[2] Useful dummy data share the following characteristics:

- Dummy data contain the same list of variables.
- Dummy data have the same file and variable names as the original data.
- Variables in the dummy data should contain the same value ranges as the original data. This means that aggregation to broader categories (e.g., federal states instead of local community codes) is not possible.

Desirable is to keep as much correlation between the variables as possible. If a variable is filled after certain values of a filter question, only this should also be the case for the dummy data. If the data are longitudinal, the consistency over time should be preserved, i.e. the industry should not change each year if it does not do so in the original data. And if there is an unbalanced panel, the structure of entries and exits should be similar to the original data.

These requirements show that dummy data are different from public use files or factual anonymous data which is created for valid analyses. It is not the purpose of dummy data to generate any research results. Their only purpose is to familiarize with the data to a certain level and to check the syntax of programs to be run on the real data. Although the purpose of dummy data is quite different from the purpose of anonymized research data, similar methods can be applied to create them. Data swapping techniques were initially developed to create research data (see e.g., Dalenius and Reiss 1982, Moore 1996, Fienberg and McIn-

[2] Gomatam et al. (2005) describe the risk-utility trade-off associated with the anonymization of micro data by data swapping.

tyre 2004) but the implementation of these elaborate methods is costly.[3] Since dummy data do not have to yield valid research results, the data swapping can be simplified. In the next section we propose a simple method of data swapping that makes it really easy to create useful dummy data even for unbalanced panels.

3 Constraining the swaps

The basic approach is to swap values between subjects randomly. Advantages of data swapping are that every variable takes the same range of values in the dummy data and the actual data (the univariate distribution of every variable stays the same) and data swapping is easy to be implemented even for datasets with many variables as it can be automatised. Disadvantages are that consistency and correlation between variables and over time are completely lost if the value swapping is completely random.[4] Gomatam et al. (2005) distinguish three parameters that determine the level of risk and utility:

1) Swap rate: fraction of records to be affected by swapping,

2) Swap attributes: variables to be swapped,

3) Constraints: on the unswapped attributes.

As said before the disclosure risk for dummy data in the internet has to be zero. Therefore we choose a swap rate of 100% and we swap nearly all variables. In order to generate the kind of utility we need, we impose constraints. Departing from Gomatam et al. (2005) we do not just impose constraints between unswapped variables but also between swapped ones.

Consistency between variables within one wave

Typically, there are sets of variables with information that is closely related, e.g. filter questions:

Q1: Did you invest last year?

Q2: If yes, how much?

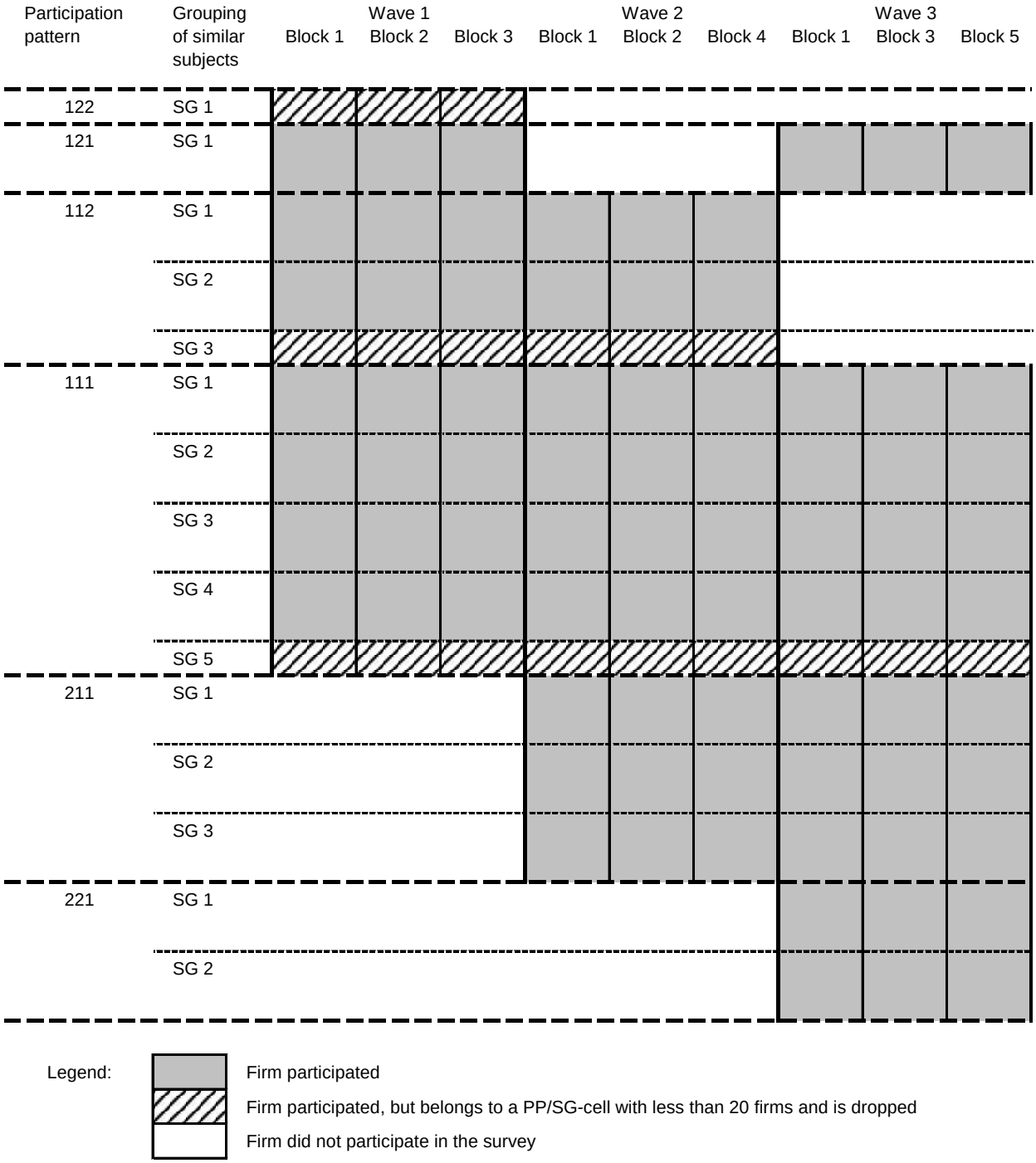
If every variable is resorted independently, a firm which says that it did not invest last year might be assigned some amount of investment. To avoid this we group all variables that are closely related and resort these complete 'blocks of variables' between subjects instead of resorting every variable separately.

Consistency over time

[3] Brand (2000) and Rosemann (2006) investigate different masking methods with regard to the anonymization of business data. Drechsler and Reiter (2009) apply multiple imputation methods to anonymize the IAB Establishment Panel. A simplification of their approach seems to be a promising alternative to create dummy data.

Panel data are usually characterized by new entries, continuers, temporary non-respondents and drop-outs. Figure 1 shows participation patterns in a typical panel data set with 3 waves.

Figure 1: How values are swapped



To explain the procedure we distinguish two types of participation:[5]

[4] See Moore (1996:4) for a list of advantages and disadvantages of data swapping.
[5] More types of participation can be defined. When we create the IAB-EP dummy data we define a third type of participation that includes observations without completed questionnaires. These are in-

1. regular participation
2. non-participation

For every subject we compute the participation pattern (PP) over time. In the example in Figure 1 the patterns are 122, 121, 112, 111, 211, and 221. In order to swap values within 'similar' subjects, we create a variable similarity group (SG) that identifies similar subjects within each participation pattern. For the IAB-EP we choose to group the establishments according to firm size so that every SG class contains 20 establishments.[6] Every observation is assigned to one particular PP/SG-cell. Establishments in PP/SG-cells with less than 20 observations are dropped.[7] These firms are shaded in Figure 1. In order to keep consistency over time we randomly resort the values of each variable block within the 20 subjects in each PP/SG-cell.

Figure 2 shows an example how the values are swapped between the firms of one PP/SG-cell. For variable block 1 firm 1 in every wave gets the values from firm 10. The values of variable block 2 are not changed at all, whereas the values for variable block 3 stem from firm 3.

Our constraints on the data swap ensure the following:

- All values of variables in one variable block stem from the same firm.
- If this variable block is included in subsequent waves, values in every wave will stem from the same firm.
- Values are swapped between 'similar' firms that belong to the same PP/SG-cell.

Note, that all variables that describe the participation pattern of the firm remain unchanged.[8]

cluded in the original data for the construction of longitudinal samples and have valid information only for a few variables.

[6] Alternatively the firms could be grouped by sales, industries or more sophisticated similarity indices that combine different dimensions.

[7] By dropping these firms from the dummy data it might happen that rare values of the original data such as certain small industries do not occur in the dummy data.

[8] In the IAB-EP these are the wellXXXX, querXXXX and panXX_XX variables.

Figure 2: Matrix of value assignment

Firm	is assigned the values from firm ...								
	Wave 1			Wave 2			Wave 3		
	Block 1	Block 2	Block 3	Block 1	Block 2	Block 4	Block 1	Block 3	Block 5
1	10	1	3	10	1	17	10	3	18
2	5	5	8	5	5	12	5	8	17
3	7	2	17	7	2	10	7	17	14
4	19	11	10	19	11	7	19	10	11
5	11	14	12	11	14	18	11	12	8
6	18	6	11	18	6	16	18	11	5
7	15	3	9	15	3	13	15	9	2
8	4	18	19	4	18	15	4	19	12
9	13	8	13	13	8	2	13	13	9
10	1	20	16	1	20	3	1	16	6
11	8	10	18	8	10	11	8	18	7
12	14	19	6	14	19	9	14	6	13
13	9	12	20	9	12	20	9	20	3
14	2	7	15	2	7	4	2	15	1
15	12	4	7	12	4	6	12	7	19
16	3	15	14	3	15	5	3	14	10
17	6	17	2	6	17	19	6	2	20
18	16	13	1	16	13	1	16	1	4
19	20	16	5	20	16	14	20	5	16
20	17	9	4	17	9	8	17	4	15

Regarding the design of the constraints two parameters determine the trade-off between risk and utility. The first parameter is the number of variable blocks. If you assign all variables to one variable block, the data are not changed at all. If you create as many variable blocks as you have variables, you keep some consistency of each variable over time but loose all consistency between different variables. In our implementation for the IAB-EP most variable blocks contain 10 to 20 variables. The second parameter is the size of the PP/SG-cells. If the size is one, nothing is changed. If it is two, you just swap values between two subjects. The larger the PP/SG-cells are, the lower the risk is. But larger PP/SG-cells are also associated with a smaller utility of the dummy data because the original correlations get weaker.

This basic concept of restricted value swapping can be supplemented by additional steps of anonymization. In the case of the IAB-EP we apply the following:

- We replace the establishment number by an artificial one.
- We draw a sample by dropping 2 out of 20 establishments in each cell (the disadvantage of drawing a sample is, that some very rare values in the original data do not occur in the dummy data any longer).
- We multiply all continuous variables with random numbers (in order to keep consistency, all variables within a variable block are multiplied by similar factors).

- We censor outliers of some continuous variables to the 90th percentile.
- For some sensitive variables such as the location of the establishment we increase the PP/SG-cell-size to 60 observations.

4 Summary

This paper describes how data swapping can be used to create time consistent dummy data for unbalanced panel data, even with a large number of survey waves and variables. The method is easy to implement, as can be seen from the program snippets in the appendix. Furthermore, the method is very flexible: the user can adjust the parameters which determine the trade-off between the utility of the dummy data and the level of confidentiality easily.

References

Brand, R (2000), Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, Beiträge zur Arbeitsmarkt- und Berufsforschung, Bd. 237, Nürnberg.

Dalenius Tore and Steven P. Reiss (1982), Data-swapping: A technique for disclosure control, Journal of Statistical Planning and Inference, 6, 73-85.

Drechsler, Jörg and Jerome P. Reiter (2009): Disclosure risk and data utility for partially synthetic data – an empirical study using the German IAB Establishment Survey, Journal of Official Statistics, 25(4): 589-603.

Fienberg, Stephen E. and Julie McIntyre (2004), Data Swapping: Variations on a Theme by Dalenius and Reiss, in: J. Domingo-Ferrer and V. Torra (Eds.), 2004, Privacy in Statistical Databases, 14-29, Springer: Heidelberg.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2005), Data swapping as a decision problem, Journal of Official Statistics, 21(4):635-656.

Moore, Richard A. (1996), Controlled data-swapping techniques for masking public use microdata sets, Statistical Research Division Report Series, RR96-04, U.S. Bureau of the Census.

Rosemann, Martin (2006), Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten, IAW-Forschungsbericht, Bd. 66, Tübingen.

Appendix: Stata code of the important steps

Step 1: Create PP/SG-cells

Participation patterns are created for each firm as described in Section 3 of this paper and firms are grouped within this participation patterns according to firm size. In contrast to Section 3, here we distinguish 3 types of participation. Our panel dataset has 14 waves.

```
foreach year of numlist 1993/2008 {
    use idnum well`year' using Original/test`year',clear
    gen     status`year' = 1 if inlist(well`year',"A","B","C","D","E","G") // firms with completed questionnaire
    replace status`year' = 2 if inlist(well`year',"H","W","X","Y","Z","")  // firms without compl. questionnaire
    assert  status`year'~=.
    drop well`year'
    sort idnum
    save data/1_b_temp_status_`year', replace
}

use data/1_b_temp_status_1993, clear            // merge status of each wave
foreach year of numlist 1994/2008 {
    sort idnum
    merge idnum using data/1_b_temp_status_`year'
    drop _merge
}

foreach year of numlist 1993/2008 {
    replace status`year' = 3 if status`year'==. // status 3 marks firms that are not included in the data
}                                               // in that year.

gen str pattern = ""
tostring status*, replace

foreach year of numlist 1993/2008 {
    replace pattern = pattern + status`year'
}

sort pattern firmsize                                   // (average) firmsize must be constant over time
by pattern: gen SGclass = int((_n-1)/20) + 1            // grouping of similar firms (here by firm size)
```

```
tostring SGclass, replace
replace SGclass= "0" + SGclass if length(SGclass)==1   // for optical reasons only
replace SGclass= "0" + SGclass if length(SGclass)==2
assert length(SGclass)==3

gen str PP_SG_cell = pattern + "_" + SGclass

bysort PP_SG_cell: keep if _N==20                // firms in cells with < 20 observations are dropped.
```

Step 2: Create the matrix of value assignment

Here the columns of Figure 2 are drawn. They determine which firm is assigned which other firm's values.

```
gen random = .
forvalues b = 1/160 {                            // 160 is the number of variable blocks we chose for the IAB-EP.
    replace randomcell = uniform()
    sort PP_SG_cell randomcell
    replace _cellblock = _n
    sort PP_SG_cell idnum
    gen long _idnumb`b' = idnum[_cellblock]      // _idnum`b' are the columns in Figure 2.
    gen _randomblock`b' = uniform()              // Some continuous variables will be added a multiplicative
                                                 // factor in Step 5. This is generated here to ensure that
                                                 // all variables of one variable block get similar factors.
}
```

Step 3: Merge the matrix of value assignment to every wave of the original data

The following steps are processed separately for every wave of the data. This can be done by a loop over the years.

```
use originaldata`year'
sort idnum
merge idnum using data_from_Step 2             // merge of _idnumb`b' _randomblock`b'
keep if _merge==3                              // observations in rare participation patterns are dropped
drop _merge

sort idnum                                     // test whether each new idnum appears exactly once in each block.
save data/temp_idnum, replace
```

```
forvalues b = 1/160 {
    preserve
        *dis `b'
        keep _idnumb`b'
        ren _idnumb`b' idnum
        sort idnum
        qui merge idnum using data/temp_idnum   // merge block-idnum with orig-idnum
        assert _merge==3                        // _merge has to correspond exactly
    restore
}
erase data/temp_idnum.dta


* Calculation of the line numbers that correspond to each new idnum in this specific wave.

forvalues b = 1/160 {
    sort _idnumb`b'
    gen _cellblock`b' = _n
}
```

Step 4: Tell Stata what to do with each variable

We have to tell the program which variables belong to the same variable block and if the variable shall only be swapped or if and what else shall be done with the variable. We do this by a simple list. In the left columns we list the variables, in the right column we tell the program what to do with it. What each coding exactly means can be inferred from the program in Step 5.

```
#d ;
global vartype
variable1           block01stet     // 'block' marks variables subject to data swapping. The number
variable2           block01stet     // after 'block' marks which variables belong to the same variable block.
variable3           block01stet     // 'stet' tells the program to add a random number.
variable4           block01
variable5           block04
variable6           block04
variable7           block04
variable8           block06
variable9           block07st90     // '90' tells the program to censor high values to the 90th percentile.
```

```
variable10          block06stet
variable11          block06stet
variable12          block08
variable13          block08stet
variable14          block08
variable15          block08stet
variable16          block08
variable17          block08stet
variable18          random          // These values are not swapped, random numbers are added only.
variable19          nothing         // Variables marked 'nothing' will not be changed.
...
;
#d cr
```

Step 5: Automated realization according to the type of the variable

After preparing the data in the foregoing steps, this step generates the dummy data. Values are changed for all duplicated observations (orignew==2).

```
expand 2                                        // duplication of every data line
bysort idnum: gen orignew = _n                  // orignew marks original (=1) and duplicated (=2) observations.

global anzvar:  word count("$vartype")

local var = 1
while `var'< $anzvar {                              // loop over all variables in $vartype
  tokenize "$vartype"
  local typ = `var' + 1
  dis in yellow "Var: ``var''" "   Typ: ``typ''"

  if "``typ''" ~= "nothing" {

    if substr("``typ''",1,5) == "block" {           // shift by block
        local z = real(substr("``typ''",6,2))
        dis "blocknumber: "`z'

        sort orignew idnum
        replace ``var'' = ``var''[_cellblock`z'] if orignew==2      // This line is the core of the program, as
```

```
                                                        // it realizes the data swapping.
                                                        // What follows below are additional simple
                                                        // masking methods, which we apply to some
                                                        // variables.

    //

    if substr("``typ''",8,2) == "st" {                                  // Continuous variables marked
        replace ``var'' =  ``var''                                      // with 'stet' are multiplicated
                          *(0.9+(_randomblock`z'[_cellblock`z']*0.2))    // with a random number, which
                          if !inlist(``var'',.,-8,-9) & orignew==2       // is constant for each variable
                                                                        // block.

        qui sum ``var'' if !inlist(``var'',.,-8,-9) & orgineu==1, de
        replace ``var'' = r(min) if !inlist(``var'',.,-8,-9)
                         & ``var'' < r(min) & orignew==2                // restriction to minimum
        if substr("``typ''",10,2) == "et" {
              replace ``var'' = r(max) if !inlist(``var'',.,-8,-9)
                         & ``var'' > r(max) & orignew==2                // restriction to maximum
        }
        if substr("``typ''",10,2) == "90" {
              replace ``var'' = r(p90) if !inlist(``var'',.,-8,-9)
                         & ``var'' > r(p90) & orignew==2                // restriction to 90% percentile
        }
        replace ``var'' = round(``var'',1)
                          if !inlist(``var'',.,-8,-9) & orignew==2       // round to integral numbers
    } // end of “st”

} // end of “block”


// multiplicative error +/- 20%, without value swapping
if "``typ''" == "random" {
    replace ``var'' =  ``var'' *(0.8+uniform()*0.4 ) if !inlist(``var'',.,-8,-9) & orignew==2
    sum ``var''              if !inlist(``var'',.,-8,-9)                        & orignew==1, de
    replace ``var'' = r(min) if !inlist(``var'',.,-8,-9) & ``var'' < r(p5) & orignew==2
                                                                                  // restr. to low percentile
    replace ``var'' = r(max) if !inlist(``var'',.,-8,-9) & ``var'' > r(p95) & orignew==2
                                                                                 // restr. to high percentile
```

```
    } // end of “random”
  } // end of “nothing”
  local var = `var' + 2   // switch to the next variable
} // end of loop over variables

keep if orignew==2                    // keeps only the duplicated observations with perturbed values
bysort PP_SG_cell: sample 18, count   // A further option is to draw a subsample as we did.
```

Imprint

FDZ-Methodenreport 3/2010

Publisher
The Research Data Centre (FDZ)
of the Federal Employment Agency
in the Institute for Employment Research
Regensburger Str. 104
D-90478 Nuremberg

Editorial staff
Stefan Bender, Britta Hübner

Technical production
Britta Hübner

Download
http://doku.iab.de/fdz/reporte/2010/MR_03-10-EN.pdf

Internet
http://fdz.iab.de/

Corresponding author:
Peter Jacobebbinghaus
The Research Data Centre (FDZ)
Regensburger Str. 104
90478 Nürnberg
Phone: 0911 / 179-1765
E-Mail: peter.jacobebbinghaus@iab.de