**README**

The purpose of this file is to explain the folder structure that led to the final result of the data processing. It will be useful both if someone wants to replicate what was done in our study and if they want to use our code to do alternative studies.

Note that we didn't put the surveys in this folder. They must therefore be placed by the researcher manually. Furthermore, in order for the code to run, the structure of the initial file names must be standardised. These should be placed in the Source/REBIDES\_Source folder. Read the TOREAD text file in the same folder for an explanation of the naming structure.

**File Folders**

Firstly, we have a general folder, which contains all the data and code that was used in this study, and which is divided into the Data, Graphs, Logs, Progs, Source and Tables folders. In the Data folder, we have the datasets used in STATA format, after they have been transformed to a certain point. In the Logs folder, we have the documentation of the STATA results window after running each of the do-files. The Progs folder contains the do-files and ADOs used. In the Source folder, we have the base data that was used, without transformation. The Tables and Graphs folders contain the tables and graphs produced in the do-files.

**Source**

V05217\_Comp\_NUTS2\_2024\_Munic: information about the NUTS II region to which the municipality belongs (available on INE).

TabelasResposta\_IEESP2023: IEESP filling tables used to answer some questions about how the survey is completed.

ExpCategoriasCITE2013: a table that distinguishes the areas of education and training in a more general sense, according to CITE 2013 (available at INE).

DGEEC\_DSEE\_DEES\_Estabelecimentos: tables with information about the address of the higher education institution.

REBIDES\_Source Folder: all available REBIDES databases, with names in the format REBIDES\_(academic year divided by "\_").

**Data**

Contract\_Source\_STATA Folder: Excel spreadsheets relating to the professor's contract, after an initial processing in STATA.

Degree\_Source\_STATA Folder: Excel spreadsheets relating to the professor's highest degree, after an initial processing in STATA.

Contract\_Data\_Tidy Folder: Excel spreadsheets relating to the professor's contract, after treatment and processing in STATA (a more developed one).

Degree\_Data\_Tidy Folder: Excel spreadsheets relating to the professor's highest degree, after treatment and processing in STATA (a more developed one).

Merged\_Data Folder: merged yearly STATA datasets from Contract\_Data\_Tidy and Degree\_Data\_Tidy.

ID\_Codes: yearly STATA datasets with an individual identifier code for each professor, created separately by the author of this dissertation.

Merged\_DataID Folder: merged yearly STATA datasets with an individual identifier code for each professor (combination of Merged\_Data with ID\_Codes)

Full\_Data: merged data from all available years, ready to use.

Establishments\_Location\_Tidy.dta: processed file with the location of all higher education institutions included in REBIDES in the last 4 years.

**Progs**

**Ado**

Each ADO contains some code that was used in the cleaning and preparing process of the data. These were used to add efficiency to the process, since the datasets are different, with different names, formats and even variables, and if we did everything in a single do-file, without this type of shortcuts, if we wanted to change a variable, or do any change at all, we would have to do this every year. In this way, we have a code type format that corresponds to a certain range of years and that is generalizable. It should be noted that within each ADO, the lines of code are accompanied by a brief comment explaining the purpose of its use. Right after this paragraph, we have a general explanation of the use of each of the ADOs but any further questions can be clarified by visiting them directly.

varscontr\* and varscontrAct - these ADOs allow us to select the information we want to use regarding contracts in a given data set. The Excel spreadsheet VARS.xlsx shows in more detail how the selection process takes place. Additionally, labels, when not necessary, are removed and organic unit codes are formatted to ensure an overall standard format.

varsdegr\* - same as above, but referring to information regarding each professor's grade. The Excel spreadsheet VARS.xlsx shows in more detail how the selection process takes place.

labelcontr - labels the variables selected on varscontr\*.

labeldegr - labels the variables selected on varsdegr\*.

categclean - as the name suggests, it “cleans” the professional categories (and working links), following what we are looking for in our study. The objective is to have a format that can be generalized from year to year, and which is also objective, indicating the most common categories and employment relationships. Given the intention to accompany the professional throughout their career, we chose to standardise the categories under their more "classic" existence concerning the profession, grouping similar positions.

estabsysclean - in a similar way to the previous one, it seeks to homogenize information but this time, regarding higher education institutions and educational subsystems. The main changes made in this ADO are those relating to the homogenisation of education subsystems and establishment codes. These are intended to allow the study of the evolution of higher education, in general, and mobility, in particular, to take place more harmoniously. This responds to two problems: in the early years of the dataset, there wasn't as much discretion when it came to defining educational subsystems, and, for example, *Universidade Católica Portuguesa* was classified autonomously; some establishments merged or integrated, and this can cause mobility without the professor changing workplace. Thus, we change the name of the subsystems so that the establishments are only distributed among universities and polytechnics, whether public or private and if two institutions have been integrated in such a way that their codes have disappeared from the dataset, we change the two old codes to the new one, for example.

hidegr\* - it deals with the homogenization of the degree obtained by the professor and ensures that we only have the highest obtained. Furthermore, we make sure that the degree information refers to the most general classification of each study cycle (first, second or third) since there is a diversity of results obtained throughout the period. The different types of bachelor, for example, are standardised and placed within the first cycle of studies. Cases that do not fit directly, because they do not correspond to any of the three cycles, or indirectly, because they represent something ambiguous, are placed in a general class called “ND”.

standartdtext\* - commands for variables standardization.

treatm\* - features code that makes some adjustments to the data set in terms of the treatment of variables. To name a few, variables relating to teaching hours, contract time, or even CNAEF areas, are dealt with here.

**Establishments\_Location.do**

This file shows the process that gave rise to the database used to discover the location of higher education institutions for the last 4 years. The lower part shows the merger between DGEEC\_DSEE\_DEES\_Estabelecimentosxlsx, V05217\_Comp\_NUTS2\_2024\_Munic.xlsx, and some information obtained from the websites of the higher education institutions when the database didn't have an organic unit.

**1.DataRead**

This file reads the initial Excel files, selects the relevant variables and cleans up small parts of them. We're referring, for example, to the creation of the variable representing the reference year when it doesn't exist.

**2.DataCleanYear**

This do-file takes the STATA files converted by 1.DataRead, and does a deeper cleaning in each of the years, standardizing variables, for example, and placing labels. It also merges the data sets on an annual basis, in particular by combining contracts with qualifications.

**3.Merge\_Append**

This is the do-file that takes the longest to run. As the name suggests, it merges each of the datasets with the file containing the professors' unique identification codes. However, that's not what makes this the heaviest file to run. This is also where the initial version of the unique identifier codes (which are in the IDs folder) is corrected, as they were not in perfect condition (however, we do not have this publicly available, due to confidentiality). Finally, this is where all the datasets are appended and where the first version of the final dataset is created.

**4.Final\_Cleaning**

It performs the final cleaning and preparation of the data, encoding some variables and correcting the codes for educational establishments or subsystems, for example. This is where the data set used in the empirical part comes from.

**VARS.xlsx**

​ This Excel shows how the datasets were assembled. Given that the REBIDES Excels had considerably different names for the variables, and sometimes had many that ended up not being used, this is a good way of demonstrating exactly what was taken from each Excel. Thus, it describes the name of the variables that were taken from each Excel, in their official format, taken directly from the DGEEC website, and the name that was placed after they were imported (to allow the data to be merged). It also describes directly which ADOs were applied to each of the Excels.

**Note:** Any folder with a “holder.txt” file had its contents removed due to confidentiality reasons. However, that data can be provided to interested researchers. To that end, please refer to [juliohenri2001@gmail.com](mailto:juliohenri2001@gmail.com).