

Data Preparation Exercise

The aim of this exercise is to prepare the tax return and audit data in order to proceed to the tax gap estimation. The data preparation is an important first step in the estimation of the tax gap and constitutes Part 1 of the Toolkit.

Exercise 1 – Stata Environment setup & data import

In this exercise, you will try to implement the First Stage of the Toolkit Part 1: Data preparation. This stage aims to ensure all necessary packages are installed and updated, prepare the STATA environment, set up directory paths, and import and check the CIT return data files.

- a) Run the code lines 23 to 28 to prepare the Stata environment.
- b) Modify line 31 to your own file directory and run line 32 to set the directory paths

```
global dir "<YOUR PATH>"
```

- c) Import the cit_return data (csv file), rename the variables region (region, isci, year) to (province, isic_code, rtn_year) respectively and save the data in Stata format with the name "cit". Browse the data to check.

Exercise 2 – Data management process for CIT data

In this exercise, you will try to implement the Second Stage of the Toolkit Part 1: Data management process for the return data.

- a) Identify the number of exact duplicates in the CIT return data. Drop exact duplicates.
- b) Taxpayers sometimes make multiple revisions of their returns. In this context, a revision is a duplicate record for the same taxpayer (tpin) and year (rtn_year). Keep the latest return revision.
- c) Tax return data can be quite large. Browse the dataset and remove any unnecessary variables. Then sort, reorder, and save the cleaned dataset

The toolkit also assists with appending datasets and duplicates from multiple files. This is not necessary in the exercise. CIT returns are typically reported annually, but PAYE and VAT data usually need to be collapsed to the annual level. The cleaned CIT data is ready for merging with the audit data. But first, the Audit data needs to be cleaned.

Exercise 3 – Data management process for Audit data

In this exercise, you will try to implement the Third Stage of the Toolkit Part 1: Data management process for the audit data.

- a) Import the audit data (csv file), rename total_tax to total_assessed, keep audits related to CIT for merging with CIT returns and save the data in Stata format with the name "audit". Browse the data to check.
- b) Investigate the Total Assess variable. Check for negative assessments. Let's assume a negative assessment means the taxpayer is compliant. Replace negative assessments with zero to indicate compliance. (You may reconsider this assumption in your context)
- c) Audits can be conducted for multiple years, but often only a total assessment is reported. Allocate total assessments across the audited years to create an annual assessment amount.
- d) Remove all exact duplicates, ensuring only unique records remain and save the cleaned audit data.

The toolkit included detailed data preparation notes for a variety of challenges in the audit data that the team have encountered. Without a consistent form, the preparation of the audit data requires careful attention and consideration.

Exercise 4 – Date formatting

In audit data, dates often appear in inconsistent formats. In this exercise, we will show an example of these date problems and how to convert them to STATA formats. In this example, you are given date_problem data with two date variables and eight observations.

- a) Import the date_problem data. Examine the assessment date variable to identify the problems.
- b) Dates with missing or incorrect DD, MM or YYYY require attention to clean. Browse the incorrect dates and attempt to correct them.
- c) Convert the corrected date variable to a Stata format and save the dataset.