# NOTE – PRODUCTIVITY CODE DESCRIPTION

This note describes the code used to estimate the numbers shown in our report on the productivity of EDBs. The code is contained in a series of R scripts. The figure below shows a model map of these various R scripts, how they fit together as well as the inputs and outputs.

Intermediate outputs have been used so that scripts early in the process do not necessarily need to be re-run. If the whole pipeline needs to be re-run, for example because there is an update to the ID data, then the scripts should be run in the correct order. Specifically, extraction scripts first, then the estimation of user cost of capital, then the analysis script and finally the index estimation and econometric estimation scripts.

It is recommended that the user tie together all R scripts in a single R project. All input files should be in a sub-folder called inputs. A sub-folder called outputs should also be created with another sub-folder called ForReport. The R scripts assume that this is the R project folder structure.

The figure below shows the process for producing the all industry analysis. The exempt/non-exempt analysis follows the same pipeline but also requires the input file EDB-Status.csv which describes whether an EBD is exempt or non-exempt.
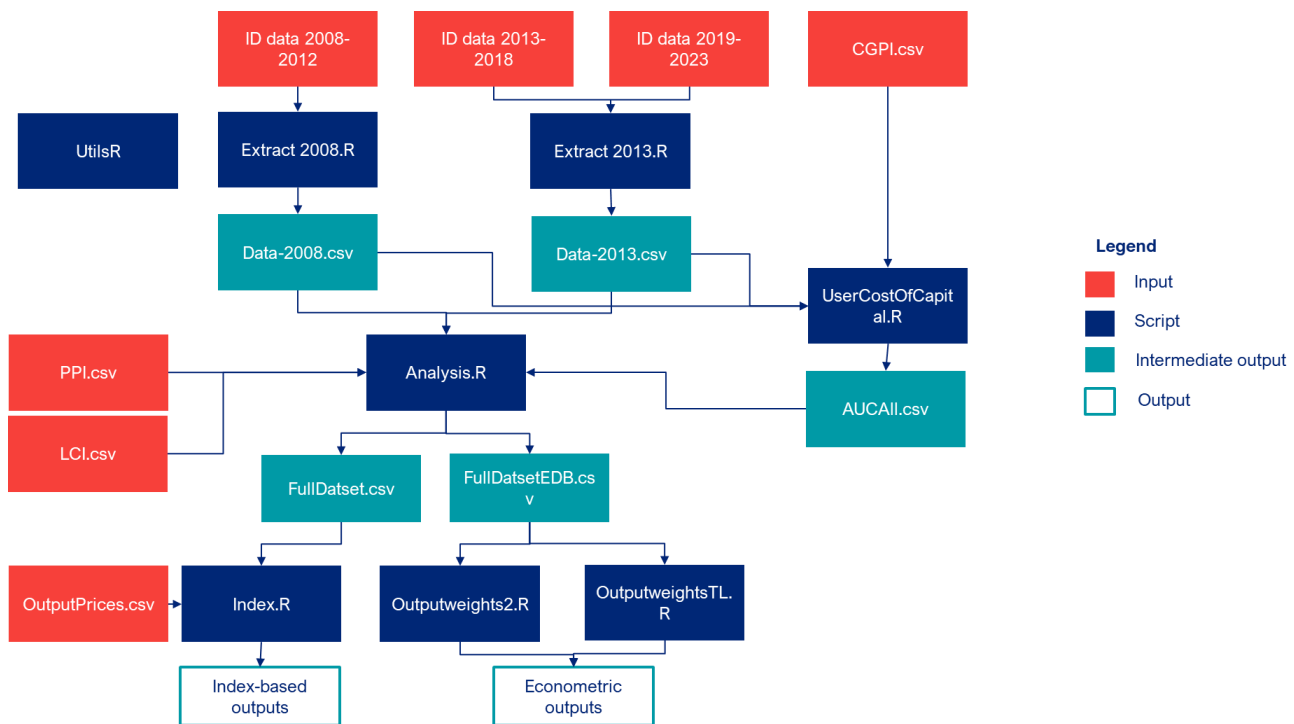
Inputs

In model map above inputs are shown as red boxes:

- **Information disclosure (ID) data**: In the figure above these are shown as the files ID data 2008-2012, ID data 2013-2018 and ID data 2019-2023. These ID data files were directly downloaded from the Commission's website.

- **Inflation data**: In the figure above these are shown as the files CGPI.csv, PPI.csv and LCI.csv. These data files have been downloaded from Stats NZ.

- **Output prices**: While the output prices are estimated using the functions in Outputweights2.R since we apply some judgement to create the ranges these are introduced by a separate file.

Scripts

There are 8 R scripts in total:

- **Extract 2008.R**: This script takes the 2008-2012 data file extracts the relevant data points required for further analysis and organises this into tidy format. The reason for a separate script for the 2008-2012 data is this is in a different format than the other data files.

- **Extract 2013.R**: This script takes the data contained in the 2013-2018 and 2019-2023 data files merges them together, extracts the relevant data points and organises this into tidy format.

- **UserCostofCapital.R**: This script calculates the annualised user cost of capital.

- **Analysis.R:** This script has four purposes. Firstly, it brings together the two intermediate outputs created by the extraction scripts. Secondly, it provides some initial analysis of the various variables. Thirdly, it calculates some other variables needed for further analysis such as ratcheted maximum demand and adjustments for inflation. Finally, it brings together these various data points into a single file with year totals for the index-based analysis and a single file split by EDB and year for the econometric methods.

- **Index.R**: This script implements the index-based methods. It first defines two functions which are then used to create the output indexes, input indexes and productivity indexes. The script is organises in that order, output indexes first by model number, then input indexes and finally productivity indexes by model number. The first function called Calc_Fisher calculates a series of index numbers for a stated set of outputs and a range of prices for these outputs. The price ranges are introduced from an input file called OutputPrices.csv. The second function called MinMax takes the output of Calc_Fisher and extracts the minimum and maximum positions in each year.

- **OutputWeights2.R**: This script implements the Cobb-Douglas econometric estimates. It defines a couple of functions. The first function is OutputWeights2 estimates a Cobb-Douglas cost function. It takes the following arguments:

    - y is the variable representing cost, for example real total cost.

    - X is the variable representing the outputs, for example circuit length and number of connections.

    - P is a variable that allows the introduction of prices of opex and capex as explanatory variables. This functionality was not used in the draft report.

    - time is a variable that allows various the function to consider various forms of the time trend. 1 if the time is introduced as an untransformed continuous variable, 2 if time is introduced as a dummy and 3 if time is introduced as an interaction term.

    - EDB is a variable determining whether the model estimates with or with EDB fixed effects.

    The second function is called reg_output. This is simply a function that takes the output of OutputWeights2() and changes the presentation.

    The remainder of the script simply implements the various model output specifications using these two functions.

- **OutputWeightsTL.R**: Identical to the script OutputWeights2.R except the functional form is translog instead of Cobb-Douglas.

- **Utils.R**: Contains two helper functions. The first for creating consistent EDB names across datasets and the second for creating relevant averages for the index-based methods.

Outputs

There are five intermediate outputs:

- **Data-2008.csv and Data-2013.csv**: The relevant extractions from the ID dataset in tidy format

- **FullDataset.csv and FullDataEDB.csv**: The relevant data used for estimation either in year sum format or year-EDB format.

- **AUCall.csv**: Estimates of the user cost of capital by year.

The code also creates a series of final output files. The key files are:

- The 12 files with the naming convention *_reg_output_*_*.csv. For example, CD_reg_output_Opex_all.csv. These contain the full econometric results.

- Table3.csv which contains a summary of the index-based method results.

- Table4.csv which contains a summary of the econometric method results.