**Weighting methodology for FBS with R/ReGenesees**

**Running instructions:**

Note: this will create weightings for two years, once for the new year’s data (sampyear=datayear) and once for the previous year’s revised data (sampyear=datayear-1).

Github Template Download

The following assumes that you have Git and Git Bash installed and setup. See instructions [here](http://saltire/orgspaces/Analytic-Professions/Analytical-Tools/Pages/version-control-guide.aspx) and [here](http://saltire/orgspaces/Analytic-Professions/Analytical-Tools/Pages/version-control-guide.aspx) if that’s not the case.

In a File browser window, navigate to the seerad/fas/raw\_data/prodYYYY folder that you want to calculate weightings for.

* Right click in the folder and choose Git Bash Here
* Enter the command:

git clone git@github.com:DataScienceScotland/FBS\_R\_Weighting.git

A folder called *FBS\_R\_Weighting* should now exist. This folder should contain:

* An *R Weighting Guidance* folder (where this document lives!)
* A SAS project called *FBS Calibration Totals*
* A SAS project called *Load new weights*
* A *Census data – workings* spreadsheet
* A Regenesees.R file, which contains functions used by R files created later.
* An “Inputs” folder.
* An “Outputs” folder.
* A couple of additional files used by R or Git – README, .gitignore, .Rhistory, FBS\_R\_Weighting.Rproj. These can be ignored.

The inputs/outputs folders will each contain a “Placeholder” text file; these are not used – they only exist as Github ignores empty folders.

There may also be other files relating to QA and canvas lists. If any of the above are missing, copy and paste the previous year’s version into this folder, and update any dates as necessary.

Creating census files

A request has to be made to the Census team to get June agricultural census data. The file the census team create is called *FBS YYYY Census Data – Weighting backdating*. YYYY here is the census year, and is one less than the FBS sampyear. So for FBS2022\_22, we use census year 2021 and for FBS2022\_21 we use census year 2020. Save a copy in the same folder as the files listed above.

* Open this file and the *Census data – workings* spreadsheet, and copy the data from the appropriate census year into the “Census Data” tab of the workings file.
* The “Total Link” and “Farm Census Link” tabs should automatically update.
* Copy the data from the Total Link tab into the CensusTotalsYYYY tab.
* Copy the data from the Farm Census Link tab into the farmcensusYYYY tab.
* Select the CensusTotalsYYYY tab and then File > Export. Save it as a csv file in the *Inputs* folder listed above. Remember that YYYY in the file you save is the census year. You’ll need to click “OK” and then “Yes”.
* Repeat for the farmcensusYYYY tab (also save as a csv in the *Inputs* folder). Remember that YYYY in the file you save is the census year.
* Repeat the previous steps for sampyear=datayear-1. Remember to use the previous census year. Alternatively, copy these from the previous year’s folder – the census data won’t have changed.

FBS Calibration Project

* Open the “FBS Calibration Totals” SAS project
* Update the datayear and sampyear in the Setup Work Area programme.
* RUN Setup Work Area programme
* RUN Census Data programme.
* RUN Read in survey data programme.
* RUN Survey Data programme.
* RUN the Calibration Totals programme.
* RUN the Macro calSet programme.
* RUN the Calibration Setup programme. This will generate FBSWTYYYY.R and FBSurveyYYYY.csv files in the Outputs folder.

R programme

In R studio, RUN the FBSWTYYYY.R file (found in the Outputs folder). This will produce FBSWTYYYY.csv and FBSWTYYYY\_Histogram.pdf in the Outputs folder.

If the Regenesees package is not available, contact the person listed at the “Request R packages that haven't been validated” step here. Note that the Regenesees.R file mentioned above and the Regenesees package are separate things.

In FBSWTYYYY.csv, the variable fbswt are the individual weights assigned to each farm.

Second sampyear

At this point repeat above (FBS Calibration SAS project and FBSWTYYYY.R, possibly preceded by creating the appropriate census data file) for the second sampyear.

Load new weights

Use the “Load new weights” SAS project to import the weights file into SAS.

* Update the datayear in Prog. 1.
* Run Prog. 1.
* Check the merged\_weights output file. Does it contain the correct number of farms for each year?
* If happy, run Prog. 2. This will overwrite the weights file in the SAS Library.
* Run Prog. 3, which will create some of the datasets stored in the agstemp library.

Updating this project

If you want to make changes to any of the programs or documents in this repository, it is best to clone a separate copy to somewhere other than the Z drive (for instance, your repos folder on your C drive). This is because some of the programs in the project create new programs, and we don’t want to upload those created programs to the Github repository.

The gitignore file has been written to try and prevent any problems which this would cause, but it’s still better to update from a fresh, separate, folder.

**How it was developed and how to calculate new weights - OLD instructions (replaced in December 2022 by the Running Instructions above).**

**Background**

In summer 2018, a project was started to look into developing a new weighting methodology for the Farm Business Survey. This was completed in February 2019 and implemented in the Annual Farm Business Income Estimates 2017-18 publication.

**Reason for project:**

As the FBS is voluntary the sample has become more unrepresentative of the population over the previous few years.

The old weighting method was based on calculating weights for groupings by type, tenure and size. The weight values were then applied in SAS using the ‘weight’ function.

This way of applying grouped weights and the unrepresentativeness of the sample caused the data from some small farms to be given too much weight.

**ONS Review (July 2018):**

RESAS asked ONS to review the old methodology and provide some suggestions for improvements. The report from ONS and the data sent to them can be found in eRDM.

Report – A25054086

Data – A25054099

**SG Review and new method (August – October 2018):**

RESAS asked SG statisticians to suggest alternative weighting methods that would be appropriate for this sample.

It was suggested that a calibration model would work for this sample. The model that was initially proposed is in eRDM

New method proposal – A25054236

**Further developments to the model (October 2018 – January 2019):**

RESAS then carried out some tests on the new model compared to the old weights method. This led to making adjustments to the model and testing these as well. Some workings from these comparisons are in eRDM.

Different model tests – A25054359

Comparison of different weights – A25054408

Weights type tenure compare – A25054419

During this process we came across an issue where the model was using the wrong area for potatoes. This was updated and so in the weights comparison spreadsheet there is a value for updated potato weight.

Note: there is still a slight anomaly with potatoes in the final weight model. The value of potatoes can sometimes look like it is being over inflated. This is largely due to the potato area (which is in the model) including the land let out for potatoes but the value does not include the money from potatoes sold from land-let area. This is mainly a problem for General Cropping farms but as potatoes are not a main source of income, it was decided that this would not have a huge impact on overall FBI. Also due to short time frame before publication further testing on whether the model could be refined was not be done.

(Potential idea for future could be to look in to the potatoes issue more and see if the model can be improved on).

**Second ONS Review (January 2019):**

ONS then reviewed the new methodology and provided some feedback. The data that was sent to ONS, ONS’ response, further comments and RESAS’ response to the ONS review are in eRDM.

Data – A25054461

Report – A25054520

Further comments – A25056326

RESAS response – A25056367

**Final model (February 2019):**

The final model that was used in the 2017-18 FBI publication was:

type + tenure + barley\_hct + oats\_hct + pota\_hct + wheat\_hct + rent\_area + tot\_area + DairyCows + BeefCows + Ewes – 1

Published figures were back-dated six years (2012-13 to 2017-18) using the new weights model.

New weights were calculated back to 2009-10 to allow for AUK data to be revised.

The reweighting files are saved here: [Z:\seerad\fas\raw\_data\prod2018\Reweighting - new methodology](file:///Z:\seerad\fas\raw_data\prod2018\Reweighting%20-%20new%20methodology)

**How to calculate calibration weights – OLD instructions (replaced by the Running Instructions above in December 2022).**

**Note**: this will have to be **run twice**, once for the new year’s data and once for the previous years revised data.

* Step 1: A request has to be made to the Census team to get June agricultural census data. This data will be used to create **farmcensusYYYY.csv**. The file the census team create is called FBS YYYY Census Data – Weighting backdating. YYYY here is the census year, and is one less than the FBS sampyear. So for FBS2022, we use census year 2021. Save a copy in the Step 1 folder.
* Step 2: Copy the relevant year of the file from the census team into the Census Data tab of the FBSYYYY weights – workings file in the Step 2 folder. The Total Link and Farm Census Link tabs will automatically update. Copy and paste the data from these into the farmcensusYYYY and CensusTotalsYYYY tabs (update the year in their names), and export these as csv files (into the relevant “FBSYYYY Sample Year YYYY” folder).
* Step 3: Data from the FBS also needs to be prepared in the format of **fbsdataYYYY.csv**. Open the Weighting Calibration SAS project file in the Step 3 folder. Update the datayear and sampyear in Prog. 0 and run Progs. 0-2 (no need to run the branches). Copy and paste Table 6 from the Results tab into last year’s fbsdataYYYY.csv file. You’ll need to do some minor editing at the top to fix the column headers and rows alignment. **Save as** and remember to change the YYYY (save in the relevant “FBSYYYY Sample Year YYYY” folder).
* Save a new version of the SAS project “FBS Calibration Totals YYYY-YY” (in the relevant “FBSYYYY Sample Year YYYY” folder).
  + Set a new file path and update the datayear and sampyear in the **Setup Work Area programme**. RUN programme.
  + RUN **Census Data programme** programme.
  + RUN **Survey Data programme** programme.
  + RUN the **Calibration Totals programme**. ~~Cali4 is the output from this that is required later on.~~
  + RUN the **Macro calSet programme**.
* The **Calibration Setup** programme will need run twice. Pay close attention to the **%let pop = 0 or 1** step. Set this to 1 first and RUN. This will generate the **FBSWTpoptemp.R** file.
* In R studio, RUN the **Regenesees.R** file to install the ReGenesees package. If the Regenesees package is not available, contact the person listed at the “Request R packages that haven't been validated” step [here](http://saltire/orgspaces/Analytic-Professions/Analytical-Tools/Pages/Analytical-Software.aspx).
* In R studio, RUN the **FBSWTpoptemp.R** file to produce the calibration target template. This will produce a file called **PoptemplateYYYY.csv**.
* ~~Open Poptemplate.csv and fill in the template using the data produced from Cali4. Save the file as~~ **~~CensusTotalsYYYY.csv~~**.
* In SAS programme Calibration Setup, set %let pop = 0 and RUN. This will generate **FBSWTYYYY.R** file.
* In R studio, RUN the **FBSWTYYYY.R** file. This will produce **FBSWTYYYY.csv** and **FBSWTYYYY­\_Histogram.pdf.**
* In FBSWTYYYY.csv, the variable **fbswt** are the individual weights assigned to each farm.
* At this point repeat above for the second sampyear.
* Add the **fa\_id** and **fbswt** variables into the **Weight file 2010-YYYY** spreadsheet and save as an updated year.
* Use the **Load new weights 2010-YYYY SAS project (update the name)** to import the weights file into SAS. The **%let datapath** and **&datapath** steps will need updated.

The table below was provided with the initial model proposed by SG statisticians and is helpful for explaining each step of the process.

| **Inputs** | **Processes** | **Outputs** |
| --- | --- | --- |
| **SAS Processes**  **- FBS Calibration.egp** |  |  |
| **Setup Work Area.sas** | define a macro variable as the path to a folder used for input and output data throughout this process | &datapath. |
| **farmcensus.csv**  Extracted from the “Individual data for weighting” workbook, “Census 2017 data” worksheet. | **Census Data.sas**  Import census data  Group categorical variables | **work.census** |
| **fbsdata.csv**  Extracted from the “Individual data for weighting” workbook, “FBS 2017 individual data” worksheet. | **Survey Data.sas**  Import survey data  Relabel categorical variables to match census values (type and tenure)  Determine cross-group samples sizes  Group categorical variables (typetenure) | **work.survey1** |
| work.census  work.survey1 | **Calibration Totals.sas**  Determine the number of holdings and the size of area variables of sub groups. These tables are used to populate the calibration target files | **work.cali4** |
|  | **MACRO calSet.sas**  A macro to produce an R script to run the calibration | **%calSet;** |
| work.census  work.survey1 | **Calibration Setup.sas**  Calculates preweights  Formats data for R procedure  Runs the %calSet macro to either:   1. Write a script to run the calibration in R 2. Write a script to produce the calibration total template | **FBSurvey.csv**  and either:   1. FBSWT.R 2. FBSWTpoptemp.R |
| **R Processes** |  |  |
| **Calibration Template**:   1. Regenesees.R 2. FBSWTpoptemp.R 3. FBSurvey.csv | 1. Installs the ReGenesees package 2. Produces the calibration target template based on the **model** and 3. the **dataset** provided | **Poptemplate.csv** |
| Poptemplate.csv  work.cali4 | (In Excel) take values from the SAS dataset produced earlier to populate the template file. Save the resulting document as: | **CensusTotals.csv** |
| **Calibration:**   1. Regenesees.R 2. FBSWT.R 3. FBSurvey.csv 4. CensusTotals.csv | 1. Installs the ReGenesees package 2. Calibrates the data by raking based on the **model** and 3. the **dataset** provided to match 4. the population targets | **FBSWT.csv**  The weighted survey data, which can be read into SAS or another analytical package for analysis.  The variable fbswt sums to the total number of holdings in the census.  fbswt\_scale will sum to the sample size  **FBSWT\_Histogram.pdf**  A density plot of the final weight values |