Accessing the Fannie Mae and Freddie Mac Data Sets:

Proceed as follows to map the folder containing the files for Fannie and Freddie loan origination and performance data (/refm-mfe/GSE/), as well the Freddie PC pool files (/refm-mfe/freddie_pool/) to your local client (laptop or desktop):

- Open a remote desktop session at the terminal server (ts.haas.berkeley.edu). Use the following credentials: "Username" = haas\<your Haas username>, "Password" = <your Haas password>.
- Run a Linux terminal session called PuTTY, located on Start->Utilities->PuTTY->PuTTY.
- On the "Host Name" dialog box enter: research.haas.berkeley.edu. Hit "Open"
- Provide your Haas logon and password to start a terminal session.
- Enter the following Linux command at the prompt : `ln -s /bulk/data/refm-mfe refm-mfe`
- This creates a link on your home network folder to a folder named "refm-mfe". If you now open a Windows Explorer an go to your "R:" drive (Windows) or the equivalent on Finder (Mac) you should see a new folder at the root called "refm-mfe" and a subfolder called "GSE" and "freddie_pool".
- Once the folder is successfully mapped you may close PuTTY and logoff from the remote desktop session.

The Fannie and Freddie data:

The "readme" subfolders contain pdf files describing the data set layouts and the codebook to the fields. In particular, the fields "loan_id" and "loan_seq_number" are unique identifiers for Fannie and Freddie respectively. The loan acquisition/origination files are sorted by these fields in ascending order. The performance files are sorted by "loan_id"/"loan_seq_number" and "reporting_period", also in ascending order.

Note that the acquisition/origination and performance files were reformatted to csv files with double quotes as the text qualifier. In addition, the datasets were filtered to start at 2005-01-01. Finally, the "refm-mfe" folder and its subfolders are read-only.

Tips for selecting subsamples:

The performance files are large – approximately 34 GB in the case of Fannie. It is advisable not trying to load the whole dataset in a Matlab, or R session, unless you have tons of memory. You should work with random subsamples and average the results at the end.
As a result, use some other tool such as Python to read the acquisition/origination file, construct a list of "loan_id"s/"loan_seq_number"s, and generate your random subsample on that list. Now, reprocess the acquisition/origination file and process the performance file to generate your random extract for calibration and analyses.

Need help:

Please contact paulоissler@berkeley.edu .