Matlab Code: EA-MD-QD data processing

Last update: March 31, 2025

Overview

The aim of this short document is to explain the methodology used by the Matlab file *routine_data.m* to process the data contained in the EA-MD-QD dataset.

The complete procedure takes 4 steps to run, regarding:

1. Country of interest
2. Frequency of the data
3. Transformations
4. Imputation of outliers/missing values

The user is guided through all of these choices with pop-up windows appearing once the code is run. All subroutines used to perform each of the aforementioned steps is included within the master file. All subroutines employ, if not necessary otherwise, standard functions included in the basic Matlab licence.

The code also provides a default set of options which allow the user to skip all the passages above.

FIGURE 1: *Default options*

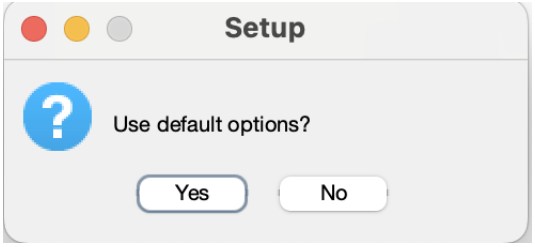


Figure 1 shows the very first pop-up window showing up once the code is run. By choosing *Yes*, the code runs automatically producing a .xlsx file containing data according to the default setup, which will be explained throughout the various sections of this short document. All pop-up windows in the following sections appear whenever default options are overridden by the user, by choosing *No*.

Country choice

If default options are not selected, the first choice the user must undertake is for which country data should be downloaded. Using *default options*, data for all countries are downloaded simultaneously.

FIGURE 2: *Country choice*

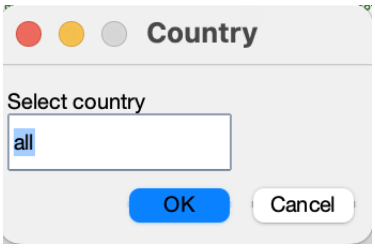


Figure 2 shows the pop-up window appearing to the user if default options are not selected. The user can choose to select all countries by manually writing *all*, or can select a specific country, or the Euro Area, by inserting the appropriate 2-digit code. The codes, in alphabetic order, are: *AT*, *BE*, *DE*, *EA*, *EL*, *ES*, *FR*, *IE*, *IT*, *NL*, *PT*.

Frequency choice

If default options are not selected, the user must further choose the frequency at which data must be delivered.

FIGURE 3: *Frequency choice*

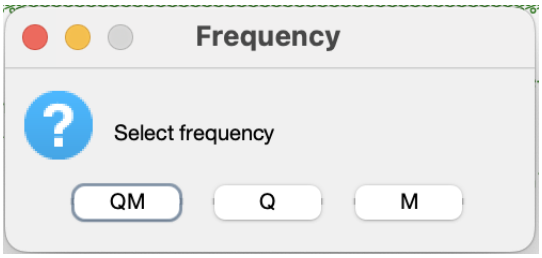


Figure 3 shows the pop-up window appearing to the user. Currently, there are three choices available:

1. **Quarterly-aggregated panel** (QM). All series included in the final dataset are at the quarterly level. Monthly series are transformed to quarterly series by standard aggregation, i.e. taking simple averages (for flows) and sums (for stocks) over the months of the corresponding quarter, provided all months related to a quarter are available as data points.

2. **Quarterly panel** (Q). Includes only the subset of data originally recorded at the quarterly frequency.

3. **Monthly panel** (M). Includes only the subset of data originally recorded at the monthly frequency.

By choosing default options, the code returns the dataset with all series, where monthly series are aggregated at the quarterly level (QM).

Transformation choice

If default options are not selected, the user must further choose the kind of transformations to apply to the data.

FIGURE 4: *Transformation choice*

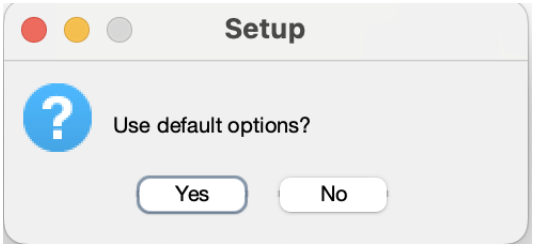


Figure 4 shows the pop-up window appearing to the user. There are two choices available:

1. **Light transformations** (*light*). All series are assumed to be integrated at most of order one. Hence, no $I(2)$ dynamics are assumed, and the heaviest transformation, for $I(1)$ series, only consists of first differences

2. **Heavy transformations** (*heavy*). Some series are treated as $I(2)$, after appropriate testing, and differenced twice. All other transformations are equivalent to the previous case.

By choosing default options, the code returns the dataset with all series transformed using the set of light transformations.

Outliers/missing values imputation

If default options are not selected, the user must finally choose whether to impute outliers and/or missing values and with which methodology.

FIGURE 5: *Imputation choice*

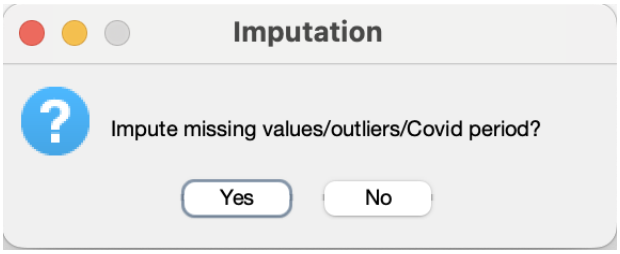


Figure 5 shows the pop-up window appearing to the user. If *Yes* is selected, another pop-up window will appear, in order to select the methodology for imputation. If *No* is selected, the code runs automatically with no imputation whatsoever.

FIGURE 6: *Methodology choice*

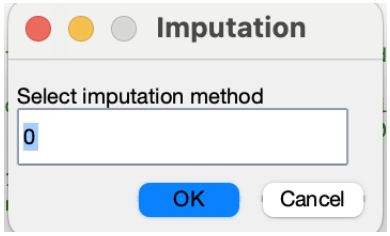


Figure 6 shows the pop-up window appearing to the user for the methodology choice. There are several choices available, which are number-coded. Methods *0* and *1* are standard, while methods *2* to *4* provide different specific methodologies to manage the outliers during the Covid period.

1. Method *0*. No adjustment for outliers, only missing values as the beginning and at the end of the sample (whenever present) are imputed (i.e., ragged edges). Imputation is based on the EM algorithm, as described in McCracken and Ng (2016).

2. Method *1*. Impute both outliers and ragged edges via the EM algorithm, as in McCracken and Ng (2016).

3. Method *2*. Impute outliers and missing values via the EM algorithm, by treating the Covid period, i.e. 2020 and 2021 (regardless of the frequency) as missing values for all real variables (column **Class** in the data description). Using this procedure, the Covid period is imputed for real variables using only the information contained in other variables, mostly financial and nominal variables.

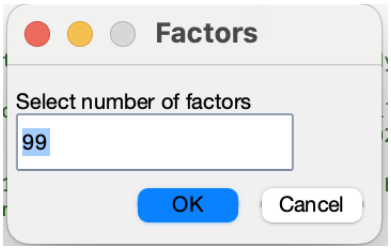
4. Method *3*. Impute outliers and missing values via the EM algorithm as in method *1*, and impute the Covid period via the Kalman Smoother, by treating 2020 and 2021 as missing values for all series.

5. Method *4*. Impute outliers and missing values via the EM algorithm as in method *1*, and impute the Covid period using principal components, where observations during the Covid period are imputed by means of the estimated common component using data up to the previous period (quarter or month).

By choosing default options, method *0* is employed, and only ragged edges are imputed. If the user wished to not perform any kind of imputation, she can simply insert an empty string (") in the dialogue box and the code will return the dataset with no imputation whatsoever.

A last pop-up box will appear to the user if any imputation is performed. As the EM algorithm is based on an underlying factor model, the number of factors used for imputation must be chosen.

FIGURE 7: *Number of factors*



By choosing $q = 99$ (default) the number of factors is automatically chosen via the criterion of Bai and Ng (2002). Otherwise, if the user has any pre-acquired knowledge on the factor structure for a specific dataset, or if he is willing to assume it, it can indicate the number of factors to be used for imputation in the appropriate dialogue box.

References

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*, 191–221.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*, 574–589.