ELSS missing data imputation.

Missing data imputation consists of imputing missing values in two phases:

1. Imputing missing values using group resemblance (imputing by level similarity)
2. Imputing missing values using additional time series recemblance (imputing by similar behavior)

and

3. Imputing missing years (interpolation and extrapolation)

1. Imputing missing values using group resemblance

Step 1:

For each country (1-61) reference group a weighted w_RANK is calculated as a weighted average of FIXED ranking positions and a VARIABLE. VARIABLE is the NIC indicator basic data (i.e. the original data with missing data points).

Table 1	Reference group	W1	RANK (country by)	w_RANK
FIXED	GDP and prouctivity	1	GDP(PPP) per capita	wR_GDP(i)
		2	GDP(PPP) per employed	
	Demographic structure	2	Level of metropolization*	wR_URB(i)
		1	Level of urbanization*	
	Economic structure	1	Service / share in laborforce	wR_ECN(i)
		2	Service / share in GDP	
		1	Industry / share in laborforce	
		2	Industry / share in GDP	
VARIABLE	NIC indicator / data	1	Current NIC indicator AVG(YYYYa-YYYYc)**	wR_IND(i)
	Totally missing NIC indicator for country		Calculated as weighted average of FIXED with some data present*	

*) Need to be calculated using different sources
Metoplization http://www.citypopulation.de
Urbanization UN and http://en.wikipedia.org/wiki/Urbanization_by_country
**) See step 3

Using inputs and weights W1 (Table 1) will produce a w_RANK index number (not necessarily integer) between 1 and 61 for each country. FIXED is assumed to be complet whereas VARIABLE represents a basic (original) NIC indicator data set (NHC_1 - NRC_12) with missing data.

Calculating RANK for FIXED is not a problem: Use data for latest year with no missing data and rank by that year (often 2014 or 2015, some times 2010 or even 2009). When data for country is totally missing (Argentina / Labor % share in service sector) use same ranking as other in same group category (Argentina / Service sector % share in GDP).

FIXED and w_RANK's are used in all imputations of missing data alike.

Step 2:

Calculate distance between countries based on FIXED, F_DIST, using weighted average (weight W2) of group ranks, w_RANK (Table 1).

Table 2	Group distance between countries	Name	W2	F_DIST
GDP and productivity	ABS[wR_GDP(country i) - wR_GDP(counry j)]	D_GDP	1	F_DIST(i,j)
Demographic structure	ABS[wR_URB(country i) - wR_URB(counry j)]	D_URB	3	
Economic structure	ABS[wR_ECN(country i) - wR_ECN(counry j)]	D_ECN	2	

By now, and using Table 2, closest neighbors by FIXED group resemblance to any country A can be selected by chosing countries B(k) with smallest F_DIST(A,B(k)), i.e. select by MIN (or rank and use lowest ranks).

Step 3:

Rank VARIABLE. Ranking countries by variable NIC indicator (original NIC data set) need to be done on a case to case basis using following guide lines:

1. Data at least for one year:
   - 1.1. Select latest and most comprehensive year(s) with data for most countries
   - 1.2. Count average for 3 consecutive years (when possible) using nearest values in average. If it is hard to find comprehensive 3 years use 2 year averages
   - 1.3. When data only for one year, use that year
   - 1.4. Calculate wR_IND by ranking available averages. wR_IND may now range from 1 to a number smaller than 61 depending on number of countries with no data at all
2. No data for country:
   - 2.1. Select 5 (five) nearest neighbors using F_DIST (Table2) and omitting countries with no data at all.
   - 2.2. Calculate missing wR_IND for country as average of its 5 nearest neighbor's wR_IND (2.1.).
3. Re rank the complete set. wR_IND now ranges from 1 - 61
4. Calculate final group distance between countries, R_DIST(I,j) (Table 3) using weights W2

Table 3	Group distance between countries	W2	R_DIST
GDP and productivity	D_GDP(I,j)	1	R_DIST(I,j)
Demographic structure	D_URB(I,j)	3	
Economic structure	D_ECN(I,j)	2	

NIC indicator / data	D_IND(I,j) = ABS[wR_IND(country i) - wR_IND(counry j)]	4	

R_DIST(I,j) is a weighted measure for group resemblance (by level) between countries i and j. R_DIST(I,j) is in fact a ***matrix*** where smaller values indicate greater structural resemblance between countries. We have chosen 1) GDP and productivity, 2) demograpic structure by level of metropolization and urbanization, 3) economic structure by shares of service and industry in laborforce and GDP ***and*** 4) NIC indicator (with the missing data) as these four together makes the basis for a structural resemblance between countries. Weights W1 and W2 are tested, but not finally optimized. They are so to say "close to be general and right".

Step 4:

Impute missing data with averages for data from 5 (five) closest neighbors.

> *We now have a complet data set,* ***wNIC(k,l),*** *k = country 1-61 and l = year 2001 - 2012 (latest year), as input for using time series resemblance in addition.*

NOTE:

> We have not yet discussed imputation of data where data is sporadic, i.e. data for some years are missing for all countries or imputation when 2014 and/or 2015 are missing. We do that in section 3.

2. Imputing missing values using additional time series recemblance

As a simple measure for time series resemblance (behavioral resemblance) we use correlation between two 2001 - 2015 (latest year) time series, i.e. r [NIC(i),NIC(j)] where **NIC(x)** is the data vector (row) 2001 - 2012 (latest year) and x =i and j denotes two countries. Wheras group resemblance (section 1) relies on levels correlation resemblance (value positive, high or close to 1) neglects levels, i.e. correlation may be high when levels differ highly: The time series just behave similarly on different levels. This must be acknowledged.

Step 5:

To balance levels: Calculate average for 2001- 2015 (latest year) for each county i, **AVG(i)** = AVG[wNIC(i,2001), …, wNIC(i,2012)].

Standardize R_DIST(i,j) by **sR_DIST(i,j)** = [61 - R_DIST(i,j)]/61. sR_DIST(i,j) will range between 0 and 1 and greatest resemblance between countries for values near 1.

Step 6:

Create **NIC(k,l)** data set by imputing missing data for country A and year Y in original data set using following procerure:

1. Calculate **rR_DIST(A,j)** = sR_DIST(A,j)*r [NIC(A),NIC(j)]
2. Select 5 (five) greatest rR_DIST(A,j), i.e. use MAX (or rank and use top 5) omitting country A (with the missing data)
3. Calculate first estimate **eNIC(A,Y)** for missing data as average using wNIC(j,Y) and j = selected countries (from 2.)
4. Calculate general average **gAVG(A**) for selected countries j using AVG(j) and j = selected countries (from 2.)
5. Impute **NIC(A,Y)** = eNIC(A,Y)*[AVG(A)/gAVG(A)]

3. Imputing missing years

First impute missing values in years for wich (some) data exist using procedures in section 1 and 2. E.g. for NHC_8_1 and NHC_8_2 treat the data set as consisting of years 2000, 2003, 2006, 2009, 2012 and 2015 and impute missing data for these years applying procedures decribed in section 1 and 2.

Of essence is to first make existing year’s complet.

Missing year(s)

As a first estimate TREND or GROWTH has several shortcomings. They are linear which produces strange jumps.

We use cubic spline curves to both interpolate (calculate missing values in between) and extrapolate (calculate missing first 2001 and 2015 year values). EXCEL add in to do the job is available at “XlXtrFun™ Extra Functions for Microsoft Excel” http://www.xlxtrfun.com/XlXtrFun/XlXtrFun.htm

Step 7:

First: Missing 2001 and/or 2015 values are calculated as average (SPLINE, TREND).

Then: Missing years in between is interpolated using SPLINE.

NOTE: This section (3) may improve. We are satisfied with the interpolation part, but the extrapolation 2001/2015 yet lack real connection to NIC. But: This is how it is done for now.