Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Imputation for missing data through artificial intelligence[1]

Byeungchun Kwon,
Bank for International Settlements

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Imputation for missing data through artificial intelligence

Heuristic & Machine learning approach to impute missing values (Test case with macroeconomic time series from the BIS Data Bank)

Byeungchun Kwon[1]

Abstract

The paper presents the new paradigm of missing data imputation method, the heuristic and machine learning imputation (HMLI), and experimentally compares 6 popular imputation methods through the macroeconomic time series from BIS Data Bank. HMLI is one of non-linear regression models. The main difference is it is based on the genetic search and the support vector machine (SVM) algorithm. HMLI consists of two parts: the best dependent variables selection and the non-linear regression. To verify the robustness of HMLI, the paper measures RMSE between predicted missing values and actual values for HMLI and 6 popular imputation methods. I tested 3,070 times for macroeconomic time series. The result shows that HMLI RMSE is the lowest about 10 percent missing data rate and second lower RMSE about 40 and 70 percent missing data rates. In this paper, I test macroeconomic times series with single frequency only, it needs to test various time series types which are different frequencies, trend and seasonality patterns.

Keywords: Heuristic search, Machine learning, Artificial intelligence, Imputation

JEL classification: C61, C82

[1] Bank for International Settlements, Byeungchun.Kwon@bis.org

Contents

1. Introduction

Time series data is widely used in various research fields and there are countless econometric functions and programs to analyse time series data. But most of mathematical analysis methodologies require complete time series data which means that all observations are filled with figures and it is not allow missing values. But missing observations are easily found in time series due to many reasons. Financial time series doesn't generate numbers on weekends and public holidays. In longitudinal studies, observations missing is usually happen. To put time series data in econometric functions, missing observations should be substituted reasonable figures.

In popular statistical tools like R and Stata, many imputation libraries exist based on various methodologies from simple interpolation to Kalman filter. In the Steffen et al., 2018, they compare 6 well known imputation methods in R for univariate time series.[2] To compare the imputation methods, it measured the root mean square error (RMSE) between actual values and predicted values. RMSE result shows that methods are so far apart statistically. It wonders these methods give us optimum imputation values about univariate time series. If we adopt the evolutionary process to impute missing values, Can we get more precise values?

This paper develops new imputation program based on the evolutionary process and machine learning algorithm, Heuristic and Machine Learning Imputation (HMLI). This is one of usual regression models so, it is composed of independent variable and dependent variables. But, HMLI does not know which dependent variables are the best set to impute missing values in independent variable. To find optimum dependent variables, HMLI introduces the genetic algorithm, one of heuristic search methods. This algorithm imitates natural evolutionary process so, we expect it can find one of best dependent variables sets through an iteration. Once this algorithm selects dependent variables, the model regresses dependent variables to an independent variable and predicts missing values. Regression method of HMLI is the support vector machine (SVM). SVM is one of the most efficient machine learning algorithm, which is mostly used for pattern recognition since its introduction in 1990s.

To verify HMLI model robustness, the paper measures RMSE between predict missing values and actual values from HMLI and 6 traditional imputation methods about 3,070 macroeconomic time series. All of the data are monthly frequency and retrieved from BIS Data Bank. It shows that RMSE from HMLI is lowest about 10 percent missing data rate and second lower RMSE about 40 and 70 percent missing data rates.

[2] Imputation methods

aggregate	Replacing NA with the overall mean
structTS	Filling NA through seasonal Kalman filter
locf	Last observation carried Forward; replacing NA with most recent non-NA value
approx	Replacing NA with linear interpolation
irmi	Iterative Robust Model-Based Imputation; filling NA through autoregressive imputation
interp	Linear interpolation for non-seasonal series. If seasonal series, a robust STL decomposition proceeded

This paper is structured as follows: section 2 explains HMLI model structure; section 3 shows the model computing process; section 4 shows experiment; and finally, section 5 includes key concluding remarks.

2. HMLI structure

As general regression model, HMLI structure is also composed of dependent variables and independent variable. The independent variable is the time series which has missing observations. HMLI model has two differences with the traditional regression model; dependent variable selection and regression line fitting.

In the traditional regression model, researchers normally select dependent variables to explain the model itself statistically and predict an independent variable correctly. But it is difficult to find best dependent variables for various reasons. If computer algorithms find optimum combination of dependent variables and predict independent variable with low error rate, it could be a perfect solution for imputation.

This paper applies the heuristic search (genetic algorithm) to select dependent variables. Heuristic search is a rule of thumb technique to find a solution more quickly when traditional approach is limited. For examples, total number of combinations to choose 6 time series from 10,000 time series is about 13×10^{20}. It is too big to compute all combinations within limited time. In this case, heuristic search could be a solution to approximate the exact solution.

In the model, genetic algorithm is implemented as heuristic search algorithm. Genetic algorithm imitates the natural selection process of Charles Darwin about natural evolution such as inheritance, mutation and crossover. A set of dependent variables is same as one chromosome and single dependent variable in the chromosome is a gene. Bad prediction means the dependent variables set can be regarded as recessive and will be disappeared in next iteration.

For the regression line fitting, HMLI model implements the support vector machine (SVM) which is one of the most efficient machine learning algorithm, which is mostly used for pattern recognition since its introduction in 1990s[3]. Once SVM calculates optimum parameters in the model, it predicts missing observations using actual dependent variables.

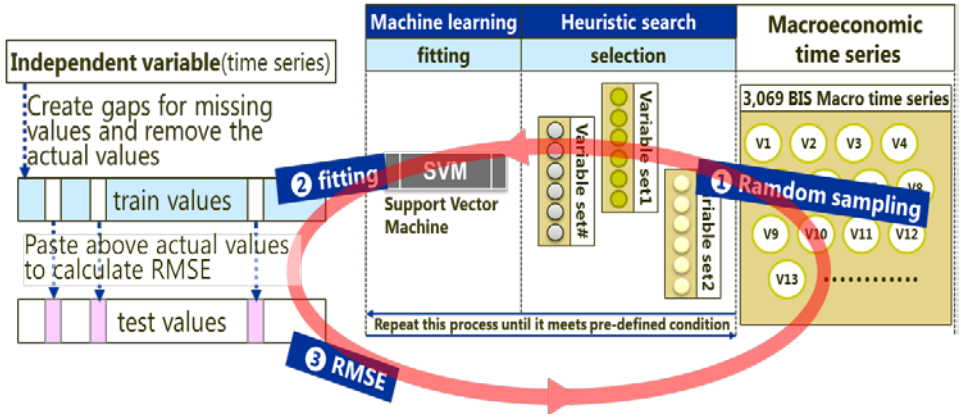
This heuristic search and machine learning combination repeats dependent variables selection, regression line fitting, missing observation prediction and performance review through the root mean square error (RMSE) between actual and prediction values. Termination condition of the repetition can be either RMSE satisfaction or iteration number.

[3] Boon Giin Lee, Teak Wei Chong, Boon Leng Lee, Hee Joon Park, Yoon Nyun Kim, Beomjoon Kim, "Wearable Mobile-Based Emotional Response-Monitoring System for Drivers", Human-Machine Systems IEEE Transactions on, vol. 47, no. 5, pp. 636-649, 2017.

1

Box: HMLI structure flow chart

As HMLI iterates this step 100 times, RMSE generally gets smaller and converged on certain number. It is heavily depending on initial values about which times series are selected to dependent variables, number of gaps and positions.



3. HMLI computing process

HMLI computing process consists of 9 steps. By iterating these steps, HMLI implements an evolutionary process. This iteration programming consumes computing resources heavily.

3.1 Steps for time series data pre-processing

STEP 1 (Data pre-process): If time series value range is wide, SVM objective function might be not working properly and calculation performance of normalization values is much faster[4]. Normalization is needed both independent and dependent variables. Once normalization is done, it starts to add gaps in independent variable. This added gaps replace actual values in time series and it will use as testing values. To decide which dates become gaps in time series, the model uses the exponential distribution and λ is a missing rate.

STEP 2 (Train and test data separation): In the step 1, the actual values, which are replaced with gaps, will be used to test values. Once HMLI model generates prediction values, RMSE between test and predicted values is a factor about performance of dependent variables. Except testing values in dependent variables, the other values is used as a train data.

STEP 3 (Sampling dependent variables): It is a random sampling process to pick a certain number of dependent variables from total variables. In this paper, it

[4] Ioffe, Sergey; Christian Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift"

randomly picks 6 time series for dependent variables from 3,069 macroeconomic time series and repeat it 10 times to create 10 sets of dependent variables.

3.2 Steps for evolutionary process

STEP 4 (Regression): Through the train data, SVM calculates best fitting curve. By putting test data into this fitting curve, SVM generates predicted values.

STEP 5 (RMSE calculation): It calculate RMSE between prediction values from SVM fitting curve and actual values replaced with gaps on STEP 1. Low RMSE means that dependent variables are good to predict gaps.

STEP 6 (Selection process): About 10 sets of dependent variables, this step calculates RMSE and ranks the sets. Based on this rank, it removes 5 lower ranked sets.

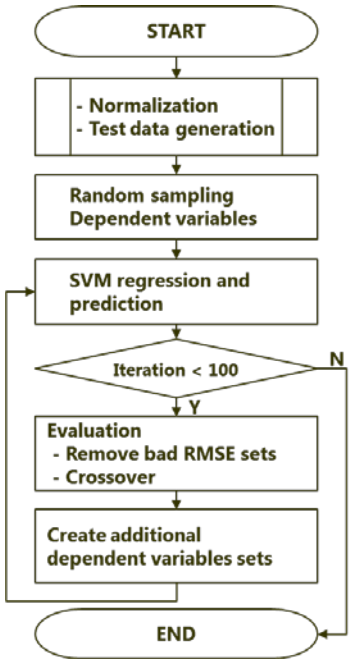
STEP 7 (Dependent variables crossover): Extract unique variables from top 5 ranked sets and generate 2 new dependent variables sets. If it cannot generate new dependent variables set, skip this process.

STEP 8 (Additional dependent variables sampling): For next iteration, 10 dependent variables sets are required. Including top 5 ranked sets and 2 crossover sets, create additional sets until it becomes 10 sets.

STEP 9 (Repetition): It repeats n times from step 4 to step 8. In this paper, Number n sets up 100. Normally, an iteration ends when it satisfies a certain condition. For example, if RMSE is less than a certain number, it stops the iteration. But this paper uses an iteration number as a termination condition because it reaches convergence level within 100 iterations.

2

Box: HMLI process diagram



4. Experiment

To verify HMLI model usability, this paper compares HMLI with six popular imputation methods. Experiment data is macroeconomic time series from BIS Data Bank. As a result, HMLI model shows very low RMSE and it is one of best imputation models.

4.1 Data and methods

4.1.1 Experimental data set

Experiment data is 3,070 macroeconomic time series from BIS Data Bank. It retrieves 26,193 monthly time series from the BIS Data Bank. To remove similar and complete time series, it extracts 3,070 time series which means that correlation coefficient is less than 0.97 between each other.

From 3,070 time series, it chooses a time series for independent variable and 3,069 time series are dependent variables set. Therefore, number of experiments is 3,070. Time period of 3,070 time series is from January 2010 to December 2017. Because it is monthly frequency, each time series has 96 observations.

4.1.2 Missing data rate

3,070 macroeconomic time series are complete data so, it doesn't have gap in the time series. To test traditional imputation methods and HMLI, it replaces actual values to gaps in an independent variable. Replaced actual values is used to calculate RMSE with predicted values for gaps through SVM.

Gaps are generated based on the exponential distribution and λ is missing data rate. So, bigger λ creates more gaps. In this experiment, it use three missing data rates; 0.1, 0.4 and 0.7. Average number of gaps 96 observations is 9.13 gaps for 0.1 missing rate, 31.52 gaps for 0.4 missing rate and 48.74 gaps for 0.7 missing rate.

Because the exponential distribution randomly pick gaps, it repeats 3 times for each missing rates. Therefore one independent variable is tested 9 times for each imputation methods.

4.1.3 Traditional imputation methods

In statistical packages, many imputation libraries exist and it can classify two types. So some libraries don't support both data types.

- Univariate time series imputation; single time series
- Multivariate time series imputation; panel data

There are many time series data analysis libraries in R. And this paper uses six imputation methods which are na.aggregate, na.locf, na.StructTS, na.approx methods of zoo library and na.interp method of forecast library and ar.irmi of VIM and customized function of Steffen, et al., 2015.

4.2 Experiment design and platform

About six traditional imputation methods, it executes 9 imputations, which are 3 different missing rate and 3 different random seed, for one variable. Total number of experiments is 165,780 which consists of

- 6 imputation methods × 3,070 variables × 3 missing rates ×3 random seeds

In case of HMLI method, it iterates 100 times for evolutionary process per one independent variable and each iteration has 10 different dependent variables sets. Total number of experiment is 27,630,000 which consists of

- 3,070 independent variables ×10 dependent variables sets ×100 iterations ×3 missing rates ×3 random seeds

The paper uses the open source R script developed by Steffen, et al., 2015 for traditional imputation methods experiment. And HMLI model is developed by parallel Python script. Both scripts are executed on Intel i7 CPU (8 cores) and it took 9 hours to finish the experiment.

4.3 Result

HMLI evolutionary process performance

While HMLI model repeats an evolutionary process, it expects mean square error (MSE) between actual and predicted values get smaller. In figure 1 MSE plot of the iterations for a time series used in this paper shows MSE decrease.

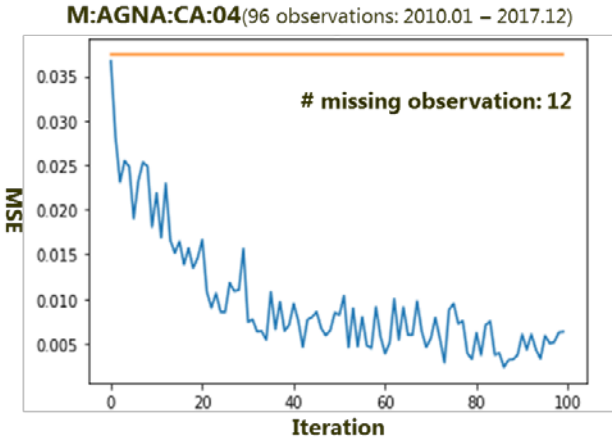


Figure 1: MSE plot for a macroeconomic time series

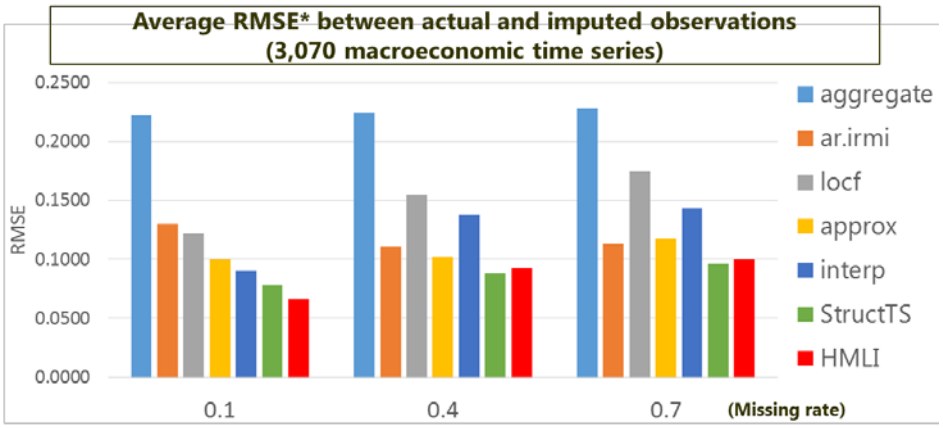
In figure 2 MSE plot of the iterations for 3,070 time series used in this paper shows average MSE of the time series also decrease as the iteration goes by which means the evolutionary process works well in HMLI.

Figure 2: Average MSE plot for 3,070 macroeconomic time series

Comparison of traditional methods

To measure HMLI performance, this paper compares RMSE between HMLI and 6 traditional imputation methods result about 3,070 time series, 3 different missing rates and 3 random seeds. About 6 traditional imputation methods, this paper uses forecast and zoo libraries in R because these libraries are popular to pre-process and analysis time series data.

In figure 3 average RMSE imputation results for macroeconomic time series shows that HMLI and StructTS are the best overall results.



Missing rate	aggregate	ar.irmi	locf	approx	interp	StructTS	HMLI
0.1	0.2221	0.1301	0.1218	0.0998	0.0901	0.0781	**0.0658**
0.4	0.2247	0.1107	0.1552	0.1020	0.1384	**0.0880**	**0.0924**
0.7	0.2280	0.1130	0.1749	0.1175	0.1432	**0.0961**	**0.1001**

Figure 3: Average MSE plot for macroeconomic time series and 3 missing rates

5. Conclusions

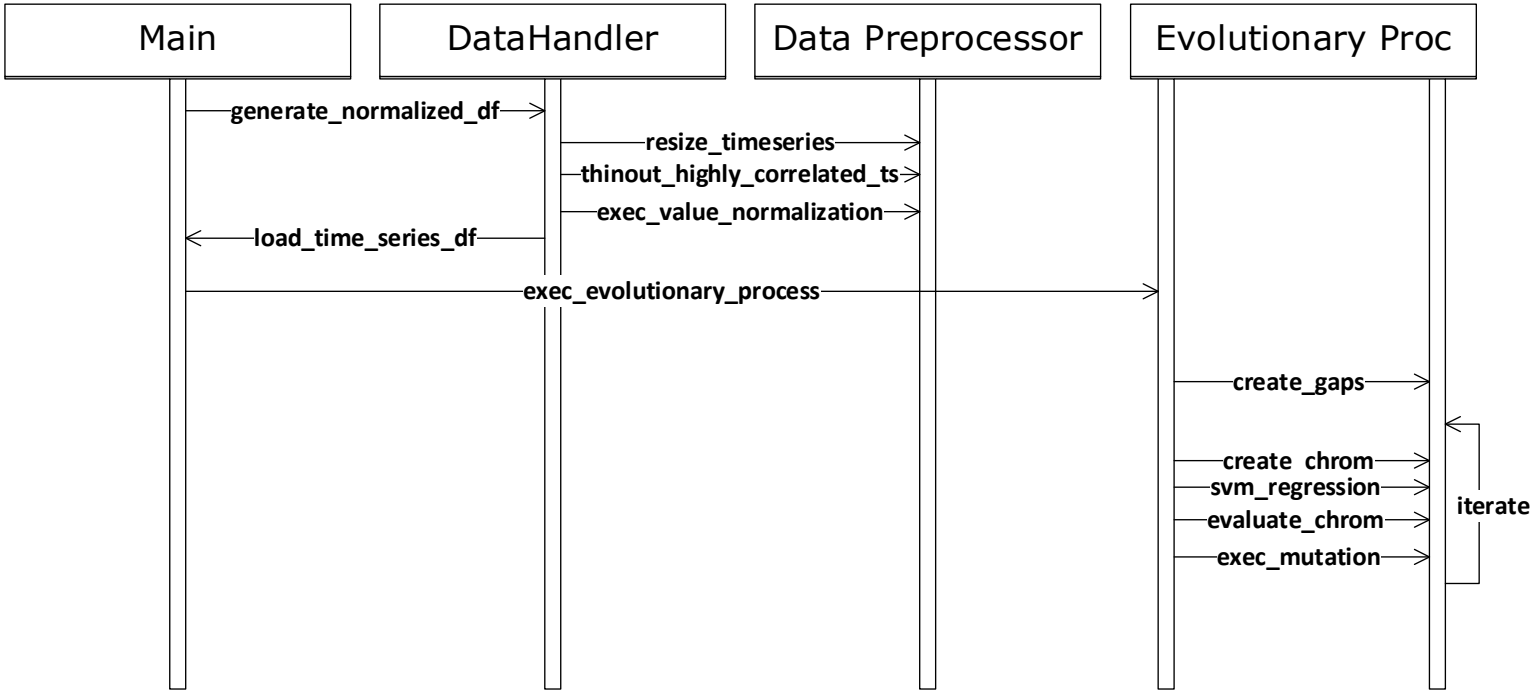
This paper describes HMLI model and compares traditional methods. HMLI result shows it is a competitive solution about missing observation imputations. In this paper, it skipped a model calibration procedure and machine learning methods comparison. It expects that the model performance can be improved through missed procedures.

The one thing we have to know is that HMLI model is very complicated and use much computing power than traditional methods because of the evolutionary process. If the other methods are able to find a reasonable solution, HMLI is not compatible any longer.

HMLI model design can have a wide application in many fields of econometric modelling like time series forecasting, macroeconomic time series now-casting or low frequency to high frequency series benchmarking. But HMLI model design cannot be a solution for the research areas require explanation or causality. Selected dependent variables through heuristic algorithms are random variables and not related to independent variable at all.

HMLI model is implemented by Python language and it is open script on the website. Also the experiment result is published. The data used in the paper cannot be share because of access permissions. But any time series data can be tested on HMLI model.

Annex 1 – HMLI model[5] sequence diagram



[5] Source code: https://github.com/byeungchun/HeuristicImputation

References

Ioffe, Sergey, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", Christian Szegedy, 2015

Boon Giin Lee, Teak Wei Chong, Boon Leng Lee, Hee Joon Park, Yoon Nyun Kim, Beomjoon Kim, "Wearable Mobile-Based Emotional Response-Monitoring System for Drivers", Human-Machine Systems IEEE Transactions on, vol. 47, no. 5, pp. 636-649, 2017

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J., "Comparison of different Methods for Univariate Time Series Imputation in R", CoRR, abs/1510.03924., 2015

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Imputation for missing data through artificial intelligence[1]

Byeungchun Kwon,
Bank for International Settlements

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

BANK FOR INTERNATIONAL SETTLEMENTS

Imputation for missing observation through Artificial Intelligence

A Heuristic & Machine Learning approach

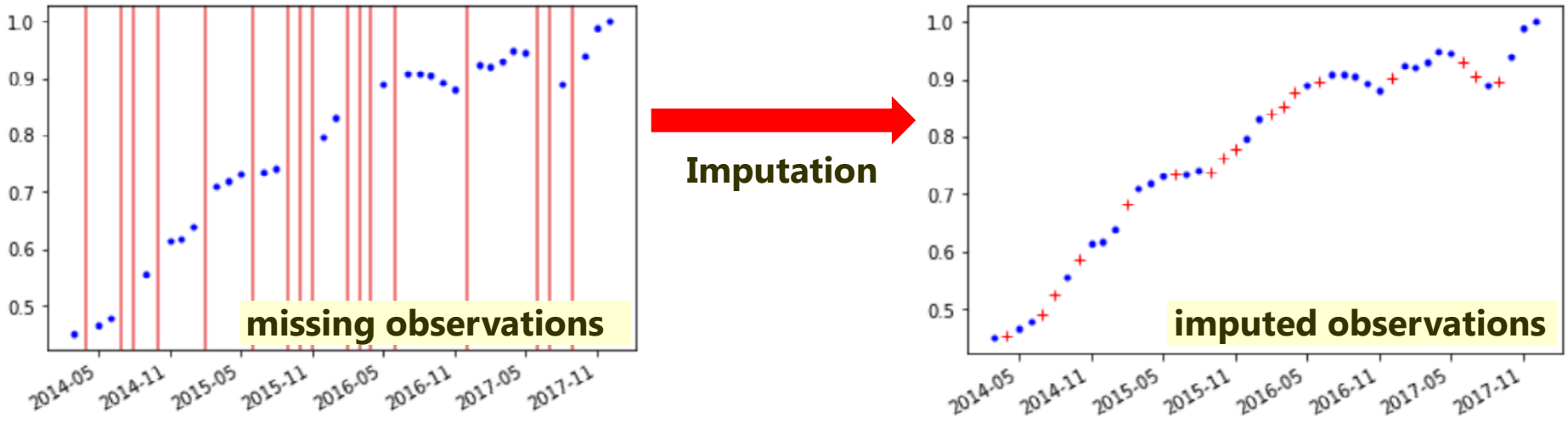
(Test case with macroeconomic time series from the BIS Data Bank)

Byeungchun Kwon

Bank for International Settlements

Missing observation imputation in univariate time series

- To impute missing observations in univariate time series, statisticians mainly use Interpolation, Moving Average, LOCF (Last Observation Carried Forward), Seasonal Decomposition, Kalman Smoothing and etc.

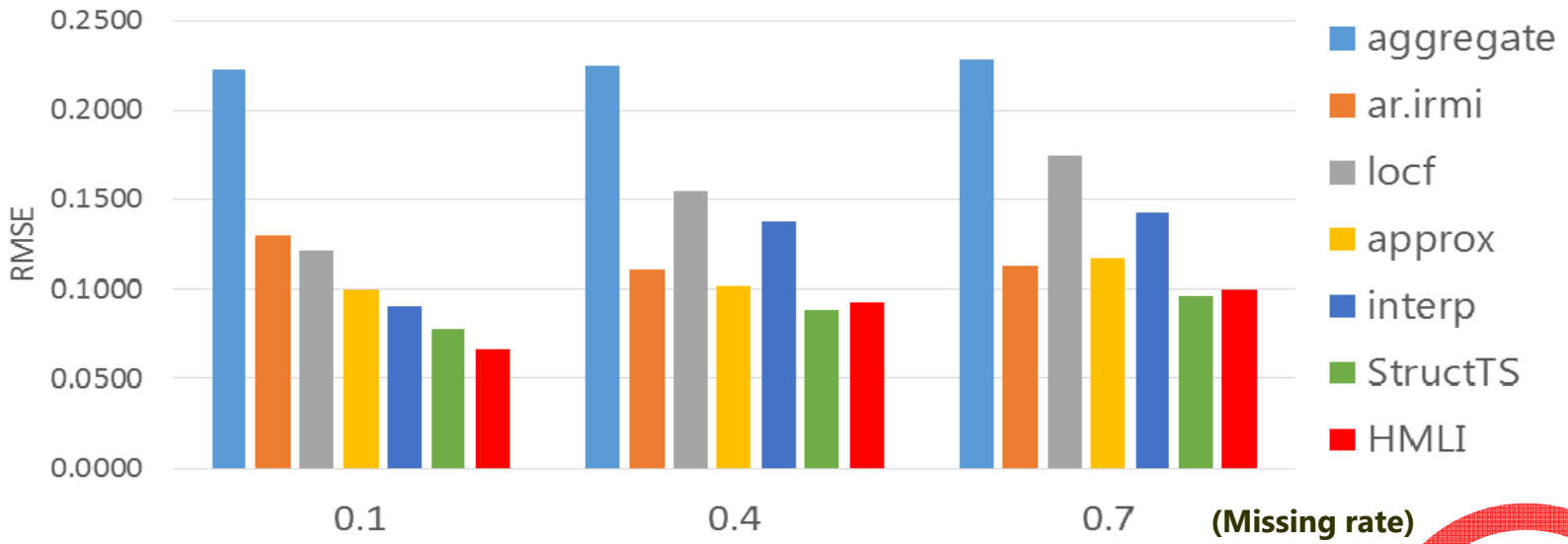


- How precise are the results? Is this the best method?

→ Let's build an Artificial Intelligence model and let's compete with traditional models

Average RMSE* between actual and imputed observations (3,070 macroeconomic time series)

* **RMSE**: Root-Mean-Square Error



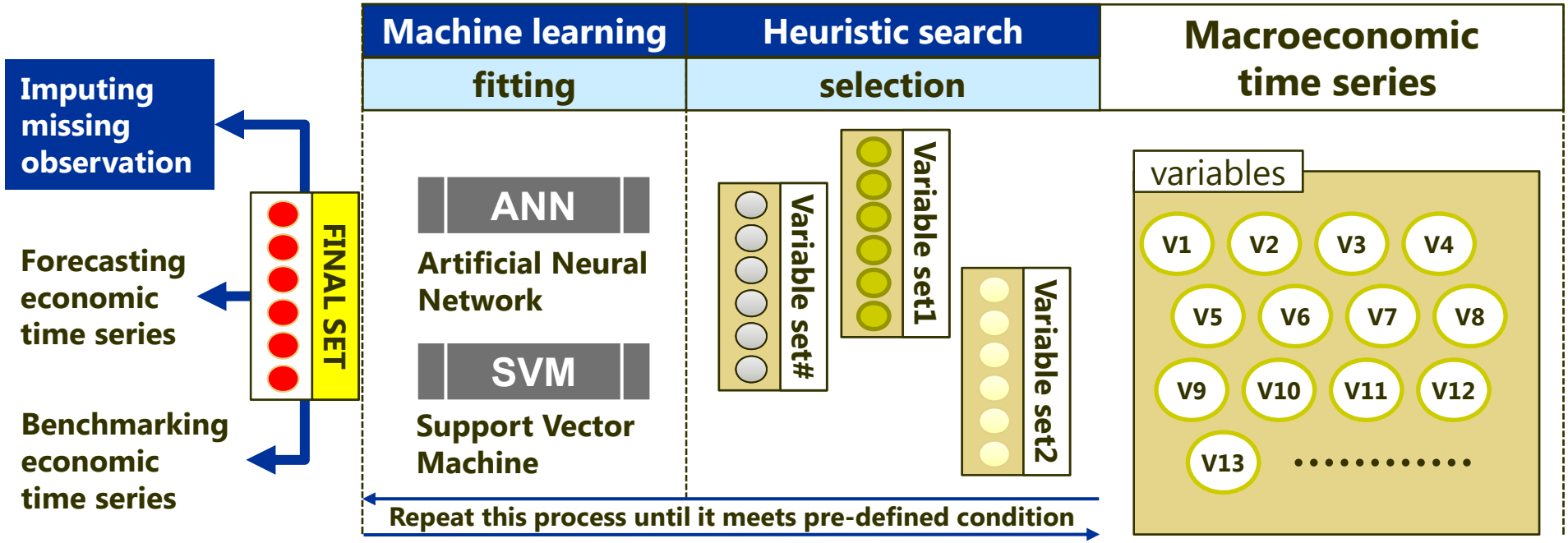
Missing rate	aggregate	ar.irmi	locf	approx	interp	StructTS	HMLI
0.1	0.2221	0.1301	0.1218	0.0998	0.0901	0.0781	**0.0658**
0.4	0.2247	0.1107	0.1552	0.1020	0.1384	**0.0880**	**0.0924**
0.7	0.2280	0.1130	0.1749	0.1175	0.1432	**0.0961**	**0.1001**

**** Comparison of different Methods for Univariate Time Series Imputation in R, Steffen Mortiz, Oct 2015***

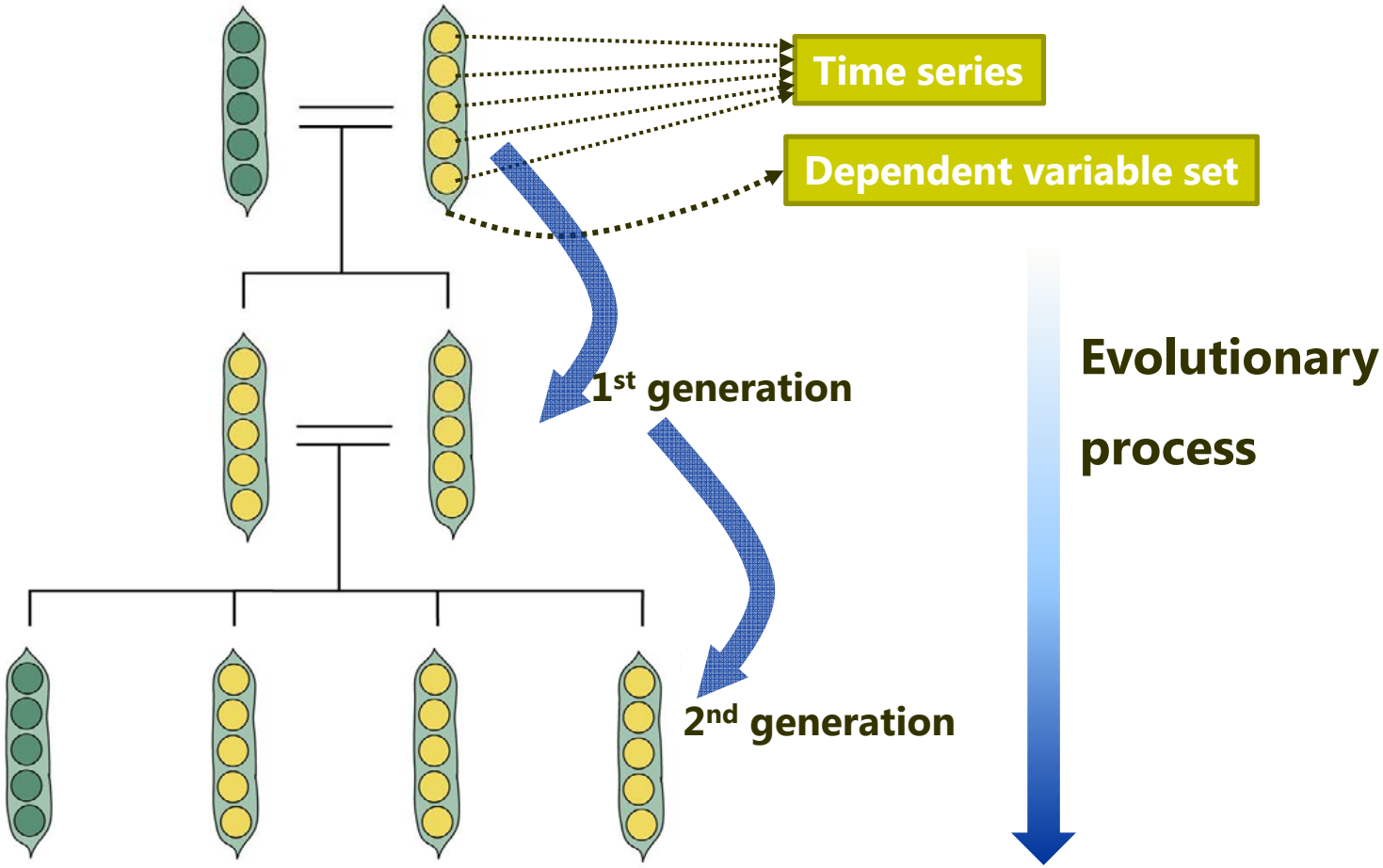
- aggregate: replacing NA with the overall mean
- structTS: filling NA through seasonal Kalman filter
- locf(Last observation carried Forward): replacing NA with most recent non-NA value
- approx: replacing NA with linear interpolation
- irmi(Iterative Robust Model-Based Imputation): filling NA through autoregressive imputation
- interp: linear interpolation for non-seasonal series. If seasonal series, a robust STL decomposition proceeded

HMLI (Heuristic & Machine Learning Imputation) structure

- HMLI is a nonlinear regression model
- Heuristic method selects dependent variables without manual intervention
- Machine Learning method estimates parameters in the model



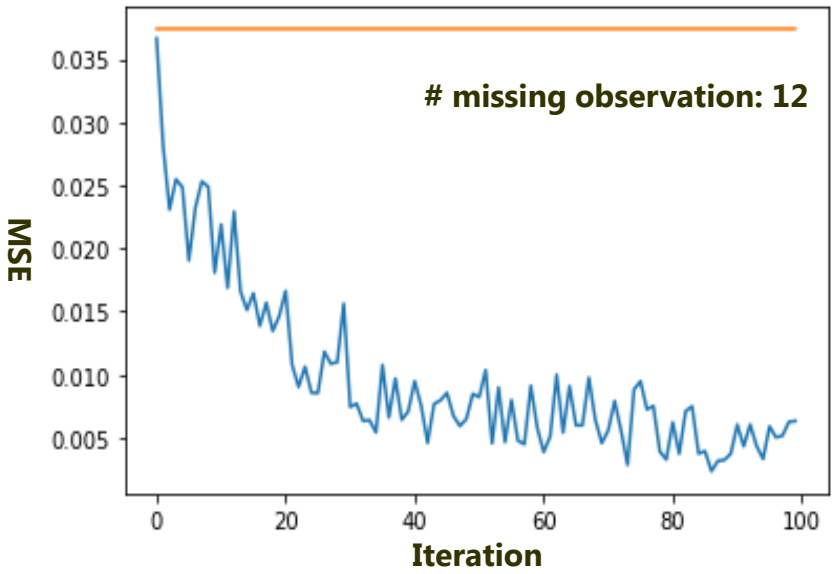
HMLI process – Idea from Mendelian Genetics



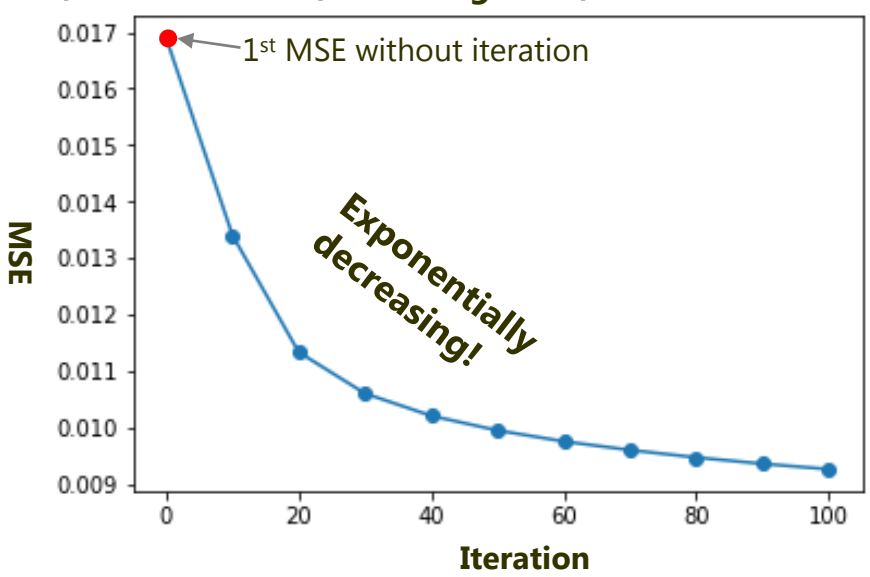
Adaptation in Natural and Artificial Systems, Holland, 1975
Natural Computing Algorithm, Barbazon et al., 2015

Mean square error (MSE) by iteration

M:AGNA:CA:04(96 observations: 2010.01 – 2017.12)



3,070 time series, 3 missing rates, 3 random seeds



Average MSE for 27,630 experiments

iteration	0	10	20	30	40	50	60	70	80	90	100
MSE	0.0169	0.0134	0.0113	0.0106	0.0102	0.0099	0.0097	0.0096	0.0095	0.0094	0.0093

HMLI process

Pre-processing: create gaps in a complete time series

(Number of gaps are decided by the exponential distribution and λ is missing rate)

Jan-17	Feb-17	Mar-17	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17	Oct-17	Nov-17	Dec-17
0.011885	0.017447	0.019291	0.011446	0.004332	0	0.007348	0.007055	0.011885	0.004332	0.007055	0.017447

Jan-17	Feb-17	Mar-17	Apr-17	May-17	Jun-17	Jul-17	Aug-17	Sep-17	Oct-17	Nov-17	Dec-17
NA	NA	0.019291	0.011446	NA	0	0.007348	NA	0.011885	0.004332	0.007055	0.017447

STEP1: remove gaps from the time series

Mar-17	Apr-17	Jun-17	Jul-17	Sep-17	Oct-17	Nov-17	Dec-17
0.019291	0.011446	0	0.007348	0.011885	0.004332	0.007055	0.017447

STEP2: (sampling) pick 6 time series from 3,070 for dependent variables and repeat this process 10 times

Set #1 V1 V5 V33 V114 V555 V1116 **Set #2** V100 V455 V1333 V3114 V3555 V4116

Set #10 V1 V5 V33 V114 V555 V1116

STEP3: SVM regression and predict gaps(missing observations)

	RMSE	ranking
SET #1	0.004	1
SET #2	0.019	10
⋮		
SET #10	0.010	5

STEP4: calculate RMSE* between the actual and predict observations

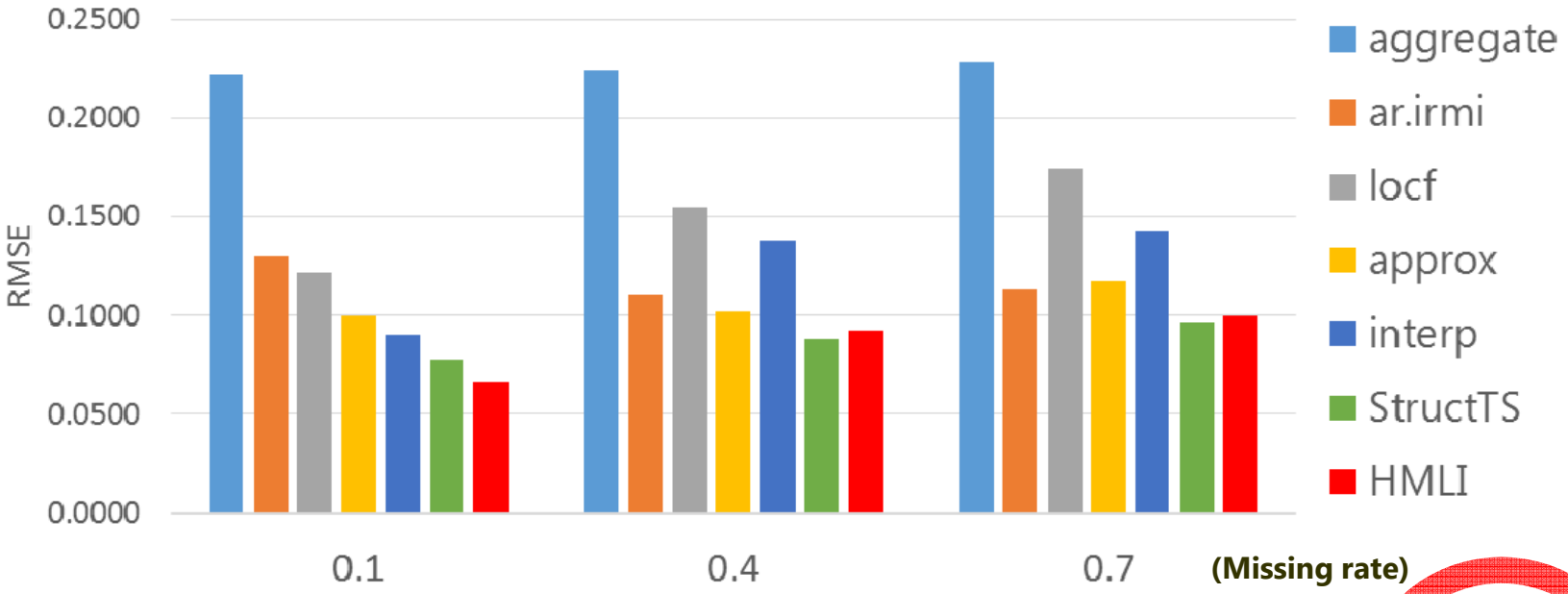
STEP5: remove 5 lower ranked sets

STEP6: generate 2 new sets through top 5 sets

STEP7: redo STEP2, but repeat 3 times to generate 3 sets

STEP8: iterate 100 times from STEP3 to STEP7

Average RMSE* between actual and imputed observations (3,070 macroeconomic time series)



Missing rate	aggregate	ar.irmi	locf	approx	interp	StructTS	HMLI
0.1	0.2221	0.1301	0.1218	0.0998	0.0901	0.0781	**0.0658**
0.4	0.2247	0.1107	0.1552	0.1020	0.1384	**0.0880**	**0.0924**
0.7	0.2280	0.1130	0.1749	0.1175	0.1432	**0.0961**	**0.1001**

**** Comparison of different Methods for Univariate Time Series Imputation in R, Steffen Mortiz, Oct 2015***

- aggregate: replacing NA with the overall mean
- structTS: filling NA through seasonal Kalman filter
- locf(Last observation carried Forward): replacing NA with most recent non-NA value
- approx: replacing NA with linear interpolation
- irmi(Iterative Robust Model-Based Imputation): filling NA through autoregressive imputation
- interp: linear interpolation for non-seasonal series. If seasonal series, a robust STL decomposition proceeded

Findings

- HMLI is one of the best solutions to impute missing observation from macroeconomic time series
- Heuristic & machine learning combination is effective in a complex space

Follow-up tasks

- Parameter calibration – number of dependent series, iteration, cutoff rate and etc.
- Test various time series data sets: different frequencies and pattern (trend, seasonality)
- Apply other machine learning functions like CNN(Convolutional Neural Networks)

Additional info

- HMLI is a Python script program and it is free. Please find the script on https://github.com/byeungchun/HeuristicImputation
- Also, experimental results are shared on this site