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)

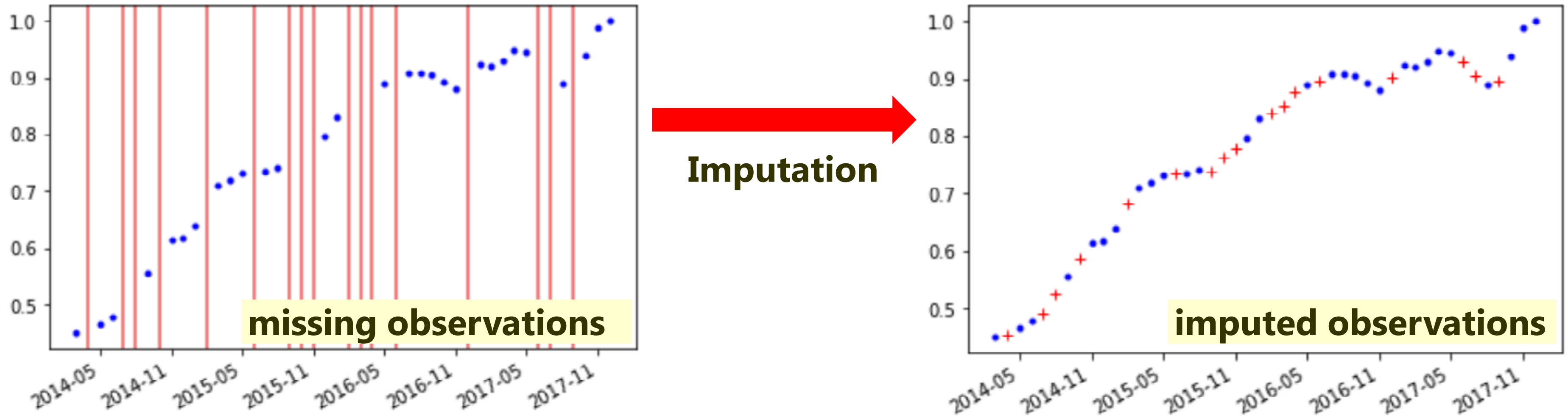
Byeungchun Kwon

Bank for International Settlements

Disclaimer: The views expressed in the presentation are those of the author and do not necessarily reflect those of the 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.

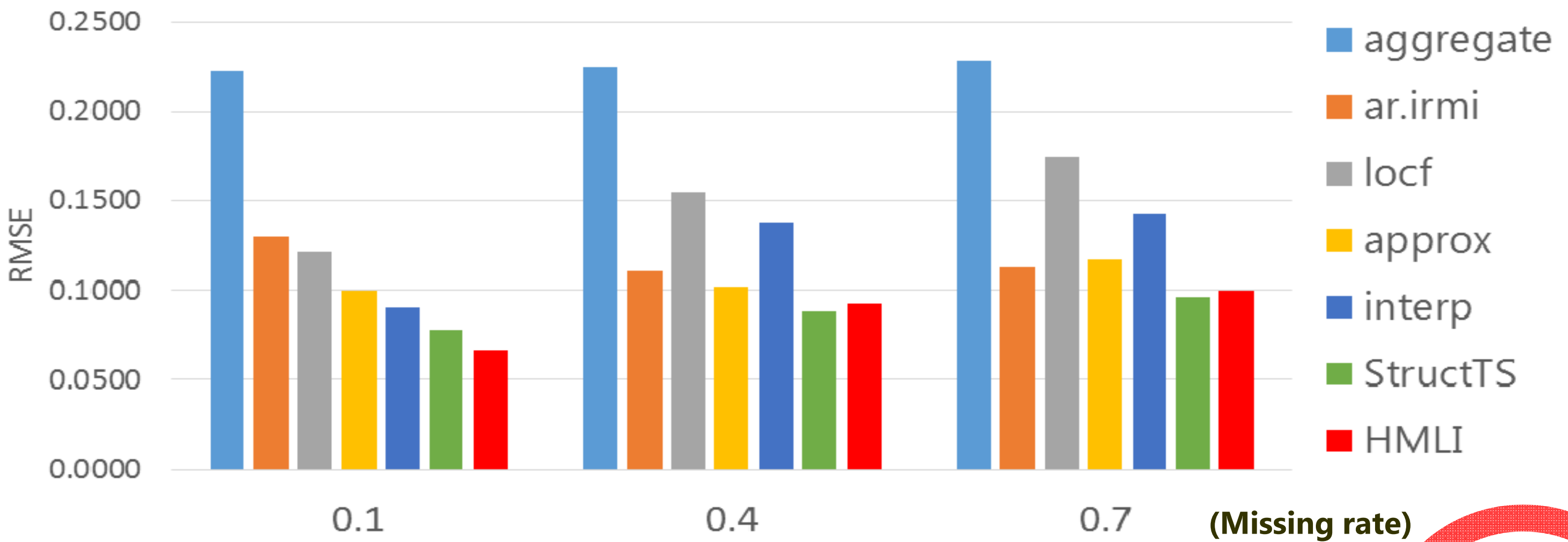


- How precise are the results? Is this the best method?

→ 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)

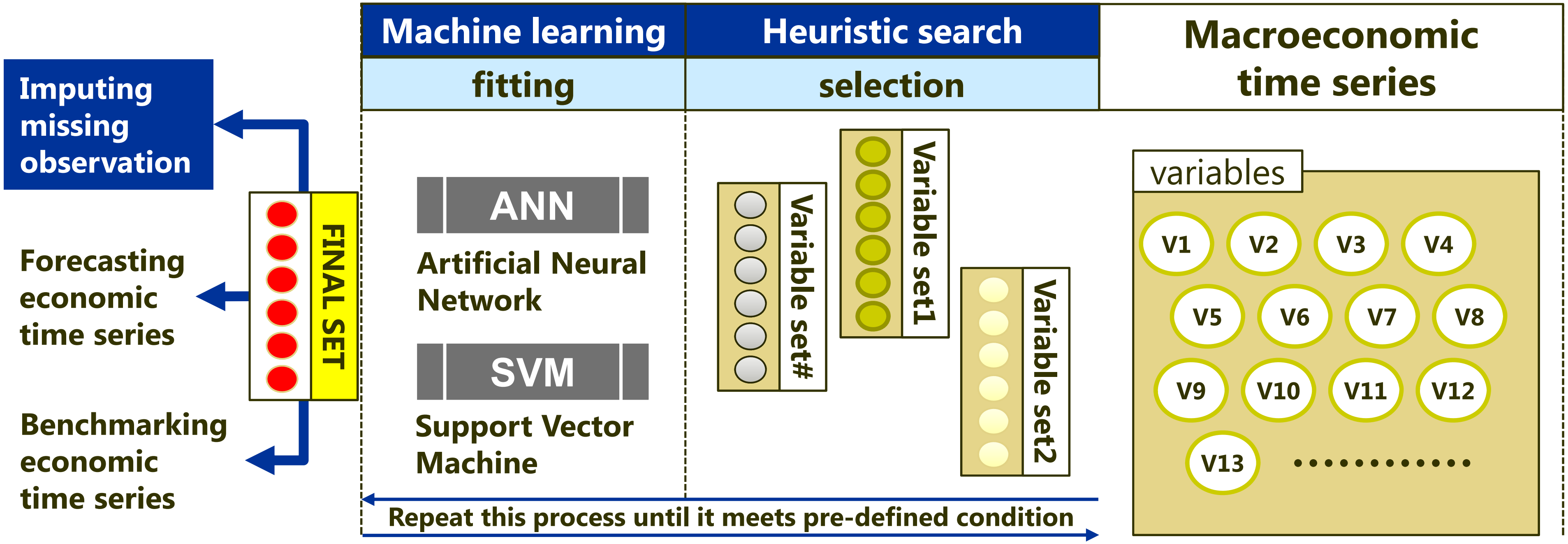
Missing rate	aggregate	ar.irmi	locf	approx	interp	StructTS	HMLI
0.1	0.2221	0.1301	0.1218	0.0998	0.0901	0.0781	**0.0658**
0.4	0.2247	0.1107	0.1552	0.1020	0.1384	**0.0880**	**0.0924**
0.7	0.2280	0.1130	0.1749	0.1175	0.1432	**0.0961**	**0.1001**

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

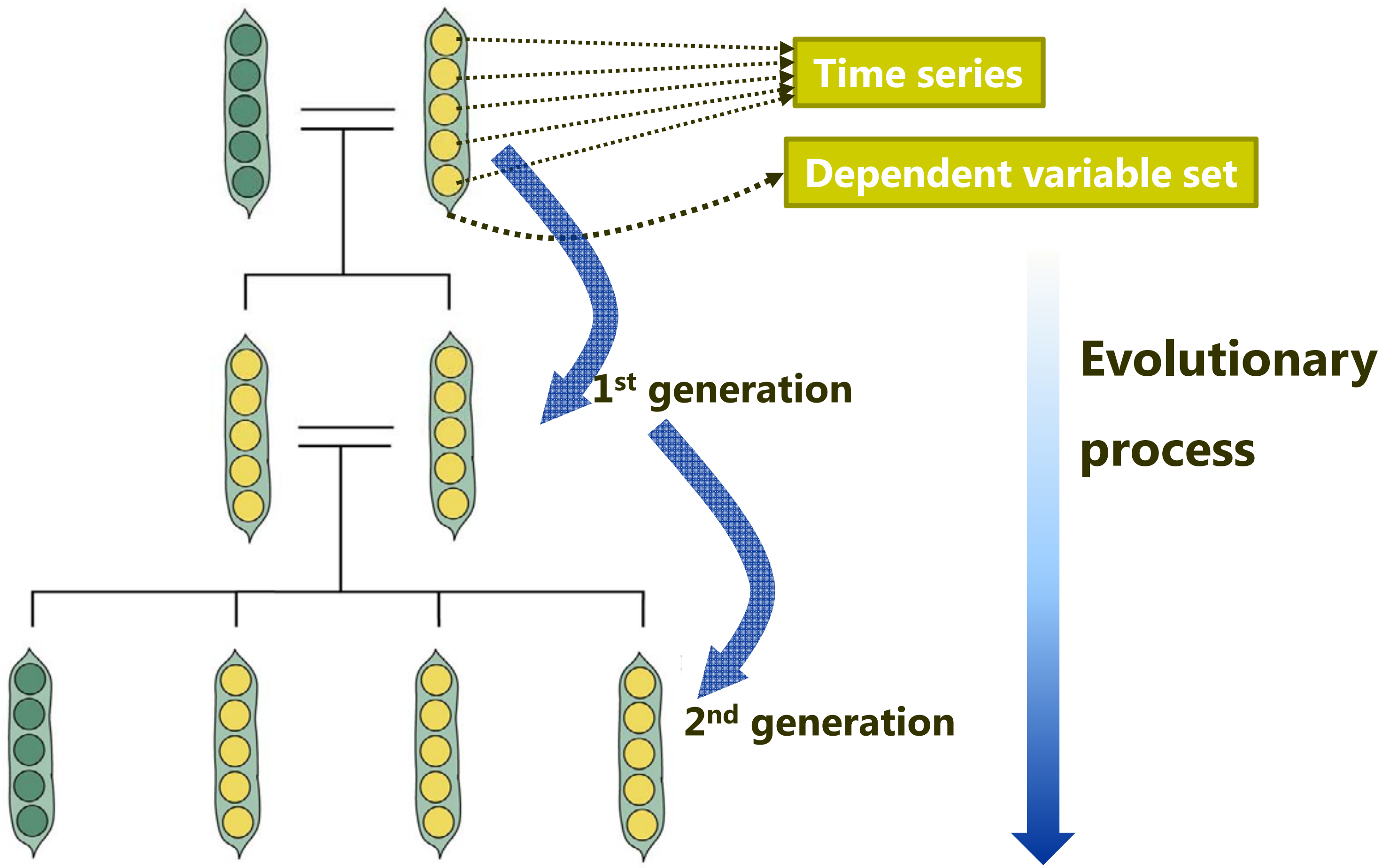
- aggregate: replacing NA with the overall mean
- structTS: filling NA through seasonal Kalman filter
- locf(Last observation carried Forward): replacing NA with most recent non-NA value
- approx: replacing NA with linear interpolation
- irmi(Iterative Robust Model-Based Imputation): filling NA through 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

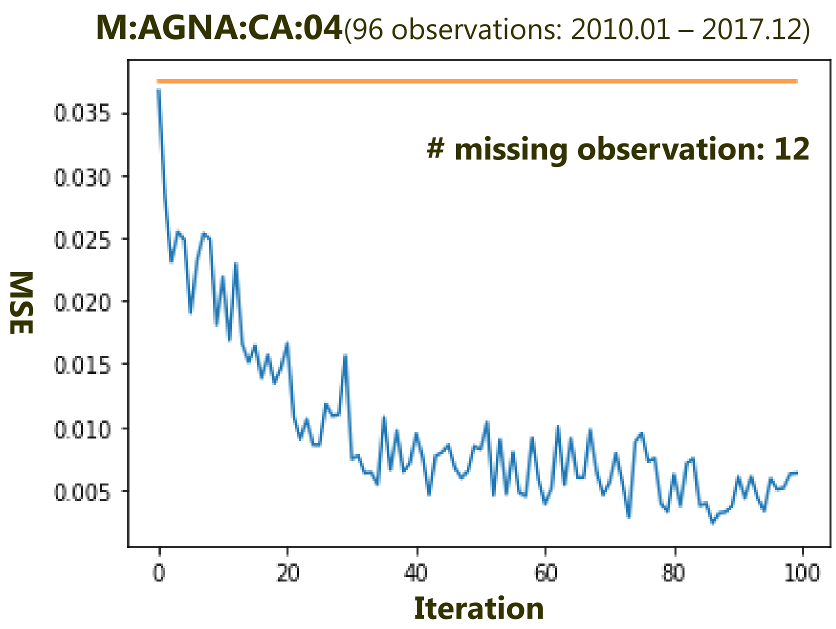


HMLI process – Idea from Mendelian Genetics



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

Mean square error (MSE) by iteration



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

Dependent variable candidates

V1 V2 V3 V4 V5 V6 V7 V8 • • • V3070

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

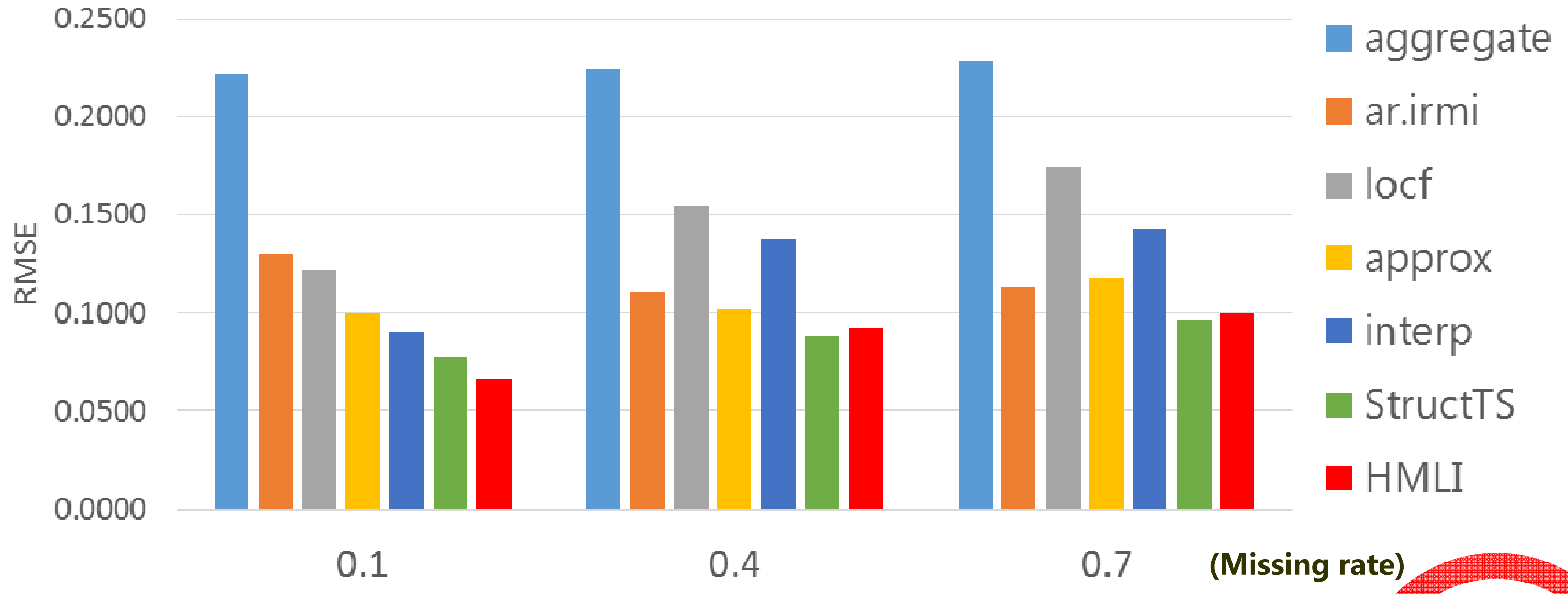
STEP5: remove 5 lower ranked sets

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