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.

![Missing observations and imputed observations](image)

- 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

<table>
<thead>
<tr>
<th>Missing rate</th>
<th>aggregate</th>
<th>ar.irmi</th>
<th>locf</th>
<th>approx</th>
<th>interp</th>
<th>StructTS</th>
<th>HMLI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>0.1</strong></td>
<td>0.2221</td>
<td>0.1301</td>
<td>0.1218</td>
<td>0.0998</td>
<td>0.0901</td>
<td>0.0781</td>
<td><strong>0.0658</strong></td>
</tr>
<tr>
<td><strong>0.4</strong></td>
<td>0.2247</td>
<td>0.1107</td>
<td>0.1552</td>
<td>0.1020</td>
<td>0.1384</td>
<td><strong>0.0880</strong></td>
<td><strong>0.0924</strong></td>
</tr>
<tr>
<td><strong>0.7</strong></td>
<td>0.2280</td>
<td>0.1130</td>
<td>0.1749</td>
<td>0.1175</td>
<td>0.1432</td>
<td><strong>0.0961</strong></td>
<td><strong>0.1001</strong></td>
</tr>
</tbody>
</table>

*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
- ar.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

<table>
<thead>
<tr>
<th>Machine learning fitting</th>
<th>Heuristic search selection</th>
<th>Macroeconomic time series</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td></td>
<td>V1 V2 V3 V4 V5 V6 V7 V8</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td></td>
<td>V9 V10 V11 V12 V13 .........</td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Variable set 1
Variable set 2
Variable set #

Repeat this process until it meets pre-defined condition
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


# missing observation: 12

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

Average MSE for 27,630 experiments

<table>
<thead>
<tr>
<th>iteration</th>
<th>0</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>0.0169</td>
<td>0.0134</td>
<td>0.0113</td>
<td>0.0106</td>
<td>0.0102</td>
<td>0.0099</td>
<td>0.0097</td>
<td>0.0096</td>
<td>0.0095</td>
<td>0.0094</td>
<td>0.0093</td>
</tr>
</tbody>
</table>
**Pre-processing: create gaps in a complete time series**
(Number of gaps are decided by the exponential distribution and λ is missing rate)

<table>
<thead>
<tr>
<th></th>
<th>Jan-17</th>
<th>Feb-17</th>
<th>Mar-17</th>
<th>Apr-17</th>
<th>May-17</th>
<th>Jun-17</th>
<th>Jul-17</th>
<th>Aug-17</th>
<th>Sep-17</th>
<th>Oct-17</th>
<th>Nov-17</th>
<th>Dec-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.011885</td>
<td>0.017447</td>
<td>0.019291</td>
<td>0.011446</td>
<td>0.004332</td>
<td>0</td>
<td>0.007348</td>
<td>0.007055</td>
<td>0.011885</td>
<td>0.004332</td>
<td>0.007055</td>
<td>0.017447</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Jan-17</th>
<th>Feb-17</th>
<th>Mar-17</th>
<th>Apr-17</th>
<th>May-17</th>
<th>Jun-17</th>
<th>Jul-17</th>
<th>Aug-17</th>
<th>Sep-17</th>
<th>Oct-17</th>
<th>Nov-17</th>
<th>Dec-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>NA</td>
<td>NA</td>
<td>0.019291</td>
<td>0.011446</td>
<td>NA</td>
<td>0</td>
<td>0.007348</td>
<td>NA</td>
<td>0.011885</td>
<td>0.004332</td>
<td>0.007055</td>
<td>0.017447</td>
</tr>
</tbody>
</table>

**STEP1: remove gaps from the time series**

<table>
<thead>
<tr>
<th></th>
<th>Mar-17</th>
<th>Apr-17</th>
<th>Jun-17</th>
<th>Jul-17</th>
<th>Sep-17</th>
<th>Oct-17</th>
<th>Nov-17</th>
<th>Dec-17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.019291</td>
<td>0.011446</td>
<td>0</td>
<td>0.007348</td>
<td>0.011885</td>
<td>0.004332</td>
<td>0.007055</td>
<td>0.017447</td>
</tr>
</tbody>
</table>

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

**STEP3: SVM regression and predict gaps (missing observations)**

<table>
<thead>
<tr>
<th>RMSE</th>
<th>ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>SET #1</td>
<td>0.004</td>
</tr>
<tr>
<td>SET #2</td>
<td>0.019</td>
</tr>
<tr>
<td>SET #10</td>
<td>0.010</td>
</tr>
</tbody>
</table>

**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**
**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