Addressing the Log Representativeness Problem using Species Discovery

Martin Kabierski, Markus Richter, Matthias Weidlich
Process Analysis in the Wild

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
The Representativeness Problem

**GENERATES**

Does the log represent the generative system well?

*Trustworthiness*

*Reliability*

*Confidence*

**ELLCITS CHANGES**

**Event log**

Is a sample of generative system

- Incompleteness
- Noise
- Erroneous logging
- Process drifts
- ...

Generative system

Addressing the Log Representativeness Problem using Species Discovery

5th International Conference on Process Mining

Rome, Italy, 24.10.2023
The many Faces of Representativeness

“A good sample will be representative in the sense that characteristics of interest in the sample can be estimated with a known degree of accuracy.”

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
The many Faces of Representativeness

„A good sample will be representative in the sense that characteristics of interest in the sample can be estimated with a known degree of accuracy.“
The many Faces of Representativeness

Functions on Event Logs
Estimating Function Convergence based on statistical tests
- Bauer, Senderovich, Gal, Grunske, Weidlich, How much Event Data is Enough? A statistical framework for process discovery., CAiSE 2018
- Bauer, van der Aa, Weidlich, Estimating Process Conformance by Trace Sampling and Result Approximation., BPM 2019
- Bauer, van der Aa, Weidlich, Sampling and approximation techniques for efficient process conformance checking., Information Systems, 2022

Completeness
Does the log contain all values present in the generative system?
Assumptions: no Drifts & no Errors in the Log
A different View: Biodiversity Analysis

Bird Populations Are in Meltdown
[wire.com, 20.06.2023]

With an estimated 5.5 million species, insects are the most diverse group of animals on the planet. More than one million have been named by scientists — and many more have yet to be discovered.
[Florida Museum]

[...] we estimate the total tree species richness at global, continental, and biome levels. Our results indicate that there are ~73,000 tree species globally, among which ~9,000 tree species are yet to be discovered
[Gatti et. al, The number of tree species on Earth, 2022]
A different View: Biodiversity Analysis

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
Species Richness Curves

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
The Bernoulli Space Model

Sampling Process
• i.i.d. sampling
• species have unknown, fixed observation probability
• one observation may contain multiple species

Chao2-Estimator [A. Chao. 1984]
\[ S_{Chao2} \approx S_{obs} + \frac{Q_1^2}{2Q_2} \]

Species Completeness
„What fraction of all species have we observed?“

Species Coverage
„How much of the probability space do the unobserved species cover?“

Sampling Extrapolation
„How much longer do we need to observe birds until we observe new species?“
From Birds to Event Log Species

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
From Birds to Event Log Species

Addressing the Log Representativeness Problem using Species Discovery
5th International Conference on Process Mining
Rome, Italy, 24.10.2023
Estimating Richness on complete Logs

Python-based Implementation of metrics and log species

Calculated all proposed metrics on public event logs
BPI-2012, BPI-2018, BPI-2019, Sepsis Cases

Considered log species
activities, df-relations, trace variants, activities + execution times

Table II: Species richness estimation, coverage, and completeness for four event logs, and seven species definitions.

<table>
<thead>
<tr>
<th>Log</th>
<th>Species Def.</th>
<th>(S_{obs})</th>
<th>(S_{est})</th>
<th>(Q_1)</th>
<th>(Q_2)</th>
<th>(Cov_{obs})</th>
<th>(Com_{obs})</th>
<th>(l_{.99})</th>
<th>(l_{.95})</th>
<th>(l_{.90})</th>
<th>(l_{.80})</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPI-2012</td>
<td>(\zeta_{act})</td>
<td>24</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{df})</td>
<td>149</td>
<td>161</td>
<td>7</td>
<td>2</td>
<td>0.999</td>
<td>0.925</td>
<td>46435</td>
<td>9577</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{ev})</td>
<td>4336</td>
<td>30346</td>
<td>3727</td>
<td>267</td>
<td>0.715</td>
<td>0.143</td>
<td>406521</td>
<td>259527</td>
<td>196219</td>
<td>132912</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{t1})</td>
<td>958</td>
<td>2816</td>
<td>535</td>
<td>77</td>
<td>0.996</td>
<td>0.340</td>
<td>190458</td>
<td>117290</td>
<td>85779</td>
<td>54267</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{t5})</td>
<td>487</td>
<td>1164</td>
<td>268</td>
<td>53</td>
<td>0.998</td>
<td>0.418</td>
<td>134446</td>
<td>81196</td>
<td>58263</td>
<td>35329</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{t80})</td>
<td>210</td>
<td>288</td>
<td>74</td>
<td>53</td>
<td>0.999</td>
<td>0.729</td>
<td>45666</td>
<td>23401</td>
<td>13812</td>
<td>4223</td>
</tr>
<tr>
<td></td>
<td>(\zeta_{te2})</td>
<td>112</td>
<td>112</td>
<td>2</td>
<td>5</td>
<td>0.999</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Estimating Richness on Log samples

(f) $S_{obs}$ for $\zeta_{df}$

(d) $Com_{obs}$ for $\zeta_{df}$

(b) $Cov_{obs}$ for $\zeta_{df}$
Estimating Richness on Subprocesses

Control-Flow based Splitting:

Does log completeness differ for different phases of the process?

<table>
<thead>
<tr>
<th>Spec.</th>
<th>Log</th>
<th>$S_{obs}$</th>
<th>$S_{est}$</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Cov_{obs}$</th>
<th>$Com_{obs}$</th>
<th>$l_{.90}$</th>
<th>$l_{.90}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\zeta_{act}$</td>
<td>Pre</td>
<td>11</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>15</td>
<td>15</td>
<td>2</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\zeta_{df}$</td>
<td>Pre</td>
<td>87</td>
<td>118</td>
<td>8</td>
<td>1</td>
<td>0.999</td>
<td>0.737</td>
<td>13811</td>
<td>4148</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>88</td>
<td>113</td>
<td>9</td>
<td>0</td>
<td>0.997</td>
<td>0.778</td>
<td>3426</td>
<td>897</td>
</tr>
<tr>
<td>$\zeta_{tv}$</td>
<td>Pre</td>
<td>298</td>
<td>735</td>
<td>180</td>
<td>37</td>
<td>0.828</td>
<td>0.405</td>
<td>10427</td>
<td>4550</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>467</td>
<td>2955</td>
<td>393</td>
<td>31</td>
<td>0.514</td>
<td>0.158</td>
<td>22735</td>
<td>10926</td>
</tr>
<tr>
<td>$\zeta_{t1}$</td>
<td>Pre</td>
<td>1041</td>
<td>1673</td>
<td>513</td>
<td>208</td>
<td>0.939</td>
<td>0.622</td>
<td>4699</td>
<td>1720</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>2392</td>
<td>13238</td>
<td>2558</td>
<td>195</td>
<td>0.646</td>
<td>0.181</td>
<td>18811</td>
<td>8980</td>
</tr>
<tr>
<td>$\zeta_{t30}$</td>
<td>Pre</td>
<td>127</td>
<td>191</td>
<td>41</td>
<td>13</td>
<td>0.995</td>
<td>0.665</td>
<td>5821</td>
<td>2010</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>1120</td>
<td>2566</td>
<td>639</td>
<td>141</td>
<td>0.889</td>
<td>0.436</td>
<td>7392</td>
<td>3170</td>
</tr>
<tr>
<td>$\zeta_{te2}$</td>
<td>Pre</td>
<td>132</td>
<td>144</td>
<td>14</td>
<td>8</td>
<td>0.998</td>
<td>0.917</td>
<td>1963</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Post</td>
<td>127</td>
<td>159</td>
<td>29</td>
<td>13</td>
<td>0.994</td>
<td>0.799</td>
<td>2717</td>
<td>638</td>
</tr>
</tbody>
</table>
Estimating Richness on Subprocesses

**Attribute-based Splitting:**

Does log completeness differ for different patient groups?

<table>
<thead>
<tr>
<th>Spec.</th>
<th>Log</th>
<th>$S_{obs}$</th>
<th>$S_{est}$</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>Cov$_{obs}$</th>
<th>Cov$_{est}$</th>
<th>l$_{99}$</th>
<th>l$_{99}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\zeta_{act}$</td>
<td>$&lt; 60$</td>
<td>15</td>
<td>16</td>
<td>2</td>
<td>1</td>
<td>0.998</td>
<td>0.938</td>
<td>547</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>16</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$\zeta_{df}$</td>
<td>$&lt; 60$</td>
<td>101</td>
<td>113</td>
<td>15</td>
<td>9</td>
<td>0.994</td>
<td>0.894</td>
<td>444</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>133</td>
<td>158</td>
<td>16</td>
<td>5</td>
<td>0.998</td>
<td>0.842</td>
<td>3675</td>
<td>631</td>
</tr>
<tr>
<td>$\zeta_{tv}$</td>
<td>$&lt; 60$</td>
<td>171</td>
<td>1946</td>
<td>158</td>
<td>7</td>
<td>0.287</td>
<td>0.088</td>
<td>11309</td>
<td>5539</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>707</td>
<td>8185</td>
<td>659</td>
<td>29</td>
<td>0.202</td>
<td>0.086</td>
<td>42374</td>
<td>20763</td>
</tr>
<tr>
<td>$\zeta_{t1}$</td>
<td>$&lt; 60$</td>
<td>907</td>
<td>3574</td>
<td>706</td>
<td>93</td>
<td>0.704</td>
<td>0.254</td>
<td>3636</td>
<td>1694</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>2858</td>
<td>10127</td>
<td>2138</td>
<td>314</td>
<td>0.793</td>
<td>0.282</td>
<td>12019</td>
<td>5543</td>
</tr>
<tr>
<td>$\zeta_{t30}$</td>
<td>$&lt; 60$</td>
<td>420</td>
<td>906</td>
<td>254</td>
<td>66</td>
<td>0.893</td>
<td>0.464</td>
<td>1703</td>
<td>718</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>1081</td>
<td>2316</td>
<td>599</td>
<td>145</td>
<td>0.942</td>
<td>0.467</td>
<td>6786</td>
<td>2857</td>
</tr>
<tr>
<td>$\zeta_{te2}$</td>
<td>$&lt; 60$</td>
<td>149</td>
<td>171</td>
<td>25</td>
<td>14</td>
<td>0.989</td>
<td>0.871</td>
<td>509</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>$\geq 60$</td>
<td>200</td>
<td>225</td>
<td>26</td>
<td>13</td>
<td>0.997</td>
<td>0.889</td>
<td>2018</td>
<td>114</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

Conclusion

• representativeness of event logs
• log completeness estimation using species richness estimation
• well-known event logs are incomplete in many dimensions
• enables assessment of log quality and analysis confidence

Future Work

• dropping error- & drift assumptions
• automatically detect incomplete sub processes
• evaluation against ground truth dataset

Thank you for your Attention!

martin.kabierski@hu-berlin.de
Credits

Icons taken from:
Freepik
Smashicons from www.flaticon.com
Vitaly Gorbachev from www.flaticon.com