SpeciAL4PM: Species Analysis of Event Logs for Process Mining

Martin Kabierski[1,2], Christian Imenkamp[3], Agnes Koschmider[3] and Matthias Weidlich[1]

[1]*Humboldt-Universität zu Berlin*
[2]*Weizenbaum-Institut*
[3]*University of Bayreuth*

Abstract

We present *SpeciAL4PM* (**Speci**es **A**nalysis of Event **L**ogs **for P**rocess **M**ining), a Python library for the analysis and visualization of event logs that incorporates notions of biodiversity research. Under this view, *SpeciAL4PM* enables the quantification of the completeness of event logs and the estimation of the diversity of the system from which the log originates. We supplement *SpeciAL4PM* with a web-based implementation, named *SpeciAL4PM-live* , that facilitates the use of the library without the need for programming, thereby supporting users in the exploration of their event data.

Keywords

Event Log Analysis, Log Completeness, Log Diversity, Process Mining

1. Introduction

Event logs, whether they are created by the execution of process-centric information systems or by simulation of process models, build the foundation of process analysis [1]. Yet, these event logs are only samples of the underlying (information) system that generated them. Hence, they cannot be assumed to be complete with respect to the recorded behavioural characteristics a-priori. This incompleteness, influenced by the size of the event log and the diversity of the behavioural characteristics of interest, may skew any analysis proportional to the diversity of these characteristics, i.e., incomplete logs may yield false insights. Thus, to draw trustworthy conclusions, one shall quantify (i) the completeness of an event log with respect to the relevant characteristics and (ii) the expected diversity of them, independent of the given log.

In recent work [2, 3], we showed how to quantify both completeness and diversity by employing biodiversity estimators. Those treat event logs as samples of observed species that are obtained under appropriate sampling models, and estimate properties of the species population therefrom. In particular, we considered diversity under different behavioural abstractions, i.e.,

Proceedings of the Best BPM Dissertation Award, Doctoral Consortium, and Demonstrations & Resources Forum co-located with 22nd International Conference on Business Process Management (BPM 2024), Krakow, Poland, September 1st to 6th, 2024.

*Corresponding author.
martin.kabierski@hu-berlin.de (M. Kabierski); christian.imenkamp@uni-bayreuth.de (C. Imenkamp); agnes.koschmider@uni-bayreuth.de (A. Koschmider); matthias.weidlich@hu-berlin.de (M. Weidlich)
0000-0002-9852-7489 (M. Kabierski); 0009-0007-4295-1268 (C. Imenkamp); 0000-0001-8206-7636 (A. Koschmider); 0000-0003-3325-7227 (M. Weidlich)

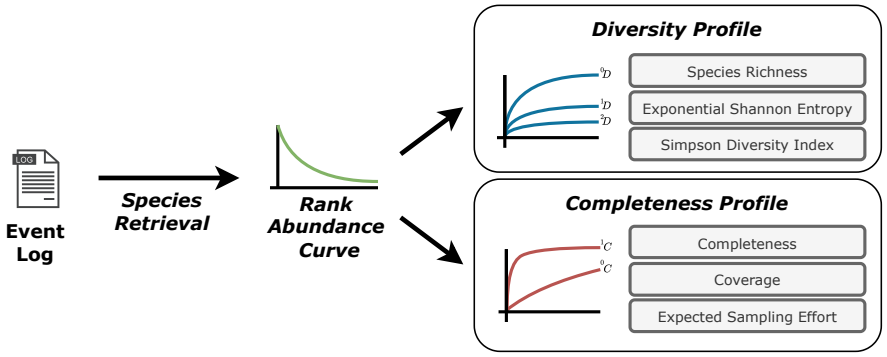


Figure 1: The functionality offered by *SpeciAL4PM* .

log species, using a diversity profile based on asymptotic Hill numbers [4]. It captures properties of the observed distribution of species, based on their occurrence frequencies in the log. We also presented a completeness profile to quantify log completeness in absolute and relative terms.

To facilitate the analysis of event logs when considering them as samples of some population of species, in this work, we propose *SpeciAL4PM* (**Speci**es **A**nalysis of Event **L**ogs **for P**rocess **M**ining), a Python library for quantifying and visualizing the completeness profile and the diversity profile of an event log. We complement the library with *SpeciAL4PM-live* , a web-based tool for the main functions of *SpeciAL4PM* , which allows a user to upload an event log and explore the completeness and diversity estimates, thus enabling an easy-to-use, coding-free application of the respective measures. With *SpeciAL4PM* , a user can assess the completeness and diversity of an event log, incorporate these factors in their event log analysis, and be confident that obtained insights are supported by sufficiently complete data.

In the remainder, in Section 2, we discuss the features of *SpeciAL4PM* and explain how the library is used. Then, in Section 3 we discuss *SpeciAL4PM-live* , before discussing the availability and maturity of the tool in Section 4. Lastly, we conclude in Section 5.

2. SpeciAL4PM

SpeciAL4PM enables the calculation and visualization of species abundance curves, completeness profiles and diversity profiles for different species definitions, as illustrated in Figure 1 and proposed in [3]. Currently, *SpeciAL4PM* supports the analysis of event logs in .xes-format. Below, we outline the functionality in more detail.

Species Retrieval A species retrieval function captures the behavioral properties of interest of the event log per trace. These properties serve as the basis for the following species-based analysis tasks. Currently, *SpeciAL4PM* supports observed activities, directly-follows relations, n-grams and trace variants, but users can provide their own species definitions as well. Furthermore, users can specify multiple species definitions per log, facilitating the efficient computation and comparison of different behavioural properties.

Rank Abundance Curves A rank abundance curve visualizes the distribution of retrieved species in the event log and provides a graphical representation of distribution characteristics and the diversity of the event log. It serves as the basis for following species analysis tasks.

Diversity Profile A diversity profile summarizes the diversity of an event log and the estimated diversity of the complete system. Currently, *SpeciAL4PM* captures the diversity profile of an event log using observed Hill numbers and estimated asymptotic Hill numbers [4]. Hill numbers are a set of measures parameterized by a diversity order q, that quantifies different aspects of an event log's diversity. Intuitively, the larger q, the more emphasis is put on the most frequent species. The Hill number of order $q = 0$ equals *species richness*, i.e. the number of observed distinct species in the event log and the number of estimated distinct species in the system. Hill numbers of order $q = 1$ and $q = 2$ correspond to the exponential of Shannon Entropy and the Inverse of Simpsons Diversity Index, two commonly used diversity measures. These can also be quantified for the observed event log and for the complete system correcting the estimate for unobserved species.

Completeness Profile Based on the rank abundance curve, a completeness profile quantifies the completeness of the event log for different dimensions. In particular, *SpeciAL4PM* allows for the quantification of the event log's *completeness*, *coverage* and *expected sampling effort*. *Completeness* quantifies the fraction of expected species in the system, that are observed in the event log, while *coverage* quantifies the probability space of all species, that the observed species in the log take up. Both measures capture completeness in absolute and relative dimensions. Lastly, *SpeciAL4PM* allows the estimation of the expected additional sampling effort needed until the log reaches a target *completeness*. All measures, sample-based and estimated, are implemented using the common sampling models, i.e. the abundance data model and the incidence data model, as described in [3], and the expected difference between both models is quantified.

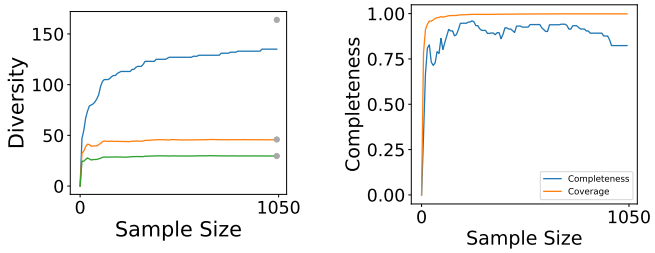
In Figure 2, we illustrate how to calculate and assess the profiles and curves for the publicly available *Sepsis Cases event log* [5] using *SpeciAL4PM* . First, an estimator object is created (line 9), specifying after how many traces the proposed metrics shall be updated. Then, two species retrieval functions are registered (lines 12-13), one considering activities per trace, and one considering directly-follows relations per trace, before the profiles and rank abundance curves for both species retrieval functions are computed (lines 6 and 16). Finally, the obtained measures can be printed (line 19), saved to a pandas data frame (line 21) or visualized (lines 23-25). The obtained profiles for the directly-follows species are illustrated in Figure 3. Additionally, *SpeciAL4PM* supports the considerations of only a subset of the proposed measures and additional visualization functions not shown here.

3. SpeciAL Online

Additionally, for ease of use, we present *SpeciAL4PM-live* , a web-based tool for the main functionalities of *SpeciAL4PM* , which can be used for quantifying species information of an event log without the need for programming. Users can upload an event log, for which then

```python
from special4pm.estimation import SpeciesEstimator
from special4pm.species import retrieve_species_n_gram
from special4pm.visualization import plot_rank_abundance, plot_completeness-profile, plot_diversity_profile

#import event log using pm4py
log = pm4py.read_xes("Sepsis_Cases_-_Event_Log.xes")

#create estimator
estimator = SpeciesEstimator(step_size=100)

#register different species definitions
estimator.register("1-gram", partial(retrieve_species_n_gram, n=1))
estimator.register("2-gram", partial(retrieve_species_n_gram, n=2))

#calculate species-based diversity and completeness
estimator.apply(log)

#print all profiles for all registered species definitions
estimator.print_metrics()
#or save profiles to a Pandas DataFrame
df = estimator.to_dataFrame()
#or visualize different aspects of event log
plot_rank_abundance(estimator, "2-gram")
plot_diversity_profile(estimator, "2-gram")
plot_completeness_profile(estimator, "2-gram")
```

Figure 2: Example Code for analyzing an event log using multiple species definitions.



(a) Diversity Profile. (b) Completeness Profile.

Figure 3: Diversity and Completeness Profile of *Sepsis Cases*.

profiles and abundance curves for different species retrieval functions are calculated and shown, as illustrated in Figure 4.

4. Availability

SpeciAL4PM is distributed via the Python Package Index[1] and can be installed from the command line interface. *SpeciAL4PM-live* can be accessed using a web browser.[2] Furthermore, the source code for both versions is available on GitHub[3] under the MIT license. Lastly, we provide a screencast showcasing the usage of both *SpeciAL4PM* and *SpeciAL4PM-live* .[4] We intend to update both *SpeciAL4PM* and *SpeciAL4PM-live* as new use cases for the species-based analysis of event logs and process data emerge.

[1] https://pypi.org/project/special4pm/
[2] https://martinkabierski.shinyapps.io/special4pm-live/
[3] https://github.com/MartinKabierski/SpeciAL-core, https://github.com/MartinKabierski/SpeciAL
[4] https://youtu.be/HVtvLeQ8cQI

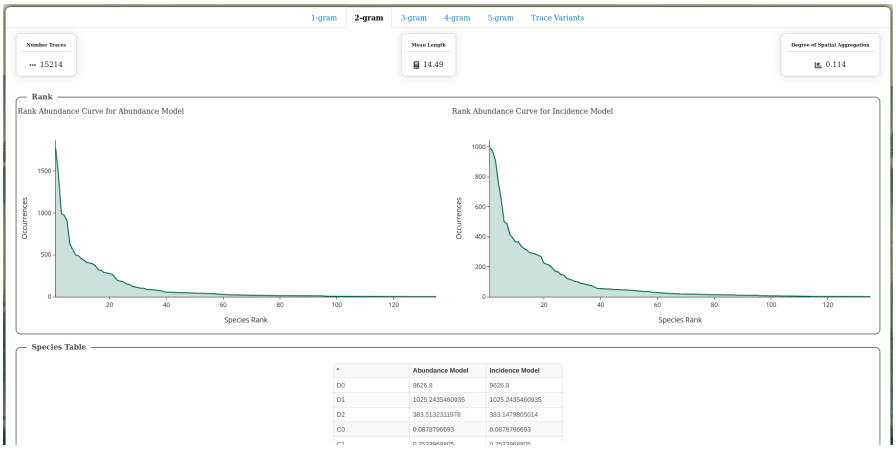


Figure 4: A partial view of *SpeciAL4PM-live* after the *Sepsis Cases* event log has been uploaded.

5. Conclusion

In this demo, we propose *SpeciAL4PM* , a library for the analysis of event logs that adopts notions from biodiversity research. The library enables the retrieval of species from an event log, the subsequent calculation of diversity profiles and completeness profiles, and the visualization of species distributions and profiles.

Acknowledgments

This work was partly supported by the German Federal Ministry of Education and Research (BMBF), grant number 16DII133 (Weizenbaum-Institute). This work received funding by the Deutsche Forschungsgemeinschaft (DFG), FOR 5495, grant 496119880. The responsibility for the content of this publication remains with the authors.

References

[1] W. Van Der Aalst, Process mining, Communications of the ACM 55 (2012) 76–83.
[2] M. Kabierski, M. Richter, M. Weidlich, Addressing the log representativeness problem using species discovery, in: 2023 5th International Conference on Process Mining (ICPM), IEEE, 2023, pp. 65–72.
[3] M. Kabierski, M. Richter, M. Weidlich, Quantifying and relating the completeness and diversity of process representations using species estimation, Available at SSRN 4790484 (2024).
[4] A. Chao, N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, A. M. Ellison, Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies, Ecological monographs 84 (2014) 45–67.
[5] F. Mannhardt, et al., Sepsis cases-event log, Eindhoven university of technology 10 (2016).