Addressing the Log Representativeness Problem using Species Discovery

Martin Kabierski, Markus Richter, and Matthias Weidlich
Department of Computer Science, Humboldt-Universität zu Berlin, Germany
martin.kabierski | markus.richter | matthias.weidlich@hu-berlin.de

***Abstract*—The analysis of event logs has become a staple in the context of business process management. Insights gained from such an analysis serve to monitor and improve the business processes that generated the logs. Yet, any event log is merely a sample of the past and possible behaviour of a business process, which raises the question of log representativeness: To which extent does the log capture the characteristics of the process that are relevant for the analysis? In this paper, we propose to answer this question using estimators from biodiversity research. Interpreting log representativeness as the completeness regarding distinct properties of a process, we show how to estimate the number of properties often leveraged in process mining in some unknown population. Applying the estimators to real-world event logs, we highlight potential issues in terms of result trustworthiness, also attributing these issues to particular parts of a process.**

I. Introduction

Event logs that are recorded during the execution of business processes provide ample opportunities for process monitoring and improvement [1]. At the same time, it is widely acknowledged that, quoting the Process Mining Manifesto, *"event logs contain only sample behaviour, [so that] they should not be assumed to be complete."* [2]. That is, a business process may be thought of as a generative system, for which an event log corresponds to a *sample* of the behaviour that was shown, or could have been shown, by the system [3]. Yet, an event log may not be a trustworthy representation of a sample, but only *include* some sample behaviour of the system, due to the inherent uncertainty of event log construction [4] as well as data quality issues [5]. Moreover, for reasons of computational efficiency, some process mining techniques may be executed solely for an explicitly selected sub-log of a given event log, see [6]. Hence, the analysis is conducted for a sample of an event log, which itself denotes a possibly erroneous sample of the possible system behaviour, as visualized in Fig. 1.

Both notions of sampling induce the question of representativeness: Given a population and a sample of it, i.e., a log of a system, or a sub-log of a log, to what degree does the sample capture or preserve relevant characteristics of the population. Such insights are valuable for two reasons:

- Result confidence: The representativeness of a sample may be seen as a confidence measure for analysis results.
- Sample construction: The representativeness of a sample may guide and terminate a sampling procedure.

In either case, representativeness depends on the characteristics of the population that are deemed important and, hence, shall be preserved in the sample.

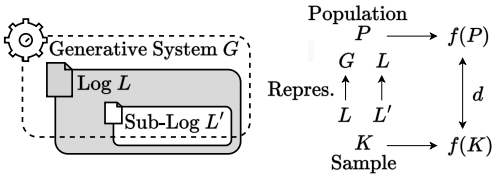


Figure 1: Representativeness problems in process mining.

Put differently, representativeness of a sample K regarding a population P can be assessed by the difference $d(f(K), f(P))$ in the results obtained with some analysis function f for the whole population and the sample, respectively. If this distance is zero, the sample perfectly represents the population regarding the characteristics that are important for the analysis function.

In general, the distance $d(f(K), f(P))$ cannot be computed, either because the population is unknown or too large to process. Hence, a common interpretation of representativeness of sample K to population P adopts set-semantics: It is based on the idea to estimate the *completeness* of the sample K to the population P in terms of the containment of the important characteristics induced by f. While other interpretations are also possible, e.g., capturing representativeness based on distributions over K and P, in the remainder, we adopt this set-based notion.

With systems, logs, and sub-logs representing behaviour, a simple instantiation of the problem would be to consider the system as the (potentially unbounded) set of all possible traces of the process, and the (sub-)logs as finite sets of traces. Then, for the case of process model discovery, the important characteristics to retain are the sets of (start, end, intermediate) activities, and the induced directly-follows relation, which corresponds to existing notions of log completeness [7]. The analysis of activities that are most often involved in conformance violations, in turn, may induce a different notion of representativeness, since traces without violations are not relevant for such an analysis.

In this paper, we formulate the question of representativeness in process mining as the problem of species discovery in biodiversity research. Behavioural characteristics of a system or a (sub-)log can be seen as species, for which the richness, i.e., the total number of species in the population, is not known and shall be estimated. While the behavioural characteristics of the generative system are never known in process mining, they may be determined for a log. Yet, even in the latter case, the estimation is valuable since the representativeness of a sub-log may be estimated *without* the need to process the entire log.

Table I: An example log with 30 traces of seven trace variants. Events are assigned a duration; their unique identifier is omitted.

	Trace	Trace variant	Count
t_1	⟨(A,1),(B,7),(C,1),(D,3),(E,3)⟩	⟨A,B,C,D,E⟩	10
t_2	⟨(A,2),(B,9),(C,2),(D,2),(E,4)⟩	⟨A,B,D,C,E⟩	8
t_3	⟨(A,1),(F,2),(F,4),(G,14),(E,7)⟩	⟨A,F,G,E⟩	5
t_4	⟨(A,1),(B,2),(D,4),(C,1),(E,3)⟩	⟨A,F,F,G,E⟩	3
t_5	⟨(A,1),(A,2),(E,9)⟩	⟨A,A,D,C,E⟩	2
	...	⟨A,A,E⟩	1
t_{30}	⟨(A,2),(B,6),(D,2),(C,1),(E,3)⟩	⟨A,A,E,A⟩	1

Our contribution in this paper is twofold. First, we formulate log representativeness in process mining, interpreted as completeness of behavioural characteristics, as a species discovery problem, which enables us to adopt species richness estimators. Second, we instantiate this problem with specific definitions of species that are motivated by common process mining tasks.

Experiments using real-life event logs indicate that for most notions of species, the logs cannot be assumed to be complete representations of the generative system, indicating that they lack critical information needed for trustworthy conclusions. We show that the proposed measures for log completeness differ for different parts of the process they capture, which enables a fine-grained differentiation of log completeness.

In the remainder, we first introduce basic terminology (§II). We then introduce the log representativeness problem (§III), explain how species richness relates to it, and instantiate estimators for process mining. Following a discussion of assumptions and limitations (§IV), we report on experiments (§V), review related work (§VI), conclude the paper (§VII).

II. Background

A. Event logs

In this paper, we rely on a relational model for event logs. Let $\mathcal{I}$ and $\mathcal{A}$ be finite sets of identifiers and activities, respectively. By $\mathcal{E}$, we denote the universe of events, where each *event* $e \in \mathcal{E}$ has a unique *identifier*, denoted by $e.id \in \mathcal{I}$, and represents the execution of an *activity*, written as $e.act \in \mathcal{A}$. Events may contain additional *data attributes* $\mathcal{D} = \{D_1, \ldots, D_p\}$ of domain $\mathrm{dom}(D_i)$, $1 \leq i \leq p$ and we write $e.D_i \in \mathrm{dom}(D_i)$ for the value of attribute D_i in event e.

A sequence of events $t = \langle e_1, \ldots, e_n \rangle \in \mathcal{E}^*$ is called a *trace* of length n. We write $t.e_i$ to refer to the i-th event of trace t. Two traces t_1 and t_2 of length n are of the same *trace variant*, if they represent an equivalent sequence of executed activities, i.e., $\forall\ 1 \leq i \leq n : t_1.e_i.act = t_2.e_i.act$. An *event log* then is a set of traces, $L \subseteq 2^{\mathcal{E}^*}$. We refer to Table I for an example event log containing 30 traces. While we omit the event identifiers, each event carries its duration as an attribute. The traces can be partitioned into seven trace variants.

B. Biodiversity Concepts

We summarize basic concepts from biodiversity research based on [8]. Let P denote a population, characterized by a sample K of observed individuals of size $k = |K|$. Each individual may belong to one or multiple species. We denote as the species richness S_P the total number of species present in P and as S_{obs} the number of species observed in a sample. It holds that $S_{obs} \leq S_P$, i.e., the number of observed species may never exceed the species richness.

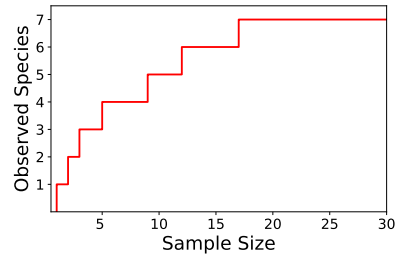


Figure 2: Collectors curve of trace variants of Table I.

Assume that we sequentially analyze each individual in sample K. Then, plotting the total number of observed species for an increasing sample size yields the collectors curve. Treating each trace variant as a species, a possible collectors curve when sampling the log from Table I is shown in Fig. 2. When increasing the sample size, such a curve will asymptotically reach S_P.

In biodiversity research, one prominent model describing the process of collecting individuals for species analysis is the Bernoulli Product Model [9]. That is, when collecting species, instead of collecting individuals one by one, sampling units are adopted that may collect multiple individuals (and thus multiple species) at once. Furthermore, only the presence or absence of a species (incidence), and not the count of species in a unit (abundance) is recorded.

Now, let S_{obs} be the number of species observed in a set of k sampling units. We define W_{ij} with $i \in \{1, 2, \ldots, S_P\}$ and $j \in \{1, 2, \ldots, k\}$ as the incidence matrix, i.e., $w_{ij} = 1$, if sampling unit j contains species i, and $w_{ij} = 0$ otherwise. We denote as $Y_i = \Sigma_{j=0}^{n} w_{ij}$ the number of sampling units belonging to species i. Then, a species i is undiscovered, if $Y_i = 0$. We denote as $Q_k = \Sigma_{i=0}^{S_p} I(Y_i = k)$ the incidence frequency count of k, i.e. the number of species that have k incidences. Then, Q_0 is the undiscovered species count, Q_1 is the singleton species count, and Q_2 is the doubleton species count. Let the probability of incidence of a species i be denoted as p_i, be independent of other species' incidences, and the same for all sampling units. Then, each w_{ij} in W_{ij} is a Bernoulli random variable with $P(w_{ij}) = p_i$ and the number of sampling units belonging to species i follows a binomial distribution with $P(Y_i = y_i) = \binom{k}{y_i} p_i^{y_i} \cdot (1 - p_i)^{n - y_i}$.

C. Estimating Species Richness

In practice, S_P is unknown and needs to be estimated based on information gathered through a sample. For the Bernoulli Product Model, this estimation task is phrased as: Given the observed species count S_{obs} and the corresponding incidence matrix W_{ij}, estimate $S_P \approx S_{est} = S_{obs} + Q_0$. Good and Turing [10] showed that occurrence frequency q_0 for the next individual to belong to a previous unseen species is expected to be close to the probability of seeing a species exactly once.

Based thereon, the Chao2 estimator [11] yields an estimate S_{est} of species richness S_P in the Bernoulli Product Model:

$$\hat{S}_{Chao2} \approx \begin{cases} S_{obs} + Q_1^2 / (2Q_2) & \text{if } Q_2 > 0 \\ S_{obs} + Q_1 (Q_1 - 1) / 2 & \text{if } Q_2 = 0 \end{cases} \quad (1)$$

The estimator yields a lower bound on the total number of species in the population, and is an unbiased point estimator for cases, where singleton and undetected species have the same occurrence frequency [12]. Yet, it is based only on the singletons and doubletons encountered during sampling. The estimator evaluates to 0, if no singletons are recorded in the sample. This is a valid assumption, given that the model assumes an unbounded number of samples that belong to a set of S_{est} species. As such, each species is expected to be eventually recorded at least two times, which is then considered to be a valid stopping criterion for sampling with the goal of species completeness [11].

D. Estimating Sample Completeness and Coverage

While Eq. 1 gives an estimate for $S_P \approx S_{est}$, it does not quantify how representative the seen species and incidence counts are in terms of (i) *completeness*, i.e., how many of all species have been discovered; and (ii) *coverage*, i.e., how much of the probability space, is covered by the discovered species.

The estimated sample completeness is defined as the ratio of the observed species and the estimated total number of species:

$$\hat{Com}_{obs} = \frac{S_{obs}}{S_{est}}. \quad (2)$$

As S_{est} is a lower bound on the absolute number of species in the population, $\hat{Com}_{obs}$ serves as an upper bound of the true *completeness*. In the Bernoulli Product Model, the sample coverage denotes the probability mass covered by the incidence probabilities of all species observed in the current sample. It is estimated as [14]:

$$\hat{Cov}_{obs} \approx 1 - \frac{Q_1}{\Sigma_{i=1}^{S_P} Y_i} \left(\frac{(n-1) Q_1}{(n-1) Q_1 + 2Q_2} \right) \quad (3)$$

Let $\hat{Com}_{obs} \leq g \leq 1$ be the desired species completeness. Chao et al. [15] showed that the required number of additional samples l_g to reach completeness g is estimated as:

$$l_g \approx \frac{\log \left(1 - \frac{n}{n-1} \frac{2Q_2}{Q_1^2} (g S_{est} - S_{obs}) \right)}{\log \left(1 - \frac{2Q_2}{(t-1)Q_1 + 2Q_2} \right)} \quad (4)$$

III. The Log Representativeness Problem

As illustrated in Fig. 1, any event log L denotes a sample of the behaviour of a generative system G, i.e., the business process. In practice, a log L may also include behaviour not present in the population, i.e., the generative system G. Below, we first neglect this aspect and return to it again, when discussing assumptions and limitations. Moreover, the log L may be further sampled, which yields a sub-log L' that is used for analysis. For either notion of a sample, the question of representativeness emerges.

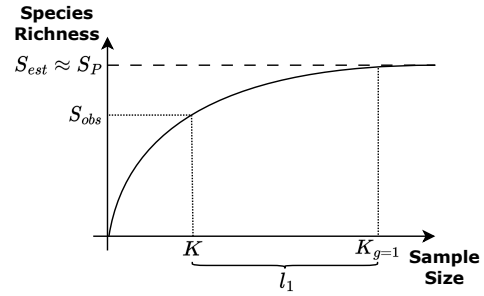


Figure 3: A schematic view on a collectors curve.

Since the behaviour of the generative system G is only known through L, the representativeness of L cannot be quantified directly. Yet, even when considering a sub-log L', an estimation of its representativeness is useful, as it avoids the need to compare L' against the entire log L as part of a sampling procedure.

One may approach the question of log representativeness with guarantees for aggregates over samples, i.e., the law of large numbers and the central limit theorem. Both notions provide guarantees on the accuracy of an aggregation function applied to a sample of increasing size, if sampling is unbiased. Yet, the given guarantees relate *solely* to the aggregate, not making any statement on the existence of individual elements in the sample. Hence, sample statistics are insufficient to uniquely describe a dataset, as illustrated by the well-known Anscombe's quartet [17] or the datasaurus dozen [18].

In process mining, common analysis tasks require information on the presence of discrete behavioural characteristics. For instance, to discover a process model or check conformance between a log and a model, information on the presence of executions of specific activities and their ordering is required. As such, statistical guarantees on the accuracy of an aggregate computed over a sample are insufficient. Rather, we strive for a notion of representativeness that is based on the *completeness* and *coverage* of an event log, or a sub-log, regarding such behavioural characteristics, which is introduced next.

A. Modelling Log Analysis as Species Discovery

To assess log representativeness, we adopt the viewpoint of species discovery. Sampling from a generative system or from an event log will select certain characteristics, i.e., species. Then, the completeness and coverage of a log or a sub-log is traced back to species richness: We compare the observed number of characteristics against their estimated total number in the population (Eq. 2) and estimate the sample coverage based on the presence of singletons and doubletons (Eq. 3).

Following the above idea, the definition of the species corresponds to the behavioural characteristics that are important for some analysis. Recall that $\mathcal{E}^*$ denotes the universe of all possible traces and let Ω be the domain of some behavioural characteristics. Then, we capture possible species definitions by a *species retrieval function* $\zeta : 2^{\mathcal{E}^*} \rightarrow 2^{\Omega}$ that assigns a set of species to each possible trace.

Before turning to possible instantiations for a species retrieval function in process mining, we illustrate its application in

the evaluation of log representativeness. Assume that we sequentially evaluate each trace of a sample K, i.e., a log or sub-log. We then retrieve the set of species this trace belongs to, and update the observed species count S_{obs}. This procedure yields a collectors curve, for which Fig. 3 presents a generalized version. As we see many new species in the beginning, S_{obs} increases fast, while slowing down the longer we keep analyzing traces. Eventually we have analyzed all traces of the sample, which is denoted by the first dotted vertical line in Fig. 3. At this point, we obtained a first final number of observed species, S_{obs}. It enables us to use the Chao2 estimator $\hat{S}_{Chao2}$ (Eq. 1) to estimate the total number of species expected in the population as S_{est}. Based thereon, we can derive the following conclusions:

- Completeness: If $S_{obs} = S_{est}$, we consider the sample K (corresponding to a log L or sublog L') to be complete with respect to the population P (corresponding to the system G or log L, respectively) and the species retrieval function ζ. Otherwise, if $S_{obs} < S_{est}$, we quantify the completeness by the ratio of observed and total estimated species (Eq. 2).
- Coverage: If K is incomplete, we compute the total incidence probabilities, thereby shedding light on the importance in terms of the probability mass of the supposedly missing species (Eq. 3 with $n = k = |K|$).
- Sample extension: If K is incomplete, and the coverage suggests that important species are missing, we estimate the number of additional traces that need to be evaluated to reach a specific species completeness g (Eq. 4), illustrated for $g = 1$, i.e. total completeness, and the second dotted vertical line in Fig. 3, for which l_1 more traces are needed.

B. Defining Species Retrieval Functions

As detailed above, the evaluation of log representativeness in terms of *completeness* and *coverage* relies on the definition of a species retrieval function ζ that maps a set of species to a trace. Therefore a trace resembles one sampling unit, containing multiple species. It thereby determines on what basis the representativeness of a log shall be assessed. Inspired by common process mining tasks, we define the following instantiations of this function and always illustrate them with the example trace $t_3 = \langle$(A,1),(F,2),(F,4),(G,14),(E,7)$\rangle$ taken from Table I.

ζ_{act}: The set of species may be defined as the set of activities of the trace: $\zeta_{act}(\langle e_1, \ldots, e_n \rangle) = \bigcup_{1 \leq i \leq n} \{e_i.act\}$. For instance, we obtain $\zeta_{act}(t_3) = \{A, E, F, G\}$.

ζ_{tv}: The trace variant of a trace may be considered as a species: $\zeta_{tv}(\langle e_1, \ldots, e_n \rangle) = \{\langle e_1.act, \ldots, e_n.act \rangle\}$. For our example, this yields $\zeta_{tv}(t_3) = \{\langle A, F, F, G, E \rangle\}$.

ζ_{df}: Many process mining algorithms exploit ordering relations defined over the events or activities, respectively. That is, the directly-follows relation over activities may induce species: $\zeta_{df}(\langle e_1, \ldots, e_n \rangle) = \bigcup_{1 \leq i < n} \{(e_i.act, e_{i+1}.act)\}$. For a trace of our example, this yields $\zeta_{df}(t_3) = \{(A, F), (F, F), (F, G), (G, E)\}$.

The definition of species is not limited to the control-flow perspective, but may also incorporate data assigned to events. To illustrate this flexibility in the definition of log representativeness, we consider two species retrieval functions that are based on durations of events (i.e., activity executions).

$\zeta_{t\lambda}$: To facilitate process mining tasks that incorporate temporal information, the species definition may include event durations. To cater for values of attributes with fine-granular or continuous domains, some aggregation into value bins may be adopted, though. Let $\lambda \in \mathbb{N}$ be a parameter for the granularity with which durations shall be considered. Then, assuming that the attribute $d \in \mathcal{D}$ denotes the duration of an event, a species definition may contain pairs of activities and aggregated durations: $\zeta_{t\lambda}(\langle e_1, \ldots, e_n \rangle) = \bigcup_{1 \leq i \leq n} \{(e_i.act, \lambda \cdot \lceil e_i.d/\lambda \rceil)\}$. As an example, for one-minute intervals, we obtain $\zeta_{t1}(t_3) = \{(A, 1), (F, 2), (F, 4), (G, 14), (E, 7)\}$, while a more coarse-granular aggregation into five-minute intervals would yield $\zeta_{t5}(t_3) = \{(A, 5), (F, 5), (G, 15), (E, 10)\}$. In the latter case, both events related to activity F are considered to be of the same species.

$\zeta_{te\lambda}$: As an alternative approach to consider durations in the definition of species, we also exemplify an instantiation that is based on an exponentially scaled aggregation of the respective values. It caters for the long-tail distributions often observed for activity durations and, using the above notations, is defined as: $\zeta_{te\lambda}(\langle e_1, \ldots, e_n \rangle) = \bigcup_{1 \leq i \leq n} \{(e_i.act, \lambda \cdot \lceil \log_\lambda(e_i.d) \rceil)\}$. For our example trace, the species would be derived as $\zeta_{te2}(t_3) = \{(A, 2), (F, 2), (F, 4), (G, 16), (E, 8)\}$.

IV. Discussion

Next, we review the assumptions imposed by the proposed estimation of representativeness.

Logs are noise- and error-free: An event log (and hence, a sub-log) may contain noisy and erroneous data. A common solution to remedy this issue is to filter out infrequent behaviour [19], assuming that noise can be characterized based on occurrence frequencies. However, removing rare events or traces before assessing the representativeness of a sample, would effectively remove species from the population. It would reduce S_P and, therefore, increase Com_{obs} and Cov_{obs}, i.e., leads to higher estimates for completeness and coverage. To counter this effect, noise filtering that is not entirely based on occurrence frequencies should be adopted.

Independence of species: Under the Bernoulli Product Model, the incidence probabilities of species are assumed to be independently distributed. Yet, once the generative system is exposed to concept drift, previously unobserved species influence the estimated species richness. Similarly, activity executions, and thus species definitions, may be correlated in a process. However, both phenomena are not problematic per se if the representativeness of the complete log is to be assessed, since it is based on the stationary incidence probabilities of the complete log. To avoid premature termination of a sampling procedure when using representativeness as a stopping criterion, however, trace selection shall be randomized and the independence assumption needs to be verified.

Assumptions of the richness estimation: The Chao2 estimator assumes that singleton and doubleton species do not account for the majority of species in the population. For a population with many very rare species, richness estimates may become prohibitively inaccurate. In that case, estimators that are tailored for heavily-skewed distributions could be employed. While these come with other limitations, we opted for the Chao2 estimator, as it is applicable for the Bernoulli Product Model and is parameter-free.

Species richness as the sole criterion: Species richness, i.e. the total number of species in the population, and the respective completeness measures, do not give an all-encompassing view of the population. Rather, they may be combined with information on species abundance, i.e. the total number *per* species in the population. We consider this idea to be a promising direction for future work.

V. Experimental Evaluation

We conducted experiments using four publicly available event logs to answer the following questions:

- How do species retrieval functions influence the representativeness of the given logs in terms of completeness, coverage, and the effort to improve on completeness?
- How do the notions of completeness and coverage behave when sampling sub-logs.
- Can representativeness be attributed to a certain part of the process, thereby pointing to well-represented parts?

Below, we first describe our setup (§V-A), before turning to each of the above questions (§V-B – §V-D).

A. Experimental Setup

We implemented the proposed estimation and extrapolation methods in Python based on the pm4py-framework [20]. The scripts, as well as the evaluation results, are publicly available.[1]

The four logs used for the experiments are:

- *BPI-2012* [21], a log of a loan application process of a Dutch financial institute (n=13087, 4336 trace variants).
- *BPI-2018* [22], a log of a process for subsidies in agriculture (n=43809, 28489 trace variants).
- *BPI-2019* [23], a log of a purchase order handling process (n=251734, 11973 trace variants).
- *Sepsis Cases* [24], a log of patient pathways in the emergency department (n=1050, 846 trace variants).

These logs have been chosen as they differ significantly in size and complexity of the underlying processes.

As species retrieval functions, we considered all functions introduced in §III-B, i.e., species based on activities (ζ_{act}), trace variants (ζ_{tv}), directly-follows relations (ζ_{df}), uniform durations ($\zeta_{t1}, \zeta_{t5}, \zeta_{t30}$) and exponential durations (ζ_{te2}).

For each log and species retrieval function, we exhaustively selected random traces until the complete log was drawn. We repeated this procedure 200 times per log-function combination, to simulate the sample-based nature of the estimation process, and recorded the following statistics, after each five traces:

- S_{obs}, the number of observed species.
- Q_1, the number of singleton species.
- Q_2, the number of doubleton species.
- S_{est}, the estimated number of species (see Eq. 1).
- Com_{est}, the estimated sample completeness (see Eq. 2).
- Cov_{est}, the estimated sample coverage (see Eq. 3).

For completely sampled logs, we quantified l_g (see Eq. 4), the additional traces that need to be sampled, to achieve completeness of $g \in \{0.8, 0.9, 0.95, 0.99\}$.

B. Representativeness of Complete Logs

Table II shows the values obtained for our seven species retrieval functions and four event logs. Here, for all logs, for all but the simplest species definition ζ_{act}, the logs cannot considered to be complete (Com_{obs}). However, the coverage values Cov_{obs} indicate that the observed species make up a significant part of the population.

Moreover, there are large differences in completeness for different species retrieval functions. For ζ_{df}, completeness is estimated to range from 0.828 to 0.925, with *Sepsis Cases* showing the lowest and *BPI-2012* showing the largest value. We note that larger sample sizes do not necessarily imply higher species completeness. For instance, consider the control-flow-centric species definitions for *BPI-2012* and *BPI-2018*. Completeness of *BPI-2012* is higher, despite smaller sample sizes. This can be attributed to *BPI-2012* showing less variation in the values, so that smaller samples are sufficient to represent a larger fraction of expected species.

Turning to species retrieval functions with uniform durations, coverage and completeness increase when increasing time intervals. This is expected, since larger intervals decrease the number of species.

The effort expected to improve completeness turns out to be prohibitively large in some cases. For *BPI-2018* and ζ_{df}, to reach 99% completeness from the estimated completeness of 85%, further 138955 traces would be required, i.e. the log would need to be 4.17 times larger as available. However, since the coverage is almost perfect, we can expect noise in the data to have artificially increased the estimated species richness.

Regarding our first evaluation question, we conclude that no available log can be considered as complete for almost all notions of species. However, the coverage values show that the missing species often only account for a small part of the population. Also, we observe that coverage and completeness effectively relate a log to the variability of the generative system, thereby providing an assessment of representativeness that is more suitable than the sample size alone.

C. Representativeness of Sub-Logs

Next, we turn to the assessment of completeness and coverage when sampling a sub-log. Due to space constraints, we focus on the results for ζ_{act} and ζ_{df} for *Sepsis Cases*, while all results are available online in the mentioned repository. Fig. 4 shows the mean completeness, mean coverage, and the mean collectors curves, all with their 95% confidence intervals over 200 experiment runs. In addition, Fig. 4 includes the values of

[1] https://scm.cms.hu-berlin.de/richtmqf/gt-sampling

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

Log	Species Def.	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.95}$	$l_{.90}$	$l_{.80}$
BPI-2012	ζ_{act}	24	24	0	0	1.0	1.0	-	-	-	-
	ζ_{df}	149	161	7	2	0.999	0.925	46435	9577	-	-
	ζ_{tv}	4336	30346	3727	267	0.715	0.143	406521	259527	196219	132912
	ζ_{t1}	958	2816	535	77	0.996	0.340	190458	117290	85779	54267
	ζ_{t5}	487	1164	268	53	0.998	0.418	134446	81196	58263	35329
	ζ_{t30}	210	288	74	53	0.999	0.729	45666	23401	13812	4223
	ζ_{te2}	112	112	2	5	0.999	1.0	-	-	-	-
BPI-2018	ζ_{act}	41	41	0	0	1.0	1.0	-	-	-	-
	ζ_{df}	619	721	86	36	0.999	0.856	138955	54739	18468	-
	ζ_{tv}	28489	409861	26634	930	0.302	0.070	2843643	1834038	1399225	964412
	ζ_{t1}	177724	380855	97438	23369	0.942	0.467	363183	216193	152888	89583
	ζ_{t5}	82965	151461	40593	12028	0.975	0.548	281769	162793	111553	60313
	ζ_{t30}	31741	49856	12824	4539	0.992	0.637	222341	122740	79844	36948
	ζ_{te2}	787	880	82	36	0.999	0.894	117825	37525	2942	-
BPI-2019	ζ_{act}	42	42	1	0	1.0	1.0	-	-	-	-
	ζ_{df}	538	614	75	37	0.999	0.876	641935	231311	54465	-
	ζ_{tv}	11973	51137	9030	1041	0.964	0.234	4736737	2979535	2222749	1465964
	ζ_{t1}	200937	346431	95716	31484	0.931	0.580	1430214	814357	549122	283887
	ζ_{t5}	89981	128326	30490	12122	0.978	0.701	1075518	565989	346548	127106
	ζ_{t30}	32084	45235	9731	3600	0.991	0.709	1146500	598929	363103	127278
	ζ_{te2}	576	602	39	29	0.999	0.957	249031	-	-	-
Sepsis	ζ_{act}	16	16	0	0	1.0	1.0	-	-	-	-
	ζ_{df}	135	163	17	5	0.999	0.828	5117	2246	1010	-
	ζ_{tv}	846	9618	784	35	0.252	0.088	53025	34116	25972	17828
	ζ_{t1}	3326	11190	2462	385	0.806	0.297	14265	8866	6541	4215
	ζ_{t5}	2229	5490	1450	322	0.886	0.406	9648	5846	4209	2571
	ζ_{t30}	1181	2607	648	147	0.949	0.453	9255	5533	3930	2327
	ζ_{te2}	202	228	24	11	0.998	0.886	2792	950	156	-

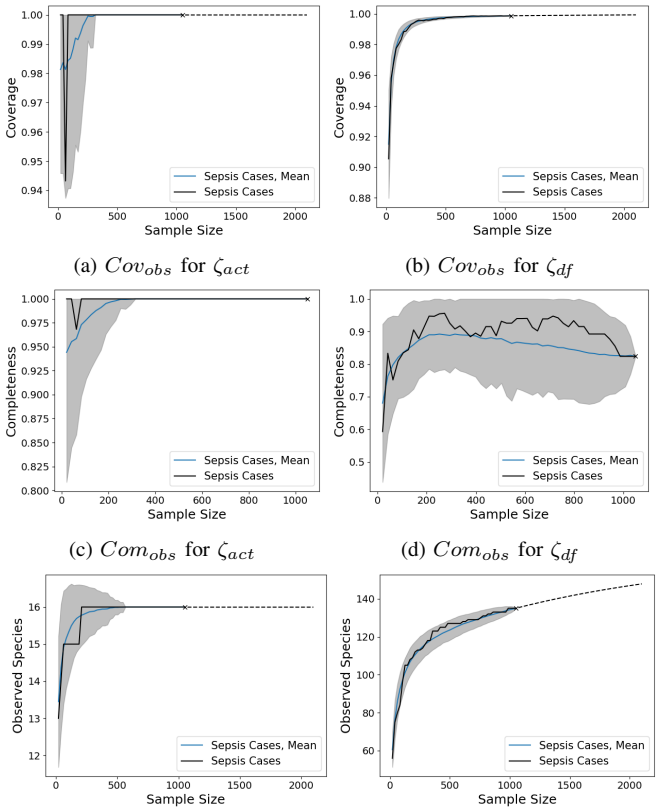
a sequential evaluation of log without any randomization (black line), and the species richness is extrapolated up to double the log size using the approach presented in [14].

Independent of the species retrieval function, coverage converges to a stable value rather quickly. Given that the most common species are expected to be drawn rather early and often, they will make up large portions of the probability mass. This, in turn, enables a fast assessment of the remaining species. The trends are more mixed for completeness. For ζ_{act}, completeness converges quickly, whereas for ζ_{df}, completeness is never reached and subject to large fluctuations. The respective collectors curve (Fig. 4f) shows that up until the processing of the complete log, new species are discovered steadily. In fact, S_{est} increases steadily, indicating that the true number of species is even higher than the final estimate. Since Com_{obs} provides an upper bound for the actual completeness, the actual value can be assumed to be even lower than the final estimate of 0.828. This is in line with *Sepsis Cases* being a rather unstructured log, so that new behaviour could indeed be expected when increasing sample size.

Note that, while completeness cannot be assumed, the unobserved species make up only roughly 1% of the total volume in the population most of the time. Hence, these species denote the long tail of the underlying distribution and could be categorized as noise.

D. Representativeness for Process Parts

Finally, we consider the analysis of species richness for different parts of a process. In practice, certain parts of a process may show more variability than others. Hence, a log may not be able to reliably represent the diversity of some part, while it is a reasonable representation for other parts.



(a) Cov_{obs} for ζ_{act} (b) Cov_{obs} for ζ_{df}

(c) Com_{obs} for ζ_{act} (d) Com_{obs} for ζ_{df}

(e) S_{obs} for ζ_{act} (f) S_{obs} for ζ_{df}

Figure 4: Mean sample coverage, completeness and collectors curves ζ_{act} & ζ_{df} on *Sepsis Cases* obtained on 200 repetitions.

Table III: Species richness estimation, coverage, and extrapolation for two phases of *Sepsis Cases*, pre- and post-admission.

Spec.	Log	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.90}$
ζ_{act}	Pre	11	11	1	0	1.0	1.0	-	-
	Post	15	15	2	0	1.0	1.0	-	-
ζ_{df}	Pre	87	118	8	1	0.999	0.737	13811	4148
	Post	88	113	19	7	0.997	0.778	3426	897
ζ_{tv}	Pre	298	735	180	37	0.828	0.405	10427	4550
	Post	467	2955	393	31	0.514	0.158	22735	10926
ζ_{t1}	Pre	1041	1673	513	208	0.939	0.622	4699	1720
	Post	2392	13238	2058	195	0.646	0.181	18811	8980
ζ_{t30}	Pre	127	191	41	13	0.995	0.665	5821	2010
	Post	1120	2566	639	141	0.889	0.436	7392	3170
ζ_{te2}	Pre	132	144	14	8	0.998	0.917	1963	-
	Post	127	159	29	13	0.994	0.799	2717	638

Table IV: Species richness estimation, coverage, and extrapolation for *Sepsis Cases*, split by the patients' age (< 60, ≥ 60).

Spec.	Log	S_{obs}	S_{est}	Q_1	Q_2	Cov_{obs}	Com_{obs}	$l_{.99}$	$l_{.90}$
ζ_{act}	< 60	15	16	2	1	0.998	0.938	547	35
	≥ 60	16	16	0	0	1.0	1.0	-	-
ζ_{df}	< 60	101	113	15	9	0.994	0.894	444	17
	≥ 60	133	158	16	5	0.998	0.842	3675	631
ζ_{tv}	< 60	171	1946	158	7	0.287	0.088	11309	5539
	≥ 60	707	8185	659	29	0.202	0.086	42374	20763
ζ_{t1}	< 60	907	3574	706	93	0.704	0.254	3636	1694
	≥ 60	2858	10127	2138	314	0.793	0.282	12019	5543
ζ_{t30}	< 60	420	906	254	66	0.893	0.464	1703	718
	≥ 60	1081	2316	599	145	0.942	0.467	6786	2857
ζ_{te2}	< 60	149	171	25	14	0.989	0.871	509	51
	≥ 60	200	225	26	13	0.997	0.889	2018	114

For this experiment, we leverage the *Sepsis Cases* log, which captures patient pathways in an emergency department, from arrival and initial examinations until an admission to care followed by further activities. Since treatments and examinations differ for patients that have been admitted and for those who just arrived and wait for admission, we split each trace based on the activity that signifies the admission (*Admission ot IC/NC*). This results in one log of 1050 traces of all patients until admission (*Pre*), and a second log of size 810 of all admitted patients (*Post*). For both parts of the process, Table III presents the results on log representativeness. Again, completeness and coverage is reached only for the simplest retrieval function ζ_{act}. However, we note clear differences in the results for both process parts. While both logs have similar species completeness Com_{obs} regarding their directly-follows relations (0.737 and 0.778), the *Post*-log needs 240 fewer traces to achieve it. Hence, the additional sample sizes expected to improve completeness differ as well.

For all other species retrieval functions, the species contained in the *Pre*-log are generally more representative, especially for the duration-related functions. Here, completeness values range from 0.622 and 0.181 for ζ_{t1} up to 0.917 and 0.799 for ζ_{te2}, for either log, respectively. Thus, the 1080 traces recording the pre-admission part better represent durations, than the 840 traces do for the post-admission part. The reason being not only the sample size, but also the fact that admitted patients usually stay for significantly longer periods of time, thus effectively increasing the number of species.

Furthermore, we repeated the experiment splitting the *Sepsis Cases* log based on the patients' age, separating those older than 60 years (827 traces) from the younger ones (223 traces). Table IV shows that, despite the size differences, both Com_{obs} and Cov_{obs} yield similar values for all species retrieval functions. This indicates that for both parts, the sub-logs can be assumed to be of similar representativeness. On the flipside, we conclude that for patients of high age, the estimated number of species, and thus the variability in the recorded behaviour, was noticeably higher than for younger patients. Yet, the larger sample size covered for this effect. Again, this emphasizes that sample size alone is not a suitable indicator for representativeness.

We conclude that our approach can reveal how well a log represents different parts of a process. The proposed measures therefore help to relate the variability in different phases of a process to the data collected in a sampled log.

VI. Related Work

The term representativeness has been assigned different meanings, for instance in qualitative research [25] and general statistics [26]. The question on the size of a sample to reliably observe an effect, arguably a criterion for representativeness, was also studied extensively, see [27].

In process mining, sample-based analysis has attracted attention recently. Incremental sampling methods have been proposed for process discovery and conformance checking, which stop sampling once the estimated probability of discovering new elements falls below a threshold [28] or the aggregated values of interest for analysis are estimated to have converged [29], [6]. Other work focuses on the selection of traces from a log to minimize the distance to the overall log [30] or to maximize the utility of the sample for a certain analysis question [31]. The accuracy of sampling strategies was explored for process discovery in [32] and by measures for under- and over-sampled behaviour in [33]. While these measures compare a sampled sub-log and a log, they do not aim at inferring insights on the generative system.

Turning to the relation between a log and a generative system, it was proposed to quantify the completeness probability of a log for workflow nets [34] and based on an approximation of the ratio of seen and all direct successorship relations [35]. In [36], log completeness was also framed as a species discovery problem that is addressed with several richness estimators. Despite the conceptual similarity with our work, the approach was designed specifically for process discovery and adopted a multinomial model with each trace being of exactly one species. In contrast, we provide a generic formulation of log representativeness that enables instantiations with various species definitions, an assessment of parts of processes, and may guide the construction of a sample with a certain completeness. Finally, properties of event logs and have been linked to quality guarantees of process discovery algorithms in [37].

Further estimators for species richness have been proposed in biodiversity research, e.g., abundance-based coverage [38] and incidence-based coverage [39]. However, it is well-established that species richness captures only aspects of a population [8]. Multiple views on population diversity may be generalized using Hill numbers [14], which provides a promising direction for future research on log representativeness.

VII. Conclusion

In this paper, we addressed the question of log representativeness in process mining. Specifically, we showed how log representativeness based on the presence of behavioural characteristics can be viewed as a species discovery problem, and be evaluated with completeness and coverage measures. Those are based on an estimator of species richness, which we instantiated for several definitions of species for event logs. Our experiments on four real-world event logs illustrated that the logs cannot be assumed to be complete regarding behavioural characteristics commonly used in process mining tasks.

In future work, we intend to complement incidence species counts and incorporate species abundance into the notion of log representativeness. The handling of noise in event logs is another direction for further research. Lastly, we want to further investigate the capabilities for process part analysis.

Acknowledgements. This work received funding by the Deutsche Forschungsgemeinschaft (DFG), grant 421921612.

References

[1] W. M. P. van der Aalst, *Process Mining - Data Science in Action, Second Edition*. Springer, 2016.

[2] W. Van Der Aalst, A. Adriansyah, A. K. A. De Medeiros, F. Arcieri, T. Baier, T. Blickle, J. C. Bose, P. Van Den Brand, R. Brandtjen, J. Buijs *et al.*, "Process mining manifesto," in *BPM 2011 Workshops.*. Springer, 2012, pp. 169–194.

[3] J. C. A. M. Buijs, B. F. van Dongen, and W. M. P. van der Aalst, "Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity," *Int. J. Cooperative Inf. Syst.*, vol. 23, no. 1, 2014.

[4] K. Diba, K. Batoulis, M. Weidlich, and M. Weske, "Extraction, correlation, and abstraction of event data for process mining," *WIREs Data Mining Knowl. Discov.*, vol. 10, no. 3, 2020.

[5] S. Suriadi, R. Andrews, A. H. M. ter Hofstede, and M. T. Wynn, "Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs," *Inf. Syst.*, vol. 64, pp. 132–150, 2017.

[6] M. Bauer, H. van der Aa, and M. Weidlich, "Sampling and approximation techniques for efficient process conformance checking," *Inf. Syst.*, vol. 104, p. 101666, 2022.

[7] S. J. Leemans, D. Fahland, and W. M. Van Der Aalst, "Discovering block-structured process models from event logs containing infrequent behaviour," in *BPM 2013 Workshops*. Springer, 2014, pp. 66–78.

[8] R. K. Colwell *et al.*, "Biodiversity: concepts, patterns, and measurement," *The Princeton guide to ecology*, vol. 663, pp. 257–263, 2009.

[9] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino, "Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages," *Journal of plant ecology*, vol. 5, no. 1, pp. 3–21, 2012.

[10] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3-4, pp. 237–264, 1953.

[11] A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of statistics*, pp. 265–270, 1984.

[12] A. Chao and R. K. Colwell, "Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling," *Statistics and Operations Research Transactions*, pp. 3–54, 2017.

[13] A. Chao, C.-H. Chiu, R. K. Colwell, L. F. S. Magnago, R. L. Chazdon, and N. J. Gotelli, "Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on good-turing theory," *Ecology*, vol. 98, no. 11, pp. 2914–2929, 2017.

[14] A. Chao, N. J. Gotelli, T. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison, "Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies," *Ecological monographs*, vol. 84, no. 1, pp. 45–67, 2014.

[15] A. Chao, R. K. Colwell, C.-W. Lin, and N. J. Gotelli, "Sufficient sampling for asymptotic minimum species richness estimators," *Ecology*, vol. 90, no. 4, pp. 1125–1133, 2009.

[16] S. L. Rasmussen and N. Starr, "Optimal and adaptive stopping in the search for new species," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 661–667, 1979.

[17] F. J. Anscombe, "Graphs in statistical analysis," *The american statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[18] J. Matejka and G. Fitzmaurice, "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing," in *CHI 2017*, 2017, pp. 1290–1294.

[19] S. J. van Zelst, M. Fani Sani, A. Ostovar, R. Conforti, and M. La Rosa, "Filtering spurious events from event streams of business processes," in *CAiSE 2018*. Springer, 2018, pp. 35–52.

[20] A. Berti, S. J. Van Zelst, and W. van der Aalst, "Process mining for python (pm4py): bridging the gap between process-and data science," *arXiv preprint arXiv:1905.06169*, 2019.

[21] B. van Dongen, "Bpi challenge 2012," 2012. https://data.4tu.nl/articles/dataset/BPI_Challenge_2012/12689204/1

[22] B. van Dongen and F. F. Borchert, "Bpi challenge 2018," 2018. https://data.4tu.nl/articles/dataset/BPI_Challenge_2018/12688355/1

[23] B. van Dongen, "Bpi challenge 2019," 2019. https://data.4tu.nl/articles/dataset/BPI_Challenge_2019/12715853/1

[24] F. Mannhardt *et al.*, "Sepsis cases-event log," *Eindhoven University of Technology*, vol. 10, 2016.

[25] G. Gobo, "Sampling, representativeness and generalizability," *Qualitative research practice*, vol. 405, p. 426, 2004.

[26] W. Kruskal and F. Mosteller, "Representative sampling, i: Non-scientific literature," *International Statistical Review*, pp. 13–24, 1979.

[27] G. D. Israel, "Determining sample size," 1992.

[28] M. Bauer, A. Senderovich, A. Gal, L. Grunske, and M. Weidlich, "How much event data is enough? a statistical framework for process discovery," in *CAiSE 2018*. Springer, 2018, pp. 239–256.

[29] M. Bauer, H. Van der Aa, and M. Weidlich, "Estimating process conformance by trace sampling and result approximation," in *BPM 2019*. Springer, 2019, pp. 179–197.

[30] G. Bernard and P. Andritsos, "Selecting representative sample traces from large event logs," in *ICPM 2021*. IEEE, 2021, pp. 56–63.

[31] M. Kabierski, H. L. Nguyen, L. Grunske, and M. Weidlich, "Sampling what matters: relevance-guided sampling of event logs," in *ICPM 2021*. IEEE, 2021, pp. 64–71.

[32] M. Fani Sani, S. J. van Zelst, and W. M. van der Aalst, "The impact of event log subset selection on the performance of process discovery algorithms," in *ADBIS 2019 Workshops*. Springer, 2019, pp. 391–404.

[33] B. Knols and J. M. E. van der Werf, "Measuring the behavioral quality of log sampling," in *ICPM 2019*. IEEE, 2019, pp. 97–104.

[34] K. M. van Hee, Z. Liu, and N. Sidorova, "Is my event log complete?—a probabilistic approach to process mining," in *RCIS 2011*. IEEE, 2011, pp. 1–12.

[35] H. Yang, L. Wen, J. Wang, and R. K. Wong, "Cpl+: An improved approach for evaluating the local completeness of event logs," *Information Processing Letters*, vol. 114, no. 11, pp. 607–610, 2014.

[36] J. Pei, L. Wen, H. Yang, J. Wang, and X. Ye, "Estimating global completeness of event logs: A comparative study," *IEEE Transactions on Services Computing*, vol. 14, no. 2, pp. 441–457, 2018.

[37] J. M. E. van der Werf, A. Polyvyanyy, B. R. van Wensveen, M. Brinkhuis, and H. A. Reijers, "All that glitters is not gold: Towards process discovery techniques with guarantees," in *CAiSE 2021*. Springer, 2021, pp. 141–157.

[38] A. Chao and S.-M. Lee, "Estimating the number of classes via sample coverage," *Journal of the American statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[39] R. L. Chazdon, R. K. Colwell, J. S. Denslow, and M. R. Guariguata, "Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern costa rica," 1998.