UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

A Comparative Study of Estimator Accuracy in Event Log Analysis

Prashant Bahuguna

(Matriculation Number: 985948)

Computer Science Department, Informatica

Master's Degree in Computer Science

Supervisor: Professor Paolo Ceravolo

Co-Supervisor: Professor Gabriele Gianini

October 2024 (AY 2023/24)

Contents

Chapter 1

Introduction

1.1 Background and Motivation

Event logs are essential in the modern landscape of business process management, as they capture detailed records of operational events, which include activity timestamps, sequences, and associated resources. As businesses and systems grow more complex, the ability to mine, analyze, and troubleshoot these logs becomes increasingly critical. Event logs not only enable process discovery and conformance checking but also aid in detecting inefficiencies, anomalies, and optimization opportunities within processes.

In the realm of process mining, an ongoing challenge is the Log Representativeness Problem—the extent to which an event log sample represents the complete underlying process. Event logs are often incomplete due to limitations in data collection, the inherent variability of processes, and system constraints. Hence, ensuring the representativeness of a sample is critical for deriving valid insights. Without representativeness, any conclusions drawn from the event log might lead to incorrect decisions, making it crucial to estimate the completeness of event logs.

Building upon research such as Kabierski's work on species discovery to estimate log completeness, which draws parallels between biodiversity estimation and event log analysis, this thesis aims to extend the methodology by testing additional estimators and introducing more robust evaluation metrics. In biological systems, estimating the number of species in a habitat using incomplete observations (e.g., sightings of individual species) mirrors the process of estimating unseen behaviors in a business process from limited event logs.

Previous research has primarily focused on incidence-based estimators such as Chao2. While effective under certain conditions, these estimators tend to suffer from limitations, particularly in handling skewed data or logs with a high prevalence of rare behaviors. To address these challenges, this thesis explores the applicability of both non-parametric and parametric estimators for event log analysis, with a focus on accuracy, reliability, and the introduction of new evaluation metrics.

1.2 Research Objectives

The main objectives of this thesis are as follows:

1. To evaluate the performance of various Non-Parametric Estiimators (Abundance Based) on event logs.

2. To investigate the applicability of Parametric Estimators like Naive Bias, Poisson based estimators etc.

3. To work on a methodology that tests the stability of different estimators.

1.3 Thesis Structure

- **Chapter 2: Literature Review** – This chapter reviews existing work related to the Log Representativeness Problem and how species discovery is be applied to event logs. This section covers key metrics for evaluating estimators which include Completeness, Coverage, Cross-Validation, Sampling, and the newly introduced Variance of Variance (VoV). It also includes literature about Non-Parametric Estimators (Incidence-based and Abundance-based) and Parametric Estimators, summarizing their applicability to the problem.

- **Chapter 3: Methodology** – This chapter explains the methodological approach and workflow used to conduct the experiments, including details on dataset preparation, feature extraction, sampling techniques, and the workflow for evaluating different estimators. It also outlines the steps to calculate VoV, this part investigates variance calculation for each estimator, and interpretation of the results.

- **Chapter 4: Experiment Results and Analysis** – This chapter presents the results of the experiments. It starts by identifying the observed species count (Sobs) based on different retrieval functions, followed by a comparison of various estimators (e.g., Jackknife, ACE) with Chao2. Additional sections cover the use of Sampling as evaluation metrics, offering detailed insights into their performance. The chapter concludes with an analysis of Parametric Estimators and a general discussion of the findings.

- **Chapter 5: Conclusion and Future Work** – This chapter summarizes the key findings of the thesis, highlighting the main contributions to the field.

- **Chapter 6: Reference** – This section provides a comprehensive list of all sources cited in the thesis, ensuring proper attribution to previous research.

Chapter 2

Literature Review

2.1 Event Logs

Event logs are structured data collections that records and captures sequence of events or activities occurring in a business process. Typically, each event in an event log can represent the execution of one activity in a process or the passing of something from one process state to another. All events bring information with them attached as attributes—timestamp, activity name, resources, costs, time etc. Event logs help in understanding the business as gives better insights to the operations, executions and also helpful in troubleshooting required to identify the problem and eventually come up with a solution.

- Trace: A sequence of events that describes one instance of a process.
- Event: A trace is composed of entries; each entry is an event, representing something that happened at some moment in time.

2.2 Previous Research

2.2.1 The Log Representativeness Problem

The issue of log representativeness in event log analysis has been a significant challenge in the field of process mining. One key work addressing this issue is the paper by *Kabierski et al., titled - Addressing the Log Representativeness Problem using Species Discovery."* This author proposes the use of biodiversity research estimators, which are traditionally used to provide an estimate of specie richness in the biodiversity. The paper draws a parallelism between event logs and species. Making them almost analogous to each other.

The paper seeks to answer how well and how much does a sample of an event log represent the entire business process. How much information does the sample contain, and if that is enough to make reliable predictions for the entire business process?

2.2.2 Species Definition and Event Log Correlation

When it comes to defining a specie in nature, living organisms are classified into groups depending on their structure and characteristics. The definition is pretty straight forward as there are certain classifications that include domain, kingdom, phylum, class, order, family, genus

(plural, genera), and species. However there are no such definitions that define as to what constitutes as a specie in an event log.
To solve this problem the paper comes up with a function which is called a **Specie Retrieval Function**, which maps the traces from an event log to sets of species. The species represent various aspects of the process behavior.
There are 5 distinct functions for that:

- **Activity-based Species (ζ_{act})**: This species is defined as the set of activities in a trace. It captures each unique activity that occurs in a trace as a species. *Example:* For trace $t_3 = \langle (A,1),(F,2),(F,4),(G,14),(E,7) \rangle$, the activity-based species would be:

$$\zeta_{act}(t_3) = \{A, F, G, E\}$$

- **Directly-Follows Relation-based Species (ζ_{df})**: This species is defined based on the directly-follows relation over activities in a trace. It identifies species by considering the ordering of activities. For a trace with events $e_1, e_2, ..., e_n$, this species captures pairs of activities where one directly follows the other. *Example:* For trace $t_3 = \langle (A,1),(F,2),(F,4),(G,14),(E,7) \rangle$, the directly-follows relations (species) would be:

$$\zeta_{df}(t_3) = \{(A,F),(F,F),(F,G),(G,E)\}$$

- **Trace Variant-based Species (ζ_{tv})**: This species definition considers an entire trace as a variant of species. If two traces share the same sequence of activities, they are of the same species. *Example:* For trace $t_3 = \langle (A,1),(F,2),(F,4),(G,14),(E,7) \rangle$, the trace variant species would be:

$$\zeta_{tv}(t_3) = \langle A, F, F, G, E \rangle$$

- **Uniform Duration-based Species $\zeta_{t\lambda}$**: This species type defines species based on the duration of activities, grouped into uniform bins of size λ. The duration of each activity is aggregated into bins to create species definitions. *Example:* For trace t_3, using a uniform binning of size 1 minute (ζ_{t1}) and size 5 minutes (ζ_{t5}), the species definitions are:

$$\zeta_{t1}(t_3) = \{(A,1),(F,2),(F,4),(G,14),(E,7)\}$$

$$\zeta_{t5}(t_3) = \{(A,5),(F,5),(G,15),(E,10)\}$$

- **Exponential Duration-based Species ($\zeta_{te\lambda}$)**: This species definition uses exponentially scaled bins for durations, which is more suitable for long-tail distributions (like activity durations that vary significantly). *Example:* For trace t_3, using exponential binning with $\lambda = 2$, the species would be:

$$\zeta_{te2}(t_3) = \{(A,2),(F,2),(F,4),(G,16),(E,8)\}$$

2.2.3 Research Gaps

- **Handling of Noisy Data**: The paper makes an assumption that the event logs are noise-free and error-free, Therefore no pre-processing is done on the dataset.

- **Use of Chao2**: Chao2 is heavily biased towards singletons (species occurring once) and doubletons (species occurring exactly twice) and if the number of these rare species is in the majority (skewed data), Chao2 gives inaccurate results.

- **Correlation between Activities and Events**: The paper assumes that there is no correlation between activities and that all events are independent of each other.
- **Species Abundance**: The paper only focuses on incidence based estimator.
- **Specie Retrieval Function**: Improvements in the specie definitions are required.
- **Evaluation Metric**: The paper uses completeness and coverage as a metric to analyse the results, but this seems to be inadequate since there is no way to know the complete dataset. We only have the sample to work with.

2.3 Non-Parametric Estimators

Estimators aim at building procedures to recover unknown parameters/species by analysing some measured data sampled from a large population.

Non Parametric Estimators are flexible tools that make little to no assumption about the underlying distribution of the data, which makes them more flexible. This flexibility allows non-parametric estimators to adopt to a very wide range of data structures making them extremely useful when the true shape of the distribution is unknown, uncertain or complex. They are further classified into 2 groups:

2.3.1 Incidence Based

Incidence-based estimators focus on whether or not a specie is observed in a sample. The counting is made on different sampling units without considering how often that species appears.

- **Chao2**: Chao 2 *(Chao 1987; Colwell and Coddington 1994)*, which uses occurrence data from multiple samples in aggregate to estimate the species diversity of the whole. This estimator is defined as:

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

  where:

  - S_{obs} is the number of observed species in the sample.
  - Q_1 is the number of species that are observed in exactly one sampling unit (singletons).
  - Q_2 is the number of species that are observed in exactly two sampling units (doubletons).

2.3.2 Abundance-Based Estimators

Abundance-based estimators take into account not only the presence of species but also their frequency within the sample. This approach provides a deeper understanding of species diversity by incorporating both common and rare species, offering more accurate estimates, especially in environments where species abundance varies greatly.

1. Jackknife Estimators (1st and 2nd Order)

Jackknife estimators are robust non-parametric methods that estimate species richness by systematically excluding subsets of the data. They are particularly effective when the sample size is small or when rare species are of interest.

Jackknife 1st Order The first-order Jackknife estimator focuses on singletons, species that are observed in only one sampling unit. It estimates species richness by extrapolating from the frequency of these rare species.

$$\hat{S}_{\text{Jackknife1}} = S_{\text{obs}} + Q_1 \left(\frac{n-1}{n} \right)$$

where:

- S_{obs} is the number of observed species,
- Q_1 is the number of singletons (species found in only one sampling unit),
- n is the number of sampling units.

Jackknife 2nd Order The second-order Jackknife estimator extends the first-order approach by also considering doubletons, species observed in exactly two sampling units. This method is more accurate for larger sample sizes or when greater precision is needed.

$$\hat{S}_{\text{Jackknife2}} = S_{\text{obs}} + Q_1 \left(\frac{2n-3}{n} \right) - Q_2 \left(\frac{n-2}{n} \right)$$

where:

- S_{obs} is the number of observed species,
- Q_1 is the number of singletons,
- Q_2 is the number of doubletons (species found in exactly two sampling units),
- n is the number of sampling units.

2. Jackknife Estimators with Resampling (1st and 2nd Order)

Resampling techniques can be applied to Jackknife estimators to improve robustness and reduce variance in the estimates. In this approach, data is resampled multiple times by systematically leaving out sampling units, and the Jackknife estimator is recalculated for each subset.

Jackknife 1st Order with Resampling The 1st order Jackknife is applied to each resampled dataset, and the final estimate is the average across all resamples, providing a more stable estimate of species richness.

Jackknife 2nd Order with Resampling Similarly, the 2nd order Jackknife is calculated for each resampled subset, averaging the results to produce a robust estimate, particularly useful for larger datasets.

3. ACE (Abundance-based Coverage Estimator)

The ACE estimator is designed to handle species abundance by dividing the species into two categories: frequent and rare. Rare species are those with fewer than 10 individuals in the sample. ACE is particularly useful when there is a significant proportion of rare species in the dataset.

ACE Traditional Definition The traditional ACE estimator estimates species richness by focusing on the abundance of rare species, while accounting for the sample coverage of these species:

$$\hat{S}_{\text{ACE}} = S_{\text{rare}} + \frac{Q_1}{C_{\text{ACE}}}$$

where:

- S_{rare} is the number of species with fewer than 10 individuals,
- Q_1 is the number of singletons (species observed only once),
- C_{ACE} is the sample coverage for rare species:

$$C_{\text{ACE}} = 1 - \frac{Q_1}{n_{\text{rare}}}$$

where n_{rare} is the total number of individuals in the rare species group.

ACE Simplified Definition The simplified ACE estimator reduces some of the complexity of the traditional approach by focusing primarily on the counts of singletons and doubletons. This version provides a quicker estimation while maintaining accuracy.

$$S_{\text{ACE}} = \frac{S_{\text{rare}}}{C_{\text{ACE}}} + S_{\text{common}}$$

where:

- S_{rare} is the number of species observed fewer than or equal to n times,
- S_{common} is the number of species observed more than n times,
- C_{ACE} is the sample coverage for the rare species, reflecting how well these species are sampled.

2.4 Parametric Estimators

Parametric estimators rely on a statistical relationship between historical data and certain variables to predict outcomes or parameters. For example, in project management, parametric estimating uses data from previous projects to estimate costs, time, or resources needed for new projects. This approach can be quite accurate if the current project is similar to past projects and the historical data is reliable. The method calculates the estimate by identifying key parameters and applying a mathematical formula to relate these to project characteristics (like time or cost).

Requirements for Parametric Estimators:

- **Historical Data Availability**: The most essential prerequisite is having a reliable set of historical data from similar projects or contexts. This data is used to establish relationships between variables (e.g., cost, time, or other parameters). If the data is inaccurate, outdated, or not comparable, the estimates may be misleading.
- **Correlation of Parameters**: The parameters being estimated (such as costs or time) must have a clear and measurable correlation with the variables from historical data. This correlation allows the model to generate predictions based on the identified patterns.
- **Assumption of a Known Distribution**: The data has to follow a known statistical distribution like normal, exponential or Poisson. The specific form of distribution must be chosen before in advance.
- **Fitting Model to the data**: An incorrect model i.e incorrect fitting can lead to poor performance.

2.5 Evaluation Criteria

2.5.1 Completeness

Completeness is calculated as the ratio of the number of observed species (i.e., unique behaviors or characteristics in the event log) to the estimated total number of species in the population. The formula for completeness is:

$$\hat{C}_{\text{obs}} = \frac{S_{\text{obs}}}{S_{\text{est}}}$$

Where:

- S_{obs} is the number of species (or unique event log behaviors) observed in the sample (event log).
- S_{est} is the estimated total number of species in the population, calculated using species richness estimators such as ACE, Chao2, or Jackknife.

This ratio provides an upper bound for completeness, as S_{est} gives a lower bound estimate of the total number of species.

2.5.2 Coverage

Coverage quantifies how much of the probability mass of the population is covered by the species observed in the sample. It measures the likelihood of encountering new, unseen species when additional data is collected. The formula for coverage is:

$$\hat{C}_{\text{ov}} \approx 1 - \frac{Q_1}{\sum_{i=1}^{n} Y_i}$$

Where:

- Q_1 is the number of species (event log behaviors) observed exactly once (singleton species).
- Y_i is the number of occurrences of species i in the sample.

- $\sum_{i=1}^{n} Y_i$ is the sum of the counts of all species observed in the sample.

The higher the coverage value, the more representative the sample is of the entire population, indicating that the sample captures a larger proportion of the total process behaviors.

2.5.3 Cross Validation

Cross-validation is a technique used to evaluate the performance of models, particularly in machine learning and statistics. It helps assess how well a model will generalize to an independent dataset (i.e., new, unseen data). Cross-validation is especially useful for avoiding overfitting, where a model performs well on the training data but poorly on unseen data.

A popular form is k-fold cross-validation, where the dataset is split into k subsets (or "folds"). The model is trained on k-1 folds and tested on the remaining fold. This process is repeated k times, with each fold being used as the test set once. The results are then averaged to provide an overall performance metric.

2.5.4 Sampling

The approach I used involved taking samples of the complete dataset, such as 90% and 50% of the data, and testing these samples on various estimators. I also implemented hierarchical sampling to enhance the robustness of the evaluation. The key assumption in this methodology is that the original dataset is considered to be complete. This allows the estimates from the samples to be compared against the observed species count (S_{obs}) from the original dataset. The evaluation was performed by calculating the absolute difference between the estimated species richness (S_{est}) and the observed species count (S_{obs}).

2.5.5 Variance of Variance

The idea of **Variance of Variance (VoV)** builds on the traditional variance calculation by assessing the variability of the variances themselves across different samples or sets of data. In the context of estimators, calculating the VoV can provide information about the stability and consistency of different estimators. If the value of VoV is low then it is implied that the estimator provides relatively consistent variance across different input sets, while if the value of VoV is high then it is implied that there is a higher fluctuation and potential instability in the estimates.

Chapter 3

Methodology

3.1 Workflow

In this study, the workflow followed while testing different estimators closely aligned with the methodology presented in the paper *The Log Representativeness using Species Discovery*. This approach ensured that my research remained consistent with prior studies, allowing for direct comparisons of the results.

The main components of the workflow are as follows:

- **Dataset**: The input datasets included the following: BPI-2012, BPI-2018, BPI-2019 and Sepsis.
- **Species Retrieval Function**: 7 distinct species were identified using different retrieval functions, the were:
  - Activity-based Species (ζ_{act})
  - Trace Variant-based Species (ζ_{tv})
  - Directly-Follows Relation-based Species (ζ_{df})
  - Uniform Duration-based Species ($\zeta_{t1}, \zeta_{t5}, \zeta_{t30}$)
  - Exponential Duration-based Species (ζ_{te2})
- **Estimator**: The following non-parametric estimators were tested in this thesis to evaluate species richness:
  - Jackknife (1st and 2nd order)
  - Jackknife (1st and 2nd order) with Resampling
  - ACE (Simplified) with R=5, 10
  - ACE Traditional with R=5, 10
  - Chao2
- **Completeness**: The completeness metric was calculated to evaluate the representativeness of the event logs in relation to the estimated species richness.
- **Results**: The final results of species estimation and coverage were obtained and analyzed.

This workflow ensured a systematic approach to testing and evaluating the different estimators, ensuring consistency and replicability with existing research.

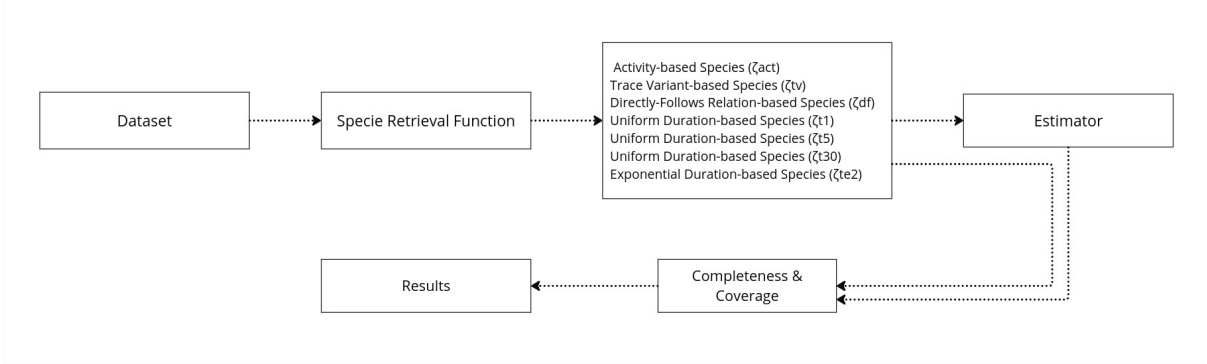

Figure 3.1: Workflow for Testing Estimators

3.2 Dataset Description and Preparation

3.2.1 Source of the Dataset

The datasets used for this thesis were publicly available event log datasets. Specifically, we utilized the following datasets:

- **BPI Challenge 2012 (BPI-2012)**: This dataset consists of event logs from a Dutch financial institute, capturing different types of cases such as loan applications and offers. The dataset includes timestamps, activity names, and resource information for each event, providing a detailed overview of the business processes.
- **BPI Challenge 2018 (BPI-2018)**: This event log dataset contains records of a purchasing process from an industrial company. The dataset includes various phases of the procurement process, such as purchase orders, goods receipts, and invoice handling, making it ideal for analyzing the efficiency of business operations.
- **BPI Challenge 2019 (BPI-2019)**: This dataset was collected from a public sector organization. It includes event logs related to the handling of permit requests. The dataset contains information about the activities involved in permit applications and processing times, which allows for detailed process mining and performance analysis.
- **Sepsis Cases (Sepsis)**: The Sepsis dataset contains event logs from a healthcare process. It tracks the treatment of patients diagnosed with sepsis, including timestamps for various medical treatments and administrative activities. This dataset provides an opportunity to analyze and improve the treatment process and patient outcomes.

3.2.2 Pre-Processing

To make sure that the results were calculated under the same conditions as those in the paper "Log Representativeness Using Species Discovery," No pre-processing of any type was applied to the dataset. This approach was taken to facilitate a more accurate comparison of results.

3.2.3 Feature Extraction

Based on different **Species Retrieval Functions**, namely ζ_{act}, ζ_{df}, ζ_{tv}, ζ_{t1}, ζ_{t5}, ζ_{t30}, and ζ_{te2}, I extracted various features from the datasets. These features were computed in accordance to

support the abundance-based estimators. As a result, seven distinct files were generated, each corresponding to a specific species retrieval function and containing the frequency and type of each specie.

3.2.4 Sampling Techniques and Data Division

The dataset was sampled in two sizes: 90% and 50% of the total dataset. Two sampling techniques were employed to divide the datasets:

- **Independent Sampling**: In independent sampling, data points (events) are sampled randomly without considering any hierarchical or structural relationships. Each sample is drawn independently from the dataset, ensuring that every data point has an equal chance of being selected.

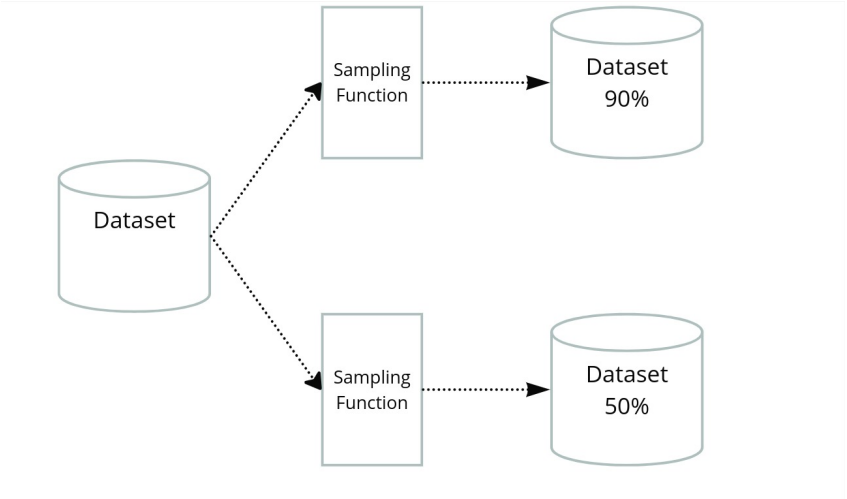

Figure 3.2: Independent Sampling

- **Hierarchical Sampling**: In hierarchical sampling, the dataset is divided into groups (hierarchies), such as event classes or process instances. Sampling is then performed within each group to preserve the underlying structure of the event log. This method ensures that important dependencies between events are maintained, providing a more representative sample for hierarchical or structured analyses.

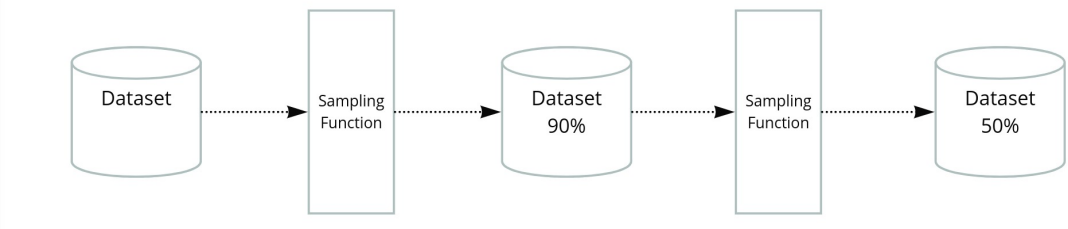

Figure 3.3: Hierarchical Sampling

Independent Sampling Use Case

- The population is relatively homogeneous.
- The primary goal is simplicity and speed of sampling.
- There's no clear hierarchical structure in the data.
- When conclusions about the population as a whole without subgroup differentiation needs to be drawn.

Hierarchical Sampling Use Case

- When the data has clear groups or levels (e.g., regions, time periods, or departments).
- When representation of all subgroups is required.
- When a more precise and stratified estimates is required.

3.3 Evaluation Methods

3.3.1 Completeness

Completeness is calculated as the ratio of the number of observed species (i.e., unique behaviors or characteristics in the event log) to the estimated total number of species in the population. The formula for completeness is:

$$\hat{C}_{\text{obs}} = \frac{S_{\text{obs}}}{S_{\text{est}}}$$

3.3.2 Using Sampled Data

The main idea of using sampled data was to calculate S_{est} of the sampled data and test it against the S_{obs} value of the complete dataset. This was done in two ways:

- **Dataset Sampling**: In this approach, the dataset was sampled in different sizes (90% and 50%), and then the features were extracted based on species retrieval functions. The estimators were applied to the sampled datasets to generate the estimated species richness (S_{est}). These estimates were then compared with the observed species richness (S_{obs}) from the complete dataset to calculate the Mean Absolute Difference (MAD).
- **Feature Sampling**: In this approach, the full dataset was used to extract features based on species retrieval functions. After feature extraction, the sampling was applied at the feature level (90% and 50%) before applying the estimators. This method allowed us to assess how reducing the feature set impacted the estimator's ability to predict species richness compared to the complete feature set.

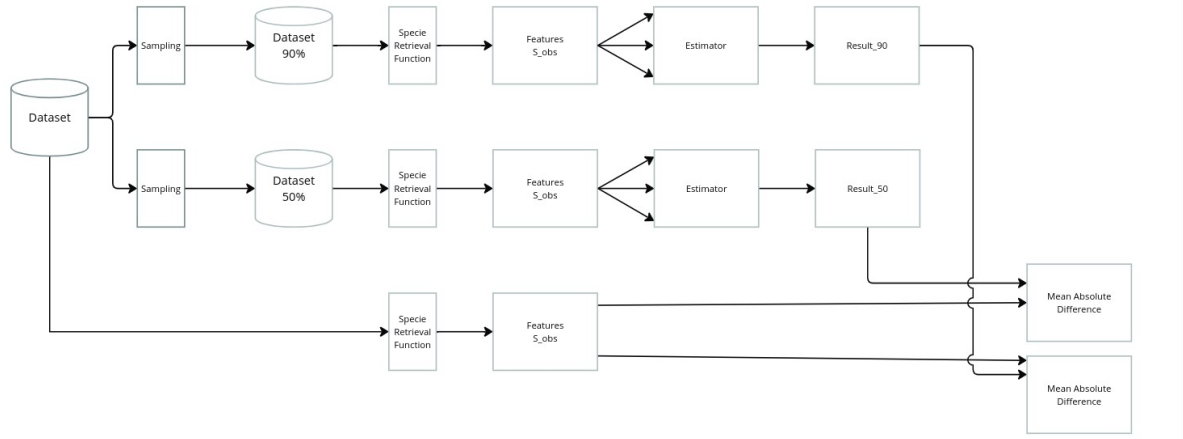

Figure 3.4: Dataset Sampling Process

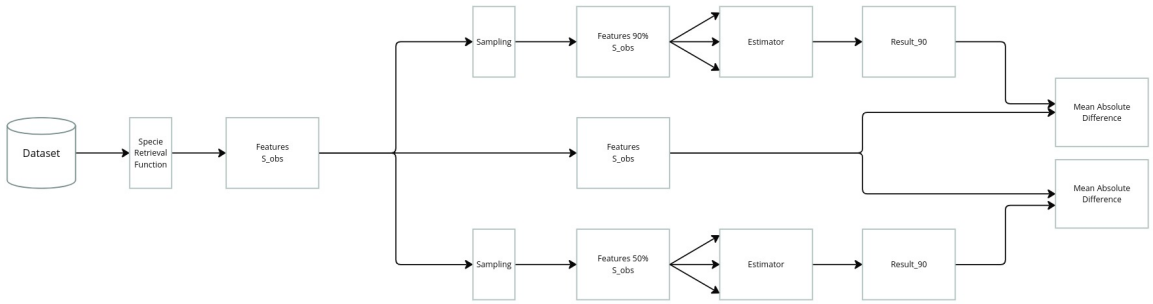

Figure 3.5: Feature Sampling Process

3.4 Steps to Calculate VoV

3.4.1 Data Preparation

The first step was to extract the feature data from the dataset using the species retrieval function.The second step is to calculate S_est using various estimators for each dataset.

3.4.2 Variance Calculation for Each Estimator

For each estimator (e.g., ”S_est_Ace_S5/10”, ”S_est_Chao2”, ”S_est_Jacknife1/2”), variance was calulated of the estimated values for all the different species definitions. The following formula was used:

$$\text{Variance} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where x_i represents the estimate values for a given species definition, and $\bar{x}$ is the mean of these values.

3.4.3 Calculate Variance of Variances (VoV)

Once the variance for each estimator had been calculated, variance of that variance was calculated across all the estimators. This provided VoV:

$$\text{VoV} = \frac{1}{k-1} \sum_{j=1}^{k} (v_j - \bar{v})^2$$

where v_j represents the variances of the different estimators, and $\bar{v}$ is the mean of those variances.

3.4.4 Display and Interpretation

The final VoV value for each dataset reflected the stability of all the different estimators that were used in this study. A low VoV indicated consistent performance across different species definitions, while a higher VoV indicated a higher variability/fluctuations and thus instability.

3.5 Conclusion

By calculating the Variance of Variance (VoV), we can gain insights into the reliability of estimators. This helps evaluate their robustness across different speciess retrieval functions.

3.6 Experimental Setup

3.6.1 Programming Environment

The experiments were conducted using **Python 3.12.7**. The following libraries were used extensively:

- **pm4py**: pm4py is a process mining library for Python implementing several process mining (PM) artifacts and algorithms.
- **Pandas**: Pandas was used extensively. Since pm4py was slow on large datasets, we decided to convert the event logs, originally in `.xes` format, into `.csv` for faster operations. Additionally we also used Pandas to create samples of and artificial data.
- **Matplotlib**: This library was used to generate various graphs for the analysis.

3.6.2 Reproducibility

The code used for this research is available on my GitHub repository: GitHub link.

Chapter 4

Experiment Results and Analysis

4.1 Finding Observed Species Count (Sobs)

Sobs was calculated in accordance to various Species Retrieval Function definitions, which were ζ_{act}, ζ_{df}, ζ_{tv}, ζ_{t1}, ζ_{t5}, ζ_{t30}, and ζ_{te2}, as defined in *this section*.

The results are displayed in the table below.

Table 4.1: Sobs for Different Logs and Species Definitions (Part 1)

Log	Species Def.	S_obs
BPI-2012	ζ_{act}	24
BPI-2012	ζ_{df}	149
BPI-2012	ζ_{tv}	4366
BPI-2012	ζ_{t1}	959
BPI-2012	ζ_{t5}	488
BPI-2012	ζ_{t30}	211
BPI-2012	ζ_{te2}	96
BPI-2018	ζ_{act}	41
BPI-2018	ζ_{df}	619
BPI-2018	ζ_{tv}	28489
BPI-2018	ζ_{t1}	177653
BPI-2018	ζ_{t5}	30330
BPI-2018	ζ_{t30}	31865
BPI-2018	ζ_{te2}	697

Log	Species Def.	S_obs
BPI-2019	ζ_{act}	42
BPI-2019	ζ_{df}	538
BPI-2019	ζ_{tv}	11936
BPI-2019	ζ_{t1}	200937
BPI-2019	ζ_{t5}	89981
BPI-2019	ζ_{t30}	32084
BPI-2019	ζ_{te2}	583
Sepsis	ζ_{act}	16
Sepsis	ζ_{df}	135
Sepsis	ζ_{tv}	846
Sepsis	ζ_{t1}	3326
Sepsis	ζ_{t5}	2233
Sepsis	ζ_{t30}	1184
Sepsis	ζ_{te2}	183

4.1.1 Analysis

I was able to replicate the results from the paper *Log Representativeness using Specie Discovery*. The accuracy of the results were form 93% to 100%. The features derived from the Species Retrieval Functions laid the benchmark for testing different estimators. The outputs from these functions generated seven distinct sets of inputs, which were then used in testing a variety of non-parametric and parametric estimators, as discussed in the following sections.

4.2 Comparing Estimators with Chao2

In this section I tested different estimators and their variants with Chao2, the metric used for evaluation was Completeness (between S_obs and S_est).

4.2.1 Jackknife vs Chao2

Two different variants of Jacknife estimator were used. For every variant, 1st and 2nd order estimates were calculated. The detailed description of the estimators can be found in *this section*.

Jackknife 1st and 2nd Order vs Chao2

In this section, we compare the Jackknife 1st and 2nd order estimators against Chao2. The results are summarized in the table below, followed by graphical analysis.

Table 4.2: Completeness Comparison for BPI-2012 and BPI-2018

Log	Species Def.	S_obs	C_Chao2	C-J1	C-J2
BPI-2012	ζ_{act}	24.0	1.000	1.000	1.000
BPI-2012	ζ_{df}	149.0	0.856	0.955	0.920
BPI-2012	ζ_{tv}	4366.0	0.146	0.538	0.376
BPI-2012	ζ_{t1}	959.0	0.336	0.641	0.490
BPI-2012	ζ_{t5}	488.0	0.410	0.644	0.500
BPI-2012	ζ_{t30}	211.0	0.725	0.738	0.647
BPI-2012	ζ_{te2}	96.0	0.821	0.932	0.873
BPI-2018	ζ_{act}	41.0	1.000	1.000	1.000
BPI-2018	ζ_{df}	619.0	0.863	0.881	0.824
BPI-2018	ζ_{tv}	28489.0	0.070	0.517	0.352
BPI-2018	ζ_{t1}	177653.0	0.466	0.646	0.509
BPI-2018	ζ_{t5}	30330.0	0.547	0.671	0.545
BPI-2018	ζ_{t30}	31865.0	0.636	0.711	0.599
BPI-2018	ζ_{te2}	697.0	0.928	0.915	0.885

- S_obs = Number of species observed in the dataset.
- C_Chao2 = Completeness score for Chao2
- C-J1/J2 = Completeness score for Jacknife 1st and 2nd order estimator.

Table 4.3: Completeness Comparison for BPI-2019 and Sepsis

Log	Species Def.	S_obs	C_Chao2	C-J1	C-J2
BPI-2019	ζ_{act}	42.0	1.000	1.000	1.000
BPI-2019	ζ_{df}	538.0	0.873	0.878	0.825
BPI-2019	ζ_{tv}	11936.0	0.236	0.571	0.414
BPI-2019	ζ_{t1}	200937.0	0.580	0.677	0.557
BPI-2019	ζ_{t5}	89981.0	0.701	0.747	0.648
BPI-2019	ζ_{t30}	32084.0	0.709	0.767	0.669
BPI-2019	ζ_{te2}	583.0	0.948	0.934	0.912
Sepsis	ζ_{act}	16.0	1.000	1.000	1.000
Sepsis	ζ_{df}	135.0	0.823	0.888	0.823
Sepsis	ζ_{tv}	846.0	0.088	0.519	0.356
Sepsis	ζ_{t1}	3326.0	0.298	0.575	0.423
Sepsis	ζ_{t5}	2233.0	0.407	0.606	0.464
Sepsis	ζ_{t30}	1184.0	0.452	0.646	0.507
Sepsis	ζ_{te2}	183.0	0.929	0.920	0.888

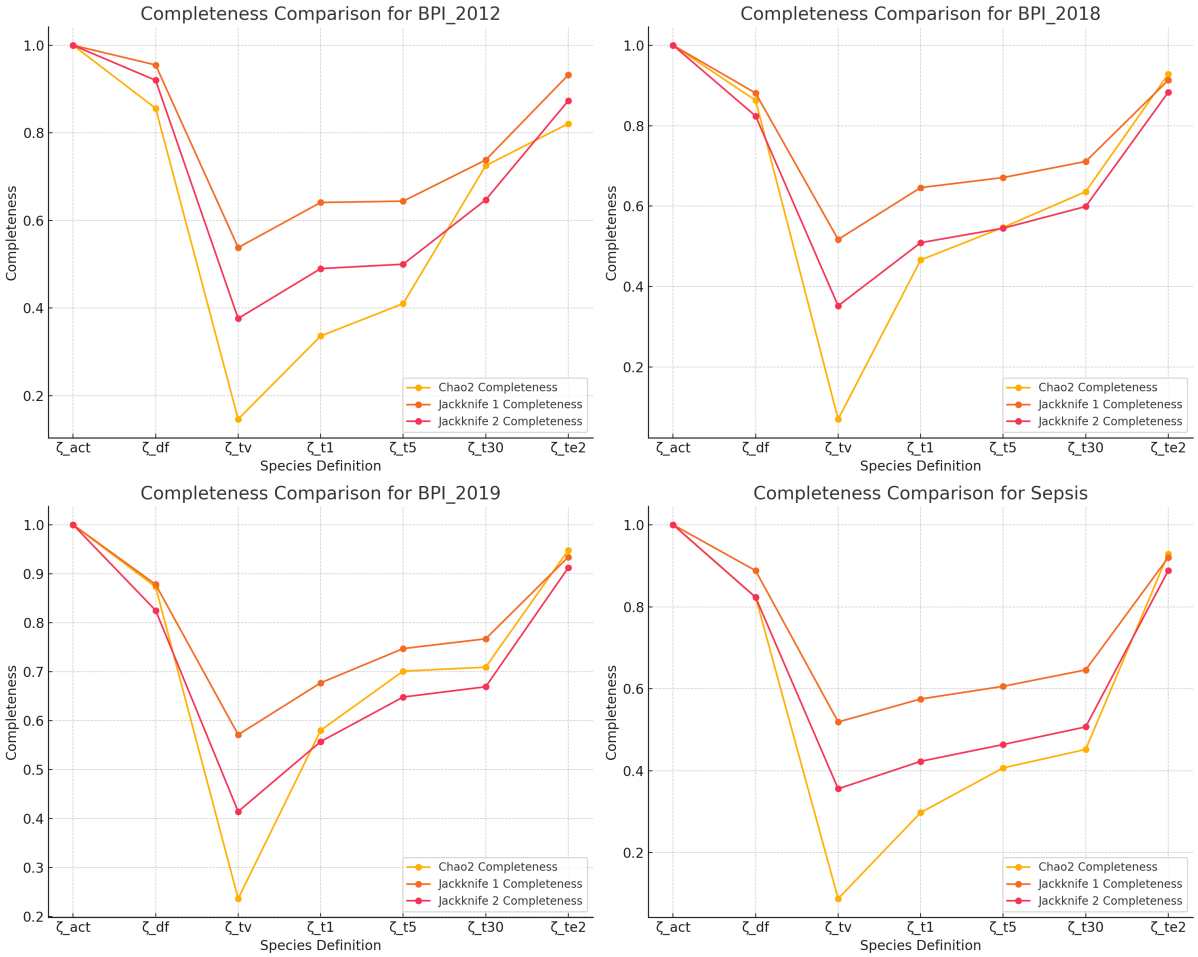

Figure 4.1: Feature Sampling Process

4.2.2 General Observation

- **Consistency Across Datasets**: All the estimators follow a consistent pattern of completeness, which infers that **the estimators are robust to different event logs but sensitive to specific definition.**
- Completeness seem to drop sharply for trace variant definition, this trend is similar for all the estimators.

4.2.3 Jackknife (Resampled) vs Chao2

In this section, I tried to implement Jackknife like a bootstrap estimator. A detailed description can be found in *this section*.

Table 4.4: Completeness Comparison for BPI-2012 and BPI-2018

Log	Species Def.	S_obs	C_Chao2	C-J1	C-J2
BPI-2012	ζ_{act}	24.0	1.000	1.043	1.043
BPI-2012	ζ_{df}	149.0	0.856	0.961	0.925
BPI-2012	ζ_{tv}	4366.0	0.146	0.538	0.377
BPI-2012	ζ_{t1}	959.0	0.336	0.642	0.491
BPI-2012	ζ_{t5}	488.0	0.410	0.646	0.501
BPI-2012	ζ_{t30}	211.0	0.725	0.740	0.651
BPI-2012	ζ_{te2}	96.0	0.821	0.941	0.881
BPI-2018	ζ_{act}	41.0	1.000	1.025	1.025
BPI-2018	ζ_{df}	619.0	0.863	0.882	0.825
BPI-2018	ζ_{tv}	28489.0	0.070	0.517	0.352
BPI-2018	ζ_{t1}	177653.0	0.466	0.646	0.509
BPI-2018	ζ_{t5}	30330.0	0.547	0.671	0.545
BPI-2018	ζ_{t30}	31865.0	0.636	0.711	0.600
BPI-2018	ζ_{te2}	697.0	0.928	0.915	0.885

- S_obs = Number of species observed in the dataset.
- C_Chao2 = Completeness score for Chao2
- C-J1/J2 = Completeness score for Jacknife 1st and 2nd order estimator (Resampled) .

Table 4.5: Completeness Comparison for BPI-2019 and Sepsis

Log	Species Def.	S_obs	C_Chao2	C-J1	C-J2
BPI-2019	ζ_{act}	42.0	1.000	1.024	1.050
BPI-2019	ζ_{df}	538.0	0.873	0.879	0.826
BPI-2019	ζ_{tv}	11936.0	0.236	0.571	0.414
BPI-2019	ζ_{t1}	200937.0	0.580	0.677	0.557
BPI-2019	ζ_{t5}	89981.0	0.701	0.747	0.648
BPI-2019	ζ_{t30}	32084.0	0.709	0.767	0.669
BPI-2019	ζ_{te2}	583.0	0.948	0.936	0.914
Sepsis	ζ_{act}	16.0	1.000	1.067	1.067
Sepsis	ζ_{df}	135.0	0.823	0.894	0.828
Sepsis	ζ_{tv}	846.0	0.088	0.520	0.356
Sepsis	ζ_{t1}	3326.0	0.298	0.575	0.423
Sepsis	ζ_{t5}	2233.0	0.407	0.606	0.464
Sepsis	ζ_{t30}	1184.0	0.452	0.646	0.507
Sepsis	ζ_{te2}	183.0	0.929	0.924	0.893

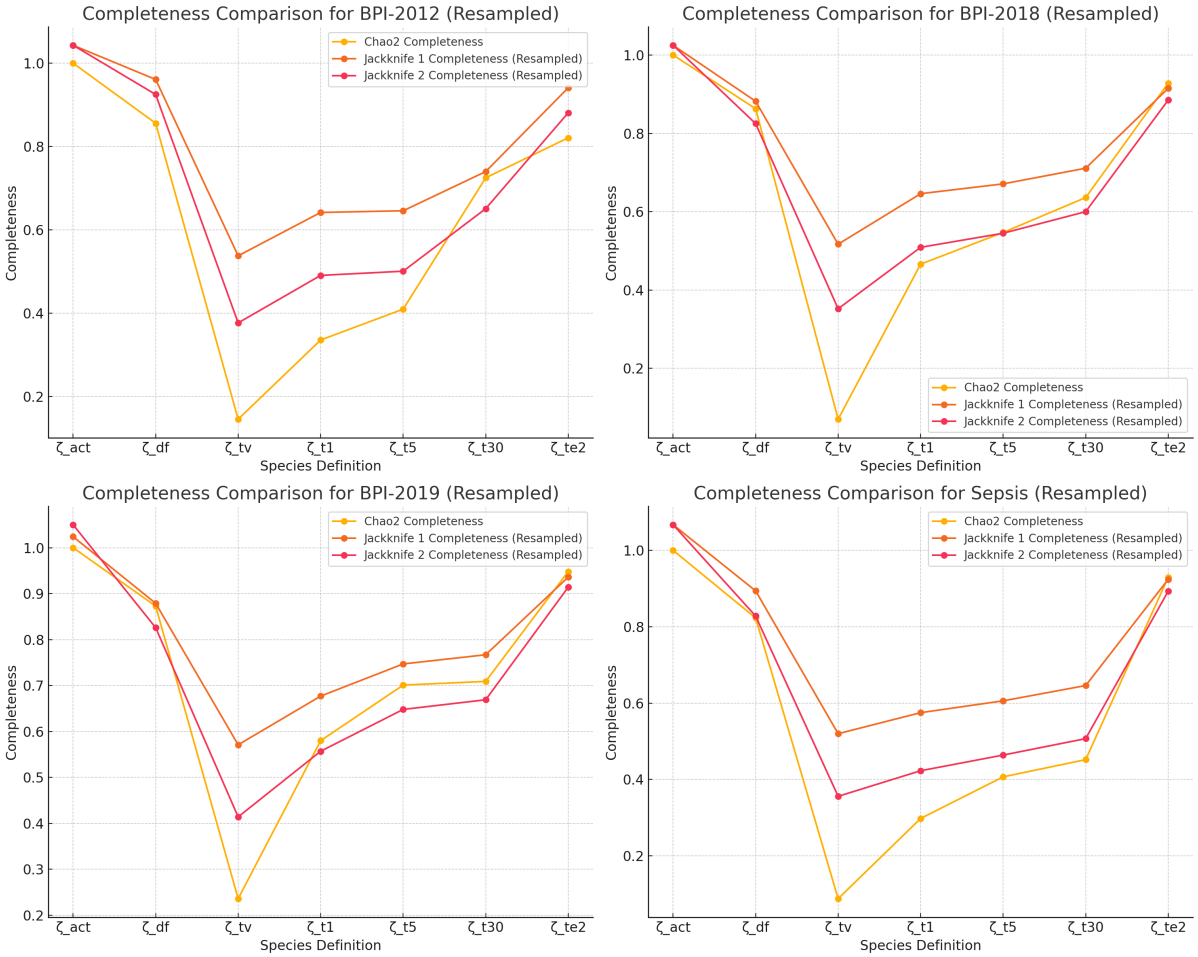

Figure 4.2: Comparison of Completeness Estimates for Jackknife (Resampled) vs Chao2

4.2.4 General Observation

- Jacknife Resampling overestimates the completeness, which means that resampling underestimates the estimation which is incorrect since the value of S_est cannot be less than S_obs.
- Below is a graph comparing the completeness of the Jacknife and Jacknife (Resampled).

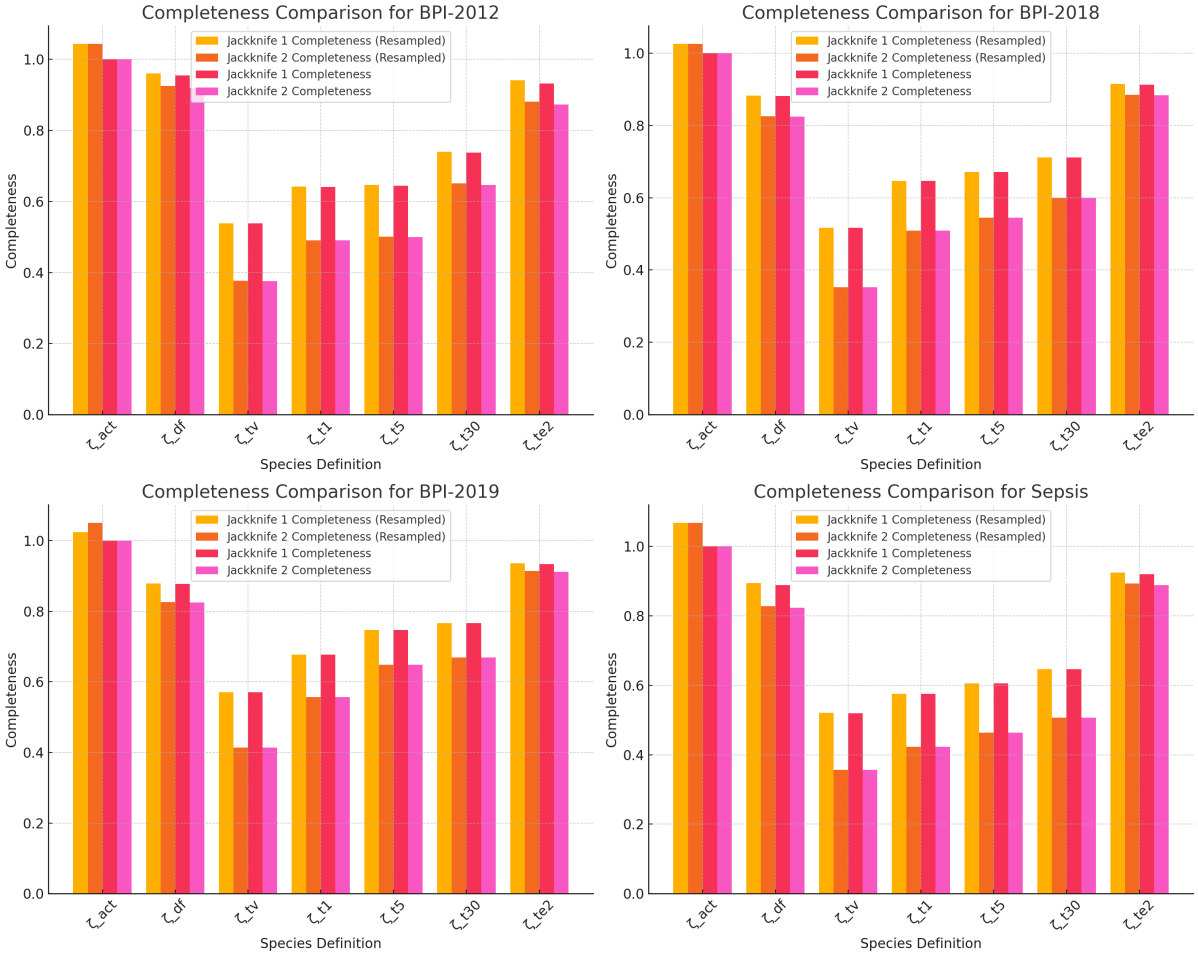

Figure 4.3: Comparison of Completeness Estimates for Jackknife (Resampled) vs Chao2

4.2.5 ACE - Abundance-Based Coverage Estimator

Two variants of the ACE estimator were used. Each variant, was tested on 2 different parameters (R= 5 & 10), were R is the threshold for rare species. A detailed description of the estimator can be found in *this section*.

Table 4.6: Completeness Comparison for BPI-2012 and BPI-2018

Log	Species Def.	S_obs	C-Chao2	C-ACE5	C-ACE10
BPI-2012	ζ_{act}	24.0	1.0	1.0	1.0
BPI-2012	ζ_{df}	149.0	0.968	0.98	0.98
BPI-2012	ζ_{tv}	4366.0	0.146	0.264	0.314
BPI-2012	ζ_{t1}	959.0	0.336	0.563	0.682
BPI-2012	ζ_{t5}	488.0	0.410	0.604	0.680
BPI-2012	ζ_{t30}	211.0	0.725	0.802	0.851
BPI-2012	ζ_{te2}	96.0	0.821	0.960	0.980
BPI-2018	ζ_{act}	41.0	1.000	1.0	1.0
BPI-2018	ζ_{df}	619.0	0.863	0.936	0.960
BPI-2018	ζ_{tv}	28489.0	0.070	0.126	0.166
BPI-2018	ζ_{t1}	177653.0	0.466	0.617	0.726
BPI-2018	ζ_{t5}	30330.0	0.547	0.689	0.766
BPI-2018	ζ_{t30}	31865.0	0.636	0.756	0.829
BPI-2018	ζ_{te2}	697.0	0.928	0.952	0.972

- S_obs = Number of species observed in the dataset.
- C_Chao2 = Completeness score for Chao2
- C-ACE5/10 = Completeness score for ACE estimator(with Rare specie = 5 and 10).

Table 4.7: Completeness Comparison for BPI-2019 and Sepsis

Log	Species Def.	S_obs	C_Chao2	C-ACE5	C-ACE10
BPI-2019	ζ_{act}	42.0	1.000	1.0	1.0
BPI-2019	ζ_{df}	538.0	0.873	0.926	0.956
BPI-2019	ζ_{tv}	11936.0	0.236	0.397	0.495
BPI-2019	ζ_{t1}	200937.0	0.580	0.709	0.791
BPI-2019	ζ_{t5}	89981.0	0.701	0.813	0.877
BPI-2019	ζ_{t30}	32084.0	0.709	0.818	0.884
BPI-2019	ζ_{te2}	583.0	0.948	0.967	0.981
Sepsis	ζ_{act}	16.0	1.000	1.0	1.0
Sepsis	ζ_{df}	135.0	0.823	0.931	0.964
Sepsis	ζ_{tv}	846.0	0.088	0.157	0.171
Sepsis	ζ_{t1}	3326.0	0.298	0.428	0.509
Sepsis	ζ_{t5}	2233.0	0.407	0.54	0.626
Sepsis	ζ_{t30}	1184.0	0.452	0.647	0.753
Sepsis	ζ_{te2}	183.0	0.929	0.963	0.979

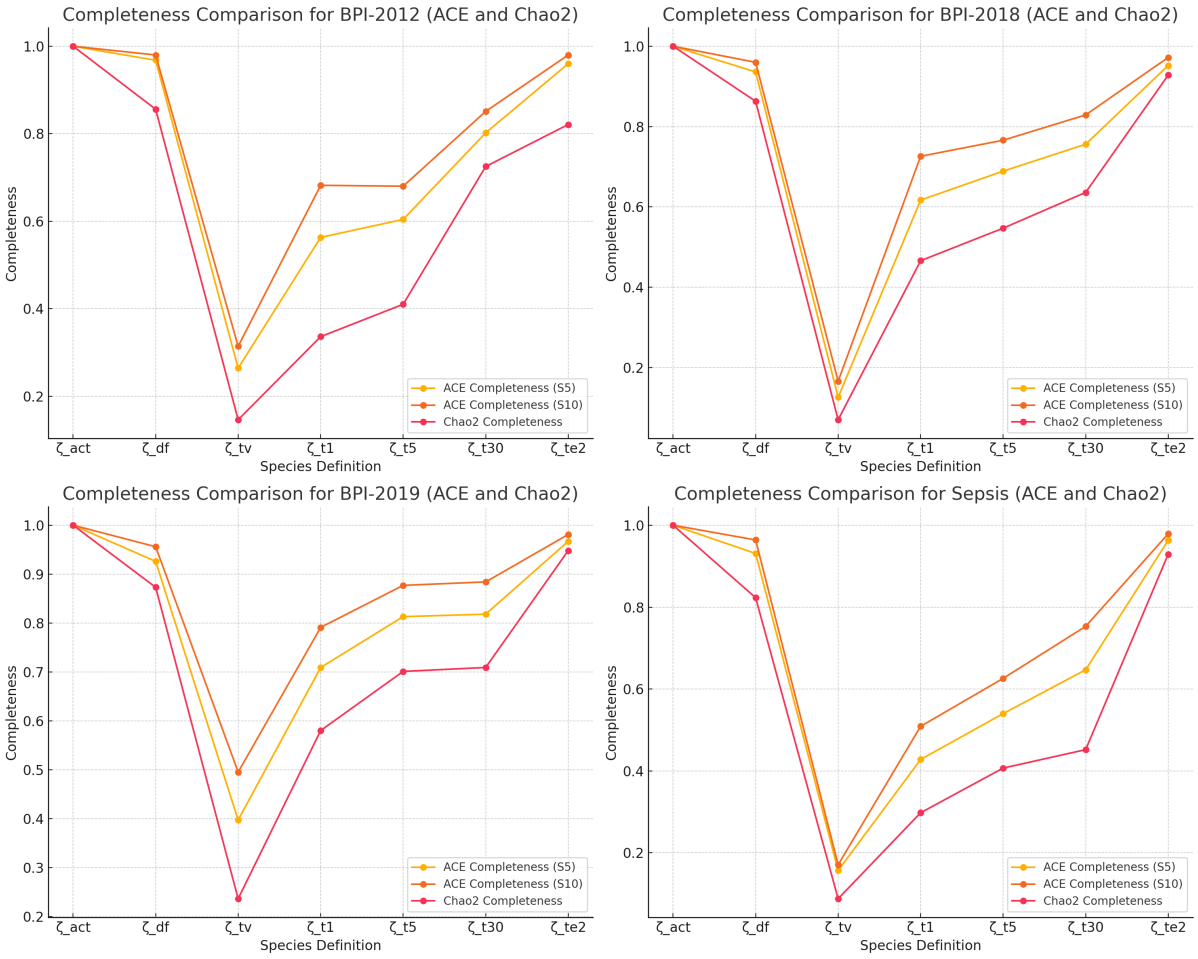

Figure 4.4: Comparison of Completeness Estimates for Jackknife (Resampled) vs Chao2

4.3 Checking the stability of Estimators using Samples

A full explanation and implementation can be found in *this section*

In this study, hierarchical sampling was applied to produce samples at 50% and 90% of the dataset. Hierarchical sampling was used in particular to ensure that each subgroup or level is adequately represented in the sample. The goal was to identify bottleneck in each estimator's effectiveness across different sample sizes. Then for each sample an estimation (S_est) was calculated, thereafter completeness was calculated, which allowed for the comparison of the estimator performance. Three key inferences could be drawn from the results:

1. If increasing the sample size does not improve completeness and remains stable, it indicates that the estimator is stable. And it produces stable estimation across different sample sizes

2. An increase in completeness with larger sample sizes suggests that the estimator performs better as more data becomes available.

3. A decrease in completeness with increasing sample size may imply that the estimator is more suited for smaller datasets.

All the estimators were subjected to sampling, and the dataset used was BPI-2018. Below are the results:

- ACE with rare species threshold set at 5

Species Def.	C(1.0)	C(0.90)	C(0.50)
ζ_{act}	1.000	1.000	1.000
ζ_{df}	0.936	0.941	0.949
ζ_{tv}	0.126	0.126	0.127
ζ_{t1}	0.617	0.617	0.618
ζ_{t5}	0.689	0.688	0.690
ζ_{t30}	0.756	0.756	0.757
ζ_{te2}	0.952	0.954	0.963

Table 4.8: Completeness Scores for ACE

  - C(1.0/.90/.50) = Completeness score after sampling 100%, 90% and 50% of the dataset.

- ACE with rare specie threshold set at 10

Species Def.	C(1.0)	C(0.90)	C(0.50)
ζ_{act}	1.000	1.000	1.000
ζ_{df}	0.960	0.962	0.969
ζ_{tv}	0.166	0.167	0.169
ζ_{t1}	0.726	0.726	0.725
ζ_{t5}	0.766	0.766	0.767
ζ_{t30}	0.829	0.829	0.829
ζ_{te2}	0.972	0.974	0.981

Table 4.9: Completeness Scores for ACE

  - C(1.0/.90/.50) = Completeness score after sampling 100%, 90% and 50% of the dataset.

- Jacknife 1st and 2nd order

Species Def.	C(1.0) J1	C(0.90) J1	C(0.50) J1	C(1.0) J2	C(0.90) J2	C(0.50) J2
ζ_{act}	1.000	1.000	1.000	1.000	1.000	1.000
ζ_{df}	0.881	0.886	0.897	0.881	0.886	0.897
ζ_{tv}	0.517	0.517	0.517	0.517	0.517	0.517
ζ_{t1}	0.646	0.646	0.646	0.646	0.646	0.646
ζ_{t5}	0.671	0.671	0.672	0.671	0.671	0.672
ζ_{t30}	0.711	0.711	0.712	0.711	0.711	0.712
ζ_{te2}	0.913	0.914	0.926	0.913	0.914	0.926

Table 4.10: Completeness Scores for Jackknife (J1 and J2)

  - C(1.0/.90/.50) = Completeness score after sampling 100%, 90% and 50% of the dataset.

- Jacknife 1st and 2nd order with resampling

Species Def.	C(1.0) J1	C(0.90) J1	C(0.50) J1	C(1.0) J2	C(0.90) J2	C(0.50) J2
ζ_{act}	1.025	1.028	1.059	1.025	1.028	1.059
ζ_{df}	0.882	0.887	0.900	0.825	0.835	0.866
ζ_{tv}	0.517	0.517	0.517	0.352	0.353	0.353
ζ_{t1}	0.646	0.646	0.646	0.509	0.509	0.509
ζ_{t5}	0.671	0.672	0.672	0.545	0.545	0.546
ζ_{t30}	0.711	0.712	0.712	0.606	0.599	0.606
ζ_{te2}	0.915	0.915	0.929	0.885	0.884	0.908

Table 4.11: Completeness Scores for Jackknife (J1 and J2)

  - C(1.0/.90/.50) = Completeness score after sampling 100%, 90% and 50% of the dataset.

- Chao2

Species Def.	C(1.0)	C(0.90)	C(0.50)
ζ_{act}	1.000	1.000	1.000
ζ_{df}	0.863	0.876	0.914
ζ_{tv}	0.070	0.070	0.069
ζ_{t1}	0.466	0.466	0.467
ζ_{t5}	0.547	0.547	0.550
ζ_{t30}	0.636	0.634	0.636
ζ_{te2}	0.928	0.926	0.946

Table 4.12: Completeness Scores Chao2

  - C(1.0/.90/.50) = Completeness score after sampling 100%, 90% and 50% of the dataset.

4.3.1 General Observation

- The general trend is that the completeness decreases as the sample size increases.
- Chao2 Estimator: Stability decreases as the sample size increases. This may be due to the fact that increasing the sample sizes also increases the number of singleton and doubleton species in the sample and it is known that the Chao2 performance decreases as the rare species count increases.

4.3.2 Gaps in this method

- Different sample sizes (eg. .75, .30) are required to confirm the hypothesis.
- More samples need to be drawn from the dataset to ensure robust results.

4.4 Parametric Estimators

Parametric estimators are widely used in biodiversity research since they are very well suited for real world applications, as they try to find relationships within the data. However, before applying parametric estimators, certain criteria must be met to avoid biased/flawed/skewed results. The most critical requirement is that the data should follow a known probability distribution.

In this part of my thesis, I tested two datasets—BPI-2012 and Sepsis, to determine if they met the criteria for applying parametric estimators. Unfortunately, both datasets exhibited significant skewness, making it difficult to apply a standard discrete probability function.

Further I applied the Shapiro-Wilk test, which tests for normality, meaning it checks whether the sample is generated from a Gaussian distribution. The results confirmed that the data did not follow a normal distribution. Additionally, I explored whether other distributions such as log-normal, gamma, or exponential might fit the data, but none of then proved suitable.

The figures below illustrate the skewness of the data after applying the Shapiro-Wilk test:

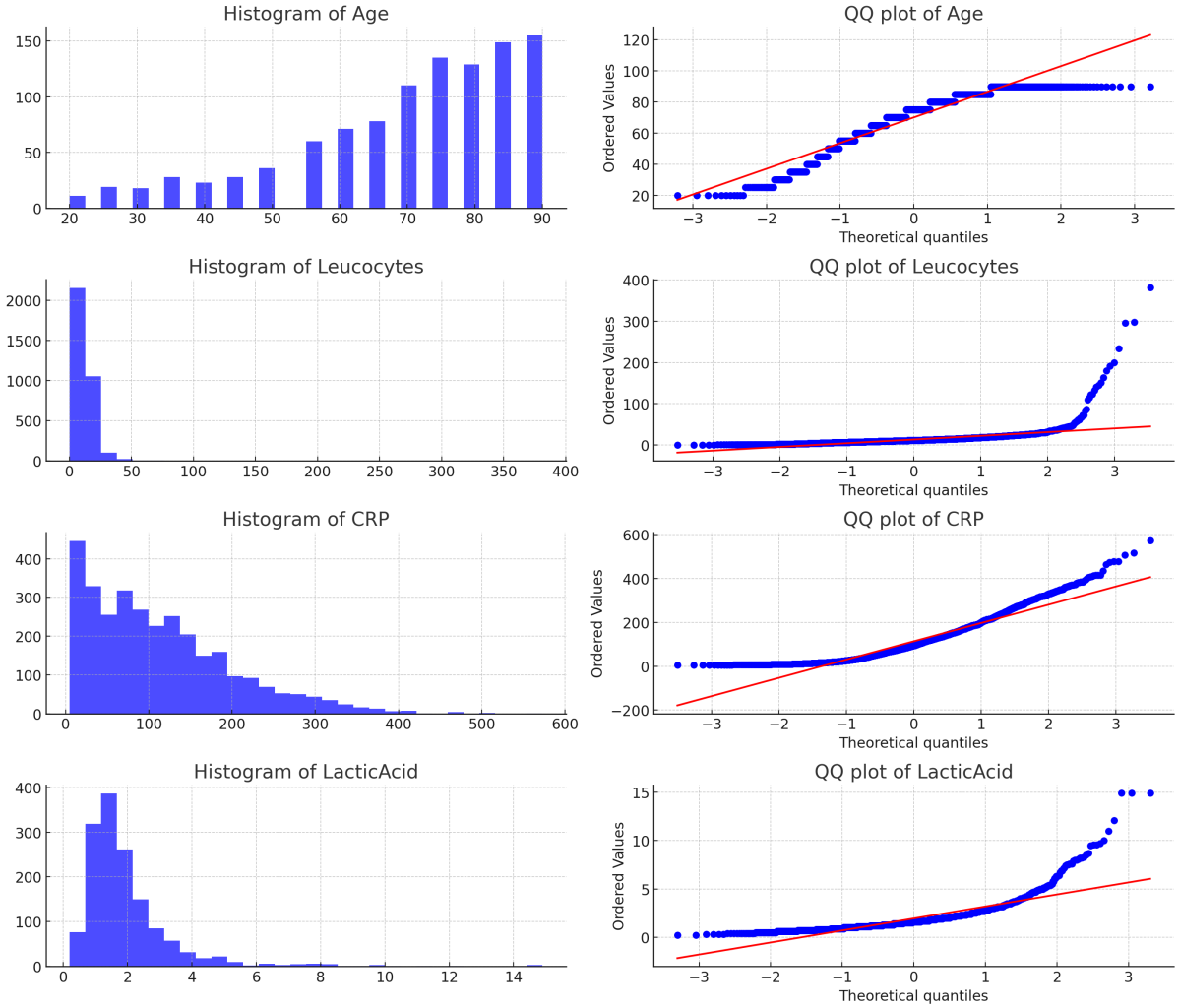

Figure 4.5: Skewness of Sepsis dataset after applying the Shapiro-Wilk test

4.4.1 Analysing fig 4.5

Skewness: All variables (Age, Leucocytes, CRP, and Lactic Acid) exhibit right skewness, where most values are concentrated at the lower end, with long tails stretching towards higher

values.
Non-Normality: The Q-Q plots show that none of these variables follow a normal distribution. The deviations from the red line (which represents a perfect normal distribution) indicate that the data points, particularly in the tails (extremes), are far from the values expected under normality.

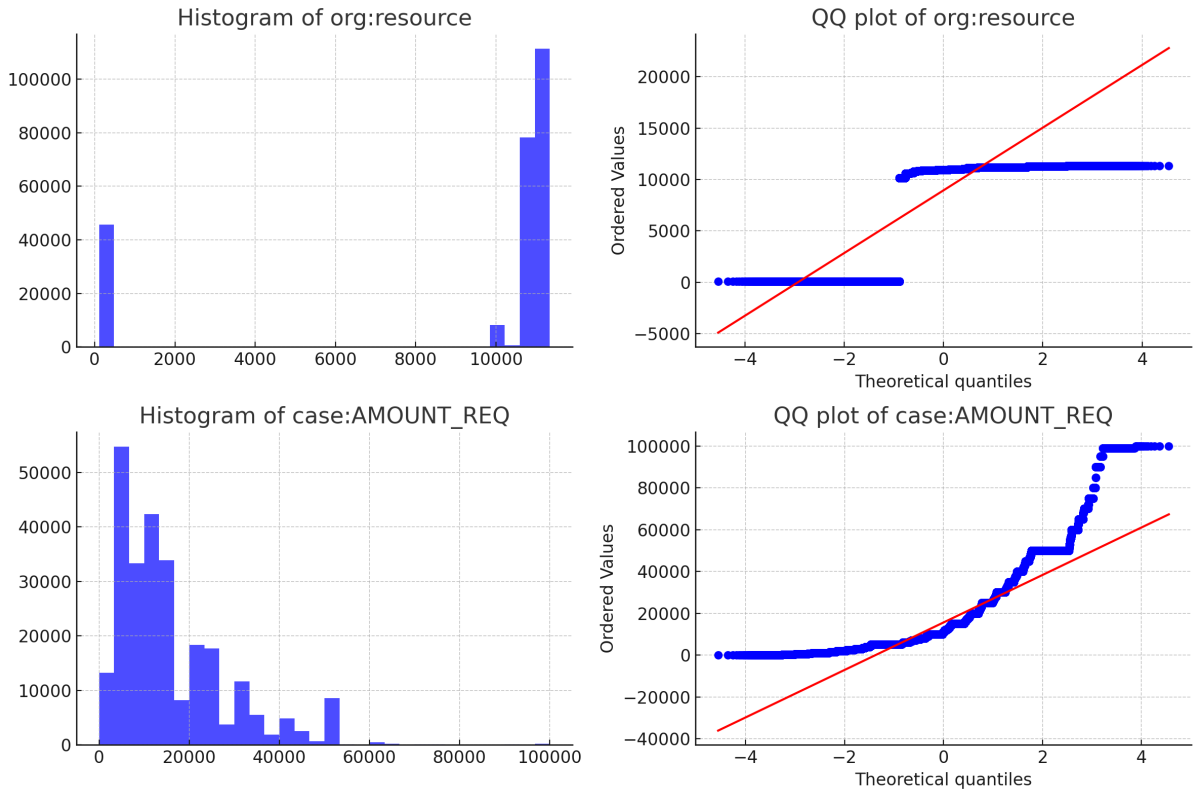

Figure 4.6: Skewness of BPI-2012 dataset after applying the Shapiro-Wilk test

4.4.2 Analysing fig 4.6

Bimodality in org: The org:resource variable exhibits a bimodal distribution, suggesting that resources are either minimally or heavily utilized, with little middle ground. This is evident from both the histogram and the Q-Q plot.
Skewness in case: The *case:AMOUNT_REQ* variable shows clear right skewness, with most data points concentrated on lower amounts and a long tail of large requested amounts.
Non-Normality: Non-Normality: Both Q-Q plots demonstrate that neither of the variables follows a normal distribution, as the data points diverge significantly from the red line, indicating violations of normality assumptions.

4.4.3 General Discussion

From the histograms and Q-Q plots of both the Sepsis and BPI-2012 datasets, it is clear that neither dataset follows a normal distribution, as demonstrated by their skewness and deviations from the expected normal quantiles.

Chapter 5

Conclusion and Future Work

5.1 Summary of Findings and Future Work

1. **Performance of Non-Parametric Estimators**: In this study, it was very evident that no single non-parametric estimator outperformed the others across all the species definitions and datasets. Most of the estimiators are particularly sensitive to rare species, especially Chao2. This study proved that Chao2 may not be the most reliable estimator, since the accuracy decreases drastically as the number of rare species (singletons and doubleton) increase. This result was quite evident in the Trace Variant Species Definition. Chao2 was outperformed by almost all the estimators.

2. **Effectiveness of Jackknife 2nd Order**: Jacknife 2nd order estimator delivered the best results overall, however it was prone to underestimating the species estimation when the size of the dataset was small, suggesting that its efficiency is dependent on size. Which further proves our point that no estimator works universally well across all the datasets.

3. **Species Retrieval Function Insights**: The idea of introducing Species Retrieval Function to event logs is insightful, novel and promising. However, it requires further refinement/improvement.

4. **Potential for Parametric Estimators**: While parametric estimators show promise in certain scenarios, they are not universally applicable. Our attempts to apply Parametric Estimators on the dataset proved to unsuccessful, since the data was heavily skewed. Future work could include inducing a certain type of transformation to the dataset.

5. **Importance of Dataset-Specific/ Domain Specific Retrieval Functions**: This study applied the same definition to the same events of the dataset ignoring other events which be the reason for low completeness. Future work could include designing retrieval functions based on the specific data properties, which could lead to more accurate estimates. Event logs from different datasets often contain events that vary in frequency, importance, or structure. For example, some events may occur rarely but are critical to understanding the complete behavior of the system, while others might be frequent but contribute less informational value. A generalized retrieval function might not differentiate between these variations, treating all events similarly

6. **Using Samples to gain further insights about the estimator**: Using samples to assess the stability an estimator, seems to work however future work could include testing es-

timators on more sample sizes. By drawing more number of samples, we can calculate Variance, Bias, Mean Squared Error, Confidence Intervals etc.

7. **Need for Robust Evaluation Metrics**: Finally, there is a clear need for improved evaluation metrics that can better assess the true performance of different estimators. The current metrics fall short in providing comprehensive insights into estimator efficacy, and future research should focus on developing more robust and reliable methods of evaluation.

Bibliography

[1] Martin Kabierski, Markus Richter, and Matthias Weidlich. *Addressing the Log Representativeness Problem using Species Discovery*. 2023.

[2] Unknown. *Bias and Confidence in Not-quite Large Samples*. The Annals of Mathematical Statistics, 1958.

[3] B. van Dongen. *BPI Challenge 2012*. 2012. Available at: `https://data.4tu.nl/articles/dataset/BPI_Challenge_2012/12689204/1`.

[4] B. van Dongen and F. F. Borchert. *BPI Challenge 2018*. 2018. Available at: `https://data.4tu.nl/articles/dataset/BPI_Challenge_2018/12688355/1`.

[5] B. van Dongen. *BPI Challenge 2019*. 2019. Available at: `https://data.4tu.nl/articles/dataset/BPI_Challenge_2019/12715853/1`.

[6] F. Mannhardt et al. *Sepsis Cases-Event Log*. Eindhoven University of Technology, vol. 10, 2016.

[7] 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, 2017, pp. 3–54.

[8] J. Stoklosa, R. V. Blakey, and F. K. C. Hui. *An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity*. Diversity, vol. 14, no. 5, 2022, p. 320. DOI: `10.3390/d14050320`.

[9] C. H. Chiu, Y. T. Wang, and A. Chao. *Good-Turing Frequency and Generalized Poisson Mixture Models for Species Estimation*. PeerJ, 2014. DOI: `10.7717/peerj.14540`.

[10] J. Hinde and C. G. B. Demétrio. *Overdispersion: Models and Estimation*. Computational Statistics & Data Analysis, vol. 27, no. 2, 1998, pp. 151–170. DOI: `10.1016/S0167-9473(98)00007-3`.

[11] M. H. Quenouille. *Approximate Tests of Correlation in Time Series*. Journal of the Royal Statistical Society, vol. 11, no. 1, 1949, pp. 68–84.

[12] S. K. Thompson. *Sampling*. 3rd ed., Wiley, 2012.

[13] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.