



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

An Integrated Framework for Estimating the Number of Classes with Application for Species Estimation

A Dissertation Submitted in Partial Fulfilment of the Requirements
for Doctor of Philosophy in Statistics

Asmaa A.M. Al-Ghamdi

Supervised by

Prof. Simon Wilson

Department of Statistics – School of Computer Science & Statistics

Trinity College Dublin – The University of Dublin

January 2021

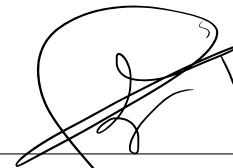
Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

The copyright belongs jointly to the 'Trinity College Dublin, the University of Dublin' and Asmaa A. M. Al-Ghamdi.

Signature

A handwritten signature in black ink, appearing to be 'Asmaa A.M. Al-Ghamdi', written over a horizontal line.

Asmaa A.M. Al-Ghamdi

October 20, 2020

Abstract

The two most common approaches for estimating the number of distinct classes within a population are either to use sampling data directly with combinatorial arguments or to extrapolate historical discovery data. However, in the former case, such detailed sampling data is often unavailable, while the latter approach makes assumptions on the form of parametric curves used to fit the discovery data, that are often lacking in theoretical justification. Instead, we propose an integrated transdisciplinary framework that dissolves the boundaries between the above two approaches. This is achieved by directly describing the sampling-discovery process in parallel with describing a co-variate latent effort process, where we have historical discovery data for the former process and some proxy data for the latent process. The linkage between these two processes allows one to form data on sampling records by forcing some constraints on how many samples were taken over time. Due to the nature of the constrained data, many inference techniques become infeasible. However, simulation-based methods such as Approximate Bayesian Computation remain available. Our proposed approach is demonstrated and analysed through many simulation experiments, and finally applied in the ecology field to estimate the number of species as an example of the number of classes problem.

Acknowledgements

Doing my PhD research has been a unique and rewarding journey. I am grateful to all the circumstances that led me to this intellectual point in my life that reflected on my academic knowledge and personality. I am grateful to my family; my precious parents and sisters whose care and support has been always infinite.

Special appreciation and gratitude go to my supervisor Prof. Simon Wilson. He has been very supportive, always responding to my questions and queries so promptly. He has been very generous in his advice and guidance until the final word of my thesis. I feel fortunate to be supervised by him.

My sincere thanks go to Prof. Mark Costello, who provided us with precious insights and real data resources to implement our proposed model. Many people have generously contributed their time and resources to accomplish the experimentation part of my thesis; I am indebted to them all.

Last, I would like to pay my gratitude and respect to my previous supervisor Dr. Brett Houlding who passed away in September 2019. He was a great source of support to me in TCD college; he was a caring supervisor who encouraged me to engage in academic activities.

Contents

Declaration	i
Abstract	ii
Acknowledgements	iii
List of Figures	viii
List of Tables	xix
List of Abbreviations	xx
List of Notations	xxii
1 Introduction	1
1.1 Generic Background	1
1.2 Problem Statement and Scope	2
1.3 Purpose and Methodology	4
1.4 Motivations within Species Context	4
1.5 Contribution of this Dissertation	8
1.6 Outline of Dissertation Chapters	8
I Theoretical Phase of The Proposed Approach	11
2 Literature Review	12
2.1 Methods of Estimating Number of Classes	12
2.1.1 Sampling-Theoretic Approach	12
2.1.2 Data-Analytic Approach	19
2.2 Methods of Estimating Number of Species	20
2.2.1 Extrapolation Based on Some Correlations	21
2.2.2 Extrapolation Based on Temporal Component	23

2.3	Bayesian Inference and Techniques	28
2.3.1	Bayesian Estimation	30
2.3.2	Bayesian Network	32
2.3.3	Bayesian Computation Techniques	34
2.3.4	Approximate Bayesian Computation	37
2.4	Chapter Summary:	41
3	The Proposed Approach	42
3.1	Model Formulation	43
3.1.1	The Sampling-Discovery Process	44
3.1.2	The Latent Effort Process	46
3.1.3	Linkage Node of the Two processes	47
3.1.4	Final Phase of the Proposed Model	50
3.2	Model Selection and Activation	51
3.2.1	Distributional Assumptions	52
3.2.2	Inferential Methods	54
3.2.3	The Adopted ABC Algorithm	58
3.3	Chapter Summary:	61
II	Implementation Phase of The Proposed Approach	62
4	Model Formulation & Simulation	64
4.1	Distributional Specifications	64
4.2	Simulation and Generic Characteristics	69
4.2.1	Fitting Artificial World-A:	74
4.2.2	Fitting Artificial World-B:	77
4.2.3	Fitting Artificial World-C:	79
4.2.4	Investigating Standardized Distance Metric:	80
4.3	Chapter Summary:	88
5	Model Performance & Evaluation	89
5.1	Model Accuracy	89
5.1.1	Accuracy in World-A Design:	90
5.1.2	Accuracy in World-B Design:	92
5.1.3	Accuracy in World-C Design:	96

5.2	Model Sensitivity	98
5.2.1	Sensitivity in World-A Design:	98
5.2.2	Sensitivity in World-B Design:	100
5.2.3	Sensitivity in World-C Design:	101
5.3	Model Additivity	103
5.3.1	Additivity in World-A Design:	107
5.3.2	Additivity in World-B Design:	116
5.3.3	Additivity in World-C Design:	125
5.3.4	Additional Attempts in World-A & World-B:	135
5.4	Chapter Summary:	137
6	Model Application on Real Data	140
6.1	Model Fitting for CoL Data	141
6.1.1	Challenges in Model specifications	142
6.1.2	Challenges in Model Fitting	143
6.1.3	Fitting Results of CoL	145
6.2	Model Chronological Analysis	146
6.3	Model Validation	149
6.3.1	Cross-sectional Validation	150
6.3.2	Longitudinal Validation	152
6.4	Model Fitting for WoRMS Data	159
6.4.1	Model Parametrizations	160
6.4.2	Fitting Results of WoRMS	161
6.5	Global Estimation	163
6.6	Comparing with Previous Work	165
6.6.1	Brief Description of NHRP Model	165
6.6.2	Our Model Results vs NHRP Model Results	166
6.7	Chapter Summary	173
7	Discussion and Conclusions	175
7.1	Summary of the Study Analysis & Synthesis	175
7.2	Main Findings of the Study Evaluation	177
7.2.1	Main findings from the artificial data experiments	178
7.2.2	Main findings from the real data experiments	180
7.3	Discussion & Suggestions for Future Work	182

A	Mathematical Equations	201
A.1	Some Equations from the Literature Review	201
A.2	Equation Explanation of the Proposed Model	202
A.3	Explanation of the Species Abundance Variation Provided by [87]	204
B	Groups of Tables and Plots	206
B.1	Simulation Results Related to Chapter 4	207
B.2	Accuracy Results Related to Chapter 5	208
B.3	Sensitivity Results Related to Chapter 5	212
B.4	Additivity Results Related to Chapter 5	214
B.5	Results Related to Chapter 6	222
B.6	Distance Metric Results	231
C	Pseudo Codes	232
C.1	Artificial-Data Algorithm	233
C.2	Real-Data Algorithm	233

List of Figures

1.1	In a broad sense, the hierarchy of the motivations for biodiversity (discovery and maintenance). It heads up with the need of knowing versus ethical responsibilities, and ends up with the basic needs of our survival, moving through the fact of the global safety issues.	5
1.2	The balance mechanism of the ecosystem that runs within higher and lower levels of the biodiversity. This process can be maintained if we keep sustaining the biodiversity.	6
1.3	The Structure of the Dissertation	10
2.1	A holistic view of the existing approaches of estimating number of classes along with the proposed approach. Most of the attempts in the species context are under the data-analytic approaches.	13
2.2	Examples of graphical models, presenting the <i>cyclic/acyclic</i> and <i>directed/undirected</i> properties of the graphical models.	33
2.3	A holistic view of the most commonly used techniques of the Bayesian computation that are used for the integration and marginalization problems.	35
3.1	The left hand side plot shows the Black-box concept of the number of kinds/classes problem in the existing literature, while the right hand side shows this concept as proposed in our framework.	42
3.2	An example of a sampling-discovery sequence, with total sampling $\sum_{j=1}^3 n_j = 30$, and total discoveries $\sum_{j=1}^3 \tau_j = 8$ over three years $T = 3$	45
3.3	A Bayesian network visually displaying the probabilistic relationships within the sampling-discovery process. The square node refers to available data, while circle nodes correspond to unknown variables.	46

3.4	A Bayesian network visually displaying the probabilistic representation of the latent effort process and its proxy. The square node refers to available data, while the circle refers to unknown data that needs to be estimated.	47
3.5	The left hand side plot shows the overlap between the integrated sampling-discovery and latent effort processes, and the set of factors they both include. The right hand side shows the black-box concept in its final stage as proposed in our framework.	49
3.6	A Bayesian network visually displaying the whole structure of the proposed model, and its probabilistic causal relationships. The model is designed under the constraints $d_i \leq \sum_{r=1}^j n_r - \delta_{i-1}$, where $d_i = \delta_i - \delta_{i-1}$. Note that as a result $t_i = j \times \mathbb{1}(\delta_i \leq \sum_{r=1}^j n_r)$	51
3.7	A chart of describing the range of model choices, which ranges from narrow (when we have actual datasets to guide the distributional assumptions) to broad (when the situation is subject to expert of the field). In the absence of concrete data (as in the middle), the theoretical logic is still good support for model choice.	53
3.8	A diagram to simplify the idea of implementing the RJ-MCMC.	55
3.9	A diagram of simultaneously implementing the ABC rejection sampling algorithm. $\mathcal{D}(\{\boldsymbol{\tau}_i^*, \mathbf{x}_i^*\}, \{\boldsymbol{\tau}, \mathbf{x}\})$ is a notation for measuring the closeness between the artificial datasets and the actual given ones.	57
4.1	The data subset from the WoRMS and the CoL inventories, after including the proxy variable. These represent the number of discovered species $\{\tau_j\}_{j=1:240}$ (grey) and the number of authors (proxy) $\{x_j\}_{j=1:240}$ (orange), from 1761 to 2000.	66
4.2	A Bayesian network displaying our proposed model within the species' distributional assumption.	68
4.3	The annual number of discoveries (Black) and the number of authors (Green) over 250 years, along with basic information about the three artificial worlds.	70

4.4	The huge difference between the Nested-loops scheme and the last three schemes in the computation time (seconds) measured on a log scale. However, there is still a considerable difference among the last three.	72
4.5	Scatter plots of the pilot experiments of the three artificial worlds, that show the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$. . .	73
4.6	Scatter plots of the main experiments of the three artificial worlds. The left panel shows the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances out of the sum $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$, while the right panel shows the correlation between distances of the number of discoveries $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$	74
4.7	Scatter plots of the resultant final sample in the three artificial worlds. The left panel shows the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances out of the sum $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$, while the right panel shows the correlation between distances of the number of discoveries $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$	75
4.8	Fitting results of the World-A.	76
4.9	Fitting results of the World-B.	78
4.10	Fitting results of the World-C.	80
4.11	Fitting results of World-A, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.	82
4.12	Scatter plots of the main experiments of fitting World-A, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.	83

4.13	Fitting results of World-B, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.	84
4.14	Scatter plots of the main experiments of fitting World-B, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.	84
4.15	Fitting results of World-C, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.	85
4.16	Scatter plots of the main experiments of fitting World-C, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.	85
4.17	Scatter plots of the main experiments of fitting the three artificial worlds (focusing on the accepted final area and ignoring the rest of x-axis and y-axis). The left panel shows a zoom-in view of the accepted final sample, while the right panel shows a zoom-out view where the thresholds lines are appeared. The green vertical line represents the 1st version threshold, while the blue vertical line represents the standardized threshold.	87
5.1	Fitting artificial World-A for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify there position relative to the true value $C = 10,000$	91

5.2	Fitting 500 artificial worlds of design World-A with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error that results in the adopted number of iterations. The left top is the distribution of the posterior means, while left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.	92
5.3	Fitting artificial World-B for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify their position relative to the true value $C = 10,000$	93
5.4	Fitting 500 artificial worlds of design World-B with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while the left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.	94
5.5	Fitting 300 artificial worlds of design World-B with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while the left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.	95

5.6	Fitting artificial World-C for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify their position relative to the true value $C = 10,000$	96
5.7	Fitting 500 artificial worlds of design World-C with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.	97
5.8	Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-A. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.	99
5.9	Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-B. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.	101

5.10	Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-C. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.	102
5.11	The resulting data of the Splitting Scheme of World-A.	107
5.12	The fitting stage in the Splitting Scheme of the two taxa in World-A.	109
5.13	The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.	109
5.14	The resulting data of the Merging Scheme - Treatment (1) of World-A design.	110
5.15	The fitting stage in the Merging Scheme - Treatment (1) of the two taxa in World-A.	111
5.16	The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.	112
5.17	The resulting data of the Merging Scheme - Treatment (2) of World-A design.	113
5.18	The fitting stage in the Merging Scheme - Treatment (2) of the two taxa in World-A.	113
5.19	The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.	114
5.20	The generated data and the results of the fitting stage of the Merging Scheme - Treatment (3) of World-A design. The left panel belongs to taxon M1_T(3), while the right panel belongs to taxon M2_T(3).	115

5.21	The aggregating stage of the Merging Scheme - Treatment (3) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.	116
5.22	The resulting data of the Splitting Scheme of World-B.	117
5.23	The fitting stage in the Splitting Scheme of the two taxa in World-B.	118
5.24	The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.	118
5.25	The resulting data of the Merging Scheme - Treatment (1) of World-B design.	119
5.26	The fitting stage in the Merging Scheme - Treatment (1) of the two taxa in World-B.	120
5.27	The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.	121
5.28	The resulting data of the Merging Scheme - Treatment (2) of World-B design.	122
5.29	The fitting stage in the Merging Scheme - Treatment (2) of the two taxa in World-B.	122
5.30	The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.	123
5.31	The generated data and the results of the fitting stage of the Merging Scheme - Treatment (3) of World-B design. The left panel belongs to taxon M1_T(3), while the right panel belongs to taxon M2_T(3).	124
5.32	The aggregating stage of the Merging Scheme - Treatment (3) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.	125
5.33	The resulting data of the Splitting Scheme of World-C.	126
5.34	The fitting stage in the Splitting Scheme of the two taxa in World-C.	127

5.35	The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.	127
5.36	The resulting data of the Merging Scheme - Treatment (1) of World-C design.	129
5.37	The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.	130
5.38	The resulting data of the Merging Scheme - Treatment (2) of World-C design.	131
5.39	The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.	132
5.40	The generated data and the fitting results of the Merging Scheme - Treatment (3) of World-C.	132
5.41	The aggregating stage of the Merging Scheme - Treatment (3), first trial, showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.	133
5.42	The aggregating stage of the Merging Scheme - Treatment (3), second trial, showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.	135
5.43	The aggregating stage of the Merging Scheme - Treatment (1) with 50% & 50% taxa of World-A, showing the final results of fitting the whole category and the combined fitting of the two taxa.	136
5.44	The aggregating stage of the Merging Scheme - Treatment (1) with 50% & 50% taxa of World-B, showing the final results of fitting the whole category and the combined fitting of the two taxa.	137
6.1	Two refined non-overlapping subsets from WoRMS (left) and CoL (right) inventory systems, over 240 years from 1761 to 2000, showing a big difference in the discovery and effort scales between the two subsets.	141
6.2	Fitting results of CoL dataset.	146

6.3	Plots of fitting results of CoL dataset over the first 150 years (Red), over the first 185 years (Blue), and over the whole period which is 240 years (Green).	148
6.4	Fitting results of CoL dataset, the top panel belongs to the original fitting, while the bottom panel belongs to the point-estimate fitting.	151
6.5	Distributions properties of the Euclidean distances (as a residual diagnostic) of the original fitting and the point-estimate fitting of the CoL dataset.	152
6.6	Fitting results of CoL, the top panel belongs to the ground-truth given C_{240}^* , while the middle and bottom panel belongs to the (FT.1) and (ST.1) given C_{150}^* , respectively. Note that the green lines in the bottom panel represent the 95% credible interval. . . .	154
6.7	Distributions properties of the Euclidean distances (as a residual diagnostic) of the ground-truth fitting given C_{240}^* , and the (FT.1) and (ST.1) fittings given C_{150}^*	155
6.8	Distributions properties of the Euclidean distances (re-computed only on the last 90 data points of the generated datasets) of the ground-truth fitting given C_{240}^* , and the (FT.2) and (ST.2) fittings given C_{150}^*	156
6.9	Fitting results of CoL, the top panel belongs to the ground-truth given C_{240}^* , while the middle and bottom panel belongs to the (FT.1) and (ST.1) given C_{185}^* , respectively. Note that the green lines in the bottom panel represent the 95% credible interval. . .	157
6.10	Distributions properties of the Euclidean distances (as a residual diagnostic) of the ground-truth fitting given C_{240}^* , and the (FT.1) and (ST.1) fittings given C_{185}^*	158
6.11	Distributions properties of the Euclidean distances (re-computed only on the last 55 data points of the generated datasets) of the ground-truth fitting given C_{240}^* , and the (FT.2) and (ST.2) fittings given C_{185}^*	158
6.12	Similarity between the WoRMS and the CoL datasets in the discovery curves.	160
6.13	Fitting results of the CoL dataset.	162

6.14	Global estimation results out of aggregating the estimation results of CoL and WoRMS datasets.	164
6.15	Our Model results (left panel) vs the NHRP Model results (right panel) of fitting the CoL and WoRMS datasets. The left panel shows the estimated posterior distributions, along with their 95% credible intervals (red lines). In the right panel, the red lines represent the annual cumulative number of discovery curves, the black solid lines represent the median fitting curves, and the black dashed lines represent the estimated 95% credible intervals.	169
7.1	The three aspects of our proposed model. The solid arrows represent the static decisions that should be taken at the beginning of the modelling process, while the dashed arrows represent the dynamic decisions that are taken for tuning the model and enhancing its performance. Note that the semi-static decisions appear in a situation where we already specified the nature and number of the thresholds but for some specific occasions we need an additional filtering step.	183

List of Tables

4.1	Parametrization set of three generated artificial worlds.	69
4.2	Artificial World-A of total number of species $C = 10,000$	76
4.3	Artificial World-B of total number of species $C = 100,000$	77
4.4	Artificial World-C of total number of species $C = 50,000$	79
5.1	Under/over-parametrization within artificial World-A.	99
5.2	Under/over-parametrization within artificial World-B.	101
5.3	Under/over-parametrization within artificial World-C.	102
5.4	Parametrization of whole category vs. two taxa of World-A.	108
5.5	Parametrization of whole category vs. two taxa of World-B.	117
5.6	Parametrization of whole category vs. two taxa of World-C.	126
6.1	Parametrization and resulting statistics of modelling CoL dataset.	145
6.2	Statistics of fitting results of CoL dataset over the first 150 years, the first 185 years, and the whole period which is 240 years.	147
6.3	Parametrization and resulting statistics of modelling WoRMS dataset.	161
6.4	Estimation results of our proposed model vs. NHRP model.	168
6.5	Examples of miscellaneous species estimates (in millions) of differ- ent taxa levels.	171
6.6	Examples of global species estimates (in millions).	172

List of Abbreviations

<i>ABC</i>	Approximate Bayesian Computation methods
<i>ABC – MCMC</i>	Approximate Bayesian Computation with Markov chain Monte Carlo sampling
<i>ABC – PMC</i>	Approximate Bayesian Computation with Population Monte Carlo sampling
<i>ABC – PRC</i>	Approximate Bayesian Computation with Partial rejection control sampling
<i>ABC – RS</i>	Approximate Bayesian Computation with rejection sampling
<i>ABC – SMC</i>	Approximate Bayesian Computation with Sequential Monte Carlo sampling
<i>ARIMA</i>	Auto-Regressive Integrated Moving Average models
<i>CCD</i>	central composite design integration method
<i>CDF</i>	Cumulative Distribution Function
<i>CoL</i>	Catalogue of Life species inventory system
<i>CV</i>	Coefficient of Variation metric
<i>DAG</i>	Directed Acyclic Graph
<i>ED</i>	Euclidean Distance metric
<i>EP</i>	Expectation Propagation method
<i>ERMS</i>	European Register of Marine Species inventory system
<i>INLA</i>	Integrated nested Laplace approximations method
<i>KLD</i>	Kullback-Leibler Divergence metric
<i>MC</i>	Monte Carlo sampling
<i>MCMC</i>	Markov chain Monte Carlo methods

<i>NHPP</i>	Non-Homogeneous Poisson Process
<i>NHRP</i>	Non-Homogeneous Renewal Process
<i>PDF</i>	Probability Density Function
<i>PMF</i>	Probability Mass Function
<i>RJ – MCMC</i>	Reversible-Jump Markov Chain Monte Carlo sampling
<i>VB</i>	Variational Bayes methods
<i>WoRMS</i>	World Register of Marine Species inventory system

List of Notations

N	Total population size of a target community
N_i	Abundance of a class i in a target community
C	Total number of classes/species
C^*	A point estimate of the total number of classes/species
D	Total number of discoveries
t_i	Time point of a discovery event of a class i
d_i	Number of inter-discovery samples until facing a new discovery event of a class i
δ_i	Location point to hold the index of a discovery event of a class i
τ_j	Number of discoveries at time unit j
n_j	Sample size drawn at time unit j
L_j	latent taxonomic efforts at time unit j
l_j	latent taxonomic efforts on a log scale, at time unit j
x_j	A proxy of latent taxonomic efforts (<i>e.g.</i> number of authors at time unit j)

Chapter 1

Introduction

1.1 Generic Background

In many life situations we would be interested in estimating the number of items within pre-specified classes and how these items are distributed among these classes, such as the number of students over grading scale, or the number of children over IQ levels. However, there are other situations where we would be interested in estimating the number of classes themselves within known or unknown sizes of population.

The problem of estimating the number of distinct classes, also known as ‘*the number of kinds problem*’, is an important statistical challenge and has been an active area of study for a long time, as first dated to the 1713 publication of Jacob Bernoulli’s *Ars Conjectandi*. At its most basic level, it is usually presented as the problem of estimating the number of different colours of balls in an urn. The question being: as balls are removed from the urn and their colours observed and recorded, how many different yet unseen colours remain?

Moving beyond this thought experiment, the problem has a broad range of practical applications. For example, in ecology, maintaining biodiversity has been a critical issue so that estimating the number of species in a given community became an important tool to serve this goal [1] [2]. In reliability, it is important to estimate the types of defect that could be faced during product testing operation. This should be important information for the quality control and reliability department in any given factory. One example of this field is software engineering, where releasing new software in the market is expected to be combined with some technical problems. Thus, estimating the kinds of bugs that

could be reported is important information in order to enhance the performance of the software [3] [4] [5]. In psycholinguistics or sociolinguistics, a linguist may be interested in estimating number of vocabularies that a famous author like Shakespeare knew. This is not because they are necessarily interested in the vocabulary knowledge of the great bard himself, but because answers to such types of questions help us to understand better how languages develop or evolve [6] [7].

As there are a variety of ways in which the number of kinds/classes problem may be posed, there are different formats of methods for attempting its solution. With reference to the urn example, differing assumptions concerning the nature of the urn and/or the form of data recorded will determine the approach taken. For example, does the urn contain an infinite number of balls, or only a finite (either known or unknown) number? In terms of possible information available, we could have access to the full sample sequence in which the i -th colour was observed, or it could possibly be that only the number of balls sampled and the number of colours within that sample is recorded, *etc.* Most methodologies, in general, can be labelled under one of two approaches, namely the ‘*Sampling-Theoretic Approach*’ which derives a solution through combinatorics and sampling theory by focusing on the ‘*sampling action*’, and the ‘*Data-Analytic Approach*’ that seeks to extrapolate data by considering the ‘*discovery curve*’ of the classes [8]. However, many of these methods are associated with assumptions that restrain them from being practical or generic approaches over different types of applications. Yet, the number of classes area is still a subject of argumentation and investigation.

1.2 Problem Statement and Scope

Looking closely at the above two approaches, we found that sampling methods tackle the problem from a cross-sectional point of view, while those considering discovery rates extend this perspective to include longitudinal analysis. Again, using the coloured balls example, the former approach will give focus to counting the sampled balls as they are drawn and their colours determined and employ results within combinatorics along with related assumptions (*e.g.* the distribution of colour abundance) to estimate the number of undiscovered colours [9].

Alternatively, methods focusing on a discovery curve assume the existence of a pattern in the colour discovery process to extrapolate the expected number of the remaining colours yet to be seen. The benefit of this is that predictions can be made without having access to the data concerning numbers of samples taken over time [10]. This pattern may be based on a temporal behaviour or a correlation with some attributes of the classes which make them more or less likely to be sampled. For example, the size of the coloured ball in which the likelihood of a ball being sampled will be affected by its size, especially if differing sizes are associated with specific colours [11] [12].

However, none of the above approaches seem to be generic or realistically practical over different applications. In the sampling methods case such detailed sampling data is often unavailable, while the discovery curve methods make assumptions on the form of parametric curves used to fit the discovery data that is often lacking in theoretical justification. An additional common drawback is that the attention in all these methods is concentrated only on one main variable, the number of classes, without considering other covariates in the background that influence the variable of interest.

In the current study we focus our attention in solving the number of kinds/classes problem in the ecology field, particularly in biodiversity. The most practical and commonly used indicator for speculating the extent of biodiversity is estimating ‘*the number of species*’ and measuring the uncertainty around it. Methods in this field covered both the sampling-theoretic approach and the data-analytic approach, and inherit the same drawbacks as mentioned above. Most of them are performed on a limited scale such as localised geographical levels or a specific taxon while, on a global scale, estimates have collectively ranged from 1.5 to over 100 million species on Earth [13]. This significant gap indicates a lack of integration and consistency in the conclusions of the estimation methods, and provokes the need to seek more applicable and adequate alternative approaches.

1.3 Purpose and Methodology

The current study proposes an integrated transdisciplinary approach for estimating the number of classes, specifically implemented for estimating the number of species. It involves formulating a novel model, along with the use of Bayesian inference, and implemented by using Approximate Bayesian Computation (ABC). It represents a framework that aims to bridge the gap between the sampling action and discovery curve approaches. In doing so, we are able to utilise assumptions that are more realistic and underpin the combinatorial arguments of the sampling methods within a restricted form of data that is seen in discovery curves.

This is achieved by introducing the concept of a domain-specific ‘*latent effort*’ behind the sampling and discovery process. While we may not be able to record precisely how many samples had been taken, we make use of available information concerning the effort put into such sampling, for example, we may have knowledge of the number of person-hours employed in the sampling-discovery and how that altered over time. This then allows us to model a system of inter-relationships occurring in the environment of the class discovery process by including a covariate - the latent effort process - that could be described by how it is related to other observable records we obtain. For example, in software reliability, a latent effort in seeking to find software bugs could be connected to the number of users or licence holders registered with the software over the time that faults are being reported [14]. In biodiversity, the latent effort in seeking to discover new species could be connected to the number of active taxonomists, *i.e.*, the number of authors involved in species description publications [13], the amount of research funding awarded, or the number of field excursions performed over time, *etc.* In general, any observable record that is somehow related to the true but latent effort process in sampling observations will be sufficient as a proxy.

1.4 Motivations within Species Context

Speculating about the extent of biodiversity is of critical importance if we wish to ensure its perseverance. It is not only to serve our planet, but also to serve human-

ity, as maintaining biodiversity is required for sustaining our existence. Human motives to maintain biodiversity vary from those of high moral levels to those of basic human needs. In fact, they can be categorized in a similar way to *Maslow's Hierarchy* of needs in his theory of human motivation [15] [16], shown in Figure 1.1.

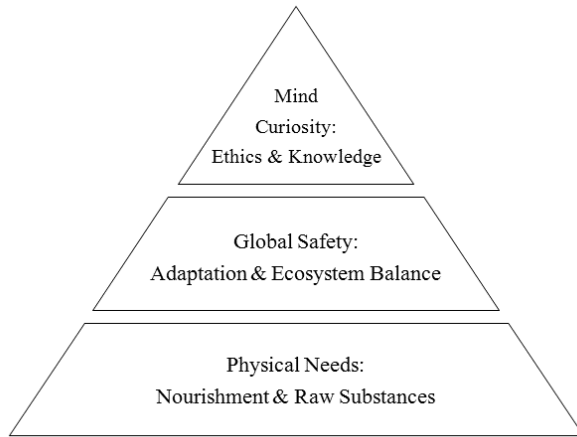


Figure 1.1: In a broad sense, the hierarchy of the motivations for biodiversity (discovery and maintenance). It heads up with the need of knowing versus ethical responsibilities, and ends up with the basic needs of our survival, moving through the fact of the global safety issues.

Starting from the higher level of the hierarchy in Figure 1.1, it may require a profound and thorough rationalisation to realize the significance of maintaining biodiversity. Biodiversity is a natural outcome of ongoing bio-interaction on Earth. Over 3 billion years since the emergence of life, the Earth as an ecosystem has been patronizing such a creative interaction. Although we as humans are the most intelligent and productive creatures on Earth, we are still miniscule in terms of this great creativity of life. Our slow research and discovery in the field and our detrimental behaviour towards the ecosystem seem to ignore the importance of biodiversity to life on Earth. A responsible reaction would instead be to save it from our own negative influence.

Ethical implications aside, biodiversity is a subject of scientific research, an interesting area to invest our rational thinking and problem-solving strategies upon. Delving into the study of biodiversity satisfies our curiosity and expands our knowledge about our surroundings. Such investment is our way to fulfil the need for knowledge and understanding (as referred to by Maslow's hierarchy [15]) as this gives us the power to adapt, develop, and control. The implications are further clarified in the other two levels discussed in the following paragraphs.

The middle level of the hierarchy, Figure 1.1, highlights how biodiversity enriches the ecosystem to make it able to adapt to changing conditions; such adaptation has been necessary since the emergence of life. In other words, biodiversity is a critical balancing mechanism of our ecosystem, involving a cyclical process that passes through genetic diversity and natural selection, as shown in Figure 1.2.

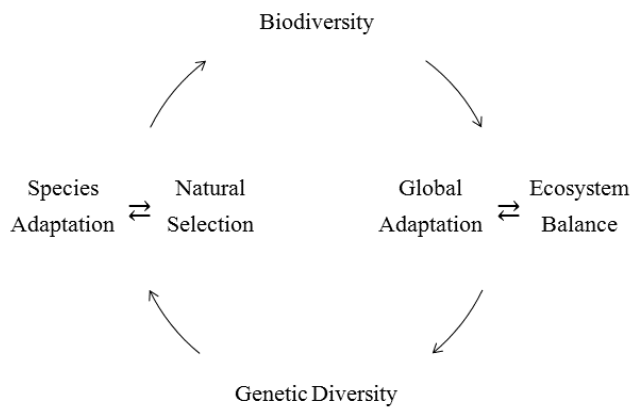


Figure 1.2: The balance mechanism of the ecosystem that runs within higher and lower levels of the biodiversity. This process can be maintained if we keep sustaining the biodiversity.

Genetic diversity empowers species with a variety of inheritable traits which allows them to adapt in adverse circumstances, thus qualifying species for natural selection. Natural selection promotes biodiversity which in turn empowers the ecosystem with adaptive features so as to maintain balance. A balanced ecosystem, in turn, represents a rich habitat and a driving force for genetic diversity. However, one species in the ecosystem, the human species, has played a negative role in the natural evolutionary cycle by accelerating deteriorative effects. Considering this state of influence, it is challenging for the ecosystem to continue in its natural rate of adaptation. It is thus crucial to preserve biodiversity, including genetic diversity, in order to ensure proper adaptation for rapid changes in the ecosystem.

An important issue affecting the ecosystem's balance is global warming caused by the ongoing increases in the levels of CO_2 in the atmosphere. This has been the case since the industrial revolution. However, the good news is that biodiversity can have a great contribution in controlling biogeochemical cycling of gases. For example, marine biodiversity aids in controlling CO_2 in the oceans.

One particular marine species, Phytoplankton (a microscopic plant-like organism that makes its own food from sunlight through photosynthesis), plays its part by removing CO_2 from the surface of the oceans by absorbing it and releasing O_2 . Moreover, when Phytoplankton die they sink to the bottom of the oceans keeping the CO_2 far away from the atmosphere. This is known as a *Biological Pump* [17]. Hence, neglecting or misbehaving in preserving marine biodiversity affects the efficiency of the Biological Pump in balancing the biogeochemical cycling of gases.

At the lowest level of the hierarchy of Figure 1.1, and taking marine biodiversity as an example, seas and oceans host enormous kinds of resources related to medication, food and nourishment, as well as raw materials obtained from coral rock and sand used for building materials *etc.* The high biochemical diversity that the marine world has, makes it a substantial source for manufacturing pharmaceuticals. For example, the oceans' sessile animals produce chemicals to defend themselves, from which pharmaceuticals can be made, as is the case with sessile land plants. As a further example, seaweed provides Polysaccharide, an important substance for humans which can also be used as food for livestock.

Concerning nourishment, finfish, shellfish, invertebrates, algae *etc* represent a great source of nutrients and protein as low calorie food for a healthy life style. It is quite ironic how a wide range of the oceans, rich in these resources, are not really utilized. Conserving marine biodiversity and promoting proper awareness offers opportunities to utilize these precious resources and potentially balance fishing practices throughout oceans. [18] [19] [1]

To conclude, every species in the biodiversity system has its important role on Earth. Hence, it is an obligation on us to maintain biodiversity. This can be achieved through a proactive approach towards species discovering, understanding, taxonomizing, and respecting their space on Earth. The current study serves in satisfying this achievement.

1.5 Contribution of this Dissertation

The contribution of the current study can be summarized in the following aspects:

- Introducing a novel framework in response to solving the number of kinds/classes problem.
- Developing the proposed model for generic implementation, and particularly within the ecological field (estimating the number of species).
- Implementing probabilistic inference for the proposed model with an Approximate Bayesian Computation approach.
- Exploring the properties of the proposed model and implementation through multiple simulation experiments, as well as model validation.
- Applying the proposed model to real ecological datasets and providing conclusions that may be useful for decision makers in the field (personal communications are on-going with Biodiversity specialists from the New Zealand National Institute of Water and Atmospheric Research).

1.6 Outline of Dissertation Chapters

The current study is divided into two parts. Part I covers the generic context and the theory of the proposed model (represented in Chapter 2 and Chapter 3), while Part II concentrates on implementing the proposed model within the species context (represented in Chapter 4, Chapter 5, and Chapter 6). The study is concluded with Chapter 7, followed by three Appendices. All are summarized as follows:

- *Chapter 2* provides a literature review of the methods used in estimating the number of classes in general, as well as the ones used specifically in estimating the number of species. This includes most of the known methods of both the sampling-theoretic and the data-analytic approaches. The chapter also explores some relevant concepts of Bayesian inference and computing techniques.

- *Chapter 3* introduces our proposed approach, explaining its model formulation, distributional assumptions, and parameter specifications. The chapter also explains the justification and the mechanism of the chosen approximate Bayesian computing techniques (ABC), how it is used in the fitting stage of our model.
- *Chapter 4* introduces the formulation of the proposed model according to the distributional assumptions of the species context. It explains the model validation through simulation experiments, generating artificial worlds and fitting them according to our model. It also illustrates the fitting efficiency of the model through using the ABC technique.
- *Chapter 5* evaluates the performance of our model and showing more aspects of its characteristics throughout different scenarios. This is done by measuring the model in terms of three main criteria: accuracy, sensitivity (robustness), and additivity.
- *Chapter 6* presents some applications of our model on real data in the biodiversity field such as the ones retrieved from the Catalogue of Life (CoL) and World Register of Marine Species (WoRMS) databases. This involves checking some criteria and making some comparisons with a related previous approach applied to the same datasets.
- *Chapter 7* summarises and concludes the study with the main findings, and highlights some suggestions for the future work.
- *Appendices A, B, and C* provides some equations of the literature and the current model, extra plots and tables related to the artificial and real data experiments, and pseudo codes, respectively.

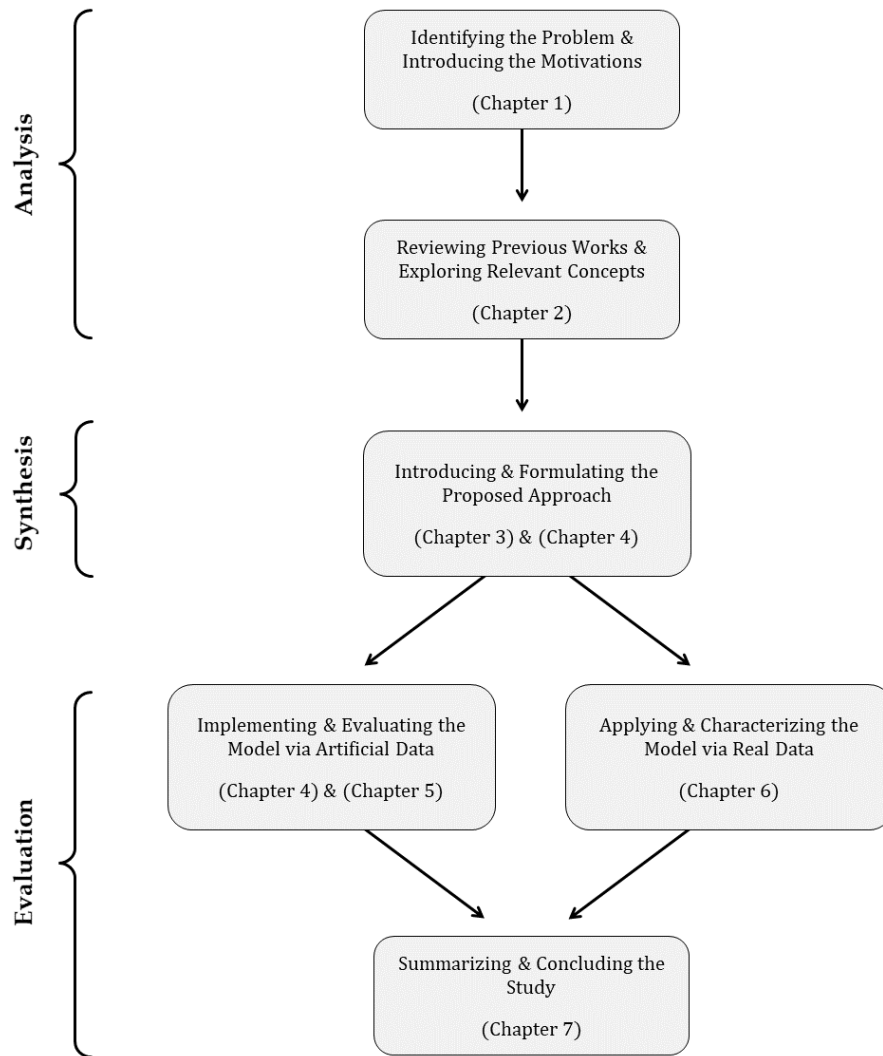


Figure 1.3: The Structure of the Dissertation

Part I

Theoretical Phase of The Proposed Approach

Chapter 2

Literature Review

As introduced in Chapter 1, most of the attempts that have been proposed for estimating the number of classes can be classified under two main approaches, the sampling-theoretic and the data-analytic. Each comes with justifications and some good properties but neither are free from drawbacks and limitations. Though, there are also other passive attempts in which scientists provided a range of estimates about the number of kinds based only on their expertise in fields such as those in biodiversity [20]. However, in the light of the current development in statistics and technology, expert guesses without solid quantitative analysis becomes outdated and unreliable. For example, in the number of species estimation, this kind of attempt lacks accuracy and results in providing a huge interval of estimates ranging from 33% underestimation to 16% overestimation [21]. The current chapter starts with exploring those active methods of estimating the number of classes in general and the number of species in particular, and ends with presenting some relevant concepts of Bayesian inference and computation techniques. It is worth presenting a holistic view of the existing approaches along with the proposed one, as is seen in Figure 2.1.

2.1 Methods of Estimating Number of Classes

In this section we illustrate two main active approaches in estimating the number of classes. The methods under these approaches are generally implemented on a variety of applications, including the biodiversity field.

2.1.1 Sampling-Theoretic Approach

This section is based on a review conducted by [8]. Before detailing the methods of this approach, we illustrate the general setting of the problem in the sampling-

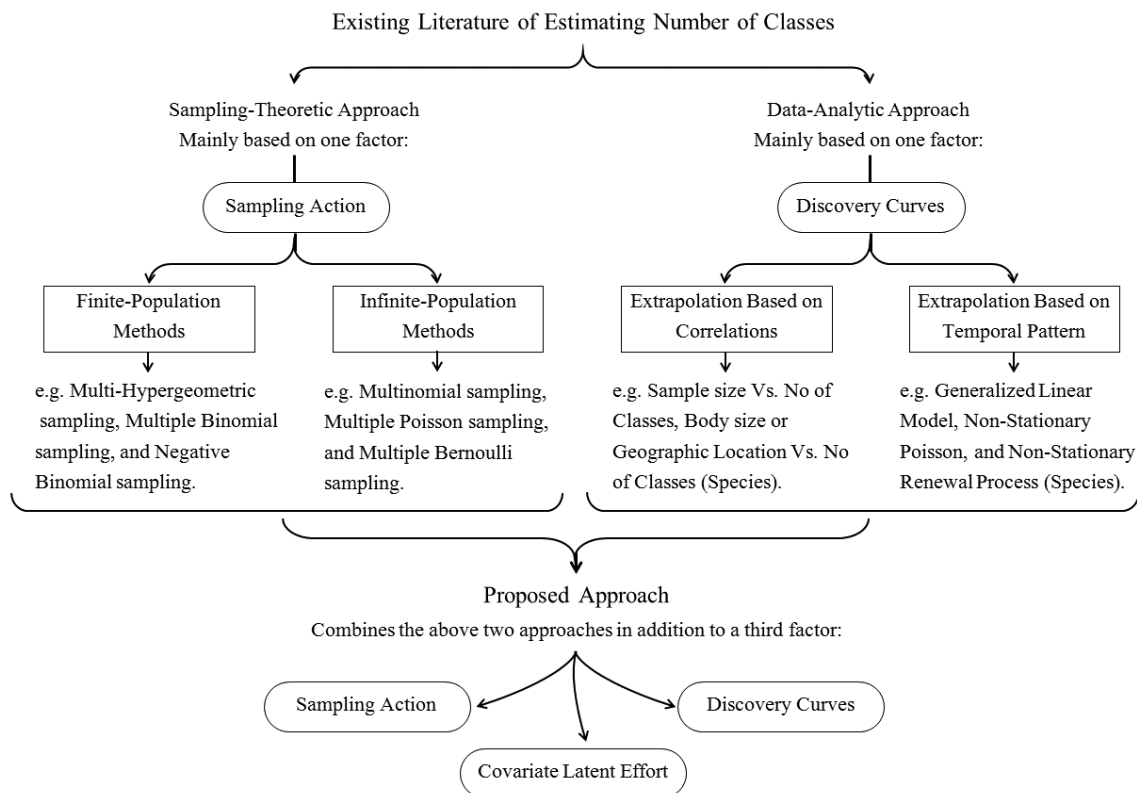


Figure 2.1: A holistic view of the existing approaches of estimating number of classes along with the proposed approach. Most of the attempts in the species context are under the data-analytic approaches.

theoretic approach. Assume that we sampled n items (specimens, in our biodiversity context) at random from a population which is partitioned into C classes (species, in the biodiversity context). Suppose that these classes have the following proportions (relative abundances, in the biodiversity context) $\mathbf{p} = [p_1, p_2, \dots, p_C]^T$, where $\sum_{i=1}^C p_i = 1$. In theory, the sample n should represent the population such that it includes all the classes in the population *i.e.* $n = \sum_{i=1}^C n_i$, where n_i is the number of the sampled items from the i th class. This theoretical random sample can be presented as a collection of sub-samples n_i and written as a random vector $\mathbf{n} = [n_1, n_2, \dots, n_C]^T$. However, in reality \mathbf{n} is not observable because not all n_i 's are expected to appear in the drawn sample. Instead, we can use an observable related random vector $\mathbf{c} = [c_1, c_2, \dots, c_n]^T$, where c_j is the number of the classes involving j items in the drawn sample. The total number of classes in the sample is $c = \sum_{j=1}^n c_j$. It is worthwhile noting that n can be represented exclusively from C ; $n = \sum_{j=1}^n j c_j$, and by definition, the probability mass function (PMF) of the observable random vector \mathbf{c} is simply the sum of the

PMF of \mathbf{n} over all points in \mathbf{n} corresponding to \mathbf{c} , *i.e.* $P(\mathbf{c}) = \sum_{\Omega} P(\mathbf{n})$, where $\Omega = \{\mathbf{n} : \#\{n_i = j\} = c_j, j = 1, \dots, n\}$. The main challenge in all the sampling-theoretic methods is that $P(\mathbf{c}) = \sum_{\Omega} P(\mathbf{n})$ does not have a closed-form expression. Hence, in light of this challenge and knowledge limitation, authors of this field approached other various ways of estimating the number of classes based on a variety of combinatorics along with some related assumptions. Some of these methods are related to the finite-population assumption such as Multi-Hypergeometric sample and Multiple Binomial sample methods, while others are related to the infinite ones such as Multinomial sample, Multiple Poisson sample, and Multiple Bernoulli sample methods. Note, some of the equations are not included within the text but they are available in Appendix A, Section A.1.

Finite-Population Methods

In this approach we assume that the size of the population, N , is known and it is partitioned into C classes. The number of items in the i th class is denoted by N_i where, $N = \sum_{i=1}^C N_i$. If we randomly sample n items without replacement from this population, then \mathbf{n} has a Multi-Hypergeometric distribution, $P(\mathbf{n}) = \prod_{i=1}^C \binom{N_i}{n_i} / \binom{N}{n}$. According to [9], only under the assumption that $n \geq M$; $M = \max_{1 \leq i \leq C} (N_i)$, a unique unbiased estimator for C exists, denoted by $\hat{C}_{Goodman1}$. As we mentioned, above the formula of $\hat{C}_{Goodman1}$ involves combinatorics using the known components N, n , and \mathbf{c} . Although this estimator is unique and a uniformly minimum-variance unbiased estimator, it suffers from several drawbacks: its variance is still very high in many cases such that estimations become useless (noted also by [22] [23]), its formula also allows for obtaining negative estimates which are meaningless, and its convergence behaviour is inconsistent and influenced by the sampling fraction [24] [25]. Independent of the above estimation efforts, an attempt (based on asymptotic behaviour) took place in [26], where the sampling fraction approaches a value $\in (0, 1)$ as N and $n \rightarrow \infty$ in a way that ensures the fraction to be within this interval. For a small sampling fraction such as 0.1, the resulting estimator, $\hat{C}_{Shlosser}$; where $\hat{C}_{Shlosser} \geq c$, performed reasonably well on simulated data, although it is a biased estimator and its variance had not been calculated. However, that simulation was not extensive enough for sufficient assessment and no formal comparison had been done with $\hat{C}_{Goodman1}$.

On the other hand, if we randomly sampled n items with replacement from this population, then \mathbf{n} has a Multiple Binomial distribution, $P(\mathbf{n}) = \prod_{i=1}^C \binom{N_i}{n_i} p^{n_i} (1-p)^{N-n}$, in which each i th class independently contributes as being in the sample n with the same probability p , *i.e.* each n_i follows a Binomial distribution with parameters (N_i, p) . Again, [9] derived an unbiased estimator, $\hat{C}_{Goodman2}$, under the assumption of knowing the value of p . Unfortunately, $\hat{C}_{Goodman2}$ shared the same drawbacks of $\hat{C}_{Goodman1}$. Another attempt proposed is, under the assumption of having small p and large N , an approximate moment method estimator but which also allows estimates less than c , like $\hat{C}_{Goodman2}$, [27]. There are suggested alternatives in which n_1, n_2, \dots, n_C are considered as *iid* Negative Binomial random variables, but they are still associated with some problems (see [28] for more details).

Infinite-Population Methods:

In this approach the size of the population is assumed to be unknown, but the population is still partitioned into C classes. We explore three main methods taking three perspective levels of analysis. Under the assumption that these partitions have the proportions $\mathbf{p} = [p_1, p_2, \dots, p_C]^T$, and randomly sample n items from this population, then \mathbf{n} has a Multinomial distribution, $P(\mathbf{n}) = \prod_{i=1}^C \binom{n}{n_1, \dots, n_C} p_i^{n_i}$. However, $P(\mathbf{c}) = \sum P(\mathbf{n})$ still does not have a closed-form expression. There are two branches of work done on this method, one assumes that the parameter vector \mathbf{p} has a pattern that can be described and used along with \mathbf{c} to estimate the total number of classes C . The other branch, does not have any assumption about \mathbf{p} but still benefits from the observed p_i 's that appear in the drawn sample along with \mathbf{c} to estimate C . The former branch is titled as a parametric model, while the latter is a non-parametric model.

In the parametric models, the first assumption starts with saying that p_i 's have a functional form that can be modelled and parametrized possibly by a small number of parameters, say ϕ . [29] suggested some models for this function and derived families of estimators for ϕ and C using the moment method. However, these estimators are complicated and their variances have not been calculated in order to assess them. Independently, [7] followed asymptotic approximations for

the p_i 's models and derived an estimator, \hat{C}_{McNeil} , which depends on an asymptotic distribution of some function. Unfortunately, this estimator is not better than the previous ones. It is also complicated and its behaviour is not well understood in the small samples context.

The second assumption proposed in the parametric models is that p_i 's have a histogram which can be approximately fitted by a probability density function depending on some parameters, say θ . [30] [31] [32] developed such a model for this function and estimated θ based on \mathbf{c} . Hence, based on θ the authors derived an estimator, \hat{C}_{Sichel} , for C . The asymptotic bias and variance of \hat{C}_{Sichel} were addressed. However, it suffered from strong bias in the small samples context. For this purpose, in [31], it was suggested that the sample size should be greater than 1500. It seemed that the second assumption estimation was more preferable than the first one in theory and in applications where there are no restrictions in specifying exact values for p_i 's.

Regarding the non-parametric models where there is no assumption about \mathbf{p} , it seems that there is no unbiased estimate that exists. In other words, having no assumption about p_i 's resulted in having unbounded bias of any estimator of C . However, an estimator, \hat{C}_{Chao1} , for the lower bound of C is obtained by [33] along with associated bootstrap confidence intervals. Later, a non-parametric estimator, \hat{C}_{Chao2} , was derived for C based on \mathbf{c} and the coverage of the p_i 's (the sum of the p_i 's that correspond to the observed classes in the sample) [34]. The later estimator is still biased but according to some empirical and theoretical evidence carried out at that time, \hat{C}_{Chao2} is preferable over \hat{C}_{Chao1} although the two estimators have not been formally compared. In the same year and also based on the coverage idea, another attempt separately took place by [35] at which the coefficient of variation (CV) of the class sizes is shown to play an important role. However, CV ends up being underestimated in this method and causing biased estimation for C when p_i 's are not equal.

In the above, the authors discussed the problem from a higher level, say n -level, at which they are concerned with describing the distribution of the total sample n as a whole (including all the classes available in the population).

In this level, a sample randomly drawn from an infinite population partitioned into C non-overlapping classes follows a C -dimensional Multinomial distribution. Later, a different approach is of discussing one sampled-class at a time, n_i -level. In this the number of items n_i of the i th class follows a Poisson distribution with mean λ_i . Since these classes are non-overlapping, we have C independent Poisson random variables so that \mathbf{n} has a Multiple Poisson distribution, $P(\mathbf{n}) = \prod_{i=1}^C e^{-\lambda_i} \lambda_i^{n_i}/n_i!$ [36]. Here also, since $P(\mathbf{c}) = \sum P(\mathbf{n})$ has no closed form expression, the efforts towards estimating C are through λ_i 's, the parameters of the distribution of \mathbf{n} . The assumption of this method is that the λ_i 's themselves constitute a random sample coming from a distribution function, say F , that has its own parameters. Hence, the estimator, $\hat{C}_{Poisson}$, can be obtained based on \mathbf{c} and estimating $1 - P_0(F)$, where $1 - P_0(F)$ is an F -mixed Poisson distribution of the class-abundance that is truncated at zero. And $P_0(F)$ is the probability that an arbitrary class will occur zero times in the drawn sample, which is governed by the distribution F that describes the λ_i 's.

[37] derived an estimate for $P_0(F)$ based on \mathbf{c} , where F represents the inverse Gaussian distribution, and estimated its parameters using the Maximum Likelihood method. This method provides a model that is considered as a special case of the one in the Multinomial sample method (done independently of the efforts in the \hat{C}_{Sichel} estimator). The authors of this method applied the Poisson-inverse Gaussian estimation of $P_0(F)$ in the ecological field on several well-known datasets (all are different groups of insects), and they noticed poor performance of their model on some datasets due to the influence of the trapping mechanism and classification errors that can distort the observed abundance counts. As a result of such errors, it may happen that no single model will emerge as the best one for a broad class of applications. In conclusion, the authors described the performance of their suggested model as follows: “*The fit of the Poisson-inverse Gaussian distribution to the several well-known data sets examined here suggests that it may be an appropriate model for species abundance data under specific conditions but will not always be the clear choice*”.

Independent of all the above, [38] gave an alternative family of estimators for $P_0(F)$ based on \mathbf{c} which seemed to be robust in keeping the variance of F

not too large, and hence used them in the $\hat{C}_{Poisson}$ estimator. The asymptotic, robustness, and efficiency of these estimators had been discussed and concluded that they can be useful but still only under very specific conditions, as it was with [37]. However, no formal comparison has been done between the two versions of $\hat{C}_{Poisson}$.

Unlike the two levels of analysis discussed so far, the final method in the infinite-population setting is on the level of the drawn items that constitute the total sample n , denoted by I_{ij} -level. Here $I_{ij} = 1$ indicates that the i th class is observed on the j th drawn of the items, where $i = 1, \dots, C$ and $j = 1, \dots, n$. The drawn items can be represented by a $C \times n$ matrix $[I_{ij}]$. It is worth mentioning that this method is based on the ‘*Capture-Recapture Sampling*’ that is commonly used in the ecological field to estimate the population size. In the current use of this method of sampling, I_{ij} represents the i th individual (an arbitrary item that represents a class i) that is caught in the j th trapping occasion, n is the trapping occasions, and the number of classes C represents the population size. Under the assumption that these draws are independent within each class with $P(I_{ij} = 1) = p_i$, then each of n_i constitutes a sum of Multiple Bernoulli random variables, $n_i = \sum_{j=1}^n I_{ij}$. Hence, $P(\mathbf{n}) = \prod_{i=1}^C \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}$, but again $P(\mathbf{c}) = \sum P(\mathbf{n})$ has no closed form expression.

Likewise the previous attempts in the same infinite-population context, [39] [40] assumed that the parameters p_i ’s themselves constitute a random sample coming from a distribution function, say F . They developed an estimator \hat{C}_{BOk} that depends on some component k , and provided a technique to choose the optimal k . They also computed the mean and variance of their estimator based on F . However, \hat{C}_{BOk} is biased and the fact that it depends on k makes it unstable because it acts as an alternating series in k . [41] proposed using an estimator \hat{C}_{Chao1} for the current model, while the same estimator was obtained in absence of any assumptions for the n_i proportions. Using a dataset provided by [42], which is conducted by a capture-recapture experiment on the taxicab population of Edinburgh (Scotland), it is found that on average \hat{C}_{Chao1} provided closer estimation to the true value of the given dataset than \hat{C}_{BOk} . In addition, through a small simulation study at which three trapping occasions are examined ($n = 5, 7, 10$),

it was found that \hat{C}_{Chao1} is better than \hat{C}_{BOk} in some situations where $n = 5, 7$, as the \hat{C}_{BOk} estimations showed severely negative bias and their coverage probabilities are much lower than the nominal level (0.95) especially when the average $p_{ij} = 0.05$. However, when $n = 10$, \hat{C}_{BOk} performed better than \hat{C}_{Chao1} . Moreover, according to [41], when n is getting larger and the average p_{ij} is also relatively large (*e.g.* when p_i 's constitute random variable coming from Uniform (0, 1) or Beta (2, 2)), \hat{C}_{Chao1} will fail to work.

2.1.2 Data-Analytic Approach

The target of this approach is mainly focused on analysing the given data to find some correlation or temporal pattern and use them for extrapolating the expected number of classes in a given population. In such a scenario, we can think of the expected number of classes as a function of some measurable variable, *i.e.* $E(c^{(x)}) = f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is some parameter vector and $c^{(x)}$ is the observed classes under the variable x such that as the value of x is asymptotically increasing, the function $f(x; \boldsymbol{\theta})$ approaches the target number of classes C , (a detailed explanation of this formulation is provided in Section 6.6.1, at which $E(c^{(x)}) = f(x; \boldsymbol{\theta})$ represents a logistic function to estimate the number of species). Note that $c = \sum_{j=1}^n c_j$ as mentioned in the previous section, and c_j is the number of the classes involving j items in a drawn sample.

Under the correlation assumption, x could be a measure of the sample size (not necessarily the count of items within a sample) where we can correlate between the sample sizes and the number of classes. This is implemented in the linguistic field where x is chosen to be the number of words in a text used for predicting the literary vocabulary. In this application, $f(x; \boldsymbol{\theta})$ is formulated (in one of the simplest forms) as a geometric series, and $\boldsymbol{\theta}$ is estimated by ad hoc methods. The target prediction is obtained as $\hat{C}_{BR} = 1/(1 - \hat{\theta})$ [43]. Another implementation is in the ecological field, in particular the paleontological research, at which a maximum amount of material (mass, volume, area, *etc.*) is examined to detect the number of species included in that material. In this application, [44] decided that x represents a mass of beach sand sampled for predicting the di-

versity of species in a given lithologic unit or fossil community. Here, $f(x; \boldsymbol{\theta})$ is formulated as a hyperbolic model (*i.e.* $f(x; \boldsymbol{\theta}) = \theta_1 x / (1 + \theta_2 x)$), and the vector $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$ is estimated by some transformation of linear regression to obtain the target prediction $\hat{C}_{DCLH} = \hat{\theta}_1 / \hat{\theta}_2$. However, according to [8], such a correlation assumption with the sample sizes leads to deriving $f(x; \boldsymbol{\theta})$ from a sampling model in a way that seems to be difficult to justify. Other types of correlation assumptions are presented in the biodiversity field and explained in the next section.

With respect to the temporal behaviour, one can consider x as a time component at which the classes are discovered. A range of examples is also available within the biodiversity context, explained in the next section.

2.2 Methods of Estimating Number of Species

As the species problem is an example of the number of kind/classes problem, the sampling-theoretic and the data-analytic approaches are still the umbrella of the methods of estimating the number of species. However, among this broad range of literature in classes/species estimation methods, we will highlight only some of the relevant methods paving the way to our proposed approach.

Under the sampling-theoretic methods, a Bayesian approach is suggested by [45] [46] in which a joint prior model, $\binom{N-1}{C-1}^{-1} f(C, N)$, is adapted for C and N_1, N_2, \dots, N_C , where $f(C, N)$ is an arbitrary distribution, $N = \sum_{i=1}^C N_i$ is a finite-population, and N_1, N_2, \dots, N_C are equally likely. Using this prior, a posterior model of C and N given c is derived from which the target estimation, \hat{C}_{Hill} , is represented by the mode of this posterior model, where $c = \sum_{j=1}^n c_j$. However, the chosen prior is non-informative and it turns out that this posterior depends only on the data through c not $\mathbf{c} = [c_1, c_2, \dots, c_n]^T$. This makes it subject to a confusion with a specified loss function [45] which was of a very different purpose.

For the infinite-population case, the same author suggested a different prior for only C to be a truncated Negative Binomial with parameter vector $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, where $\theta_1 \in (0, \infty)$ and $\theta_2 \in (0, 1]$, and the posterior model for estimating C is obtained by allowing $N \rightarrow \infty$. This work is extended by suggesting that the

prior includes a C -dimension symmetric Dirichlet density and also use the mode of the posterior model as an estimate for C , see [47] [8] for more details. The above is an addition to the attempts [36] [37] that are already explained previously in the Section 2.1.1 (in the infinite-population case), which are implemented within the species context.

With respect to the data-analytic approach, two methods are explored in the current study: extrapolation using some kinds of correlations such as species richness versus the body size or species richness versus the geographic location. The second is extrapolation using the past description of the species richness. This will be illustrated in the following subsections.

2.2.1 Extrapolation Based on Some Correlations

This section is based on a review conducted by [21]. The first method is based on the assumption that there is a relationship between the species' body size and their richness. In various marine and terrestrial taxa, it has been found that species with larger body size tend to be easily recognized and described [11] [12] [48]. Moreover, in a study on Holoplankton and fish species, it was also found that species with a larger body size, in addition to being endemic to a larger (in width and depth) geographic range, have also been described earlier [19] [49] [50]. Hence, in light of the above correlations, species of a given taxon could be represented by a frequency distribution in which having gaps in the distribution may suggest as yet undescribed species [51] [52] [53] [54]. Therefore, estimating the size of the gaps leads to estimating the number of unknown species. However, the basis of using the body size versus species richness relationships to predict the number of species is not solid enough for all species. In fact, there seems to be no relationship between them in a study done on marine metazoan species [55]. It also requires good knowledge of the taxon richness in advance which is unlikely to be available for all species but even when it is, such knowledge defeats the purpose of using such techniques. In addition, according to a study on tropical fish [50], prediction based on the body size versus species richness relationships results in poor predictors. Besides, this type of analysis needs to consider the size-variation of one

species growth in addition to the habitat-variation during the growth stages [21]. This may cause a confusion in recognizing new species as well as in modelling the distribution of the species' body sizes.

On the other hand, the second attempt neglected the species' body sizes and only considered the assumption that there is a high influence of the geographic locations on the species richness. The idea is to study species proportions of a given taxa across different areas. Hence, if there is a high correlation between these proportions, this implies that there is a high correlation between these areas so that we can extrapolate species richness from one well known area to another. However, the reality is more complicated especially when we talk about extrapolation from local to global scales. Many reasons are raised against the validity of this method. According to [56] most of these attempts suffer from weak or uneventual relationships between species proportions across different areas. A justification of this suggested by [48] that the factors influencing species richness may operate differently in different environments, and it can be seen between different regions or between local such as a continent and the whole globe. This makes species vary in their endemism to a particular habitat and consequently vary in their proportions within a given taxa between habitats. This means that the current method is not valid in this context. Moreover, from an analysis across 24 regions of the world [57] it is found that there is variability in the relative number of marine species across phyla and classes, and that is not yet clear whether it is a reflection of different species evolution influenced by the environment or just a bias in taxonomic effort and knowledge. Thus, there is a lack of clear correlations between these proportions across geographic locations which compromises using such a method.

In addition, since even within one taxa there is a variation among species in their endemism and pandemicity, there is no possibility to have sufficient knowledge about relationships among alpha, beta, and gamma biodiversity¹ neither in marine [59] nor in terrestrial [60] environments. For example, species richness may

¹Alpha diversity refers to species richness within-community such as in a habitat, while beta diversity is richness between-community excluding overlapping. Gamma diversity is the total richness such as across regional or global areas. [58]

be low in some habitats but high in the region including such habitats, because in total this region has a large variety of habitats with a little overlap in species *i.e.* this region has small alpha but large gamma diversity due to large beta diversity. On the other hand, other regions may have similar gamma diversity but large alpha and small beta diversity. Therefore, such geographical variation in these relationships challenges the attempts to predict global from local richness [61]. Another argument against the validity of the current method, is that it has been noticed, through many sampling efforts done on local scales, that there is a pattern of finding few species which are abundant while more species are rare. While the sampling methods take into account the abundance property in their local predictions, the current extrapolation method from local to global areas results in a shortage in capturing this property on the global scale. Moreover, [62] concluded that extrapolation based on the geographic location versus species richness correlation is the poorest method of estimating species richness on a wider scale.

2.2.2 Extrapolation Based on Temporal Component

Another direction for a data analytic approach is extrapolation based on the past behaviour of species richness. The main idea is to estimate the total number of species in a community by extrapolation from a temporal pattern of known species description. Species description refers to either species discovery events or species discovery rate. In such attempts, accurate estimates could be expected, but in applications it is subject to much uncertainty. Hence, many statistical methods in this field have been used to estimate species richness and assess the estimation in a variety of means.

One of these methods was done by [63], in which sighting a species has been treated as a non-homogeneous Poisson process (a type of point process). The rate parameter of this process is defined as a function which depends on two components: the first one is a constant that measures the visibility of a species, and the second is a function that captures the trend of the sighting skill and effort over time. The sighting skill/effort is defined as an exponential function with parameter β , while the species visibility is defined as realizations from an exponential distribution with parameter θ . The analysis was based on modelling the discovery

time (the time of the first sighting of a species) t_i , such that it depends on above two components (including θ and β) while following an Exponential distribution, $p(t_i)$. The maximum likelihood method was used to estimate θ and β . In light of the above setting, it is assumed that the number of discovered species follows a Binomial distribution such that the number of trials is represented by the total number of species (known n and unknown m), and the probability of success is given by the cumulative distribution of the discovery time, $P(t_i)$. Therefore, the discovery rate is fitted by the Binomial mean curve $(m+n)P(t_i)$, from which the point and interval estimations of the unknown number of species are derived.

However, although this method included an additional variable (the sighting skill/effort), rather than the discovery rate only, this variable was not assumed to be a random variable and was just represented by a mathematical function that does not capture the variability existing in reality. This leads to having a very smooth fitting curve for the discovery rate that does not account for the fluctuating behaviour occurring in the actual data. In addition, the used dataset in this work is not representative, it was very few (only 100 large marine species) and limited to only 169 years (from 1828 to 1996). Since the authors believe that their model seemed to work well, no estimation for the prediction error has been made, which is another drawback.

Moreover, the assumption of representing the data by the non-homogeneous Poisson process does not really capture the variation in the given data despite its changing parameter over time. The reason is that the Poisson process has one parameter in which the dispersion (variance) is equal to the mean. However, the changing pattern in the mean is different from the changing pattern of the dispersion of the discovery process. This is influenced by many factors such as changes in the speed of the ecosystem evolution, developing technology, availability of taxonomists *etc.* Hence, it is normal to observe under-dispersion and over-dispersion² in the behaviour of the discovery process.

In the same year, other work was published by [64] in which the discovery

²The under-dispersion occurs when the variance is less than the mean, while it is the opposite for the over-dispersion case.

process was also considered as a point process, but treated in a different way. The analysis was based on the discovery dates, since each discovery event is mainly recognized by its discovery date. The non-homogeneous Renewal process was the assumption for the given data to overcome some drawbacks of implementing the non-homogeneous Poisson process. The benefits of using the non-homogeneous Renewal process is to allow, in principle, the inter-discovery time to be represented by any distribution that has support $[0, \infty)$. However, for mathematical convenience the Gamma distribution was used here since it is a generalisation of the Exponential distribution and has two parameters θ_1 and θ_2 that are enough to capture the mean and the under-dispersion/over-dispersion behaviour. The Renewal process was modelled with mean parameter changing over time and it was represented as a logistic function. The logistic function is linked to the Gamma parameters, and it depends on two parameters ψ_1 and ψ_2 that control the rate of discovery. The logistic curve was used to fit the given data and to predict the remaining number of species yet to be discovered. A Bayesian approach along with MCMC sampling was used to estimate the above parameters and do the required prediction.

The fitted curve was based on the cumulative data, which was still very smooth and did not capture the fluctuating behaviour occurring in the actual data. However, adopting the Bayesian approach allowed for estimating the uncertainty of the prediction by giving a posterior distribution for the remaining number of species yet to be discovered. Despite the fact that the data used were verifiable (the first list of marine species in Europe [65]), this list was limited to Europe and to the marine species, hence the performance of the proposed method on the global scale was not known at that time. In addition, the analysis has been made separately on four big taxa, thus the performance of this method is also not known in the case of taxa aggregation.

Two years later, a completely different method was published by [10] though the form of the data being used was still the discovery dates. In this work, the expected number of species remaining to be discovered at time t is considered as some fraction k of the number of species that have not yet been discovered at previous time $t - 1$. This fraction represents the discovery curve and depends

on several interacting factors such as visibility of new species, discovery effort, expertise in identifying new species, and the proportion of habitat remaining unexplored. The value of k decreases and increases according to the dominant factor, for example, if most of the easily visible species have already been discovered and the visibility of new species becomes less, then k will decrease, on the other hand, if the discovery efforts increased within a restricted area of unexplored habitat, consequently k will increase. However, k has been taken as a constant for simplicity, under the assumption that over time new species yet to be discovered is eventually decreasing regardless of the variation in the above factors. A generalized linear model was formulated for the species remaining to be discovered at time t regressed on the cumulative discoveries up to time $t - 1$, with an error term assumed to be a quasi-Poisson. The parameters of the model were estimated using maximum likelihood, and from which confidence intervals were derived.

This study concluded that unless the inventory of a given community is nearly complete, estimating the total number of species is associated with very large margins of error, and using the discovery curves of such incomplete inventory for estimation is largely futile. However, the resulting margins of error were large according to the suggested model of this study. Although this paper did not take into account the effect of the discovery effort by considering k as constant, it emphasized that in addition to the discovery curve being governed by the remaining number of species to be discovered, the discovery curve is also governed by the variability of the discovery effort. Moreover, it is indicated that even for completed inventories, what is required is only an effort in a variety of means to discover new species.

Later in [21], the work published by [64] was applied on a wider collection of marine datasets, and was contrasted with other related approaches, though not with formal statistical inference. Then in [13], the work on the same model was extended globally to include terrestrial species in addition to marine ones. The real contribution of the later paper is involving the number of authors along with the discovery curve. The number of authors is used as a proxy for the latent taxonomic effort, which is an important indicator for estimating the total number of species on the globe. For example, a high trend for the marine discovery rate

is generally noticed after about 1950, while the trend for the terrestrial discovery rate was stationary. Simultaneously, the average number of authors describing species each year was increasing in both cases reflecting an increased amount of effort. For the terrestrial species, this might be an indicator that it is getting harder to discover new species possibly due to fewer of them remaining to be discovered or they were being localized such that in each year there are still some localities that need to be explored. On the other hand, in the marine case, this might be an indicator that there is still great opportunity to find new species in the oceans and implementing more efforts probably will help to discover most of them. However, the latent taxonomic effort was contrasted with the discovery curve only by general observation and description, not through a formal statistical perspective, unlike the current proposal which introduces a statistically rigorous framework by including this factor.

In conclusion, the field of estimating the number of classes in general and species in particular has witnessed two main approaches: the sampling-theoretic and the data-analytical. This is in addition to the subjective attempts (*e.g.* expert opinion) that do not seem to be reliable, because experts may not have sufficiently complete knowledge of all the available species to provide reasonable estimates. However, the two approaches are not without some shortcomings as shown above.

The sampling-theoretic methods mainly relied on the sampling action which required assumptions about the class-abundance and detailed sampling data. Therefore, such methods are highly influenced by the sampling mechanism and classification errors that affect the observed abundance counts and therefore affect the accuracy and the consistency of the estimation. In addition, such detailed sampling data is often not available especially in large-size communities that include a variety of classification groups (*e.g.* in the ecological field we might be interested in the animal kingdom where our target community includes insects, reptiles, and birds groups). Hence, employing such methods at a global scale seems to be unrealistic.

Similarly, the data-analytical methods that are based on some correlations also suffer from some subjective assumptions such as that one ecosystem is similar

to another and thus can be used for extrapolation of the species richness *etc.* On the other hand, the advanced contribution offered by the statistical modelling is implemented by the data-analytical methods that are based on a temporal component. These methods relied on studying the cumulative pattern of a given discovery dataset, at which the given data is treated as a noisy process that is assumed to follow some parametric process (*e.g.* a non-homogeneous Renewal process). The goal is to fit a mean curve for the given data, and the fitting process is formulated by some representative function (*e.g.* logistic function). However, these methods seem to suffer from some drawbacks and limitations. In all these methods, the process of the curve fitting is still data-specific *i.e.* very limited to the undertaken dataset and cannot be generalized to another datasets that might have different patterns and therefore different fitting models.

Instead, we introduce a novel approach of estimating the number of classes/species that provides an integrated framework using Bayesian modelling. Our proposed modelling is not exclusively subject to the class-abundance influence (which is a drawback of the sampling-theoretic methods), but includes other variables in context. Moreover, our proposed modelling is not formulated as a data-specific problem (which is a drawback of the data-analytical methods), it is formulated as a flexible framework that can be adjusted to take into account a variety of datasets. Our proposed model does not only account for a point and interval estimations, but also accounts for a posterior probability distribution that provides a better measurement of the estimation accuracy.

2.3 Bayesian Inference and Techniques

Statistical inference is the process of making statistical propositions about a population using a dataset sampled from that population. In most cases the statistical proposition is operationally defined by an unobserved quantity about which we wish to learn. It is worth distinguishing between two types of these unobserved quantities: first, the potentially observable ones such as predictions (future observations of a given process), and second, the quantities that are not directly observable such as parameters governing a hypothetical process leading to the

current observed dataset. For instance, when applying our proposed model on the biodiversity field, the number of species along with how they abound represents a parameter that contributes to governing the underlying process leading to the observed number of discoveries.

While statistical inference makes a wide range of types of conclusions, Bayesian inference makes its conclusions in terms of probability statements. The probability statements are basically conditional on the observed dataset. There are some situations where these probabilities are also conditional on observable values of covariates such as the number of authors in the biodiversity field. The main component in the Bayesian inference is the posterior probability which, proportionally, can be derived from the following formula using Bayes' rule:

$$\begin{aligned}
 P(H | \mathbf{x}) &= \frac{P(\mathbf{x}, H)}{\int_H P(\mathbf{x}, H) dH} \\
 &= \frac{P(\mathbf{x} | H) \times P(H)}{\int_H P(\mathbf{x} | H) \times P(H) dH} \\
 \Rightarrow P(H | \mathbf{x}) &\propto P(\mathbf{x} | H) \times P(H)
 \end{aligned} \tag{2.1}$$

where, ' H ' stands for a hypothesis or a belief around the unobserved quantity that is subject of the inference, ' \mathbf{x} ' refers to the observed dataset used to make the inference, $P(H)$ is the prior probability which expresses our hypothesis/belief before using the observed dataset. Such a prior hypothesis can be based on either a previous estimation or an expert opinion. $P(\mathbf{x}|H)$ is the likelihood of observing the dataset under the umbrella of the given hypothesis, $P(\mathbf{x}, H)$ is the full probability model that joins all the observable and unobservable quantities (dataset and given hypothesis, in the simplest case), and finally $P(H | \mathbf{x})$ is the posterior probability of updating our hypothesis/belief after observing the dataset. Note that since ' \mathbf{x} ' is observed and considered as fixed values and ' H ' is unobserved and it is the random quantity, the likelihood $P(\mathbf{x} | H)$ can be expressed as a function of the random quantity denoted by $\mathcal{L}(H | \mathbf{x})$. Hence, the posterior probability in equation (2.1) can be re-written as:

$$P(H | \mathbf{x}) \propto \mathcal{L}(H | \mathbf{x}) \times P(H) \tag{2.2}$$

Bayesian inference of a given hypothesis can be generally summarized in three steps as follows:

1. Setting up a full probability model of the observed dataset and any other unobserved quantity in problem context, $P(\mathbf{x}, H)$.
2. Deriving posterior probability, $P(H | \mathbf{x})$, by conditioning on the observed dataset and incorporating any prior knowledge about the given hypothesis.
3. Calculating some statistical properties and evaluating the quality of the fitting.

However, in complex cases setting up a full probability model is not an easy step and needs the aid of clarification such as employing Bayesian Network. Moreover, in most applications deriving the posterior probability model involves marginalization of the joint full probability model which in turn includes high dimensional integration. This complexity forces us to approximate the posterior probability model by some numerical techniques instead of computing it analytically. [66]

2.3.1 Bayesian Estimation

Since our target of study - the number of classes C - is of the second type of the unobservable quantity, we focus our attention on how to estimate a parameter. Typically, focus is on two types of estimation: point estimation and interval estimation. However, in both cases we need to consider the posterior probability distribution of our targeted parameter from which we derive our estimates.

Every point estimation is associated with a risk that should be minimized over all possible values in the realm of the targeted parameter. This risk is defined as the expected loss when we use a specific estimator for our parameter, $E[L(C, \hat{C})]$, where $L(C, \hat{C})$ is the loss function that measures the distance between the targeted parameter C and its estimator \hat{C} . Thus, a point estimator can be expressed in its simplest case as the following:

$$\hat{C} = \arg \min_{\hat{C}} \int L(C, \hat{C}) \times P(C | \mathbf{x}) dC \quad (2.3)$$

There are different forms of the loss function from which we can derive different forms of estimators. In our study, we are interested in two types of the central tendency estimators, the Bayesian posterior mean and mode, which can be derived from the quadratic and hit-or-miss loss functions, respectively.

$$\hat{C} = \arg \min_{\hat{C}} \int (C - \hat{C})^2 \times P(C | \mathbf{x}) dC \quad (2.4)$$

Minimizing the above integral can be derived by first using simple derivative with respect to \hat{C} :

$$\frac{\partial}{\partial \hat{C}} \int (C - \hat{C})^2 \times P(C | \mathbf{x}) dC = \int -2(C - \hat{C}) \times P(C | \mathbf{x}) dC \quad (2.5)$$

Then, setting the above result to be equal to zero:

$$\begin{aligned} & \int -2(C - \hat{C}) \times P(C | \mathbf{x}) dC = 0 \\ \Rightarrow & \int 2\hat{C}P(C | \mathbf{x}) dC = \int 2CP(C | \mathbf{x}) dC \\ \Rightarrow & \hat{C} \int P(C | \mathbf{x}) dC = \int CP(C | \mathbf{x}) dC \\ \Rightarrow & \hat{C} = \int CP(C | \mathbf{x}) dC, \end{aligned} \quad (2.6)$$

which is the mean and is called the ‘*minimum mean square error*’ estimator, \hat{C}_{MMSE} .

On the other hand, if we wish to use the hit-or-miss loss where $L(C, \hat{C}) = 1$ if $|C - \hat{C}| \geq \varepsilon$ and zero otherwise, our Bayesian posterior mode is defined as follows:

$$\begin{aligned} \hat{C} &= \arg \min_{\hat{C}} \left[\int_{-\infty}^{\hat{C}-\varepsilon} 1 \times P(C | \mathbf{x}) dC + \int_{\hat{C}+\varepsilon}^{\infty} 1 \times P(C | \mathbf{x}) dC \right] \\ &= \arg \min_{\hat{C}} \left[1 - \int_{\hat{C}-\varepsilon}^{\hat{C}+\varepsilon} 1 \times P(C | \mathbf{x}) dC \right] \end{aligned} \quad (2.7)$$

where, minimizing this formula can be simplified to just maximizing the above integral. Hence, for small ε and smooth $P(C | \mathbf{x})$ the maximum of the integral occurs at the maximum of $P(C | \mathbf{x})$ which is the mode and called the ‘*maximum a posteriori*’ estimator, \hat{C}_{MAP} .

With respect to the interval estimation, Bayesian inference introduces the concept of the ‘*credible interval*’ within which the value of a targeted parameter falls, with a particular probability $(1 - \alpha)$. This interval is defined on the realm of our targeted parameter as a subset from the support (domain) of its posterior probability distribution function,

$$\int_a^b P(C | \mathbf{x}) dC = 1 - \alpha \quad (2.8)$$

where, α is called the ‘*significance level*’ which is the probability of committing a Type I error³. Note that α defines the amount of uncertainty, and contributes in fixing the bounds $[a, b]$ of the credible interval.

There are several ways to specify a and b . The usual practice is to specify their values such that the probability of being below a is as likely as being above b . Such interval is called ‘*central interval*’, which is in most cases (including our current case) ensures that our credible interval will include all the central tendency measures. However, in some situations, especially when the distribution is skewed, it would be interesting to study the narrowest interval which involves those values of highest probability density (including the mode), such an interval is called ‘*highest posterior density interval*’ and denoted by (HPDI). [66]

2.3.2 Bayesian Network

A Bayesian Network is a combination of elements of both graph and probability theories. In graph theory, a graphical model is used as a tool to visually illustrate a set of variables and their dependencies/conditional dependencies. It is composed of a set of nodes and edges connecting them, the nodes represent the variables while the edges describe relationships among them. According to the nature of the relationships, edges could be cyclic⁴ or acyclic and directed or undirected. If there is a causal relationship between given two variables, the edge will be directional from the cause variable to the effect variable. On the other hand, if there

³ A Type I error occurs when the researcher rejects a null hypothesis when it is true.

⁴ A cycle exists in a graph, if we can start from a given node and travel along with a set of edges and arrive back to the starting node. Note that travelling must be in the correct direction if the edges are directed.

is just a correlation between the two variables, then the edge will be undirected.

Figure 2.2 shows four types of graphical networks where plots (a), (b) and (c) represent acyclic graphs, while plot (d) for example includes a cycle among the variables X, Y, and Z. On the other hand, plots (a) and (c) are directed graphs showing causal relationships, while plots (b) and (d) are undirected ones. With respect to the dependencies and conditional dependencies, we can say, for example, that the variables X and Y are conditionally independent given Z in plots (a) and (b), while these three variables are dependent on each other in plot (d). However, the variable U is conditionally independent of the set $\{X, Z\}$ given the variable Y in both plots (c) and (d).

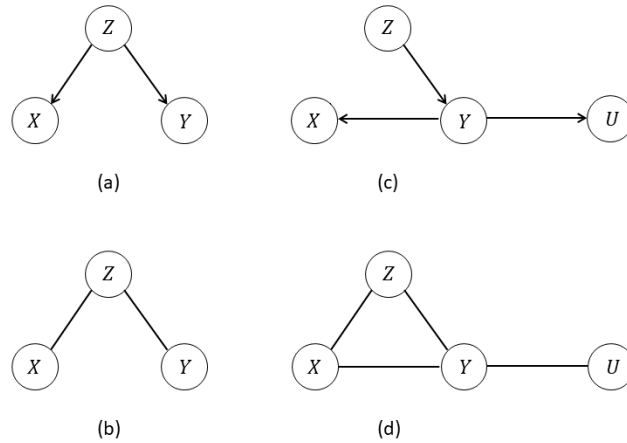


Figure 2.2: Examples of graphical models, presenting the *cyclic/acyclic* and *directed/undirected* properties of the graphical models.

The probability theory part is incorporated in the graphical network when there is a probability model for each variable in the graph. These probability models are defined according to the situation of the edges among their variables. For example, in Figure 2.2, there are joint probability models defined on plots (a) and (c) and can be respectively expressed as $P(X, Y, Z) = P(Z)P(X|Z)P(Y|Z)$ and $P(X, Y, Z, U) = P(Z)P(Y|Z)P(X, U|Y) = P(Z)P(Y|Z)P(X|Y)P(U|Y)$. However, in the undirected graphs, we can only proportionally factorize the joint probability model into individual potential functions, $\phi(\cdot)$'s, corresponding to each clique⁵. For example, in Figure 2.2 - plots (b) and (d), the joint probability

⁵ A clique, in an undirected graph, is a subset of adjacent nodes such that for every two nodes in this subset there exists an edge connecting them.

models can be expressed as $P(X, Y, Z) \propto \phi(X, Z)\phi(Y, Z)$ and $P(X, Y, Z, U) \propto \phi(X, Y, Z)\phi(Y, U)$, respectively.

‘*Bayesian network*’ is a specific type of probabilistic graphical model in which the nodes (variables) have probability distributions and the edges (relationships) are directed and acyclic. This makes it a useful tool to illustrate our proposed model. A Bayesian network is also called a directed acyclic graph (DAG). However, the undirected and cyclic graphs are such models that obey Markov property indicating that any variable is conditionally independent of any other variables given its neighbours. These models are called ‘*Markov Random Fields*’, but they are out of the context of the current study. [67]

2.3.3 Bayesian Computation Techniques

Bayesian computation mainly revolves around integrations used for computing marginals or expectations of distributions, usually a posterior or a posterior predictive. In most cases where the complexity, high dimensionality, and the absence of closed-form expressions exist, these computations start to be cumbersome and unable to be implemented analytically. Alternatively, numerical or distributional approximation techniques can help to approximate such distributions. Figure 2.3 shows a holistic view of the most used methods in this field. Since the number of classes or species problem represents a parameter in our model, we are only interested in the computations of the posterior distribution and the joint posterior distribution which are all about the parameters. Note that the joint posterior distribution is the full model, and the posterior distribution is usually a marginal of the full model computed with respect to a specific parameter or subset of parameters.

In the distributional approximation methods, the main purpose is to approximate a given posterior model by some simpler parametric distribution, or factorize it to a collection of parametric distributions. These techniques differ according to a spectrum of complexity and multiple dimensionality of the targeted model and the possible number of modes included. It is mainly based on finding

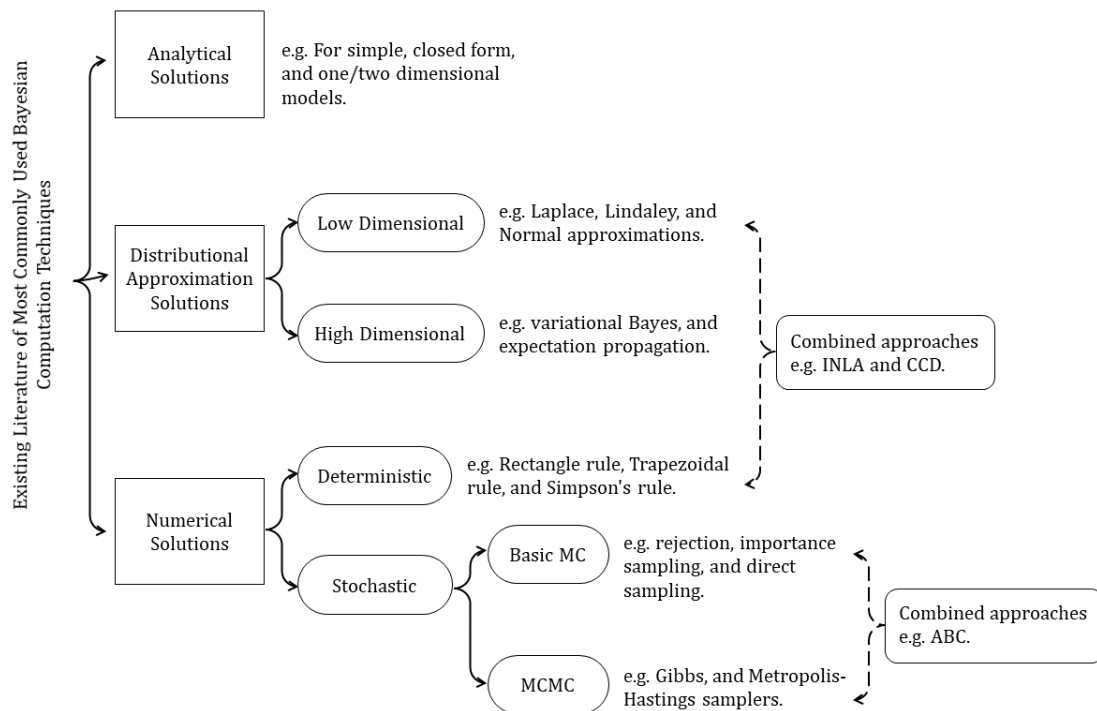


Figure 2.3: A holistic view of the most commonly used techniques of the Bayesian computation that are used for the integration and marginalization problems.

the modes as a way to begin mapping the posterior density/mass, and using some asymptotic expansions for the given integrals. Laplace, Lindale, and Normal approximations are examples of these methods [68] [66]. However, in the high dimensional cases, the above examples are not sufficient to solve the approximation problem and there is a need for iterative approaches such as the variational Bayes (VB) and the expectation propagation (EP) algorithms [66].

When the complexity of the integration problem prevents us from implementing the distributional approximation methods, due to being a computationally prohibitive or a costly process, we can find an escape in the numerical solutions. Numerical methods can be classified into deterministic and stochastic algorithms. The deterministic ones, also called quadrature, are based on evaluating the integrands on a finite set of points using some quadrature integration rules which are deterministic in nature, and then combining these evaluations through a weighted sum to be an approximation of the targeted integral. The quadrature rules are derived by constructing interpolating functions (polynomials) that are easy to integrate. These rules vary according to the order/degree of the polynomi-

als. For example, the rectangle rule is based on a constant interpolating function (a polynomial of degree zero), the trapezoidal rule is based on a straight line interpolating function (a polynomial of order 1), while Simpson's rule is based on a polynomial of order 2. All these rules are titled under the Newton-Cotes quadratures [69] [66]. There are other rules such as Chebyshev and Clenshaw-Curtis rules that come under the Gaussian quadrature, or the Gauss-Hermite rule which is a result of integration of a product of some other functions; see [69] [68] for more details. These methods are usually performed for one-dimensional integrals, and they are required to be iterated in the case of multiple dimensions. However, when the dimensionality become higher, it becomes difficult to implement the deterministic techniques and result in a low accuracy of approximation. A better solution is to turn to the stochastic techniques.

Stochastic methods, also called Monte Carlo methods, are a collection of simulation algorithms based on iterated sampling techniques. Here, the marginalizations or the integration of the targeted model is formulated as an expectation, which is in turn, approximated by averaging over randomly drawn samples from the targeted model. This approximation is supported by the law of large numbers and the central limit theorem [70]. The stochastic methods can be divided into Basic Monte Carlo (MC) and Markov chain Monte Carlo (MCMC) techniques. Examples of the the Basic Monte Carlo techniques are rejection sampling, importance sampling, and the direct simulation which is based on pseudo random generators and some transformation functions. These techniques are used in the low dimensional problems where it is easy to produce independent samples, while the MCMC are adapted for high dimensional and more complex cases where dependence is included in the sampling as a way to improve simulation process. However, the dependence will be disposed of by filtering the generated samples in such a way that their distribution asymptotically approaches the targeted posterior distribution. Gibbs and Metropolis-Hastings samplers are examples of MCMC [66].

There are also combined approaches of all or some of the above techniques. Integrated nested Laplace approximations (INLA) and central composite design integration (CCD) are examples of such approaches that combine the distribu-

tional approximation methods and deterministic numerical solutions [66]. Another interesting approach, which is based on a set of stochastic techniques, is the approximate Bayesian computation (ABC). It solves another challenge which is the absence of a tractable or analytic/closed form likelihood.

2.3.4 Approximate Bayesian Computation

In Bayesian inference, we start with specifying a prior model of the target parameter and a likelihood model from which the dataset is assumed to be sampled under this target parameter. However, in the high dimensional and complicated models such as the ones represented by Bayesian network, it is the opposite as we have the joint model at first and it should be possible to factorize it into the prior and the likelihood. However, in most cases, deriving the likelihood involves complex high-dimensional integrals and numerically evaluating them multiple times makes them computationally challenging. The approximate Bayesian computation is one of the approaches that is described as a likelihood-free inference approach and used in such complex situations, especially when MCMC cannot be implemented as it is the case of our proposed model (details will be explained in Section 3.2.2).

The basic idea of the ABC approach is to approximate the posterior model $P(C|\mathbf{x})$ by generating samples proportionally from a joint model using an accept-reject decision rule (recall that \mathbf{x} denotes the observed dataset). First, we generate several proposals for C , namely $C_1^*, C_2^*, \dots, C_m^*$. Then, simulated datasets $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_m^*$ are generated from a joint model $\mathcal{M}(\mathbf{x}^*|C)$ for each C_i^* . A detailed clarification of the ABC algorithm for the i th sample is provided below:

1. Draw a proposal value C_i^* from a prior model of C .
2. Draw a simulated dataset \mathbf{x}_i^* , called ‘*replicated data*’, from a joint model $\mathcal{M}(\mathbf{x}^*|C = C_i^*)$.
3. Compute a distance metric, denoted by $\mathcal{D}(\mathbf{x}_i^*, \mathbf{x})$, that measures the distance between the replicated and the observed datasets. This metric acts as a decision rule where the lower distance indicates high closeness. For computational ease, this metric is usually defined between sufficient⁶ summary statistics of the

⁶A sufficient statistic provides as much information about the target parameter as the whole data set itself.

observed dataset and the replicated one.

4. Accept C_i^* if $\mathcal{D}(\mathbf{x}_i^*, \mathbf{x}) \leq \epsilon$, and reject otherwise, where ϵ is a pre-specified threshold.

By the end of the above algorithm, the output is a set of accepted samples $(C_1^*, \mathbf{x}_1^*), (C_2^*, \mathbf{x}_2^*), \dots, (C_s^*, \mathbf{x}_s^*)$ that satisfies the constraint $\mathcal{D}(\mathbf{x}_i^*, \mathbf{x}) \leq \epsilon$. The ABC-approximation of the joint distribution of this accepted samples can be expressed as $P_\epsilon(C, \mathbf{x}^* | \mathbf{x}) := P(C, \mathbf{x}^* | \mathbf{x}^* \in \Omega_\epsilon(\mathbf{x}))$, where $\Omega_\epsilon(\mathbf{x}) = \{\mathbf{x}^* \in \chi : \mathcal{D}(\mathbf{x}^*, \mathbf{x}) \leq \epsilon\}$ and χ is the observation space [71]. The above approximated distribution is called the ‘ABC - joint posterior’, and by applying Bayes’ rule, it can be factorized as,

$$\begin{aligned} P_\epsilon(C, \mathbf{x}^* | \mathbf{x}) &\propto \Pr_\epsilon(\mathbf{x} | C, \mathbf{x}^*) \times \mathcal{M}(\mathbf{x}^* | C) \times P(C) \\ &= \Pr(\mathbf{x}^* \in \Omega_\epsilon(\mathbf{x}) | \mathbf{x}^*) \times \mathcal{M}(\mathbf{x}^* | C) \times P(C) \end{aligned} \quad (2.9)$$

where $\Pr(\mathbf{x}^* \in \Omega_\epsilon(\mathbf{x}) | \mathbf{x}^*)$ is a kernel function that represents the probability of the event $\mathbf{x}^* \in \Omega_\epsilon(\mathbf{x})$. In our case, this kernel is defined as an indicator function for the set $\Omega_\epsilon(\mathbf{x})$ that assigns a value of 1 when $\mathcal{D}(\mathbf{x}^*, \mathbf{x}) \leq \epsilon$ and 0 otherwise, *i.e.*

$$\Pr(\mathbf{x}^* \in \Omega_\epsilon(\mathbf{x}) | \mathbf{x}^*) = \mathbb{1}_{\Omega_\epsilon(\mathbf{x})}[\mathbf{x}^*] = \begin{cases} 1 & , \text{ if } \mathcal{D}(\mathbf{x}^*, \mathbf{x}) \leq \epsilon \\ 0 & , \text{ otherwise} \end{cases} \quad (2.10)$$

Under this kernel, the probability is reduced to either being 1 or 0, simply indicating whether the condition $\mathcal{D}(\mathbf{x}^*, \mathbf{x}) \leq \epsilon$ has been met or not [71] [74]. However, more generally one can define a kernel that allows this probability to take other values in $[0, 1]$, for example making it a decreasing function of distance.

By plugging in equation (2.10), equation (2.9) can be re-written as,

$$P_\epsilon(C, \mathbf{x}^* | \mathbf{x}) \propto \mathbb{1}_{\Omega_\epsilon(\mathbf{x})}[\mathbf{x}^*] \times \mathcal{M}(\mathbf{x}^* | C) \times P(C) \quad (2.11)$$

Therefore, the ABC-approximation of the target posterior of C is defined by,

$$\begin{aligned} P(C | \mathbf{x}) &\approx P_\epsilon(C | \mathbf{x}) = \int_{\mathbf{x}^*} P_\epsilon(C, \mathbf{x}^* | \mathbf{x}) d\mathbf{x}^* \\ &\propto \left[\int_{\mathbf{x}^*} \mathbb{1}_{\Omega_\epsilon(\mathbf{x})}[\mathbf{x}^*] \times \mathcal{M}(\mathbf{x}^* | C) d\mathbf{x}^* \right] \times P(C) \end{aligned} \quad (2.12)$$

where the quantity between the brackets in equation (2.12) represents the ABC-approximation of the likelihood *i.e.*

$$\mathcal{L}_\epsilon(C | \mathbf{x}) = \int_{\mathbf{x}^*} \mathbb{1}_{\Omega_\epsilon(\mathbf{x})}[\mathbf{x}^*] \times \mathcal{M}(\mathbf{x}^* | C) d\mathbf{x}^* \quad (2.13)$$

ABC algorithms can take many forms. The simplest of them is the one explained above and it is called the ‘*ABC rejection sampling*’ (ABC-RS). However, this algorithm might be inefficient when the threshold value ϵ is too small so that the rejection rate becomes too high, which means ϵ should be small enough but in a degree that allows the data to provide information. The convergence of the ABC approximation can be explained through equation (2.13) as follows:

If $\epsilon \rightarrow \infty$, this implies that $\mathcal{D}(\mathbf{x}^*, \mathbf{x}) \leq \infty$ is always true which results in,

$$\begin{aligned} \mathcal{L}_\epsilon(C | \mathbf{x}) &= \int_{\mathbf{x}^*} (1) \times \mathcal{M}(\mathbf{x}^* | C) d\mathbf{x}^* \\ &= 1 \end{aligned} \quad (2.14)$$

Hence, substituting the result of (2.14) into equation (2.12) concludes that,

$$P_\epsilon(C | \mathbf{x}) \propto [1] \times P(C)$$

which implies that the ABC - posterior converges to the prior as $(\epsilon \rightarrow \infty)$ *i.e.*

$$P_\epsilon(C | \mathbf{x}) \xrightarrow{\epsilon \rightarrow \infty} P(C)$$

On the other hand, if $\epsilon \rightarrow 0$, this implies that $\mathcal{D}(\mathbf{x}^*, \mathbf{x}) = 0$ *i.e.* $\mathbf{x}^* = \mathbf{x}$ and therefore, the approximated likelihood will approach the exact likelihood,

$$\begin{aligned} \mathcal{L}_\epsilon(C | \mathbf{x}) &= \int_{\mathbf{x}^*} (1) \times \mathcal{M}(\mathbf{x}^* = \mathbf{x} | C) d\mathbf{x}^* \\ &\propto \mathcal{M}(\mathbf{x} | C) \end{aligned} \quad (2.15)$$

Thus, substituting the result of (2.15) in equation (2.12) concludes that,

$$P_\epsilon(C | \mathbf{x}) \propto [\mathcal{M}(\mathbf{x} | C)] \times P(C)$$

which implies that the ABC - posterior converges to the exact target posterior as $(\epsilon \rightarrow 0)$ *i.e.*

$$P_\epsilon(C | \mathbf{x}) \xrightarrow{\epsilon \rightarrow 0} P(C | \mathbf{x})$$

It is worth mentioning that the rejection rate is not only affected by the threshold ϵ , but also by the choice of the prior model; a prior with considerably broad support also results in high rejection rate. Thus, the support of $P(C)$ and ϵ should be chosen carefully, and this can be achieved through pilot experiments. For $P(C)$, we can run multiple pilot simulations (with fewer iterations) of suggested upper bound values of C . This is done until it appears that the right tail of the approximate posterior distribution lies within the upper bound. This gives us an indication that we reached the maximum possible value of C and there is no point for exploring larger values in the prior support. After we specify the right prior support, we can choose the value of ϵ based on the accepted posterior distribution in the pilot experiments (more details on choosing ϵ are explained in Section 3.2.3).

Another form of ABC is the ABC-MCMC algorithm. Briefly, MCMC sampling is a process that filters proposed values for the target parameter to end up with a sample of values assumed to be from the desired posterior distribution. The most popular example is Metropolis-Hastings at which an ‘*acceptance probability*’ needs to be determined by the likelihood which is itself an obstacle in our case. However, the acceptance probability of the ABC-MCMC is modified to include the decision rule $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$ and, therefore, skip computing the likelihood as follows:

$$\mathcal{A}(C_{i-1}, C_i^*) = \begin{cases} \min \left(1, \frac{P(C_i^*)/Q(C_i^*|C_{i-1})}{P(C_{i-1})/Q(C_{i-1}|C_i^*)} \right), & \text{if } \mathcal{D}(\mathbf{x}_i^*, \mathbf{x}) \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \quad (2.16)$$

where, $Q(\cdot)$ is a proposal kernel distribution for the posterior, which should be easier to deal with. Now, the modified acceptance probability involves two steps of filtering. First, C_i^* is used to sample \mathbf{x}_i^* from the joint model, and when $\mathcal{D}(\mathbf{x}_i^*, \mathbf{x}) \leq \epsilon$, then C_i^* is initially accepted to go through the next step. In the next step, $\mathcal{A}(C_{i-1}, C_i^*)$ is computed and if it is greater than a randomly sampled uniform[0,1] value, then C_i^* is finally accepted where $C_i = C_i^*$ with probability $\mathcal{A}(C_{i-1}, C_i^*)$, otherwise no move is taken and $C_i = C_{i-1}$ with probability $1 - \mathcal{A}(C_{i-1}, C_i^*)$.

However, ABC-MCMC algorithm could be worse than the simpler ABC-RS if the chain is trapped in low-probability regions of the proposed kernel at which the rejection rate becomes extremely high. Also, proposing a kernel can be a highly challenging step by itself in our context. Moreover, the ABC-MCMC algorithm cannot be parallelized, unlike the ABC-RS which can be easily implemented simultaneously. This makes the ABC-RS algorithm a plausible solution for our proposed model where the parallelization approach is essential in light of the complexity and high dimensionality our model. There are other algorithms such as ABC Partial rejection control (ABC-PRC), ABC population Monte Carlo sampling (ABC-PMC), and ABC Sequential Monte Carlo sampling (ABC-SMC) which are based on the ‘*particle filtering*’⁷ concept. However, research in this area are still ongoing and stimulated by practical challenges, and as with any developed technique, there are pros and cons and possible limitations for these algorithms. For more details about the ABC algorithms see [71] [72] [73] [74].

2.4 Chapter Summary:

We started this chapter with exploring the literature of the active methods in estimating the number of classes as well as of estimating the number of species. We saw that these methods can be classified into two main approaches, the sampling-theoretic and data-analytic. However, most of the advanced statistical contributions, in the species context, are under the data-analytic approach. An interesting one is provided by [13] which introduced the number of authors as a proxy of a covariate factor, but not through a formal statistical perspective, unlike the current proposal which introduces a statistically rigorous framework using Bayesian inference by including this factor. The rest of the chapter covered some related Bayesian concepts: Bayesian estimation, Bayesian Network, and Bayesian computation techniques. We ended this chapter by illustrating the approximate Bayesian computation (ABC) which is used to implement our proposed framework.

⁷A way of simultaneously dealing with a large pool of proposal parameter values, called ‘*particles*’, instead of one at a time.

Chapter 3

The Proposed Approach

As mentioned in Chapters 1 and 2, our proposed approach is comprehensive in that it acts as a framework which integrates two main approaches in the number of kinds problem by considering a covariate influence, and lends itself to different applications. To simplify the idea of our proposal, let us think of the number of kinds problem as a black-box system which has an input and output, where the output is obviously the number of kinds/classes. While the previous methods of the two distinct approaches consider the number of discoveries as an input only, our proposal takes the effort employed on making discoveries (which should be - to some extent - available information for us) as another input to that black-box. Figure 3.1 shows this simple idea that will be clarified and developed in more details in the following sections. The sampling-theoretic approach utilizes only combinatorics assumptions and the data-analytic approach utilizes only correlations/temporal assumptions on the input (the number of discoveries) to model what is inside the box in order to estimate the output (the number of kinds/classes). On the other hand, our proposal integrates both assumptions in a flexible comprehensive way on both inputs (the number of discoveries and the

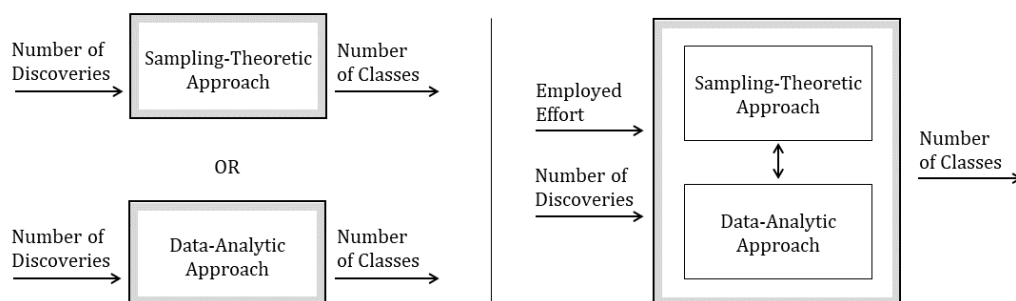


Figure 3.1: The left hand side plot shows the Black-box concept of the number of kinds/classes problem in the existing literature, while the right hand side shows this concept as proposed in our framework.

employed effort) to model what is inside the black-box using Bayesian modelling and computations, in order to estimate the target output.

In this instance, we formulate a model that directly describes the process of class sampling-discovery in parallel with a covariate effort process which is described by being ‘*latent*’. The reason for this description is that in most cases our main process, sampling-discovery, is unlikely to be completely explained by the underlying covariate process. It is worth mentioning that the two concepts, ‘*latent*’ and ‘*effort*’, are widely used in many branches of science. Intelligence and verbal ability are examples of the latent variables in psychology and linguistics, respectively [75]. On the other hand, voltage as introduced by the *Maxwell analogy* is considered as an effort variable in physics and electrical engineering [76]. In addition, as the case of our model, the two concepts together are encountered in ecological and angler processes [77].

The current chapter starts with describing the structure of the proposed model and how it is built-up through linking the above two processes. Then, it moves to discuss the model properties that may be required by the context of the application, *e.g.*, distributional assumptions, parameter selection, and proper sampling methods. Finally, it ends with explaining the inferential attempts to activate the proposed model.

3.1 Model Formulation

The model is introduced as a two-step built-up structure. In the first step, we will illustrate the sampling-discovery process part, and in the second step, we will illustrate the latent effort process part. Then, we will explain how these two parts are integrated to compose our model which represents what is inside the black-box. Before explaining details about these two processes, we introduce the setting of their environment as described below:

Suppose that our targeted community consists of an unknown finite number of classes C with unknown finite abundances N_1, \dots, N_C , so that the total popula-

tion of this community is $N = \sum_{k=1}^C N_k$. Since the circumstances of the sampling experiment differ according to the targeted community and also according to the type of effort employed, it is more convenient for us as statisticians to adopt the assumption that items are sampled with replacement so that each of them is equally likely to be chosen. Hence, the probability that a sampled item will be of i -th class is $p_i = N_i/N$. This assumption helps to simplify the exposition without loss of generality.

3.1.1 The Sampling-Discovery Process

In illustrating the class sampling and discovery process, we will start with a building unit called ‘*discovery event*’. This event should be recognized by some recording manner, for example, the recording manner could be a time point such as a date at which a bug report is received in the software reliability context, or it could be a location point such as a line in a book at which a new vocabulary is found in the linguistics context. However, in our proposed model, and specifically in the species context, we choose to recognize the discovery event in two dimensions, location point and time point. This recording mechanism is explained in the following:

As a sequence of samples is taken, the C classes will eventually be discovered. Within this sequence, when an i -th class is flagged to be a new class representing a discovery event, we define δ_i to hold the index which is the location point of this new class, and t_i to hold the time point at which this class is discovered. Note that indexing the sequence of draws will begin from the first discovery, hence, $\delta_1 = 1$ indicates that the first draw is the first class to be discovered at a time point t_1 . If the second draw was a repeat of this class but the third came from a different class then we would have $\delta_2 = 3$ at a time point t_2 , *etc.* However, in most cases t_i is the only available information, while δ_i is not and needs to be estimated. For notational convenience, we define another variable $d_i = \delta_i - \delta_{i-1}$, which is the number of draws between successive discovery events, namely the ‘*inter-discovery sample*’. Note that as we considered $\delta_1 = 1$, then $d_1 = 1$ as well. Figure 3.2 shows an example of the sampling-discovery sequence of a total sample

size $\sum_{j=1}^T n_j = 30$, and total discoveries $\sum_{j=1}^T \tau_j = 8$ over three years $T = 3$, where n_j is the sample size within a time unit j , and τ_j is the number of discoveries within the same time unit j (the latter two variables will be explained in the next sections).

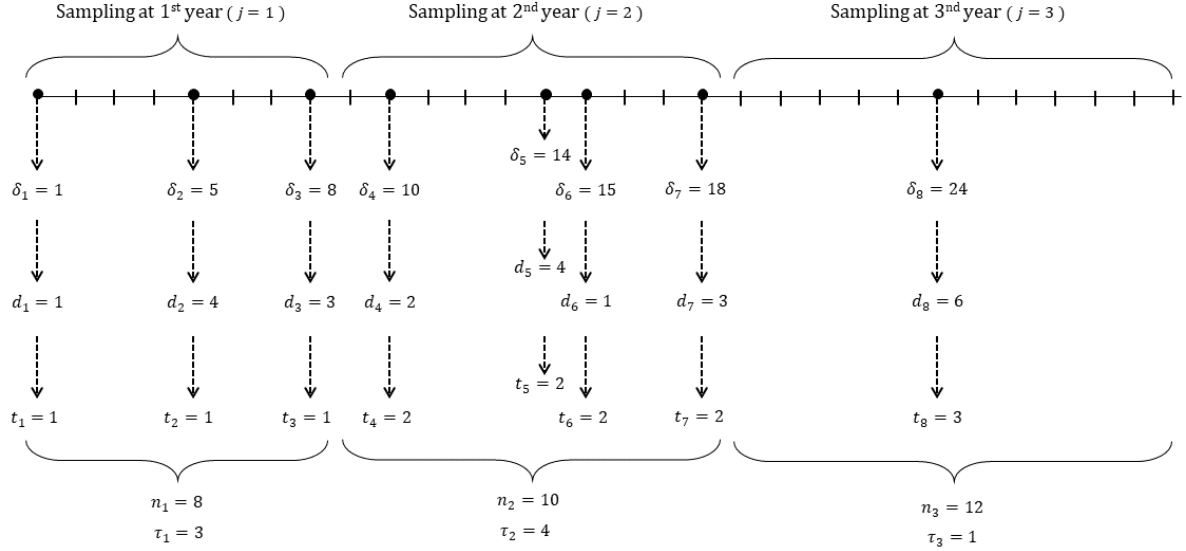


Figure 3.2: An example of a sampling-discovery sequence, with total sampling $\sum_{j=1}^3 n_j = 30$, and total discoveries $\sum_{j=1}^3 \tau_j = 8$ over three years $T = 3$.

In such a setting, d_i acts as a Geometric random variable where the inter-discovery sample represents the number of drawn items until meeting a discovery. This process is governed by the number of classes C and their abundances N_1, \dots, N_C , so that the probability of success in meeting a new class i determined by the abundance of that class and the abundances of the remaining classes yet to be discovered. The mathematical representation of this probability is $(1 - \sum_{k=1}^{i-1} N_k/N)$, where N_1 is the abundance of the first sampled item which is already discovered and came from a known class, N_2 is the abundance of the second sampled item that is also already discovered and from a known class, and so on N_{i-1} is the abundance of the last known sampled class just before i -th class. Note that if, for example, the first and second items came from the same known class, then we should consider that $N_1 = N_2$ in calculating the probability of success.

However, the extent to which we can know the class-abundances is limited to

only specifying their probability distribution which varies according to the context of the application in use. For generality, we will assume that N_k follows some probability distribution F_N parametrized by θ_N which could be a representation for a single or vector of parameters. Thus, by expanding our view from just one discovery event to a sequence of these discovery events until we get D total number of discoveries, then we have $\{d_i\}_{i=1:D}$ that can be described as a stochastic process following a joint Geometric distribution conditioned on the knowledge of the distribution of the class-abundances $\{N_k\}_{k=1:C}$ up to time t_D ,

$$P(d_{2:D}|N_{1:C}) = \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \quad (3.1)$$

Figure 3.3 shows Bayesian network representation for the conditional relationships within the sampling-discovery process. Up to this point, we explained one aspect of recognizing the discovery events which is the location dimension, $\{d_i\}_{i=1:D} = \{\delta_i - \delta_{i-1}\}_{i=1:D}$. The time dimension $\{t_i\}_{i=1:D}$ is a following incidence to the location dimension *i.e.* there is no t_i unless we have δ_i . However, d_i or δ_i are not the only elements that contribute in formulating t_i , hence, we will discuss the details about formulating t_i when we introduce the linkage node between the sampling-discovery process and the latent effort process.

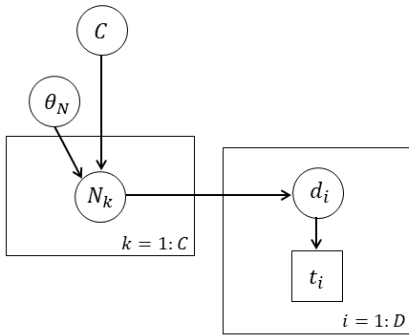


Figure 3.3: A Bayesian network visually displaying the probabilistic relationships within the sampling-discovery process. The square node refers to available data, while circle nodes correspond to unknown variables.

3.1.2 The Latent Effort Process

It is a fact that there is an effort employed in the sampling and discovery process, which we should utilize in the number of kinds/classes problem. Therefore, our

proposed framework assumes that this underlying effort comes as a temporal latent process $\{l_j\}_{j=1:T}$ evolving over discrete units of time j and co-varying with the sampling-discovery process, where T being the time length in which the whole sampling has occurred. Although this process is latent, it can be characterized by some probability distribution F_l with parameter(s) θ_l suggested by the context of the application. However, being unable to measure this process directly, obliges the use of a measurable proxy through which we can infer about $\{l_j\}_{j=1:T}$. Hence, this proxy, denoted by $\{x_j\}_{j=1:T}$, represents now available information for us to be used as an input for our black-box of the number of kinds/classes problem. Figure 3.4 shows Bayesian network representation for the conditional relationships within the latent effort process.

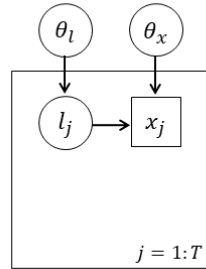


Figure 3.4: A Bayesian network visually displaying the probabilistic representation of the latent effort process and its proxy. The square node refers to available data, while the circle refers to unknown data that needs to be estimated.

3.1.3 Linkage Node of the Two processes

To incorporate the latent effort process to the sampling-discovery process, we introduce a linkage node which represents the sizes of the samples, n_j 's, within which the discoveries are revealed. These samples are assumed to be collected within each time unit j as a result of the effort employed. There are two perspectives in achieving this linkage, one is from the sampling-discovery point of view, while the other is from the latent effort point of view.

In the first perspective, we can see that each sample size n_j is built-up by the discovery events and the inter-discovery samples that are accumulated by time j . Consequently, this provides an upper bound on the possible values of $\{d_i\}_{i=1:D} = \{\delta_i - \delta_{i-1}\}_{i=1:D}$ in that we must have the constraints, $\delta_i \leq \sum_{r=1}^j n_r$, which contributes to shape the natural mechanism of the geometric sampling and discovery process but within the collected samples n_j 's. The inequality simply

states that the index of the current draw in which the i -th class is discovered must be less than or equal to the total number of samples accumulated up-to the time unit j . In this instant, as δ_i is constrained by $\sum_{r=1}^j n_r$, at the same time t_i is constrained by j in that it is formulated as a following resultant:

$$t_i = j \times \mathbb{1}(\delta_i \leq \sum_{r=1}^j n_r) \quad (3.2)$$

where, $\mathbb{1}(\cdot)$ is an indicator function (1 if true, 0 if not), $i = 1, 2, \dots, D$ and $j = 1, 2, \dots, T$ such that i is a free index, while j moves from one value to another only when it violates the constraints. In other words, t_i 's are time points within each time unit j . Hence, if we, for example, found three discoveries within the first sample size n_1 in the first time unit $j = 1$, then $t_1 = t_2 = t_3 = 1$. And, if we found two discoveries within the second sample size n_2 in the second time unit $j = 2$, then $t_4 = t_5 = 2$, *etc.* Therefore, conditioned on $\{d_i\}_{i=1:D} = \{\delta_i - \delta_{i-1}\}_{i=1:D}$ and $\{n_j\}_{j=1:T}$ - through the above constraints - each time point t_i follows a Degenerate distribution that represents these constraints and can be expressed as,

$$P(t_i | d_i, n_{1:j}) = \begin{cases} 1, & \text{if } d_i \leq \sum_{r=1}^j n_r - \delta_{i-1} \\ 0, & \text{otherwise} \end{cases} \quad (3.3)$$

As we in the stage of linking the sampling-discovery and the latent effort, It should be more convenient to synchronize the evolving time of these two processes. Hence, we define τ_j to represent the number of discoveries evolving over the time unit j such that $\tau_j = [\sum_{\forall i} \mathbb{1}(j - 1 < t_i \leq j)]_{\forall j}$, where $\mathbb{1}(\cdot)$ is an indicator function (1 if true, 0 if not). This will count all discoveries to occur within each time unit j .

However, although n_j 's play important role in describing d_i and formulating t_i , in most cases n_j 's are not available information and need to be estimated which leads us to the second point of view. In the second perspective, the sample sizes can themselves be considered as '*noisy*' realisations of the underlying effort process. In other words, since the sample sizes are outcomes of a counting action, it should be more convenient to assume that $\{n_j\}_{j=1:T}$ is a Non-Homogeneous Poisson process parametrized somehow by the latent effort process $\{l_j\}_{j=1:T}$, and

can be generally expressed as follows,

$$P(n_{1:T}|l_{1:T}) = \prod_{j=1}^T [f(l_j)]^{n_j} \frac{\exp[-f(l_j)]}{n_j!} \quad (3.4)$$

where, $f(l_j)$ is some function of the latent effort and represents a parameter of modelling $\{n_j\}_{j=1:T}$.

To sum up, we have a system of two main processes linked to each other, which are the sampling-discovery and the latent effort. Under the umbrella of both processes we have four factors: the class-abundances N_k , sampling action (covered by d_i , t_i , and n_j), discovery curve τ_j , and the employed effort (the latent l_j and its proxy x_j). This umbrella reflects the integration of the sampling-theoretic and the data-analytic approaches as we can see the sampling action and the discovery curve, Figure 3.5 - left hand side. Among this set of variables, we have two of them as given information which are the number of discoveries $\{\tau_j\}_{j=1:T}$ and the proxy of the effort employed on making discoveries $\{x_j\}_{j=1:T}$. Hence, when we go back to the black-box concept of the number of kinds/classes problem, it is obvious that the main variable $\{\tau_j\}_{j=1:T}$ and the covariate $\{x_j\}_{j=1:T}$ both represent inputs to the box, as shown in Figure 3.5 - right hand side. The box itself represents the integration of the above two processes which is modelled by a Bayesian network as illustrated in the next subsection.

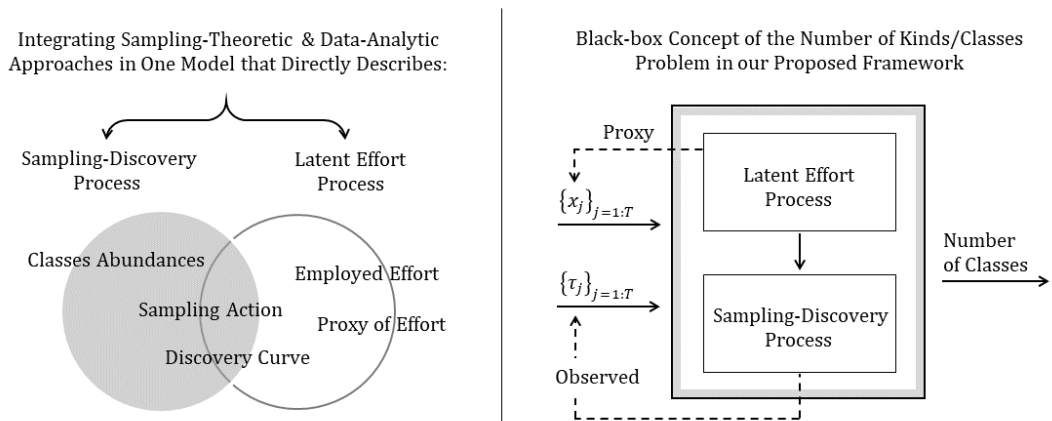


Figure 3.5: The left hand side plot shows the overlap between the integrated sampling-discovery and latent effort processes, and the set of factors they both include. The right hand side shows the black-box concept in its final stage as proposed in our framework.

3.1.4 Final Phase of the Proposed Model

In this stage we introduce the proposed integrated model as Bayesian Network that captures a system of inter-relationships occurring in the environment of the class discovery and influencing the estimation of the total number of classes C in a given community, see Figure 3.6.

By defining $\boldsymbol{\theta} := \{\theta_N, \theta_l, \theta_x\}$ and $\boldsymbol{\Psi} := \{\mathbf{N}, \mathbf{d}, \mathbf{n}, \mathbf{l}\}$, the model can be mathematically represented as a full joint probability distribution,

$$P(C, \mathbf{t}, \mathbf{x}, \boldsymbol{\Psi}, \boldsymbol{\theta}) = \underbrace{P(\mathbf{t}, \mathbf{x} | \boldsymbol{\Psi}, \boldsymbol{\theta}, C)}_{\text{Full likelihood given both the target } C \text{ and the contextual variables } \boldsymbol{\Psi}} \times \underbrace{P(\boldsymbol{\Psi}, \boldsymbol{\theta} | C)}_{\text{Distribution of the contextual variables } \boldsymbol{\Psi} \text{ given the target } C} \times \underbrace{P(C)}_{\text{Prior distribution of the target } C} \quad (3.5)$$

$$\begin{aligned} \Rightarrow P(C, \mathbf{t}, \mathbf{x}, \boldsymbol{\Psi}, \boldsymbol{\theta}) = & \\ & P(\mathbf{t} | \mathbf{d}, \mathbf{n}) P(\mathbf{x} | \mathbf{l}, \theta_x) P(\theta_x) \times \\ & P(\mathbf{n} | \mathbf{l}) P(\mathbf{l} | \theta_l) P(\theta_l) P(\mathbf{d} | \mathbf{N}, \mathbf{n}) P(\mathbf{N} | C, \theta_N) P(\theta_N) \times \\ & P(C) \end{aligned} \quad (3.6)$$

where, $\mathbf{N} = \{N_{1:C}\}$, $\mathbf{d} = \{d_{1:D}\}$, $\mathbf{t} = \{t_{1:D}\}$, $\mathbf{n} = \{n_{1:T}\}$, $\mathbf{l} = \{l_{1:T}\}$, $\mathbf{x} = \{x_{1:T}\}$. However, when the model satisfies the constraints $d_i \leq \sum_{r=1}^j n_r - \delta_{i-1}$, then $P(\mathbf{d} | \mathbf{N}, \mathbf{n}) = P(\mathbf{d} | \mathbf{N})$ and $P(\mathbf{t} | \mathbf{d}, \mathbf{n}) = 1$. Moreover, if we take the hyper-parameters set $\boldsymbol{\theta} := \{\theta_N, \theta_l, \theta_x\}$ as a fixed set, then the full joint probability distribution of our proposed model can be simplified to the following expression,

$$P(C, \mathbf{t}, \mathbf{x}, \boldsymbol{\Psi}) = P(\mathbf{x} | \mathbf{l}) P(\mathbf{n} | \mathbf{l}) P(\mathbf{l}) P(\mathbf{d} | \mathbf{N}) P(\mathbf{N} | C) P(C) \quad (3.7)$$

Therefore, given the knowledge of \mathbf{t} and \mathbf{x} , the posterior model of our target variable which is the number of classes C can be approximately derived as follows,

$$P(C | \mathbf{t}, \mathbf{x}) = \int_{\boldsymbol{\Psi}} P(C | \mathbf{t}, \mathbf{x}, \boldsymbol{\Psi}) d\boldsymbol{\Psi} \propto \int_{\boldsymbol{\Psi}} P(C, \mathbf{t}, \mathbf{x}, \boldsymbol{\Psi}) d\boldsymbol{\Psi} \quad (3.8)$$

The approximated distribution is assumed to be almost unimodal with support $C \in \{\mathbb{N} \geq D\}$, where D is the number of classes that are already discovered.

From this distribution, a point and interval estimation of C are expected to be calculated.

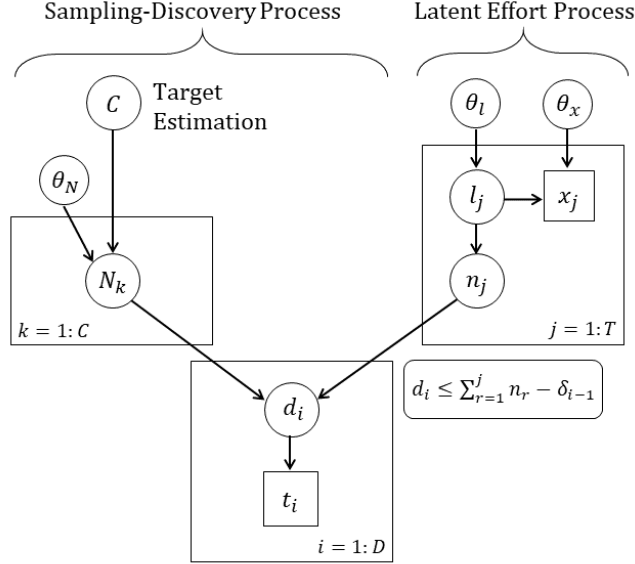


Figure 3.6: A Bayesian network visually displaying the whole structure of the proposed model, and its probabilistic causal relationships. The model is designed under the constraints $d_i \leq \sum_{r=1}^j n_r - \delta_{i-1}$, where $d_i = \delta_i - \delta_{i-1}$. Note that as a result $t_i = j \times \mathbb{1}(\delta_i \leq \sum_{r=1}^j n_r)$.

However, while C is 1-dimensional, the full joint distribution is of dimension $1 + C + D + 2T$ and subject to D constraints (when we consider \mathbf{t} and \mathbf{x} as given datasets, otherwise in the simulation experiments when they should be simulated, the dimension is $1 + C + 2D + 3T$). Since a conjugate prior is not available, the standard approach of performing inference is then to initiate some of the sampling techniques which can be challenging, as will be explained in the next section.

3.2 Model Selection and Activation

The current section discuss two aspects of the proposed model. First, the flexibility in mathematically shaping the model as will be explained in the distributional assumptions section. Second, the challenges in making this model actually work as will be introduced in the inferential methods section.

3.2.1 Distributional Assumptions

Only two out of six variables are available information for us which are the discovery time points $\{t_i\}_{i=1:D}$ (or cumulatively the number of discoveries $\{\tau_j\}_{j=1:T}$ over the time units j) and the proxy of the efforts employed in making discoveries $\{x_j\}_{j=1:T}$. This makes our proposed model subject to several distributional assumptions and parameter specifications that vary according to the field in use. However, even these two variables need to be simulated when implementing some of the inferential methods. Simulating $\{x_j\}_{j=1:T}$ requires a model suggestion which depends on the chosen proxy and guided by the pattern of the observed data, while $\{t_i\}_{i=1:D}$ is resultant from satisfying the constraints and from simulating $\{d_i\}_{i=1:D}$ and $\{n_j\}_{j=1:T}$.

On the other hand, although the joint Geometric distribution of the inter-discovery samples, $\{d_i\}_{i=1:D}$, should be a straightforward choice based on the sampling theory, it would be subject for different model choices if we assume that sampling from the C -partitioned population occurs randomly *without* replacement so that draws from this population will not be equally likely to be sampled at any point in time. In addition, the sample sizes $\{n_j\}_{j=1:T}$ do not necessarily have to follow the Non-Homogeneous Poisson process. It is still parametrized by the covariate latent effort, but it can be a process other than the Poisson such as the Renewal process or the Point process.

The covariate effort process, as described in the introduction of the current chapter, is latent and cannot be directly measured. Therefore, several assumptions might be considered for modeling it, and all of them are subject to the field convenience and the expert knowledge. One may assume independence between the events of the latent effort process, while others may assume the Markov property or even more as it is the case in ARIMA models of higher orders [78]. Regardless of the dependency issue, one may suggest the truncated Gaussian model, or the half-Gaussian model to ensure positive support since logically the effort is something that cannot be negative. Moreover, one may suggest just the Gaussian model and use some kind of transformation such as log-transformation or linear transformation to ensure positive support for the latent effort *etc.*

Regarding the distribution of class-abundances, also known as commonness and rarity distribution or count data distribution, there are several suggestions in the literature of this context. It varies according to the application in use, for example, the Lognormal, Poisson Lognormal, and Double-Geometric distributions are used a lot in the number of species problem [79], while the Zip’f distribution is more convenient to apply in the linguistics field [80]. In the software reliability field, we can see the Negative Binomial as well as the Poisson distributions [81].

Figure 3.7 shows a simple chart that describes the change in the range of choices for modelling the six variables included in our framework. The bottom level of the chart represents the observable variables where we can have actual datasets. Having available data helps to recognize patterns and therefore narrow the range of model choices. In the middle level we have the variables that are supported by the theoretical logic but no concrete datasets which may widen the range of making decision about the model choices. Finally, in the top of the chart, we have a broad range of choices to make for modelling the class-abundances and latent effort which is mainly up to the experts and the field of application.

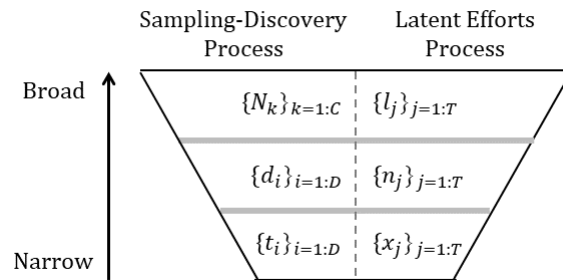


Figure 3.7: A chart of describing the range of model choices, which ranges from narrow (when we have actual datasets to guide the distributional assumptions) to broad (when the situation is subject to expert of the field). In the absence of concrete data (as in the middle), the theoretical logic is still good support for model choice.

As we are using Bayesian inference to model our proposal, the distributional assumptions are not restricted to the above six variables, but also reaches the prior distribution of C , and it is not restricted to the model choice only but also reaches the parameter specifications within a model. The prior distribution could be informative such as Poisson or truncated Poisson distributions as the

rational assumption is that C follows almost unimodal, or non-informative such as the discrete uniform distribution. With respect to the parameter specifications of $\boldsymbol{\theta} := \{\theta_N, \theta_l, \theta_x\}$, there is a flexibility to incorporate either a simple Bayesian model with fixed parameters suggested by experts or by some other means, or an empirical Bayesian model guided by estimated parameters from the given data. Another possibility is to adopt a complex hierarchical model by considering hyper-parametrization where the parameters are variables and have their own distributions. [66]

3.2.2 Inferential Methods

As C is a random variable with altering dimension, the gold-standard approach for obtaining its posterior distribution (which is a proportional to the full joint distribution) is to employ a Reversible-Jump Markov Chain Monte Carlo (RJ-MCMC) sampler [82]. In doing so, we need to derive full conditional distributions for each variable in the proposed model which requires acceptance-rejection techniques such as Metropolis-Hastings. It also requires simultaneous updating techniques such as Gibbs for the full joint distribution, see Figure 3.8.

The standard Metropolis-Hastings is used to simulate $P(\mathbf{n}|\text{rest})$ and $P(\mathbf{l}|\text{rest})$, while the reversible jump Metropolis-Hastings is used to simulate $P(C | \text{rest})$, $P(\mathbf{N} | \text{rest})$ and $P(\mathbf{d} | \text{rest})$, as \mathbf{N} and \mathbf{d} are affected by C which has altering dimension. Figure 3.8 shows which probability models (in circles) are needed to derive each full conditional distributions given the rest, for example, $P(\mathbf{N}|\text{rest})$ can be derived from the proportionality to the product of $P(\mathbf{N} | C)$ and $P(\mathbf{d} | \mathbf{N}, \mathbf{n})$. Unfortunately we now encounter a limitation in implementing these techniques in light of holding the model constraints. Even if exploring a $1 + C + D + 2T$ dimensional space sufficiently were not a limiting factor on its own, the probability of simultaneously satisfying all of the D constraints when proposing a new move/jump along the $\{d_i\}_{i=1:D}$ and $\{n_j\}_{j=1:T}$ dimensions is extremely small. This hampers the sampling convergence hence making such an approach impossible in practice.

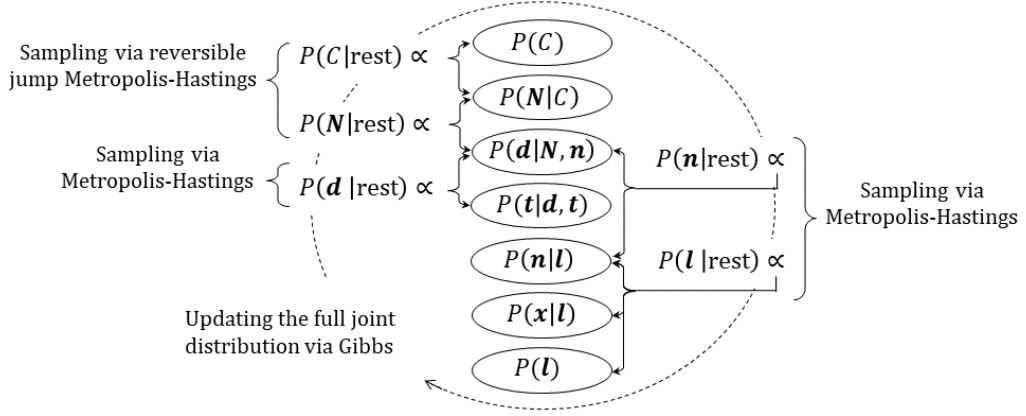


Figure 3.8: A diagram to simplify the idea of implementing the RJ-MCMC.

Another approach is to obtain a coarser approximation for the posterior distribution of C , based on iterative use of Bayes factor [83]. The idea is to suggest different values for C from its realm, ranging from D to multiple mD , where $m \in \mathbb{Z}^+$, and perform the Bayes factor which is the posterior distributions ratio for each two successive candidates. Then, the posterior can be approximated by invoking that $\sum_{r=1}^m P(C = rD | \mathbf{t}, \mathbf{x}) = 1$.

In the simplest case when using a uniform prior, the Bayes factor will be reduced to the likelihood ratio. For that we need to obtain the marginal likelihoods which is intractable in our case, and one way is to approximate it as follows,

$$\begin{aligned}
 P(\mathbf{t}, \mathbf{x} | C) &= \int_{\Psi} P(\mathbf{t}, \mathbf{x} | \Psi, C) P(\Psi | C) d\Psi \\
 &= E\{ P(\mathbf{t}, \mathbf{x} | \Psi, C) \} \\
 &\approx \frac{1}{R} \sum_{r=1}^R P(\mathbf{t}, \mathbf{x} | \Psi_r, C)
 \end{aligned} \tag{3.9}$$

where, $\{\Psi_1, \Psi_2, \dots, \Psi_R\}$ can be sampled from $P(\Psi | C)$. The above approximation (3.9) is supported by the law of large numbers, where the above estimation converges as $R \rightarrow \infty$. Such approximation is called the ‘*arithmetic mean*’ estimation of the marginal likelihood, at which $P(\Psi | C)$ represents a prior joint model of the set of the contextual variables Ψ given a proposed value of C . On the other hand, if we are interested in sampling the Ψ ’s from their posterior joint model

$P(\Psi | \mathbf{t}, \mathbf{x}, C)$, then the marginal likelihood will be approximated by the ‘*harmonic mean*’,

$$P(\mathbf{t}, \mathbf{x} | C) \approx \left(\frac{1}{R} \sum_{r=1}^R \frac{1}{P(\mathbf{t}, \mathbf{x} | \Psi_r, C)} \right)^{-1} \quad (3.10)$$

However, estimating the marginal likelihood through such sampling methods is still a difficult area due to the high variance of the estimator and its slow convergence to the true integral. Moreover, [84] conducted a comparison study of multiple computation methods for estimating the marginal likelihood, and the author concluded the following: “*The main finding is that both the method of computing the marginal likelihood as the expectation of the likelihood under the prior, as the method of computing the harmonic mean of likelihood values when sampling from the posterior, are not trustworthy...*” (see more details in [84]).

However, even if the estimation issue were not a limiting factor on its own, sampling from $P(\Psi | C)$ or $P(\Psi | \mathbf{t}, \mathbf{x}, C)$ returns us to the same limitation that we encountered in the RJ-MCMC, which is the complexity of high dimensional sampling along with satisfying the model constraints.

Subsequently, to overcome the above drawbacks we adopted the Approximate Bayesian Computation (ABC) approach (illustrated in Section 2.3.4), in particular the ABC rejection sampling algorithm. The idea of implementing this algorithm has similarities with the iterative implementation of Bayes factor in that it starts with proposing a range of values for C . However, in Bayes factor we can check only two proposed values for C at a time and we need to derive or approximate the likelihood, while in the ABC we skip obtaining the likelihood and we can check a set of proposals for C simultaneously.

The distinguishing property of this approach is that instead of iteratively estimating the posterior distribution of C by sampling through RJ-MCMC or Bayes Factors, we derive a posterior distribution approximation by sampling datasets. In the current study we adopt a parallel implementation of this approach in that; several ‘*artificial worlds*’ are generated according to our proposed framework and

a suggested range of values for C drawn from its prior distribution. Only those worlds which result in ‘artificial datasets’, $\boldsymbol{\tau}^*$ ¹ and \boldsymbol{x}^* , that are sufficiently close to the observed actual data, $\boldsymbol{\tau}$ and \boldsymbol{x} , are accepted. Hence, the empirical distribution of the proposed C^* values that correspond to the accepted artificial datasets, is taken as an approximation of the targeted posterior of C . Figure 3.9 shows a simple illustration of the parallel implementation of this algorithm.

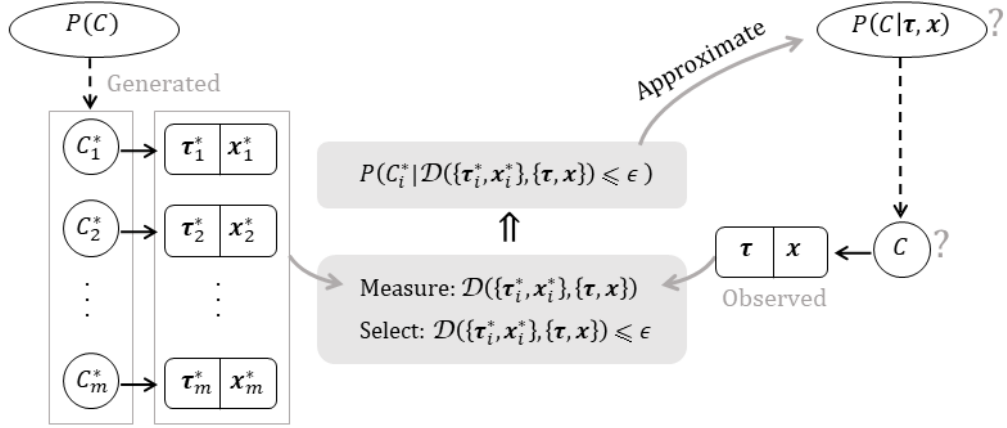


Figure 3.9: A diagram of simultaneously implementing the ABC rejection sampling algorithm. $\mathcal{D}(\{\boldsymbol{\tau}_i^*, \boldsymbol{x}_i^*\}, \{\boldsymbol{\tau}, \boldsymbol{x}\})$ is a notation for measuring the closeness between the artificial datasets and the actual given ones.

However, our implementation of the ABC rejection sampling algorithm is slightly different. As the proposed model includes two observed datasets $\boldsymbol{\tau}$ and \boldsymbol{x} , our implementation includes two filtering steps. Since $\boldsymbol{\tau}$ is resultant of almost all other variables in the model, while the covariate \boldsymbol{x} can be simulated independently of them except for one, the latent effort, the first filtering is applied only on \boldsymbol{x} , while the second filtering is applied on a combination of both $(\boldsymbol{\tau}, \boldsymbol{x})$. Note that the purpose of the first filtering is to speed up the computation time, while the second filtering is for approaching the accuracy of the estimation. Obtaining a filtering threshold requires running a pilot experiment (including 10,000 iterations) which is explained in details along with the main fitting in the next section.

¹For more convenience in the rest of the dissertation, we will use the number of discoveries $\{\tau_j\}_{j=1:T}$ instead of the discovery time points $\{t_i\}_{i=1:D}$.

3.2.3 The Adopted ABC Algorithm

ABC Algorithm of the Pilot Experiment:

1. Import the given observed data, the number of discoveries τ and the effort proxy \mathbf{x} . Note that the total number of the current discoveries D is determined by the sum of τ , and the time length T (in which the whole sampling and discovery process has occurred) is determined by the length of \mathbf{x} .
2. Determine fixed values for the hyper-parameters set $\theta := \{\theta_N, \theta_l, \theta_x\}$. Note that we choose to consider this set as fixed in the current study for the purpose of simplicity.
3. Generate a range of proposed values $C^* = \{C_1^*, C_2^*, \dots, C_{10,000}^*\}$ from a suggested prior distribution of the number of classes C (given the total number of current discoveries D). Note that each value of C^* directly feeds into generating N .
4. Within the 10,000 - iteration loop and for each value of C^* , generate values for the replicated variables $\{\tau^*, \mathbf{x}^*\}$ and for Ψ ; each from its suggested distribution (examples of suggested distributions will be introduced in the species context in the next chapter). The generation is implemented sequentially, following the arrows in the Bayesian network graph of our proposed model, such that generating N feeds into generating \mathbf{d} which in turns feeds into generating τ^* , and generating l feeds into generating \mathbf{n} and \mathbf{x}^* . However, by imposing the model constraints, \mathbf{d} is not only affected by N but also by \mathbf{n} , consequently, τ^* is not only affected by \mathbf{d} but also by \mathbf{n} .
5. Compute a distance metric between the observed dataset and the generated one for each replicated variable separately then add them *i.e.* $\mathcal{D}(\tau^*, \tau) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})$. The chosen distance metric in the current study is the ‘*Euclidean*’. It is worth mentioning that through an additional pilot experiment in the species context, we checked whether the performance of the model is affected by the distance metric choice, and we found that there are no significant differences between metrics such as *Manhattan*, *Minkowski*, *Canberra*, and the *Euclidean*, see Group B.25 of Section B.6 in Appendix B. In fact, most of these metrics are equivalently monotonic transformations except for some such as the *Maximum*.

However, the choice of the distance metric might be important according to the applied field, which is not in our case.

6. End up the loop with a list of primary distances $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})$ with their corresponding C^* values. This list is sorted ascendingly according to the distances.
7. Select the smallest 1000 distances, denoted by $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000}$, from which the threshold value is defined as $\epsilon := \max\{\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000}$, to be used in the main fitting experiment.

It is worth mentioning that the selected datasets (which have the smallest 1000 distances) represent 10% of the generated datasets, and they should be close enough to the given data if the model parameters are successfully specified. As will be illustrated through experimentation in the next chapters, we used a plot of accepted and rejected samples of \boldsymbol{x}^* and $\boldsymbol{\tau}^*$ as guidance for checking the closeness of the selected datasets. The closeness is defined by that the accepted area of the generated data should cover the given data (so that the given data is roughly placed in the middle of the accepted area), and should capture the annual pattern of the given data. For example, in the bottom panel of Figure 4.8 on page 76, we have two plots, the right-side represents the number of discoveries, while the left-side represents the number of authors. These plots show how the 1000 selected datasets (dark grey) are close to the given data (red), while the rejected datasets (light grey) are clearly far away. However, when the dark grey area of the selected datasets appear to be far away and does not cover the given data as shown in the right-side plots (a, b, c, and d) of Group B.17 in Appendix B on page 223, this is a clear indication of mis-parametrization which needs to be reconsidered.

ABC Algorithm of the Main Fitting Experiment:

1. Import the same given observed data, the number of discoveries $\boldsymbol{\tau}$ and the effort proxy \boldsymbol{x} , where $D = \text{sum}(\boldsymbol{\tau})$ and $T = \text{length}(\boldsymbol{x})$. In addition, import the threshold value $\epsilon := \max\{\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000}$ from the pilot experiment.
2. Determine the same fixed values for the hyper-parameters set $\boldsymbol{\theta} := \{\theta_N, \theta_l, \theta_x\}$ that is used in the pilot experiment.

3. Generate a range of proposed values $C^* = \{C_1^*, C_2^*, \dots, C_m^*\}$ from the same suggested prior distribution of C that is used in the pilot experiment.
4. Within a loop over a large number of iterations and for each value of C^* , generate values for the replicated variables $\{\boldsymbol{\tau}^*, \boldsymbol{x}^*\}$ and for $\boldsymbol{\Psi}$ from the same suggested distributions that are used in the pilot experiment. However, here the generation process is implemented in two steps, where the second step depends on the filtering output of the first step. In the first, the value of the latent effort \boldsymbol{l} is generated and feeds into generating a value of the proxy variable.
5. At a given index i , compute the Euclidean distance metric between the given observed dataset of the proxy variable and the generated one *i.e.* $\mathcal{D}(\boldsymbol{x}_i^*, \boldsymbol{x})$.
6. Accept \boldsymbol{x}_i^* if $\mathcal{D}(\boldsymbol{x}_i^*, \boldsymbol{x}) \leq \epsilon$, and proceed to the next step of generating $\boldsymbol{\tau}_i^*$. Reject otherwise, and go back for new generation.
7. Within the same loop and for the same value of C^* , based on the previous filtering step, continue generating the rest of the variables $\boldsymbol{\Psi}$ from the same suggested distributions that are used in the pilot experiment. Here, \boldsymbol{N} is generated and feeds into generating \boldsymbol{d} which in turns feeds into generating $\boldsymbol{\tau}^*$, also, the already generated \boldsymbol{l} proceed to feed into generating \boldsymbol{n} . Then, impose the model constraints so that \boldsymbol{d} is not only affected by \boldsymbol{N} but also by \boldsymbol{n} , and, $\boldsymbol{\tau}^*$ is not only affected by \boldsymbol{d} but also by \boldsymbol{n} .
8. End up the loop with a list of primary distances that passed the first filtering step along with their corresponding C^* values. This list is sorted ascendingly according to the distances.
9. The second filtering step is represented by selecting a sample of size ‘s’ of the smallest distance values, denoted by $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_s$, so that the empirical distribution of their corresponding C_s^* is taken as an approximation of the target posterior distribution of C , from which a point and interval estimation can be calculated. As indicated in section 3.1.4, The rational assumption is to assume that C follows almost a uni-model distribution, where $C \in \{\mathbb{N} \geq D\}$.

The pseudo code of the above algorithm is explained in (Appendix C, Section C.1, Algorithm 1). This algorithm is implemented with specified distributions and parameter-values in the ecological context to estimate the number of species in the next chapter.

3.3 Chapter Summary:

We started this chapter by describing the number of kinds problem as a black-box and we attempted to model what is inside that box. In the first section, we illustrated the structure of our proposed model which includes: the sampling-discovery process, the latent effort process, the linkage node of these two processes, and the final phase of the model. As our proposed framework is subject to modelling and parameter specifications, in the second section of the chapter, we discussed a variety of possible distributional assumptions. We also explored the challenges associated with well known inferential methods that are considered to activate the proposed model. We ended this chapter by explaining the ABC rejection sampling method which is the one found necessary to run our model.

Part II

Implementation Phase of The Proposed Approach

Evaluating our proposed model will be done through a case study in the ecology field which is estimating the number of species that exist on Earth. It is a typical indicator that is used to measure the extent of biodiversity on Earth which is of critical importance if we wish to ensure its perseverance. Though attempts to measure biodiversity extent are normally performed on a taxon or a localised geographical level, applied on a global scale, estimates have collectively ranged from 1.5 to over 100 million species on Earth, throughout the species literature [13], which is clearly a large range of estimates. These estimates are highly dependent on the methodology and data used to derive them. A common perspective of most of these attempts is that they have relied on one of two approaches, either the sampling actions or the discovery curves, neither of which considers associated factors linked to species discovery such as the discovery efforts. However, we believe that through our proposed framework a comprehensive analysis for the situation is undertaken to provide worthy and reliable estimates of species richness.

Our proposed model will be analysed, characterized, implemented, and validated through artificial and real datasets. The artificial data have been generated by the model, while the real data can be retrieved from many biodiversity information systems such as the Catalogue of Life (CoL) [85] and the World Register of Marine Species (WoRMS) [86]. Through these inventory systems, it is shown that the number of species descriptions has been growing over years. Figure II shows an example of such data that we are dealing with in the species discovery field.

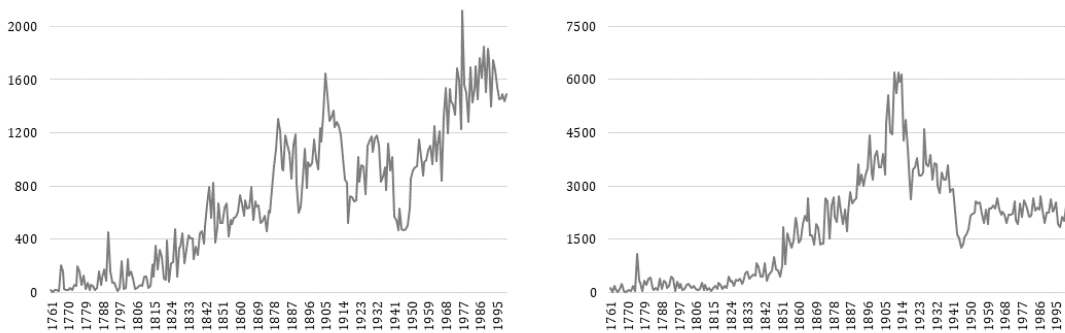


Figure II: To the left is a subset of data from WoRMS, while to right is from CoL. These represent the number of discovered species $\{\tau_j\}_{j=1:T}$ over $T = 240$ years, from 1761 to 2000.

The following three chapters concern the model formulation and activation within the species context. Several artificial worlds are generated according to the distributional assumptions related to the species field. The main artificial worlds are covered in Chapter 4 to validate our proposed model, while the rest are branches of them that are covered by Chapter 5 to evaluate the model. Finally, our proposed model is implemented on a real world of species in Chapter 6.

Chapter 4

Model Formulation & Simulation

The current chapter involves two main sections: the first section introduces the mathematical formula of our model according to the distributional assumptions of species richness and discovery, while the second section concerns the model validation through simulation experiments.

4.1 Distributional Specifications

As stated in Section 3.2.1, our model is subject to several distributional assumptions and parameter specifications. However, in the current case study we keep adopting the equally likely sampling setting. Therefore, in the sampling and discovery process, the inter-discovery samples $\{d_i\}_{i=1:D} = \{\delta_i - \delta_{i-1}\}_{i=1:D}$ are still following a joint Geometric distribution conditioned on the knowledge of the species abundances $\{N_k\}_{k=1:C}$ up to time t_D ,

$$P(d_{2:D}|N_{1:C}) = \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \quad (4.1)$$

where, based on species literature [79] [87], we choose one of the distributional assumptions of the species abundances $\{N_k\}_{k=1:C}$ which is the Log-Normal distribution with parameters $\boldsymbol{\theta}_N = \{\mu, \sigma\}$,

$$P(N_{1:C}|\mu, \sigma) = \prod_{k=1}^C \frac{1}{N_k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-1}{2} \left(\frac{\log(N_k) - \mu}{\sigma} \right)^2 \right] \quad (4.2)$$

In the latent effort process $\{l_j\}_{j=1:T}$, the annual amount of taxonomic effort is found to be more informative covariate factor in the environment of species discovery. One suggestion that seems more natural and computationally con-

venient is to assume that this process is characterized by a Gaussian Markov property, evolving smoothly over years. However, since the Gaussian distribution has infinite support $(-\infty, +\infty)$, but as the effort is something that should not be negative, it is plausible to apply the Gaussian model on a log-scale to ensure that $l_j > 0$, *i.e.* on a log-scale, $\{l_j\}_{j=1:T}$ is assumed to be a discrete Log-Gaussian Markov process. For notational convenience, we define another variable L_j such that $L_j = \log(l_j)$, so that $\{L_j\}_{j=1:T}$ itself evolves as a discrete Gaussian Markov process with mean L_{j-1} and standard deviation σ_L , where $\boldsymbol{\theta}_L = \{L_{j-1}, \sigma_L\}$,

$$P(L_{2:T}|L_1, \sigma_L) = \prod_{j=2}^T \frac{1}{\sqrt{2\pi\sigma_L^2}} \exp \left[\frac{-1}{2} \left(\frac{L_j - L_{j-1}}{\sigma_L} \right)^2 \right] \quad (4.3)$$

Note that the starting point L_1 has a high influence on the above process. However, since it has no specified prior distribution, the given joint distribution in equation (4.3) is not well-defined. Such a distribution is called an Intrinsic Gaussian Markov process (more details about such a case is discussed in [88]).

With respect to the proxy of the latent taxonomic efforts process, several choices could be made. For example, the number of authors involved in species description, the number of publications in this field, the amount of research funding awarded, or the number of excursions performed over time, *etc.* However, choosing a proxy is sometimes influenced by the availability of its data with respect to the targeted community of species. In our case study, the number of authors has some justification as a good choice due to its availability (as the number of authors is the most available information over several taxa compared to other factors such as the expeditions funding which is limitedly available to some taxa only) and due to its co-varying behaviour with the number of discoveries [13]. Although there is a high correlation between the number of publications and the number of authors [89], we believe that the latter includes a wider range of discoveries because there are some discovered species not being associated with publications, but recognized by other authorities such as museums. Figure 4.1 shows the same example of the CoL subset data and WoRMS subset data that was shown previously in Figure II in the introduction of Part II, but now with the inclusion of the number of authors $\{x_j\}_{j=1:T}$ in addition to the discoveries $\{\tau_j\}_{j=1:T}$.

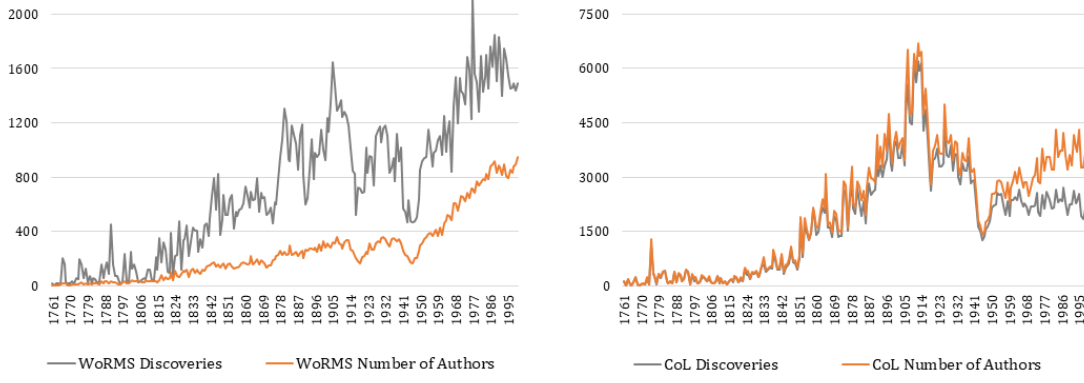


Figure 4.1: The data subset from the WoRMS and the CoL inventories, after including the proxy variable. These represent the number of discovered species $\{\tau_j\}_{j=1:240}$ (grey) and the number of authors (proxy) $\{x_j\}_{j=1:240}$ (orange), from 1761 to 2000.

We can consider the number of authors as a result of counting action, hence, we can fairly assume that it evolves as a Non-Homogeneous Poisson process with a changing rate $[\exp(L_j)/\exp(\alpha)]$,

$$P(x_{1:T}|L_{1:T}, \alpha) = \prod_{j=1}^T \left[\frac{\exp(L_j)}{\exp(\alpha)} \right]^{x_j} \exp \left[-\frac{\exp(L_j)}{\exp(\alpha)} \right] / x_j! \quad (4.4)$$

Note that $\exp(\alpha)$ is the expected number of specimens to be collected per author per year j , while $\exp(L_j)$ is the effort rate.

This leads us to characterize the distribution of the linkage node which represents the sample sizes $\{n_j\}_{j=1:T}$. As introduced in Section 3.1.3, it is more convenient to assume that the sample sizes follow a Non-Homogeneous Poisson process that is parametrized somehow by the latent effort process. However, after setting the distributional selection for the latent effort in equation (4.3), it is now clear that the distribution of $\{n_j\}_{j=1:T}$ is parametrized by the effort rate $\exp(L_j)$,

$$P(n_{1:T}|L_{1:T}) = \prod_{j=1}^T [\exp(L_j)]^{n_j} \frac{\exp[-\exp(L_j)]}{n_j!} \quad (4.5)$$

As indicated in Section 3.1.3, the discovery index δ_i is constrained by $\sum_{r=1}^j n_r$, and the discovery time point t_i is constrained by j and also by $\{n_j\}_{j=1:T}$ through $\{d_i\}_{i=1:D} = \{\delta_i - \delta_{i-1}\}_{i=1:D}$ so that $t_i = j \times \mathbb{1}(\delta_i \leq \sum_{r=1}^j n_r)$, where the distribution of each time point t_i follows a Degenerate distribution that parametrized by

these constraints $d_i \leq \sum_{r=1}^j n_r - \delta_{i-1}$, see equations (3.2) and (3.3). However, under satisfying the constraints, the distribution of $\{t_i\}_{i=1:D}$ will be $P(t_{1:D}|d_{1:D}, n_{1:T}) = 1$. It is worth remembering that, as indicated in Section 3.1.3, for notational convenience, $\tau_j = [\sum_{v_i} \mathbb{1}(j-1 < t_i \leq j)]_{v_j}$ is used instead of t_i to synchronize the evolving time of the sampling-discovery and the latent effort processes.

At this stage, we specified the species distributional assumptions of the six variables that are included in our proposed model, in equations (4.1) to (4.6). With respect to the prior distribution of C , we choose a Uniform distribution defined on the support $[D, mD]$, where m is a multiple for the total number of the discoveries, and can be selected according to the applied experiment as we will see in the artificial worlds in the next section and the real worlds in the next chapter. Therefore, by combining the equations (3.2), (3.3), and (4.1) to (4.6) along with satisfying the constraints $d_i \leq \sum_{r=1}^j n_r - \delta_{i-1}$, and fixing the hyper-parameters set $\boldsymbol{\theta} := \{\mu, \sigma, L_1, \sigma_L, \alpha\}$, the final mathematical formula of the full joint probability distribution provided by our model within the species context is:

$$\begin{aligned}
& P(C, \mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}, \mathbf{x}, \mathbf{t}) = \\
& P(C) \times P(\mathbf{N} | C) \times P(\mathbf{d} | \mathbf{N}) \times P(\mathbf{L}) \times P(\mathbf{n} | \mathbf{L}) \times P(\mathbf{x} | \mathbf{L}) \times P(\mathbf{t} | \mathbf{d}, \mathbf{n}) = \\
& \left\{ \frac{1}{mD - D} \right\} \times \left\{ \prod_{k=1}^C \frac{1}{N_k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-1}{2} \left(\frac{\log(N_k) - \mu}{\sigma} \right)^2 \right] \right\} \times \\
& \left\{ \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \right\} \times \left\{ \prod_{j=2}^T \frac{1}{\sqrt{2\pi\sigma_L^2}} \exp \left[\frac{-1}{2} \left(\frac{L_j - L_{j-1}}{\sigma_L} \right)^2 \right] \right\} \times \\
& \left\{ \prod_{j=1}^T [\exp(L_j)]^{n_j} \frac{\exp[-\exp(L_j)]}{n_j!} \right\} \times \left\{ \prod_{j=1}^T \left[\frac{\exp(L_j)}{\exp(\alpha)} \right]^{x_j} \exp \left[-\frac{\exp(L_j)}{\exp(\alpha)} \right] / x_j! \right\} \times \{1\}
\end{aligned} \tag{4.6}$$

By excluding the fixed values from the above expression and rearranging some terms, the full joint distribution can be proportionally simplified to the following,

$$\begin{aligned}
& P(C, \mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}, \mathbf{x}, \mathbf{t}) \propto \\
& \left(\prod_{k=1}^C \frac{1}{N_k} \right) \left(\prod_{j=1}^T \frac{1}{x_j!} \right) \left(\prod_{j=1}^T \frac{1}{n_j!} \right) \left\{ \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \right\}
\end{aligned}$$

$$\exp \left\{ \frac{-1}{2\sigma^2} \sum_{k=1}^C (\log(N_k) - \mu)^2 + \frac{-1}{2\sigma_L^2} \sum_{j=2}^T (L_j - L_{j-1})^2 + \sum_{j=1}^T (n_j + x_j)L_j - \alpha x_j - (1 + e^{-\alpha})e^{L_j} \right\} \quad (4.7)$$

Therefore, given the knowledge of the number of discoveries \mathbf{t} and the number of authors \mathbf{x} , the posterior model of our target variable, the number of species C , should be derived by integrating out all the contextual variables $\Psi := \{\mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}\}$,

$$P(C | \mathbf{t}, \mathbf{x}) = \int_{\Psi} P(C, \Psi | \mathbf{t}, \mathbf{x}) d\Psi \quad (4.8)$$

where, $\mathbf{N} = \{N_{1:C}\}$, $\mathbf{d} = \{d_{1:D}\}$, $\mathbf{t} = \{t_{1:D}\}$, $\mathbf{n} = \{n_{1:T}\}$, $\mathbf{l} = \{l_{1:T}\}$, $\mathbf{x} = \{x_{1:T}\}$. However, due to the complexity and high dimensionality of this model, the posterior will be approximated by using our suggested ABC algorithm (Section 3.2.3 and Algorithm 1 of Section C.1 in Appendix C). More details about moving from equation (4.6) to equation (4.7) are available in Appendix A, Section A.2. Figure 4.2, shows our proposed model with the selected distributions in the species context. These distributions will be adopted throughout all the simulation experiments in the rest of the thesis.

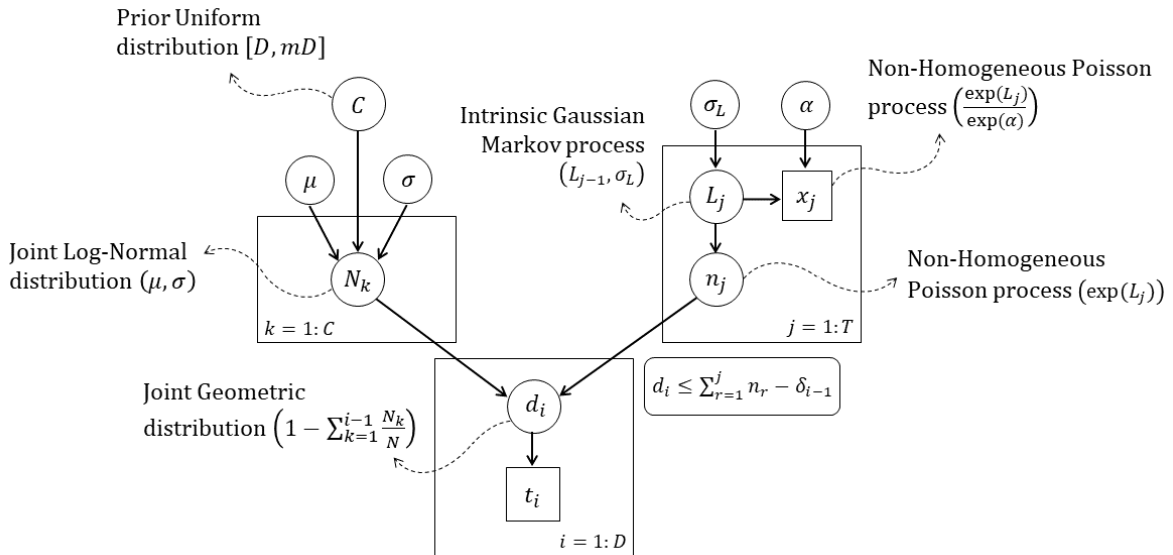


Figure 4.2: A Bayesian network displaying our proposed model within the species' distributional assumption.

4.2 Simulation and Generic Characteristics

Several simulation experiments have been conducted within the above distributional assumptions. They are used as an independent benchmark to evaluate the performance of our proposed model and describe its main properties. Three artificial worlds - namely *World-A*, *World-B*, and *World-C* - are generated according to three combinations of parametrizations and three selected values for the number of species, tabulated (Table 4.1) as follows:

Table 4.1: Parametrization set of three generated artificial worlds.

Number of Species	World-A	World-B	World-C
C	10,000	100,000	50,000

Hyper-Parameters	World-A	World-B	World-C
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	{10, 3.5}	{6, 2}
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	{7, 0.15}	{7, 0.08}
$\theta_x = \alpha$	$\log(30)$	$\log(100)$	$\log(50)$

These worlds are generated over a specified time length $T = 250$ years. The annual datasets of the number of discoveries $\{\tau_j\}_{j=1:250}$ and the number of authors $\{x_j\}_{j=1:250}$ of each world are plotted, along with recording the total number of population N , the total drawn sample n , and the current number of discoveries D that resulted from each simulation, as shown in Figure 4.3.

The three worlds are generated to represent three different patterns of the discovery effort. Although, they all start from similar angles which is high number of discoveries and low number of authors, they progress and finish differently, as shown in Figure 4.3. The logical justification of this start up pattern is to reflect the fact that with the ‘*Linnaean taxonomy*’ initiation in 1735, the already available species (which are many) are considered new as they are described for first time, while the taxonomic effort was at its beginnings. However, Figure 4.1 does not show this start up pattern in the given real data, as it is neglected in the selected subsets of the CoL and WoRMS data bases (more explanation will be in Chapter 6). In World-A, the number of discoveries and the number of authors

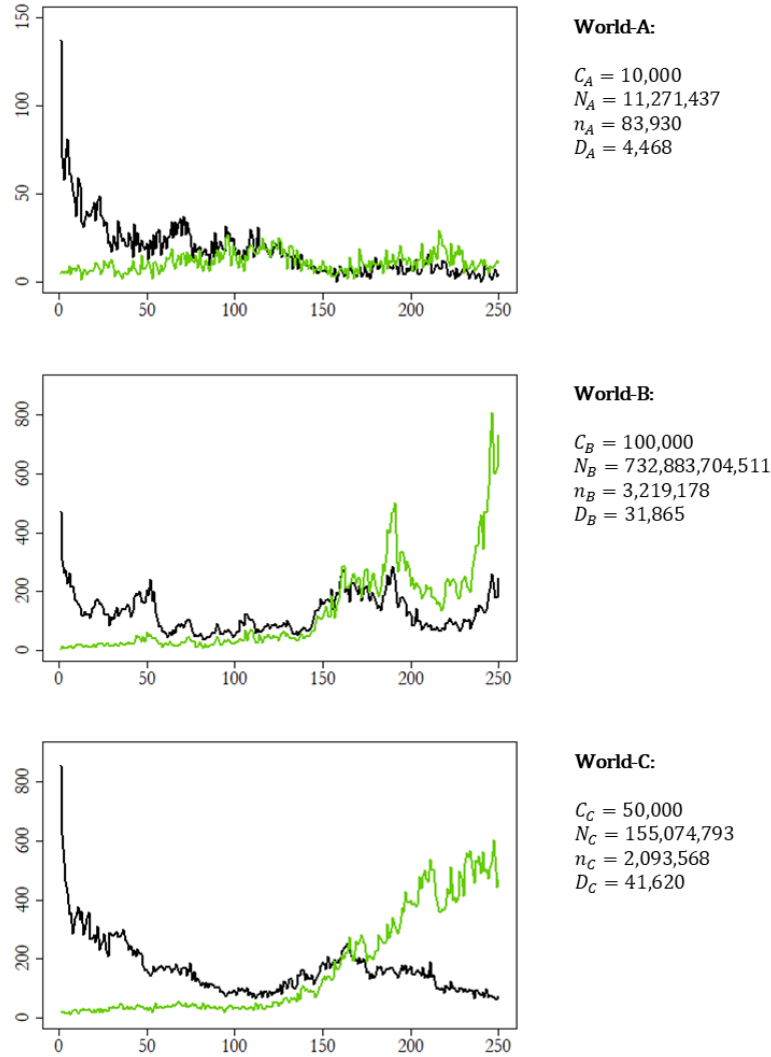


Figure 4.3: The annual number of discoveries (Black) and the number of authors (Green) over 250 years, along with basic information about the three artificial worlds.

stay adjacently stationary almost during the last 100 years with a very slight decrease at the end, while it is completely the opposite in World-B and World-C at which the number of authors is growing further from the number of discoveries. In World-B the two curves witness big fluctuations that end up with a noticeable increase in both the number of discoveries and the number of authors, while in World-C the increase in the number of authors is associated with a decrease in the number of discoveries.

We believe that these three artificial worlds reflect three distinct patterns that can fairly capture a sufficient variety of scenarios, in order to objectively evaluate our proposed model. From these three patterns, we can understand the

benefit of exploiting the latent effort process in explaining the discovery process, and therefore estimating the remaining number of species yet to be discovered. By the end of year 250, the total number of discoveries is almost half in World-A at which the number of authors and the number of discoveries are both end up with a stationary pattern. This pattern indicates that the employed efforts is just enough to keep the same level of the discovery rate, and there is a potential of revealing more discoveries if we increased the rate of the effort. This interpretation is approved by having just 44.7% total discoveries in World-A. The curves in World-B reflect that there are still more species to be discovered, as the number of discoveries increases along with the high increase of the number of authors. This interpretation is approved by the low number of total discoveries 31.9% by the end of year 250. On the other hand, World-C shows that when the number of discoveries decreases while the number of authors highly increases, this is a reflection of making a lot of discoveries and what is left is few to be discovered. This interpretation is also approved by the high number of total discoveries 83.24% by the end of year 250.

Efficiency Development:

After generating an artificial world, the generated data are fitted to the model with our suggested ABC algorithm. It is worth mentioning that the fitting performance itself went through stages of efficiency development. The first transformation was made when we adopted the ABC approach instead of the RJ-MCMC and the iterative implementation of the Bayes factor. We started the ABC coding using nested loops which took a huge amount of computation time. There were one main loop, and two sub-loops for generating δ_i and t_i . Then, we replaced the two sub-loops with small routines that accomplish tasks simultaneously, but kept the main loop, which extremely enhanced the computation time. However, we made a considerable reduction in the computation time by implementing the parallelization scheme instead of the sequential main loop. The last enhancement we considered is by including two filtering steps (as explained in Section 3.2.3 in the main fitting experiment) instead of one at which we exploited the existence of the latent effort proxy by establishing a threshold based on the generated values of the number of authors $\{x_j\}_{j=1:T}$.

Figure 4.4 shows an example of the last four stages of the efficiency development measured in seconds. This example is made on fitting World-A with only 10,000 iterations. Through several experiments we notice that the computation time is linearly increased with the number of iterations with order 10, which means when we increase the number of iterations from 10,000 to 1,000,000 the computation time will be increased by 100-times. The fitting time on the ‘*one-loop*’ code takes more than 2.5 hours with 1,000,000 iterations, while on the ‘*Parallel+2 filters*’ code takes only 0.5 hours with the same number of iterations and same world. Therefore, although in the figure it seems that the difference is little among the last three schemes, it in fact becomes huge in serious experiments when we implement higher number of iterations.

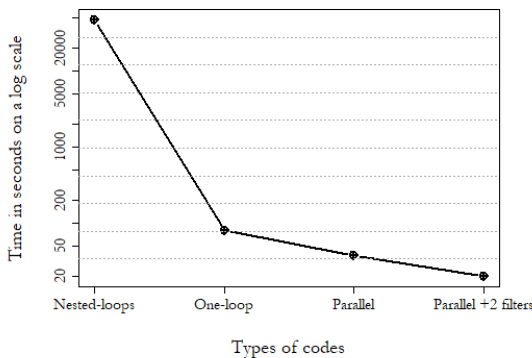


Figure 4.4: The huge difference between the Nested-loops scheme and the last three schemes in the computation time (seconds) measured on a log scale. However, there is still a considerable difference among the last three.

Threshold Justification:

The decision to include this additional filter is supported by noticing a high correlation between the distance values measured on the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distance values of adding both the number of authors and the number of discoveries $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$. Figure 4.5 shows this correlation as witnessed in the pilot experiments of the three artificial worlds.

However, this pattern does not mean that $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$ is the dominant in measuring the total distances, we still need to include $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$ in order to estimate the posterior distribution of the number of species. It is just that the

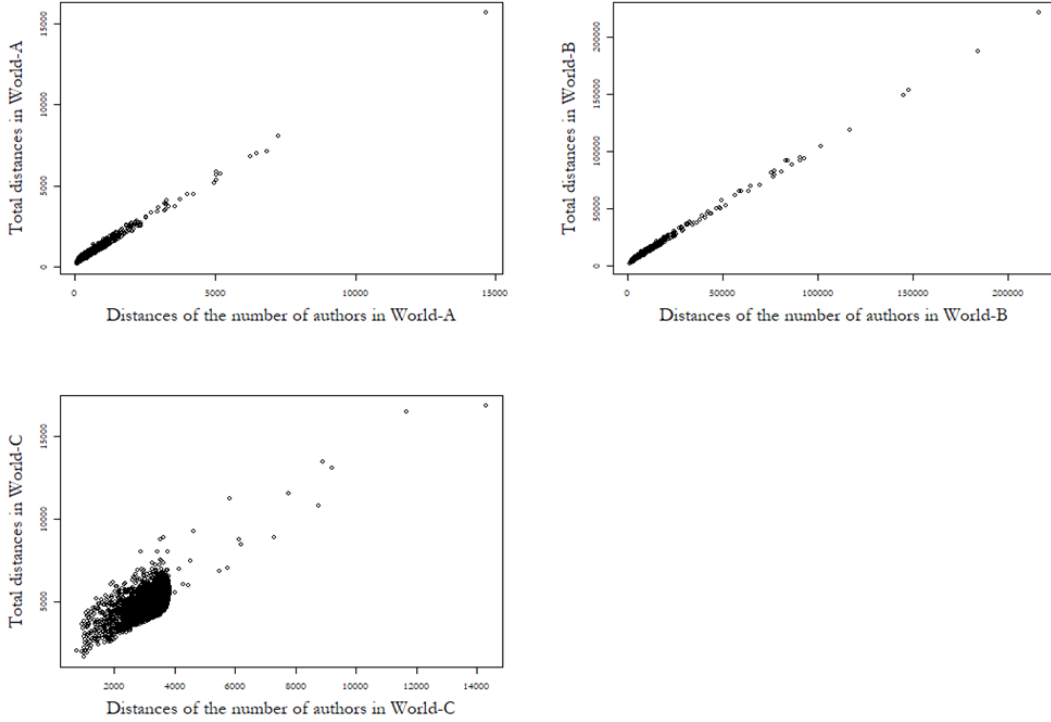


Figure 4.5: Scatter plots of the pilot experiments of the three artificial worlds, that show the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$.

scale at which $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$ varies is smaller than the scale of $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$. Moreover, this pattern is noticed in the pilot experiments that only involve 10,000 iterations, while in the main experiments where the number of iterations is 10^2 times higher, the correlation pattern becomes obvious between the number of discoveries $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$, and the total distance values $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$, as shown in Figure 4.6. Note that these main experiments are not only higher in the number of iterations but also involve the threshold that we based on $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$.

It is worth mentioning that this correlation pattern with the total distances becomes equivalent for $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$ and $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$ in the results of the final sample that is used to represent the posterior distribution of C , as shown in Figure 4.7. This sample is of size 1000 in all the three artificial worlds, as explained in the adopted ABC algorithm (Section 3.2.3 and Algorithm 1 of Section C.1 in Appendix C).

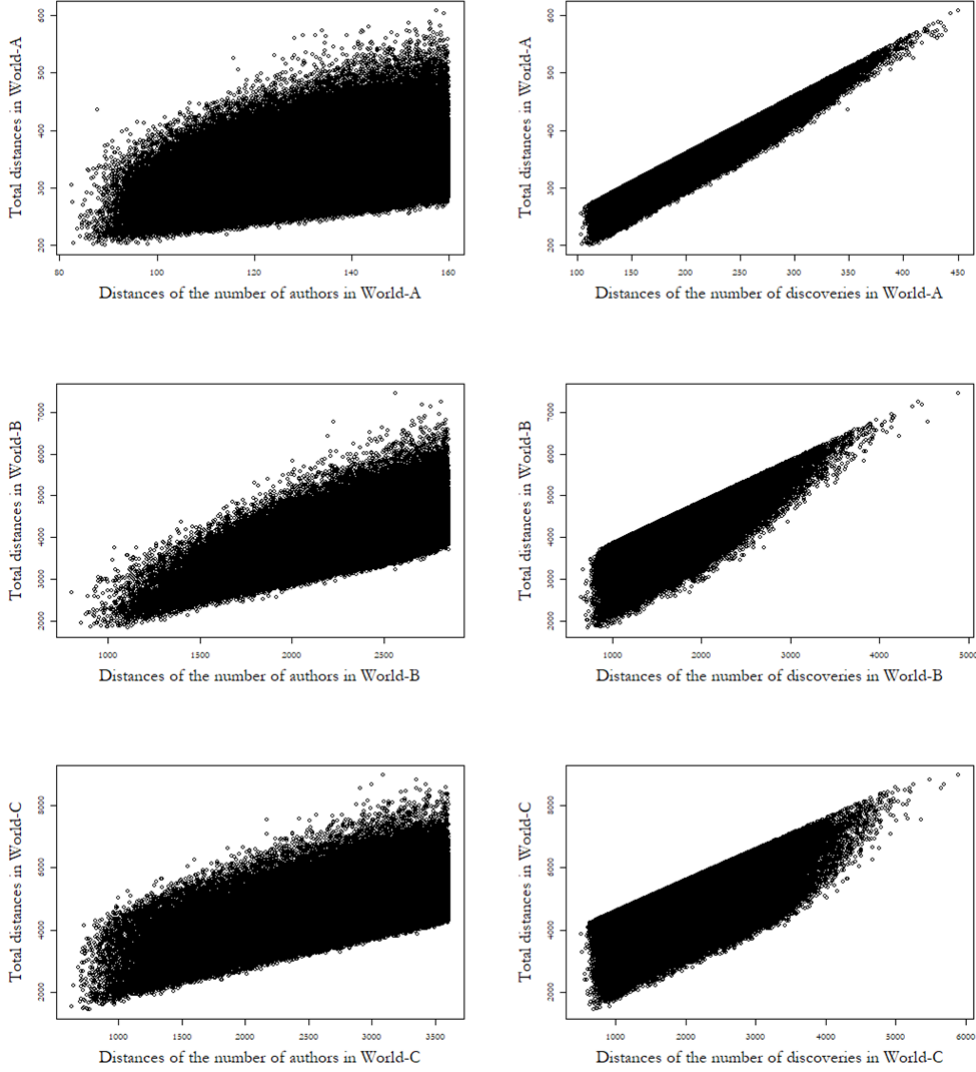


Figure 4.6: Scatter plots of the main experiments of the three artificial worlds. The left panel shows the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances out of the sum $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$, while the right panel shows the correlation between distances of the number of discoveries $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$.

In the following subsections, we explore the model main characteristics through fitting the three artificial worlds after adopting the ‘*Parallel+2 filters*’ scheme in the ABC algorithm:

4.2.1 Fitting Artificial World-A:

We fitted the given data set of the artificial World-A according to the ABC algorithm with 1,000,000 iterations. This fitting is based on using a uniform prior support at which $C \in [D, 5D]$, where $D = 4,468$ and $C = 10,000$. The upper bound of the prior support is specified through a few trials of pilot experiments,

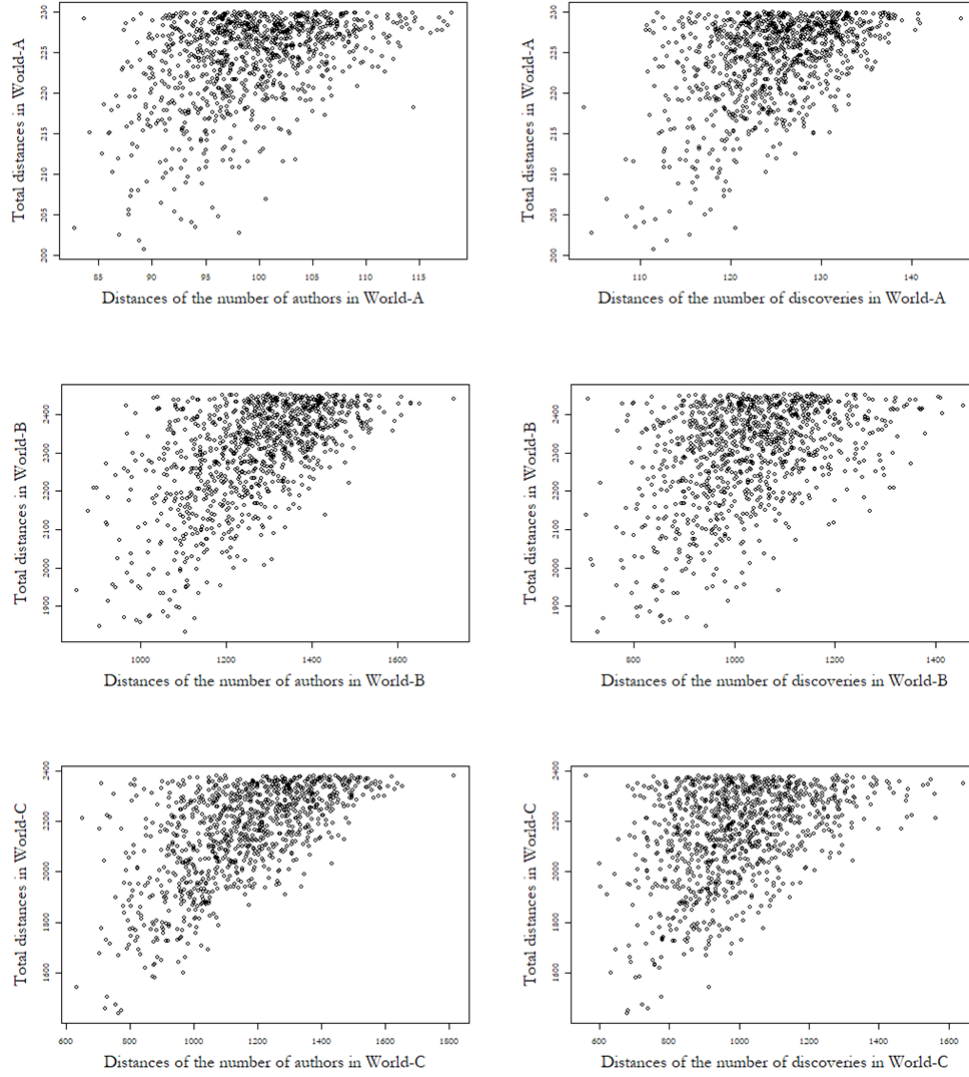


Figure 4.7: Scatter plots of the resultant final sample in the three artificial worlds. The left panel shows the correlation between distances of the number of authors $\mathcal{D}(\mathbf{x}^*, \mathbf{x})$, and the total distances out of the sum $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$, while the right panel shows the correlation between distances of the number of discoveries $\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})$, and the total distances $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x})\}$.

and then detecting when the right-side tail of the estimated posterior distribution is getting lighter. As we can see in the top-right plot of Figure 4.8, the maximum possible value of C (shown in the x-axis) is 18,000 which is less than the upper bound of the support $5D = 22,340$. This gives us an indication that there is no point to explore larger values in the C support, which explains our choice of the prior support $[D, 5D]$. The estimation of the posterior distribution of the number of species C is based on a final selected sample of size $s = 1000$ (as explained in Section 3.2.3). A point estimation (posterior mean and mode) of C is derived as well as a 95% credible interval, as shown in Table 4.2.

Table 4.2: Artificial World-A of total number of species $C = 10,000$.

World Parameterization		Fitting Results	
$C \in [D, 5D]$	$C \in [4,468 - 22,340]$	Posterior Mean	10,320
$\theta_N = \{\mu, \sigma\}$	$\{4, 2.5\}$	Posterior Mode	9,970
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	$\{5, 0.1\}$	95% Credible Interval	[7,222 - 14,163]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	1,850

Figure 4.8, the top-right plot, shows that the estimated posterior distribution of the number of species in the World-A seems to be a unimodal symmetric distribution but left-truncated which is justifiable since the support of C is already left-truncated by D *i.e.* $C \in \{\mathbb{N} \geq D\}$. The distribution is concentrated around the true value of C at which the posterior mean = 10,320 and the posterior mode = 9,970, which are very close to the true value. According to this fitting, with a probability = 0.95 the estimated value of C is greater than the current discoveries D by almost 3,000, but at most it roughly does not exceed 14,000 species.

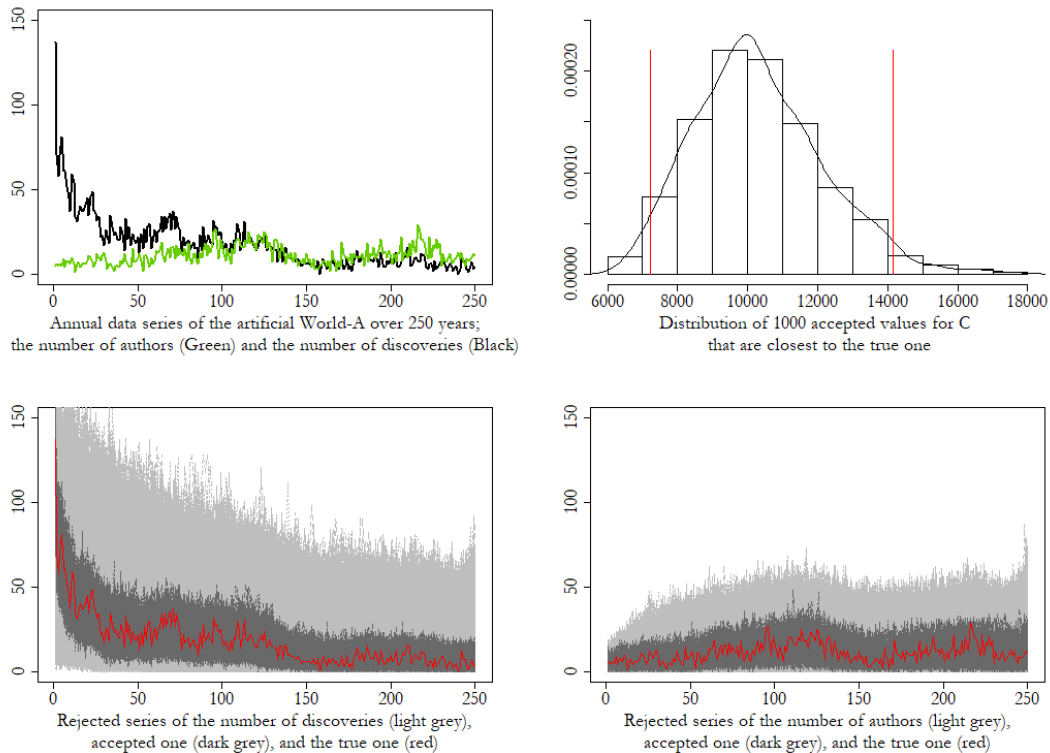


Figure 4.8: Fitting results of the World-A.

In other words and from another angle, this 95% credible interval tells us that we have at least roughly $D/2$ remaining species yet to be discovered and at

most roughly $2D$ unknown species. Therefore, having this narrow interval, along with having a stationary pattern of discoveries that is associated with a stationary curve for the authors, indicates that what we have discovered is still beyond half of the available species. If we pretended that we did not know the true value of C in the World-A, we can see that this speculation is consistent with the fact of having about 45% discoveries (which is a fact that we do not know if we are in a real world).

The bottom panel of Figure 4.8 shows about half of the 1,000,000 generated series from which only 1000 series are accepted (dark grey) which are the closest ones to the true given series (red) for both the number of discoveries $\{\tau_j\}_{j=1:250}$ and the number of authors $\{x_j\}_{j=1:250}$.

4.2.2 Fitting Artificial World-B:

We fitted the data set of the artificial World-B according to the ABC algorithm also with 1,000,000 iterations as in the first case. This fitting is based on using a uniform prior support at which $C \in [D, 8D]$, where $D = 31,865$ and $C = 100,000$. Here as well, we specified the upper bound of the prior support through multiple trials. We started with prior support $[D, 5D]$, and we found that it was not sufficient to explore the entire support of C as the estimated posterior distribution looks to have a heavy-tail or truncated on the right-side. Thus, we explored wider ranges until we found that the prior support $[D, 8D]$ is the sufficient one. The estimation of the posterior distribution of the number of species C is also based on a final selected sample of size $s = 1000$. A point estimation (mean and mode) of C is derived as well as a 95% credible interval, as shown in Table 4.3.

Table 4.3: Artificial World-B of total number of species $C = 100,000$.

World Parameterization		Fitting Results	
$C \in [D, 8D]$	$C \in [31,865 - 254,920]$	Posterior Mean	109,793
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	97,160
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[48,927 - 188,961]
$\theta_x = \alpha$	log(100)	Standard Deviation	37,053

Figure 4.9, the top-right plot, shows that the estimated posterior distribution of the number of species in the World-B seems to be also a unimodal symmetric left-truncated distribution that is concentrated around the true value of C at which the posterior mean = 109,793 and the posterior mode = 97,160. According to this fitting, with a probability = 0.95 the estimated value of C is greater than the current discoveries D by more than 17,000, and at most it does not exceed 189,000 species.

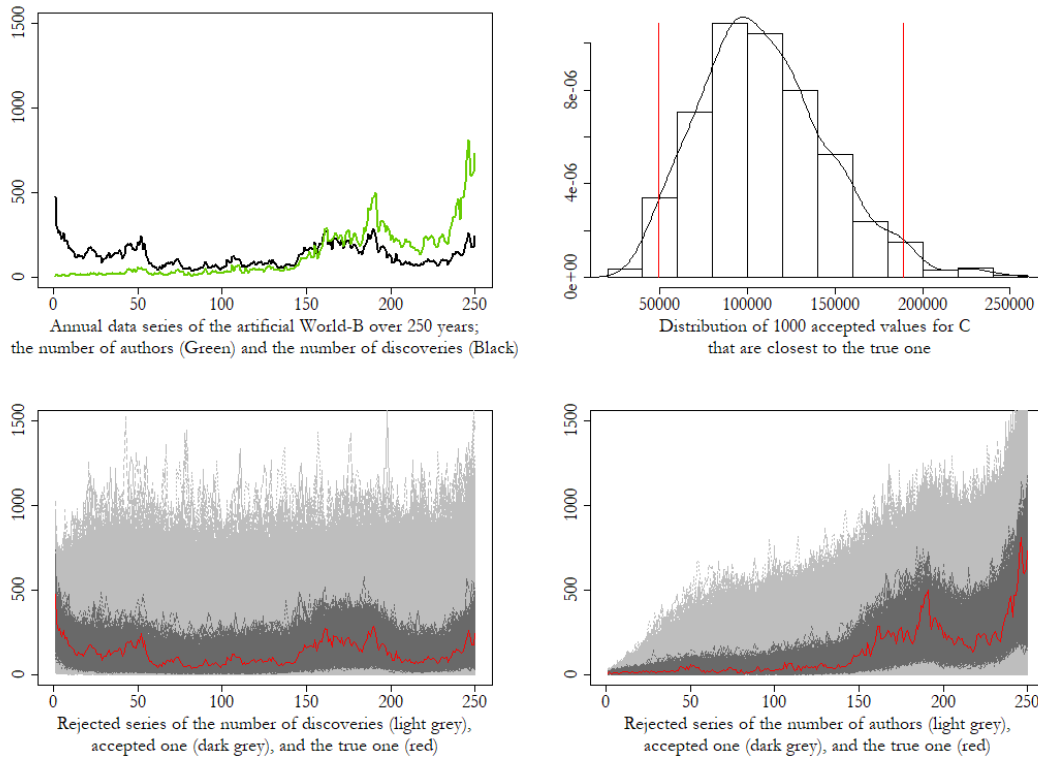


Figure 4.9: Fitting results of the World-B.

Also like World-A, this 95% credible interval of the World-B tells us that we have at least almost $D/2$ remaining species yet to be discovered. However, at the maximum limit of this interval there are roughly up to $5D$ species are still to be discovered, which reflects a very wide credible interval. Hence, having this big range, along with having an increasing pattern of the late discoveries that is associated with an increasing curve for the authors, indicate that there are still large amount of species waiting to be revealed. Again, if we pretended that we did not know the true value of C in the World-B, we can see that this speculation is consistent with the fact of having just 32% discoveries in this world.

The bottom panel of Figure 4.9 shows about third of the 1,000,000 generated series from which only 1000 series are accepted (dark grey) which are the closet ones to the true given series (red) for both the number of discoveries $\{\tau_j\}_{j=1:250}$ and the number of authors $\{x_j\}_{j=1:250}$.

4.2.3 Fitting Artificial World-C:

We fitted the given data set of the artificial World-C according to the ABC algorithm again with 1,000,000 iterations. Through several trials with different upper bounds on the support of C , this fitting is based on using a uniform prior support at which $C \in [D, 3D]$, where $D = 41,620$ and $C = 50,000$. The estimation of the posterior probability distribution of the number of species C is again based on a final selected sample of size $s = 1000$. A point estimation (posterior mean and posterior mode) of C is derived as well as a 95% credible interval, as shown in Table 4.4.

Table 4.4: Artificial World-C of total number of species $C = 50,000$.

World Parameterization		Fitting Results	
$C \in [D, 3D]$	$C \in [41,620 - 124,860]$	Posterior Mean	53,161
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	48,969
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7, 0.08}	95% Credible Interval	[42,280 - 69,765]
$\theta_x = \alpha$	$\log(50)$	Standard Deviation	7,703

Figure 4.10, the top-right plot, shows that the estimated posterior distribution of the number of species in the World-C is a right-skewed distribution, but both the posterior mean = 53,161 and the posterior mode = 48,969 are still around the true value of C . According to this fitting, with a probability = 0.95 the estimated value of C is greater than the current discoveries by just nearly 600, and at most it roughly does not exceed 70,000 species, if we consider the central credible interval (shown in red lines). On the other hand, the 95% HPD interval ranges from 41,711 (which is just about 90 species higher than what is already discovered) to 67,333 total number of species (shown in blue lines).

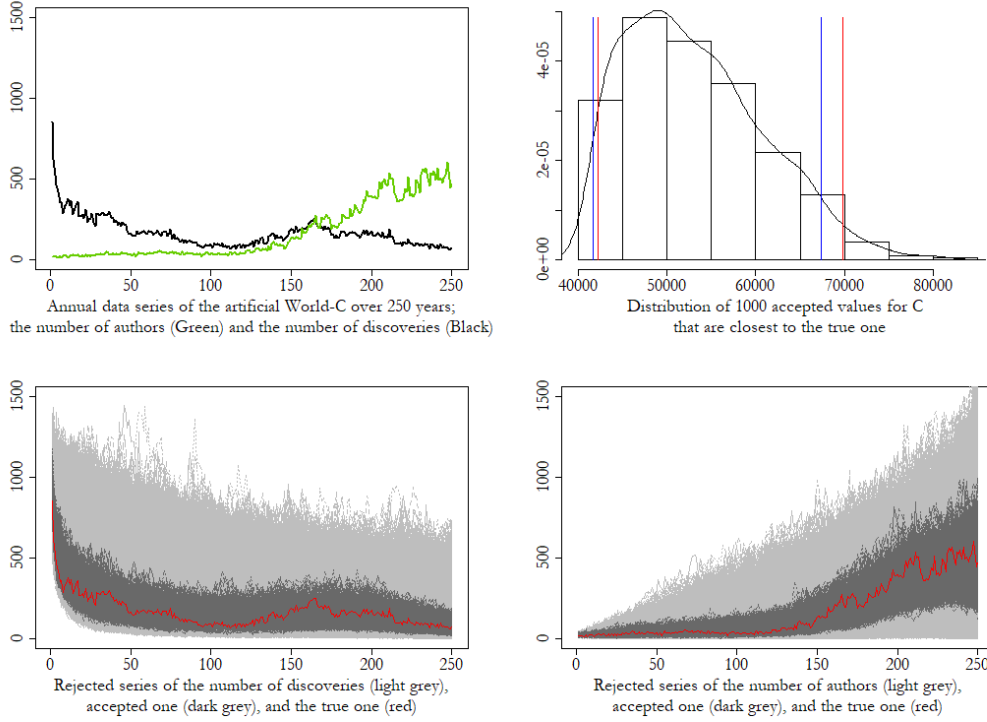


Figure 4.10: Fitting results of the World-C.

As we see, both the 95% credible intervals of the World-C are very narrow so that at the maximum limit of these intervals there are only around $2D/3$ remaining species yet to be discovered. Thus, having these small ranges, along with having a decreasing pattern of the late discoveries that is associated with an increasing curve for the authors, indicates that what is left for future discovery is just a few number of species. Again, if we pretended that we did not know the true value of C in the World-C, we can see that this speculation is consistent with the fact of having discovered 83% of all species in this artificial world.

The bottom panel of Figure 4.10 shows about half of the 1,000,000 generated series from which only 1000 series are accepted (dark grey) which are the closest ones to the true given series (red) for both the number of discoveries $\{\tau_j\}_{j=1:250}$ and the number of authors $\{x_j\}_{j=1:250}$.

4.2.4 Investigating Standardized Distance Metric:

In the adopted algorithm and within the pilot experiment (involving 10,000 iterations), after generating the annual datasets of the number of authors $[x^*]_r$ and the

number of discoveries $[\boldsymbol{\tau}^*]_r$ (where $r = 1 : 10000$, $\boldsymbol{x}^* = \{x_j^*\}_{j=1:T}$, $\boldsymbol{\tau}^* = \{\tau_j^*\}_{j=1:T}$, and $T = 250$ years), we computed our distance metric in the following steps:

First version of Distance Metric:

1. Compute distances of authors $[\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_r$.
2. Compute distances of discoveries $[\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})]_r$.
3. Add the above: $[\text{Total}]_r = [\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_r$, and sort them ascendingly.
4. Accept the smallest 1000 of them *i.e.* $[\text{Total}]_{1:1000}$. Note that the accepted maximum total distance Total_{1000} does not imply that $\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})_{1000}$ is also the maximum among the accepted authors distances, thus, we need the next step.
5. Go back to the accepted $[\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_{1:1000}$, and choose the maximum one of them as threshold *i.e.* $\text{Threshold} = \text{Max}([\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_{1:1000})$. This threshold is used in the main experiment as a first filtering step to filter the generated $[\boldsymbol{x}^*]_r$ before proceeding to generate $[\boldsymbol{\tau}^*]_r$. The reason is to speed up the code performance, because generating $[\boldsymbol{\tau}^*]_r$ is computationally demanding, while generating $[\boldsymbol{x}^*]_r$ is much easier.

The main experiment also includes computing total distances and sorting them ascendingly. The second filtering step is only accepting the smallest 1000 total distances out of 1,000,000, at which the distribution of their C 's is the estimation of our target variable. We adopted the above algorithm in fitting the three artificial worlds as we explained in the previous subsections and we found good results with efficient performance. However, we also investigated the standardized version of our distance metric, as explained next. Again, within the pilot experiment that involves 10,000 iterations and after generating the annual datasets of the number of authors and the number of discoveries, the standardized distance metric is computed as follows:

Standardized Distance Metric:

1. Derive mean and standard deviation of the distances *i.e.* $\mu_{\boldsymbol{x}} = \text{mean}([\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_r)$, $\sigma_{\boldsymbol{x}} = \text{sd}([\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_r)$, $\mu_{\boldsymbol{\tau}} = \text{mean}([\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})]_r)$, and $\sigma_{\boldsymbol{\tau}} = \text{sd}([\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})]_r)$.
2. Compute standardized distances of authors $[\mathcal{D}_{\boldsymbol{x}}^z]_r = \{[\mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})]_r - \mu_{\boldsymbol{x}}\} / \sigma_{\boldsymbol{x}}$.
3. Compute standardized distances of discoveries $[\mathcal{D}_{\boldsymbol{\tau}}^z]_r = \{[\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})]_r - \mu_{\boldsymbol{\tau}}\} / \sigma_{\boldsymbol{\tau}}$.

4. Add the above: $[\text{Total}^z]_r = [\mathcal{D}_x^z + \mathcal{D}_r^z]_r$, and sort them ascendingly.
5. Accept the smallest 1000 of them *i.e.* $[\text{Total}^z]_{1:1000}$.
6. Go back to the accepted $[\mathcal{D}_x^z]_{1:1000}$, and select $\text{Max}([\mathcal{D}_x^z]_{1:1000})$ to derive our threshold *i.e.* $\text{Threshold}^z = \text{Max}([\mathcal{D}_x^z]_{1:1000}) \times \sigma_x + \mu_x$ which is used in the main experiment as a first filtering step to filter the generated $[\mathbf{x}^*]_r$. Note that in the main experiments, the standardization treatment is applied after computing the distances.

We applied the above standardized distance metric on fitting the three artificial worlds, and we found that there is no significant difference in the results of the first version of our distance metric and the results of the standardized distance metric. For example, the top panel of Figure 4.11 shows the estimated posterior distributions of C (in Word-A) using the first version of our distance metric (left plot) and the standardized distance metric (right plot), and the bottom panel shows their box-plots and CDF's which indicate that there is no significant difference between the two distributions.

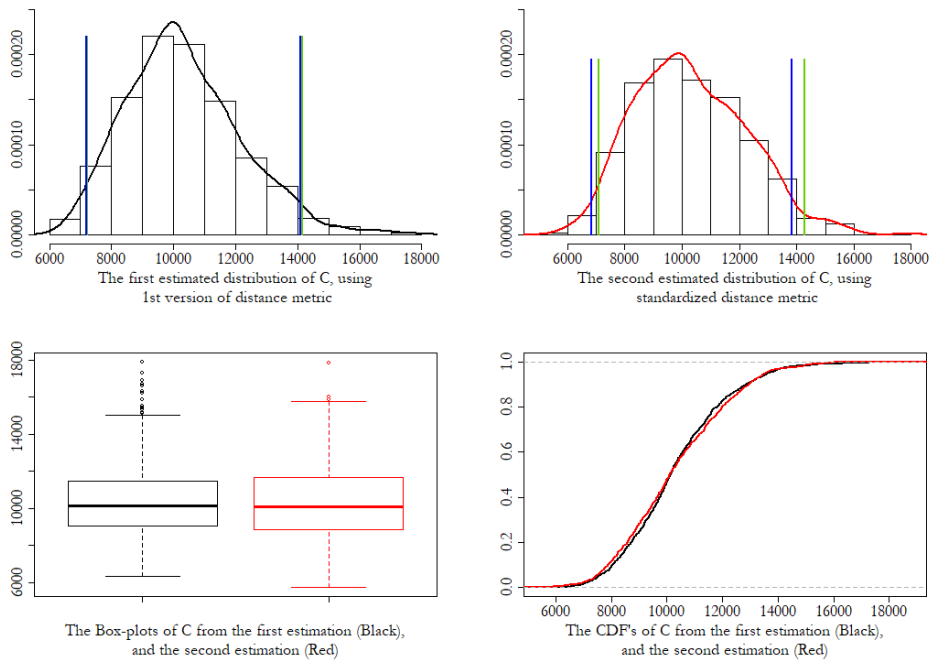


Figure 4.11: Fitting results of World-A, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.

Although there are some differences in their scatter plots, the fact that we rely only on the smallest 1000 total distances out of 1,000,000 makes the final

estimation similar, as show in Figure 4.12. We can see from the top panel of Figure 4.12 that there is a high correlation between the total distances and the discovery distances in both treatments (1st version distance metric and standardized distance metric) of the main experiments, though the correlation seems to be much higher without standardization. However, we can see in the bottom panel that this correlation becomes similar in both treatments in the final accepted sample that make the final estimation of our target. The similarity appears in having equivalent correlation pattern (*i.e.* equivalent slope) in both the discovery distances and the authors distances with the total distances.

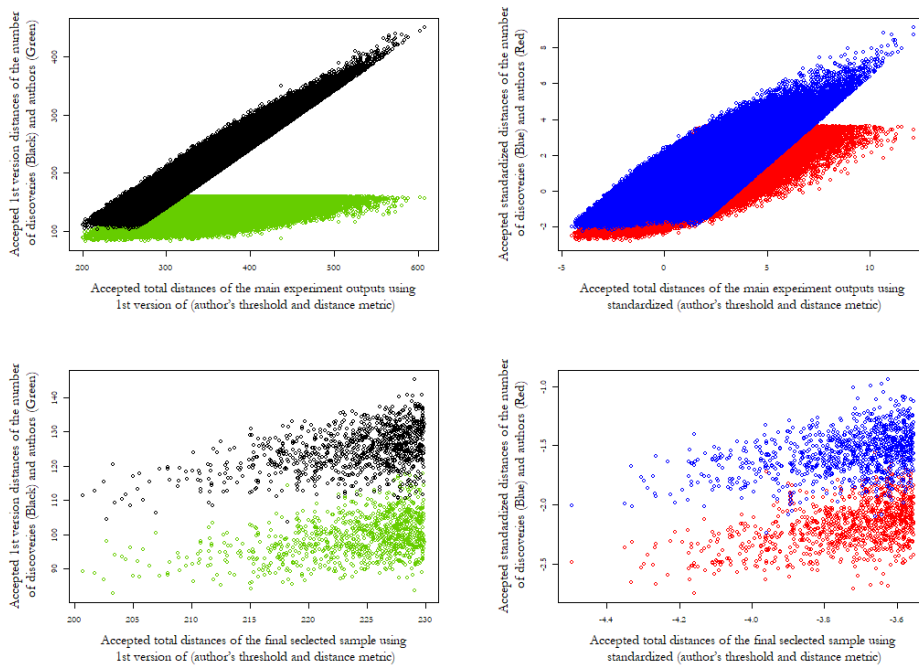


Figure 4.12: Scatter plots of the main experiments of fitting World-A, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.

We found similar resulting patterns in both World-B and World-C as shown in (Figure 4.13 and Figure 4.14) and (Figure 4.15 and Figure 4.16), respectively. This concludes that there is no significant differences in the results between using our 1st version of the distance metric and the standardized version. However, there are some differences in the computation time which justify adopting our 1st version of the distance metric in the rest of the experimentations. These differences are summarized in the next comments.

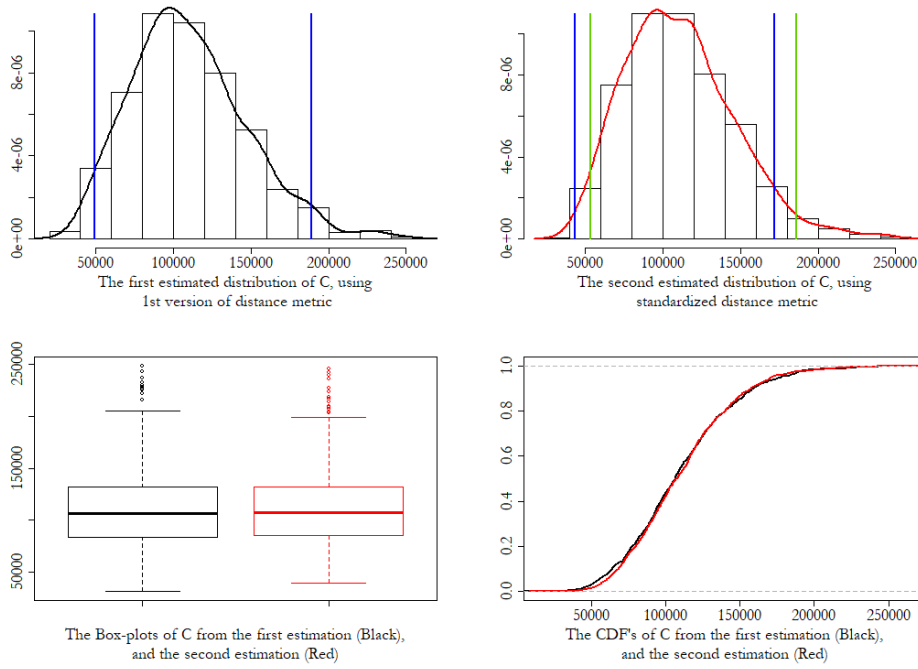


Figure 4.13: Fitting results of World-B, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.

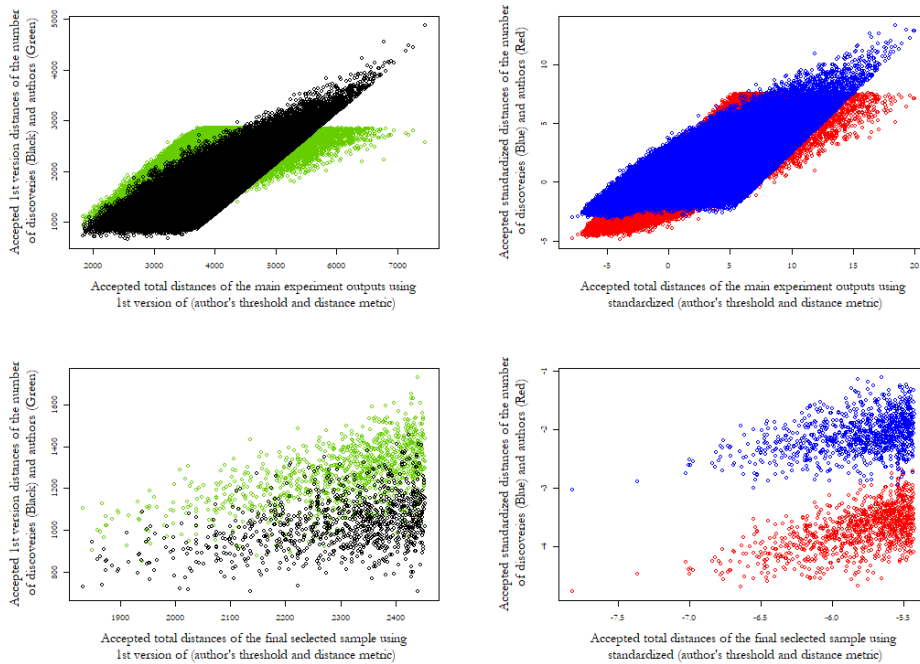


Figure 4.14: Scatter plots of the main experiments of fitting World-B, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.

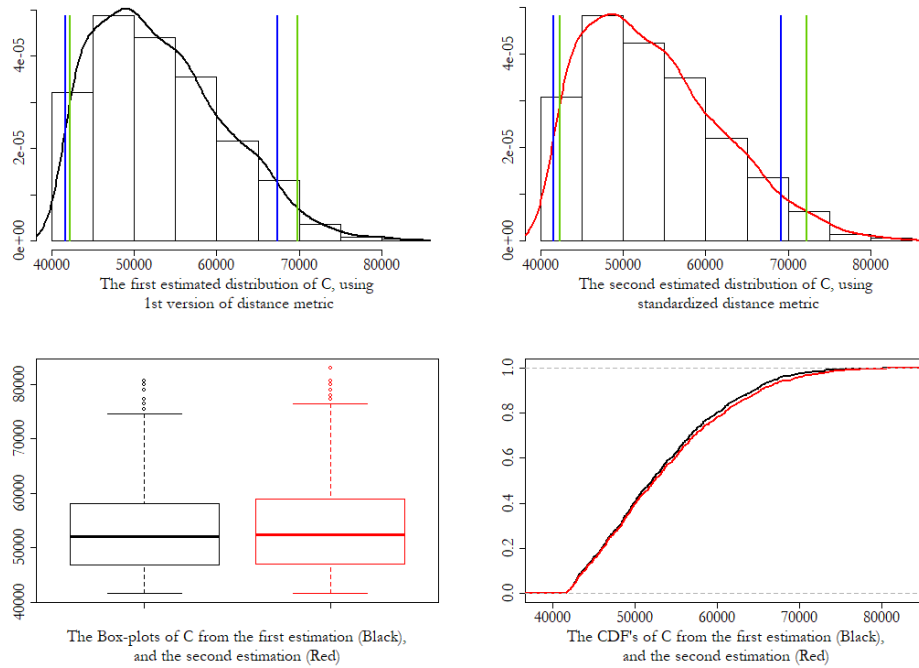


Figure 4.15: Fitting results of World-C, using the first version of distance metric versus the standardized distance metric. Note that the green vertical lines represents the 95% central credible interval, while the blue vertical lines represents the 95% HPD interval.

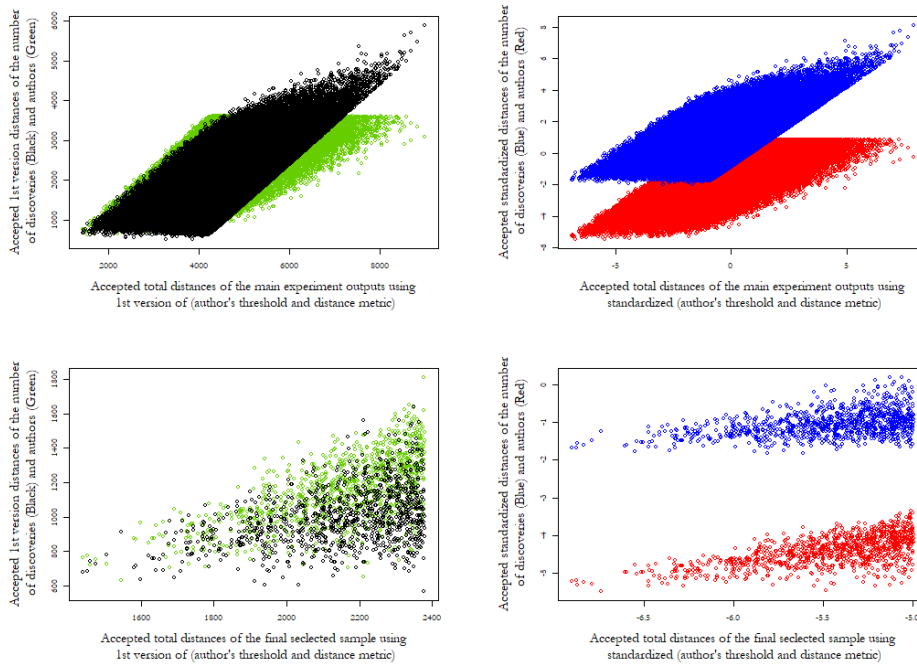


Figure 4.16: Scatter plots of the main experiments of fitting World-C, using the first version of distance metric (black for discoveries and green for authors) versus the standardized distance metric (blue for discoveries and red for authors). The bottom panel shows the scatter plots of the final samples in both treatments.

Comments on the Differences between the Two Treatments:

- In the artificial World-A, we found that the threshold of the 1st version treatment is $\text{Max}([\mathcal{D}(\mathbf{x}^*, \mathbf{x})]_{1:1000}) = 160$, while the threshold of the standardized treatment is $\text{Max}([\mathcal{D}_{\mathbf{x}}^z]_{1:1000}) \times \sigma_{\mathbf{x}} + \mu_{\mathbf{x}} = 246$ (where $\text{Max}([\mathcal{D}_{\mathbf{x}}^z]_{1:1000}) = 0.1046$). Since the standardized treatment implies a larger threshold, a larger number of the generated datasets passed this threshold (about 360,000 extra datasets compared to the 1st version treatment). Therefore, extra computation time is spent on these extra datasets with no added benefit as they are rejected in the final filtering step.
- In the artificial World-B, we also found that the threshold of the standardized treatment ($\text{Max}([\mathcal{D}_{\mathbf{x}}^z]_{1:1000}) \times \sigma_{\mathbf{x}} + \mu_{\mathbf{x}} = 6,148$) is larger than the threshold of the 1st version treatment ($\text{Max}([\mathcal{D}(\mathbf{x}^*, \mathbf{x})]_{1:1000}) = 2,853$). It also allowed for about 690,000 extra generated datasets to pass its threshold, and, therefore, extra computation time spent with no added benefit.
- In the artificial World-C, we found a special situation at which the threshold of the standardized treatment is equal to the threshold of the 1st version treatment ($\text{Max}([\mathcal{D}_{\mathbf{x}}^z]_{1:1000}) \times \sigma_{\mathbf{x}} + \mu_{\mathbf{x}} = \text{Max}([\mathcal{D}(\mathbf{x}^*, \mathbf{x})]_{1:1000}) = 3,604$). Therefore, same expected computing time for both treatments.

The interpretation of having the above fewer thresholds and shorter computation times in World-A and World-B in the 1st version distance metric is that: in these worlds where the influence of the number of authors is higher (see Figure 4.3), the correlation between the authors distances and total distances are higher in the pilot experiments, and result in usefully tight thresholds. On the other hand, in World-C where the number of authors ends up with less influence (decreasing discoveries with increasing number of authors, see Figure 4.3), the correlation between the authors distances and total distances is lower in the pilot experiment compared to World-A and World-B, and results in a threshold that is as the same as the one of the standardized treatment. However, we should emphasize on the fact that such correlations will diminish with increasing the number of iterations as we saw in the main experiments.

The right panel of Figure 4.17 shows the threshold values of the three artificial worlds within the context of the scatter plot of the authors distances $[\mathcal{D}(\mathbf{x}^*, \mathbf{x})]_r$ and discovery distances $[\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau})]_r$ in the main experiments, where $r = 1 : 1000000$. The green vertical line represents the 1st version threshold ($\text{Max}([\mathcal{D}(\mathbf{x}^*, \mathbf{x})]_{1:1000})$), while the blue vertical line represents the standardized threshold ($\text{Max}([\mathcal{D}_x^z]_{1:1000}) \times \sigma_x + \mu_x$). The left-side area of the vertical lines represents the datasets that pass through the specified thresholds. The right panel also shows the accepted final sample (black area) whereby the red diagonal line represents the maximum accepted total distance *i.e.* $(\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\mathbf{x}^*, \mathbf{x}))_{1000}$. The left panel of Figure 4.17 shows a zoomed-in view of the accepted final sample.

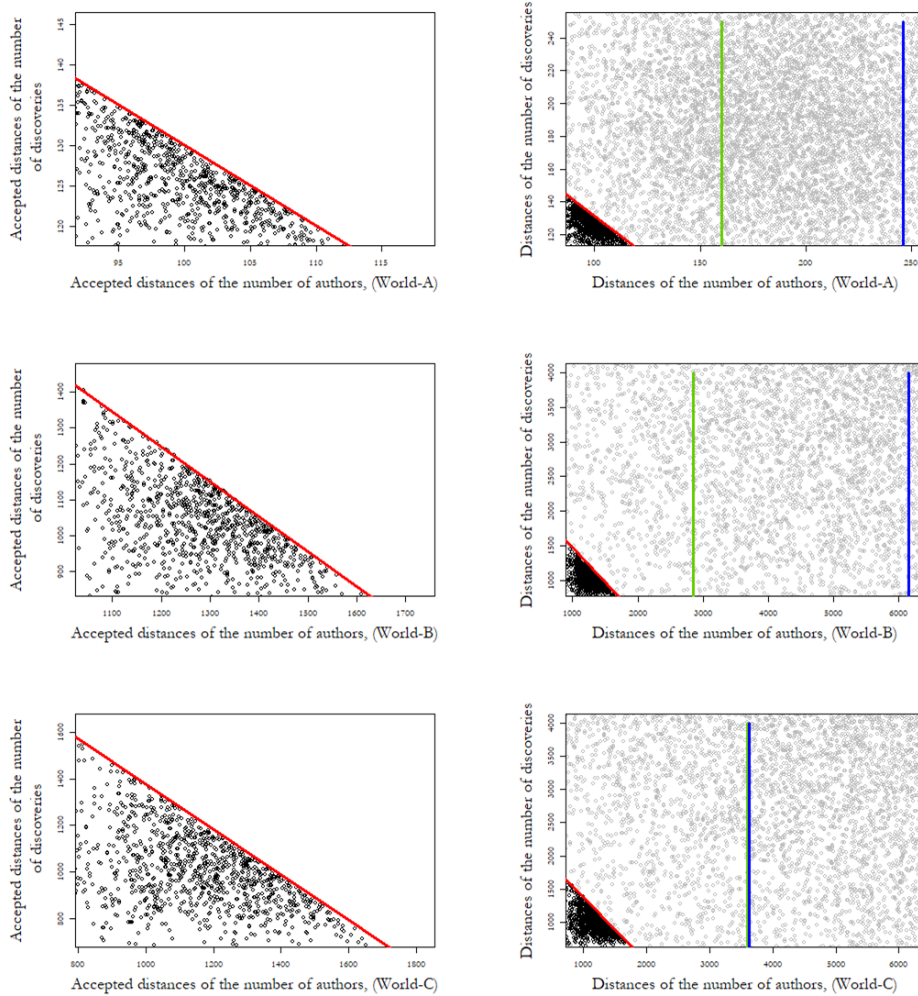


Figure 4.17: Scatter plots of the main experiments of fitting the three artificial worlds (focusing on the accepted final area and ignoring the rest of x-axis and y-axis). The left panel shows a zoom-in view of the accepted final sample, while the right panel shows a zoom-out view where the thresholds lines are appeared. The green vertical line represents the 1st version threshold, while the blue vertical line represents the standardized threshold.

In conclusion, there are no significant differences in the results with or without the standardization. If the pilot experiments show a higher correlation between the author distances and total distances, it is preferable not to standardize the distance metric to allow for having a tight threshold and therefore efficient computation time. Increasing the number of iterations in the main experiment as the one that we adopted (1,000,000) and accepting small size final sample is important for accurate results.

4.3 Chapter Summary:

We started this chapter by setting up the distributional assumptions of our proposed model within the species context, as explained in the first section. Then this model is validated, in solving the number of kind (species) problem, through three simulation experiments (generating artificial worlds and fitting them according to the proposed model), as shown in the second section, also see Group B.1 of Section B.1 in Appendix B for more details. This section also explained the efficiency development of the fitting stage through using the ABC algorithm with an additional filtering step. Evaluating the performance of the model in different scenarios and according to some criteria will be introduced in the next chapter.

Chapter 5

Model Performance & Evaluation

The current chapter is concerned with describing and evaluating the performance of our proposed model through refitting the artificial worlds (introduced in Chapter 4) in different scenarios. In three sections of the current chapter, we attempt to seek answers of three question-form criteria about our model:

- To what extent is our model accurate in estimating the number of species?
- How sensitive or robust is our model against mis-parametrization?
- Can we describe our model as being additive (either completely or partially additive)? How much difference is there between applying the model on the whole dataset, and applying the model on subsets of data then aggregating?

5.1 Model Accuracy

In this section, we are trying to answer the first question in our criteria list: *To what extent is our model accurate in estimating the number of species?* As the ABC algorithm is basically a set of Monte Carlo techniques, it is subject to the Monte Carlo error. This error is expressed in the variation of the Monte Carlo estimator that is taken across hypothetical repetitions of the simulation, where each simulation is based on the same design and number of replications. However, in our case, we are estimating a probability distribution in addition to the point and interval estimation. Therefore, we performed two types of simulation experiments in the accuracy context to explore two aspects of the Monte Carlo error. The first aspect is an error due to the taken number of replications (sample size), while the second aspect is an error due to the adopted number of iterations (simulation repetition).

In the light of checking the first aspect error, we fitted the given ‘artificial’ data set according to ABC approach (explained in Section 3.2.3) at which we fixed the number of iteration to be 1,000,000, while we checked three sample sizes: $s = 100$, $s = 1000$, and $s = 10,000$. The fitting experiment is repeated ten times for the same artificial world (same seed). On the other hand, in checking the second aspect error, we fixed the final selected sample size to be $s = 1000$, while we checked four values of iterations: $iter = 100,000$, $iter = 500,000$, $iter = 1,000,000$, and $iter = 1,500,000$. This type of experiment is repeated over 500 artificial worlds of the same design (different seeds but same parametrization). These two types of simulation experiments are implemented over the design of the three artificial worlds (introduced in Chapter 4) as illustrated in the next subsections. Note that detailed results with holistic-view plots are available in Appendix B, Section B.2.

5.1.1 Accuracy in World-A Design:

In Section 4.3.1 we saw good results in fitting the artificial World-A. However, for checking the accuracy of the estimation, that fitting is repeated ten times for the same artificial World-A, but three different sample sizes are considered, as shown in Figure 5.1. We can see that as the sample size increases from 100 to 10,000, the MC error (represented in the estimator’s standard deviation) decreases for both the estimated posterior mean and mode, and the estimated curves of the posterior probability distributions become closer. However, although these estimators seem to be more consistent with increasing the sample size, they tend to overestimate behaviour. Therefore, choosing the sample size $s = 1000$ seems to be a balanced trade-off to provide a relatively less MC error with a relatively less biased estimation.

Moving to the second type of the simulation experiments, we generated 500 artificial worlds according to the World-A parametrization design but with different generating seeds. These worlds are fitted according to the ABC algorithm and from each fitting, 500 posterior means and 500 posterior modes are derived. The distribution, average, variance/standard deviation, and the 95% credible interval of these posterior means and modes are obtained. These 500 fittings and results

are repeated four times for different number of iterations with fixed sample size.

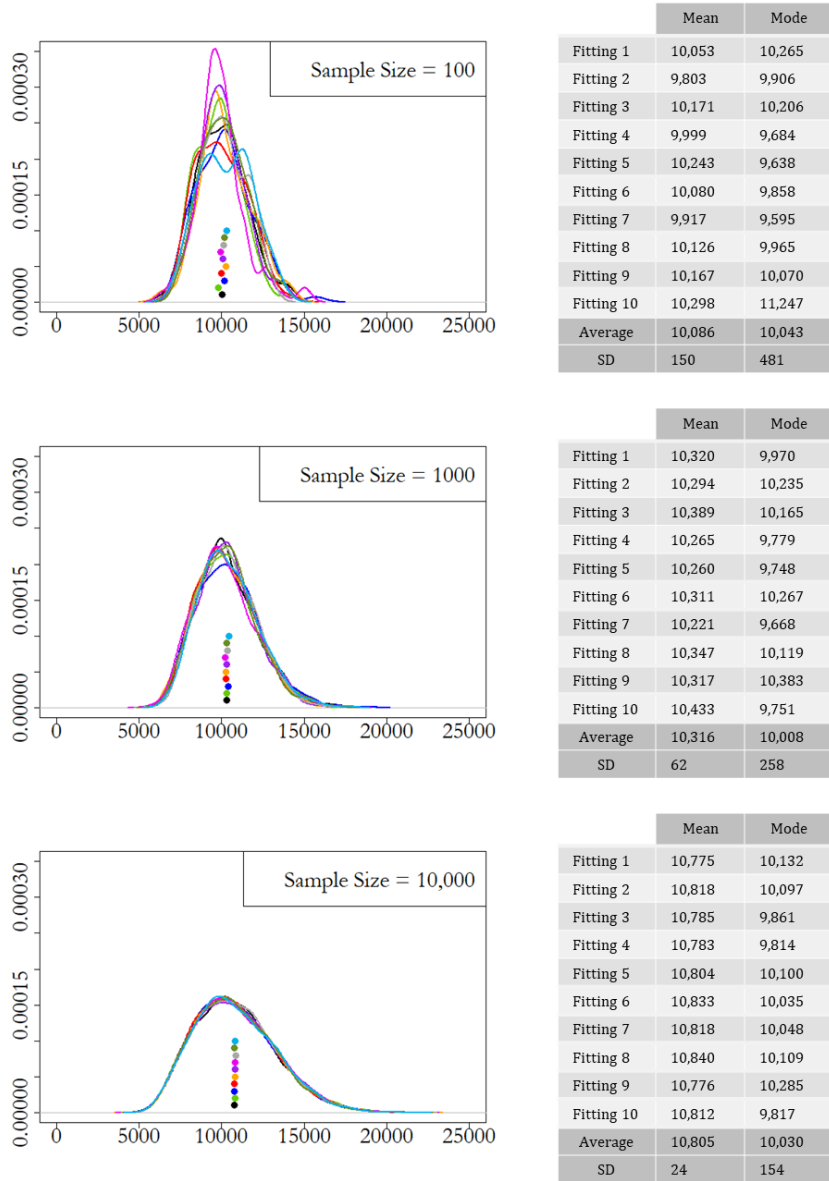


Figure 5.1: Fitting artificial World-A for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify their position relative to the true value $C = 10,000$.

Figure 5.2 shows four distributions with their 95% credible intervals of the posterior means over the four number of iterations in the left top plot, while the left bottom belong to the posterior modes. The plots also include the average of the estimators (posterior means and modes) marked in dots aligned above each other. We can see that the distributions along with their credible intervals become

tighter and concentrated around the true value of C as we increase the number of iterations. Consequently, The MC error decreases as shown in the right top plot for both the posterior means (top curve) and posterior modes (bottom curve). However, in the current design it seems that the enhancement in the estimation is not that significant after 1,000,000. Therefore, choosing the number of iterations to be 1,000,000 in the fitting stage with selecting a final sample of size 1000 seems to be sufficient to achieve a relatively less MC error with a relatively less biased estimation, where $E(C_{mean}^*) = 10,083$ with percent bias = 0.83% when we adopt the posterior mean, and $E(C_{mode}^*) = 9,763$ with percent bias = 2.4% when we adopt the posterior mode, while the true $C = 10,000$. See Group B.4, Appendix B for details.

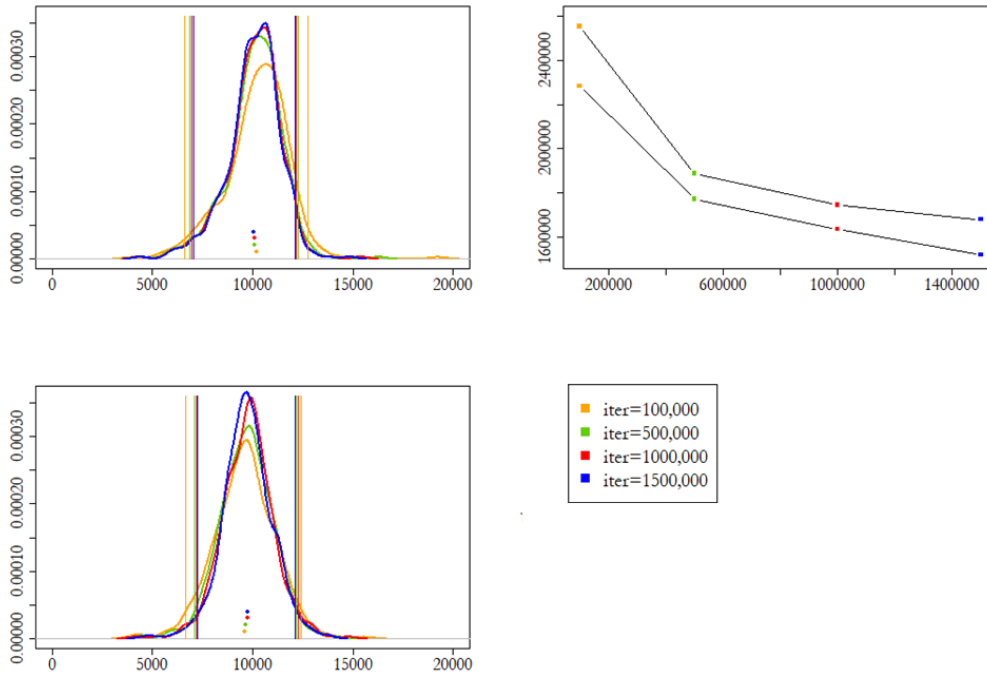


Figure 5.2: Fitting 500 artificial worlds of design World-A with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error that results in the adopted number of iterations. The left top is the distribution of the posterior means, while left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.

5.1.2 Accuracy in World-B Design:

Here as well, the artificial World-B is fitted ten times and for each fitting three different sample sizes are considered, as shown in Figure 5.3. We can see that as

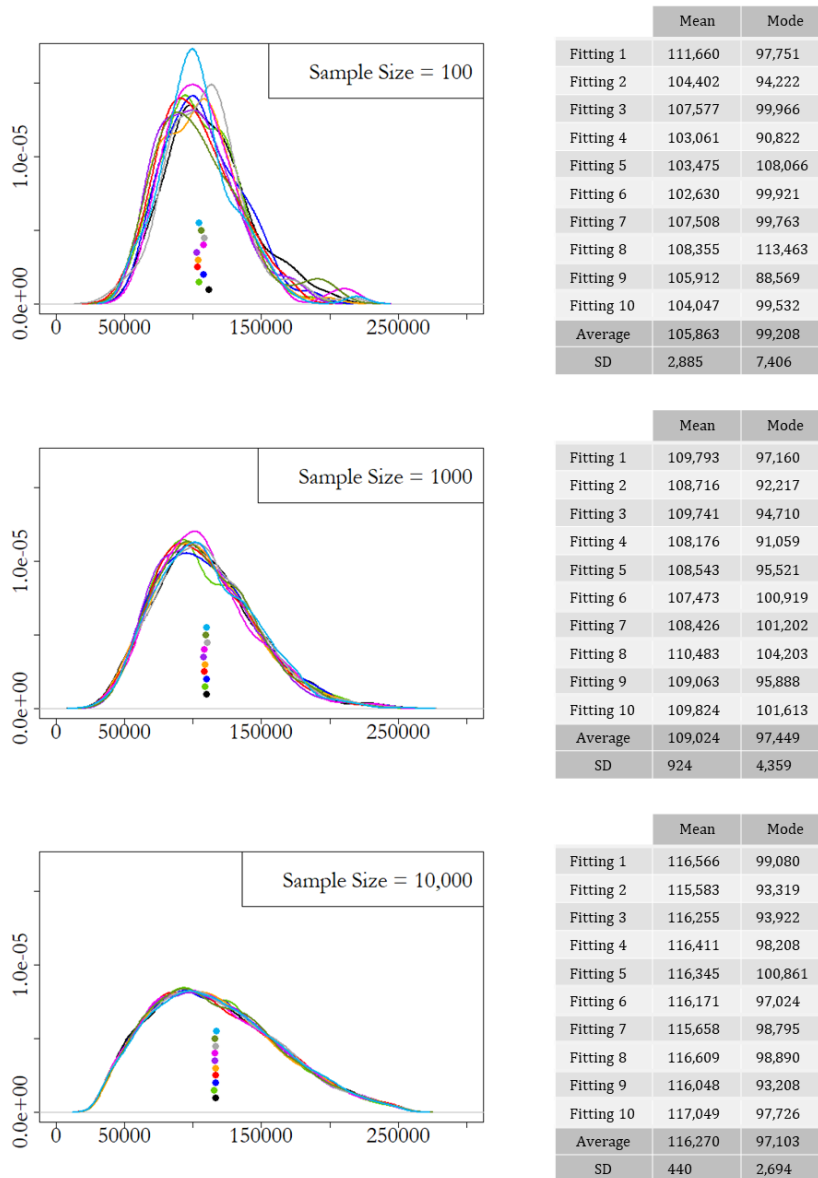


Figure 5.3: Fitting artificial World-B for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify their position relative to the true value $C = 10,000$.

the sample size increases from 100 to 10,000, the estimated curves of the posterior probability distributions become closer, and the MC error (represented in the estimator standard deviation) decreases for both the estimated posterior mean and mode. Also, these estimators tend towards overestimation as the sample size increases. Therefore, choosing the sample size $s = 1000$ still seems to be a balanced trade-off situation that provides a relatively less MC error with a relatively less biased estimation.

Moving to the second type of the simulation experiments in the accuracy context, we also generated 500 artificial worlds according to the World-B parametrization design but with different generating seeds. These worlds are fitted according to the ABC algorithm and from each fitting, 500 posterior means and 500 posterior modes are derived. The distribution, the average, the variance/standard deviation, and 95% credible interval of the posterior means and modes are obtained. These 500 fittings and results are repeated four times for different number of iterations with fixed sample size $s = 1000$, as shown in Figure 5.4 (the same plot description of World-A is also applied here).

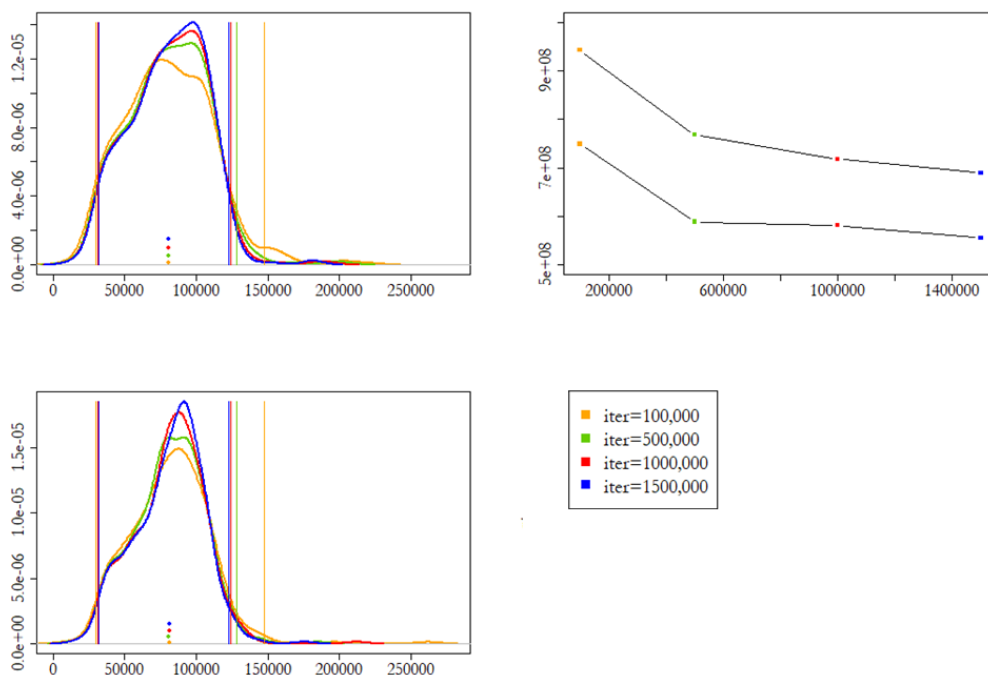


Figure 5.4: Fitting 500 artificial worlds of design World-B with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while the left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.

However, although there is a decrease in the MC error along with increasing the number of iterations, there is still a consistently some bias in the average of the posterior means and the average of the posterior modes. Hence, increasing the number of iterations in fitting World-B does not make a significant enhancement in the MC error especially after 1,000,000.

With respect to the bias issue, it can be explained by the fact that there is a high variation in the 500 generated datasets (the number of discoveries and the number of authors) under the World-B design, see Group B.3 and Group B.4 in Appendix B for details. By checking the current discoveries D of these 500 generated artificial worlds, we found that the values range between 2000 to over 50,000 discovered species with 60% coefficient of variation. However, in the artificial World-B that we already adapted in Figure 4.9, the value of the current discoveries is $D = 31,865$. Therefore, by allowing this value to vary by 20,000 species higher and lower, we found that there is 300 out of the 500 of the generated artificial worlds that have D 's ranging between 12,000 and 50,000 discovered species. By checking the distribution of posterior means and modes, it seems that the bias is significantly reduced, where $E(C_{mean}^*) = 98,125$ with percent bias = 1.9% when we adopt the posterior mean, and $E(C_{mode}^*) = 95,492$ with percent bias = 4.5% when we adopt the posterior mode, while the true $C = 100,000$. See Figure 5.5.

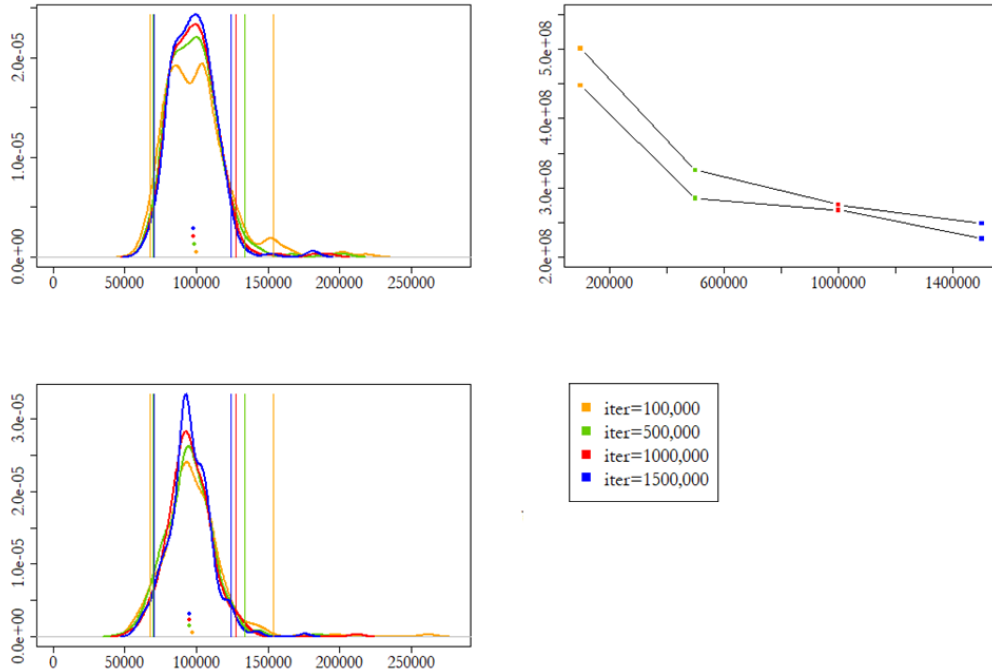


Figure 5.5: Fitting 300 artificial worlds of design World-B with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while the left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.

5.1.3 Accuracy in World-C Design:

Again, we made the same types of experiments to explore the two aspect of the MC error in World-C. Figure 5.6 shows that with increasing the sample size from $s = 100$ to $s = 10,000$, the estimated curves of the posterior probability distributions become closer, and the MC error (represented in the estimator standard deviation) decreases for both the estimated posterior mean and mode. However, the decrease in the MC error is not consistent in the case of the posterior mode, though the

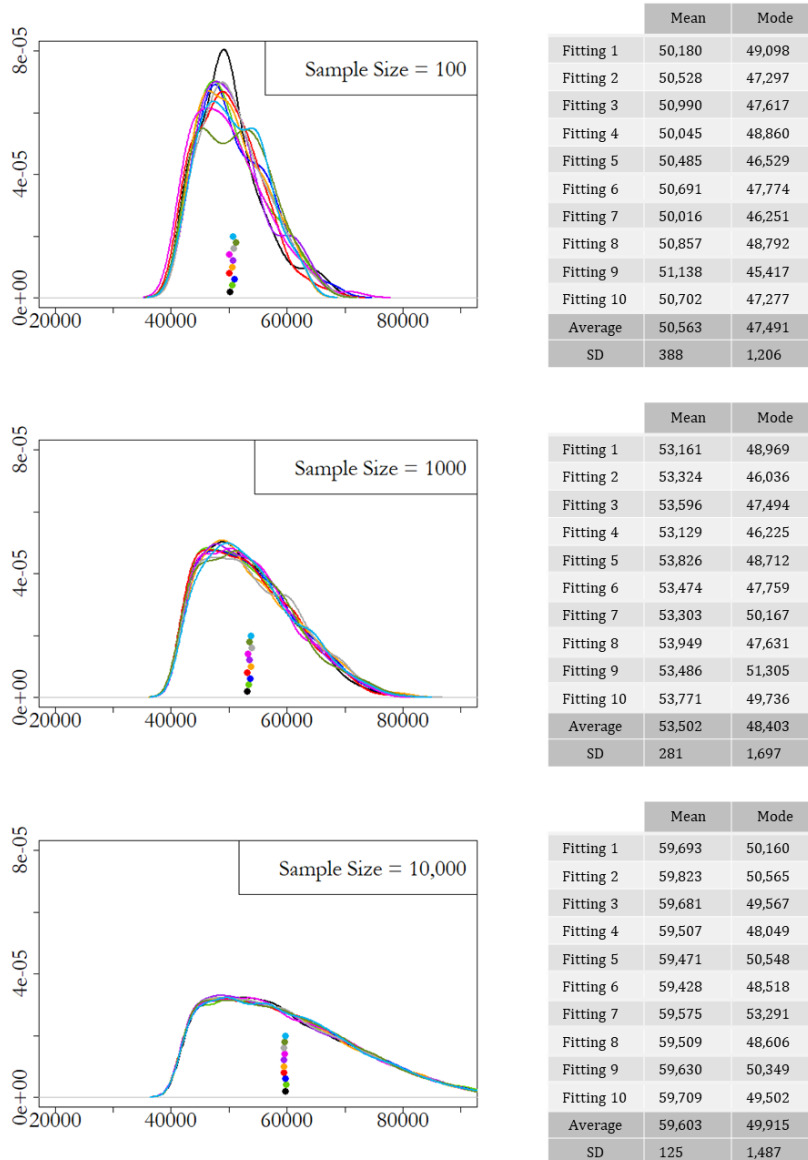


Figure 5.6: Fitting artificial World-C for 10 times with 3 different sample sizes and fixed number of iterations, to explore the first aspect of MC error due to the selected sample size. Note that each color represents a fitting trial. The circles represent the estimated means, they are aligned above each other just to clarify their position relative to the true value $C = 10,000$.

difference is small. The mean estimator slightly tends to overestimated behaviour as we increase the sample size, while it is the opposite case for the mode. Therefore, the sample size $s = 1000$ still seems to be a balanced trade-off situation that provides a relatively less MC error with a relatively less biased estimation.

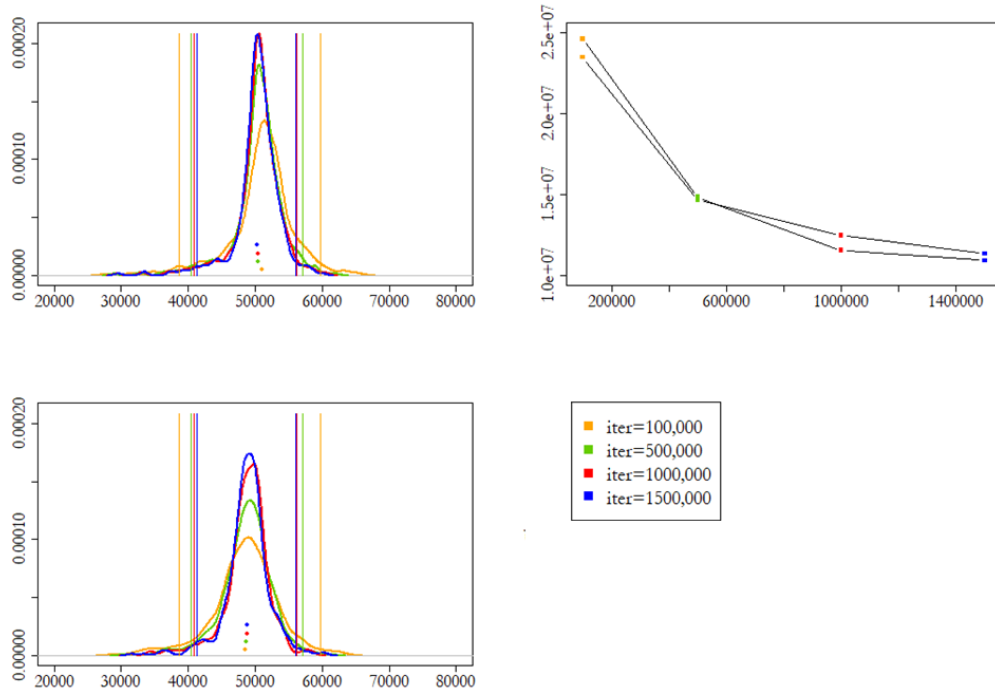


Figure 5.7: Fitting 500 artificial worlds of design World-C with 4 different simulation iterations and fixed sample size, to explore the second aspect of MC error due to the adopted number of iterations. The left top is the distribution of the posterior means, while left bottom is the distribution of the posterior modes. The right top is the changing rate in the variance of the estimators in which the top curve belongs to the posterior means, while the bottom curve belongs to the posterior modes.

Figure 5.7 shows the results of inspecting the second aspect of MC error. We can see that the distributions of the 500 posterior means and 500 posterior modes along with their credible intervals become tighter and concentrated around the true value of C , as we increase the number of iterations, which leads to decrease the MC error. Again, as the same as the previous worlds, the enhancement in the estimation is not that significant after 1,000,000. Therefore, choosing the number of iterations to be 1,000,000 in the fitting stage with selecting a final sample of size 1000 seems to be sufficient to achieve a relatively less MC error with a relatively less biased estimation, where $E(C_{mean}^*) = 50,383$ with percent bias = 0.77% when we adopt the posterior mean, and $E(C_{mode}^*) = 48,776$ with percent bias = 2.45% when we adopt the posterior mode, while the true $C = 50,000$. More details are available in Appendix B, Group B.4.

5.2 Model Sensitivity

In this section, we are trying to answer the second question in our criteria list: *How sensitive or robust is our model against mis-parametrization?* As we mentioned previously in Section 3.2.1, our proposed model is subject to parameter specifications. However, these parameters may be mis-specified. Our model includes a set of five parameters, $\theta := \{\mu, \sigma, L_1, \sigma_L, \alpha\}$, that each needs to be specified, either by a fixed one value or a range of values, or even by a probability distribution. In the current case study, we already fixed them to certain values in generating and fitting the three artificial worlds.

However, we re-fitted these worlds against different parametrizations, to check on the estimation sensitivity to changes in the parameter values. We tested one parameter at a time and fixed all the others on their true values. We believe that 20% deviation from the true parameters is enough to test this aspect of our model. Thus, for each parameter we performed a re-fitting experiment against two other values, one is 20% over-parametrization and the other is 20% under-parametrization. In doing so, we performed ten re-fitting experiments using 1,000,000 iterations each time and the sample size of 1000 for each artificial world as will be illustrated in the next subsections. Note that detailed results with holistic-view plots are available in Appendix B, Section B.3.

5.2.1 Sensitivity in World-A Design:

Table 5.1 introduces the true values for each parameter along with the chosen 20% over and under parametrization in World-A. After the re-fitting experiments, we found that there are almost two levels of sensitivity in our proposed model. Figure 5.8, shows the fitting results in estimating the true value C by the posterior mean and mode along with their 95% credible intervals against mis-parametrization of the five parameters, each at a time. Looking at the plots from bottom to up, we can see that our model is almost robust against the mean of the species abundances, μ_N , and the deviation of the latent efforts, σ_L . The point estimations (posterior mean and mode) of C seems very close to the true value, as well as

their 95% credible intervals which are within the 95% credible intervals of C .

Table 5.1: Under/over-parametrization within artificial World-A.

Parameters within their Distributions	True Value	20% Under	20% Over
$N_k \sim \text{LogNormal}(\mu, \sigma = 2.5)$	$\mu = 4$	$\mu = 3.2$	$\mu = 4.8$
$N_k \sim \text{LogNormal}(\mu = 4, \sigma)$	$\sigma = 2.5$	$\sigma = 2$	$\sigma = 3$
$L_1 \sim \text{Normal}(\mu_{L_1}, \sigma_L = 0.1)$	$\mu_{L_1} = 5$	$\mu_{L_1} = 4$	$\mu_{L_1} = 6$
$L_1 \sim \text{Normal}(5, \sigma_L)$	$\sigma_L = 0.1$	$\sigma_L = 0.08$	$\sigma_L = 0.12$
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\alpha))$	$\alpha = \log(30)$	$\alpha = \log(15.2)$	$\alpha = \log(59.2)$

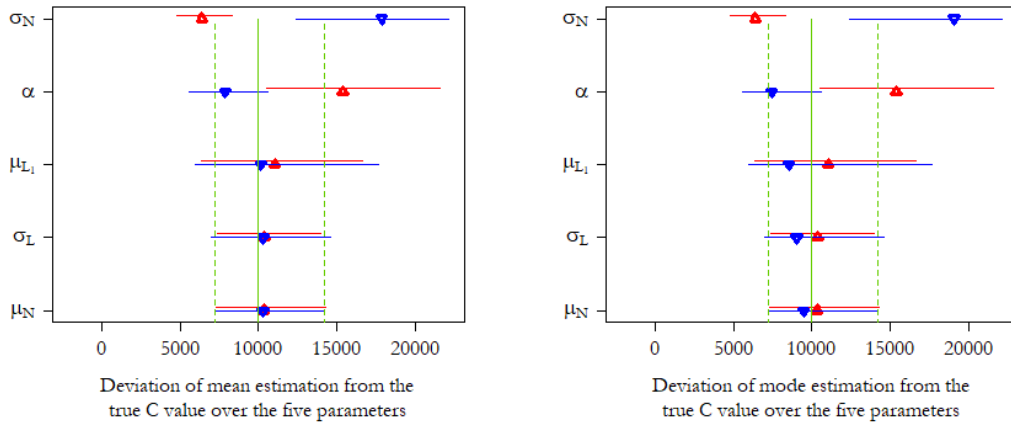


Figure 5.8: Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-A. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.

However, the mean of latent effort of the first year, μ_{L_1} , seems to have some influence on the efforts of the rest of the years so that a mis-parametrization in this value might lead to a relatively small deviation in the estimation from the true value of C . In spite of that, the point estimations (both posterior mean and mode) are still within the 95% credible interval of C , with being very close to C in the mean estimation, also, most of their credible intervals are still within the 95% credible intervals of the true value of C . Thus, we can say that our model is slightly sensitive to μ_{L_1} in World-A.

On the other hand, our model seems to be highly sensitive to two parameters in World-A, which are the variation of the species abundances, σ_N , and the expected number of specimens to be collected per author per year (on a log-scale), α . Although, with over-parametrizing α the point estimations (posterior mean and mode) of C is still within the 95% credible interval of C , the rest of estimations due to under-parametrizing α and over/under-parametrizing σ_N all exceeded the 95% credible interval of the true value of C . The interpretation of the high sensitivity of α can be explained as: increasing the value of α (with fixing all other parameters) means that the current amount of discoveries D is associated with a higher efforts than the actual one. This indicates that what is left to be discovered is few and therefore the total number of species should be under-estimated. With respect to σ_N , the higher variation in the species abundance with having the same D (in light of fixing all other parameters), indicates that it is expected to discover more species and the total number of species is assumed to be higher, and therefore it is over-estimated. Therefore, we should be careful in choosing the values of these two parameters. Finally, it is worth mentioning that there is a general behaviour of these parameters which is the negative relationship between them and the estimate of C , except for σ_N ; see Group B.6, Appendix B for details.

5.2.2 Sensitivity in World-B Design:

Table 5.2 introduces the true values for each parameter along with the chosen 20% over and under parameters in World-B. Again, after the re-fitting experiments, we found that the sensitivity of our model in World-B is similar to the one in World-A. However, as shown in Figure 5.9, it seems that the mis-parametrization of α has little influence in World-B, as we can see that the point estimations (posterior mean and mode) of the true value of C is still within the true 95% credible interval of C . In addition, the estimated credible intervals of the altered values of α covers the true C . As well as World-A, the general negative relation between the chosen values of these parameters and the estimation of C (except for σ_N) exist. See Group B.6, Appendix B for details.

Table 5.2: Under/over-parametrization within artificial World-B.

Parameters within their Distributions	True Value	20% Under	20% Over
$N_k \sim \text{LogNormal}(\mu, \sigma = 3.5)$	$\mu = 10$	$\mu = 8$	$\mu = 12$
$N_k \sim \text{LogNormal}(\mu = 10, \sigma)$	$\sigma = 3.5$	$\sigma = 2.8$	$\sigma = 4.2$
$L_1 \sim \text{Normal}(\mu_{L_1}, \sigma_L = 0.15)$	$\mu_{L_1} = 7$	$\mu_{L_1} = 5.6$	$\mu_{L_1} = 8.4$
$L_1 \sim \text{Normal}(7, \sigma_L)$	$\sigma_L = 0.15$	$\sigma_L = 0.12$	$\sigma_L = 0.18$
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\alpha))$	$\alpha = \log(100)$	$\alpha = \log(39.8)$	$\alpha = \log(251.2)$

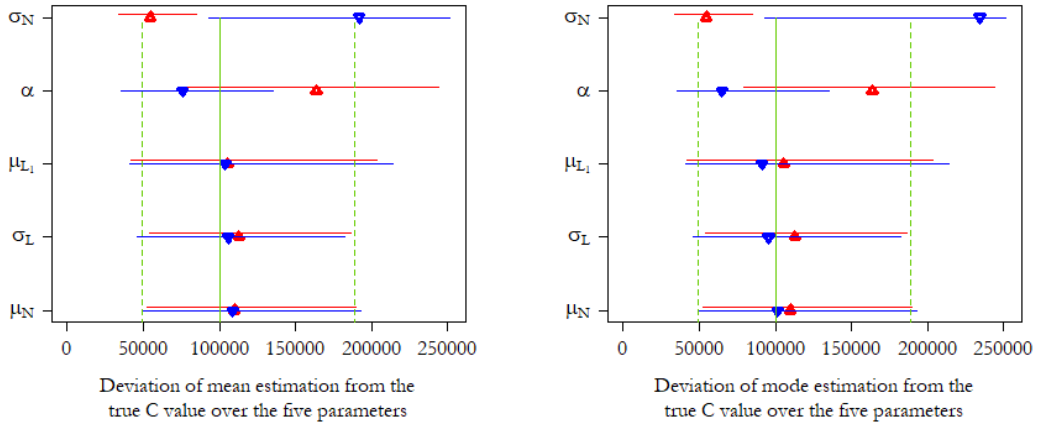


Figure 5.9: Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-B. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.

5.2.3 Sensitivity in World-C Design:

Table 5.3 introduces the true values for each parameter along with the chosen 20% over and under parameters in World-C. However, in World-C, we found a slightly different pattern of sensitivity than World-A and World-B, as shown in Figure 5.10. It seems that with a high proportion of discoveries (83%) at which the posterior distribution of C becomes skewed, the sensitivity of our model become generally less against the mis-parametrization. As we see, all the point estimations (posterior mean and mode) for all parameters are within the 95% credible intervals of the true C . In addition, all the estimated credible intervals of all parameters cover the true value of C . See Group B.6, Appendix B for details.

Table 5.3: Under/over-parametrization within artificial World-C.

Parameters within their Distributions	True Value	20% Under	20% Over
$N_k \sim \text{LogNormal}(\mu, \sigma = 2)$	$\mu = 6$	$\mu = 4.8$	$\mu = 7.2$
$N_k \sim \text{LogNormal}(\mu = 6, \sigma)$	$\sigma = 2$	$\sigma = 1.6$	$\sigma = 2.4$
$L_1 \sim \text{Normal}(\mu_{L_1}, \sigma_L = 0.08)$	$\mu_{L_1} = 7$	$\mu_{L_1} = 5.6$	$\mu_{L_1} = 8.4$
$L_1 \sim \text{Normal}(7, \sigma_L)$	$\sigma_L = 0.08$	$\sigma_L = 0.064$	$\sigma_L = 0.096$
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\alpha))$	$\alpha = \log(50)$	$\alpha = \log(22.9)$	$\alpha = \log(109.3)$

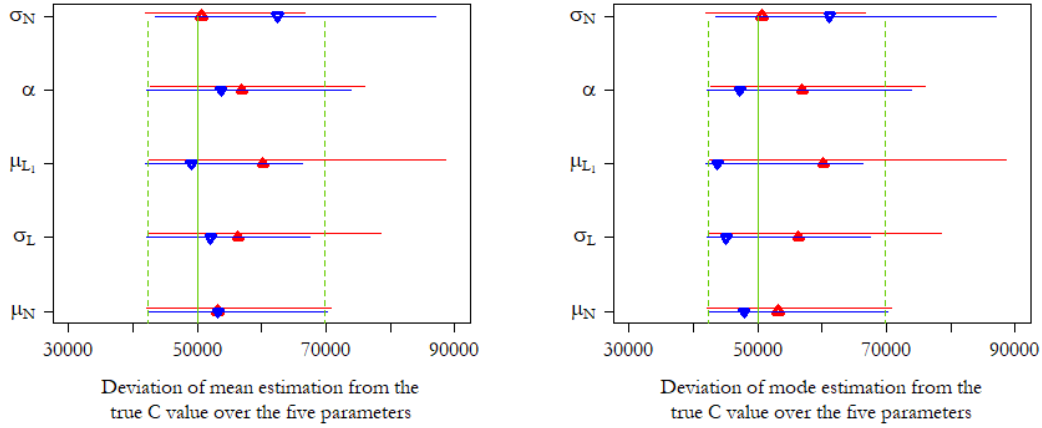


Figure 5.10: Model sensitivity in estimating the posterior mean and mode against mis-parametrization of five parameters in World-C. The solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.

However the skewness, associated with a high proportion of the total discoveries, highly influenced the log-scale parameters (μ_{L_1} and σ_L) especially with the under-parametrization case. This leads to having deviations in the sensitivity of these two parameters higher than their case in World-A and World-B. Therefore, having skewness in the posterior distribution of C should flag a caution in selecting the values of the parameters μ_{L_1} and σ_L .

5.3 Model Additivity

In cases where some estimates are already made on sub-groups of a target community, it seems a desirable property that the estimate for the total number of classes in that community should be consistent with an estimate made by aggregating these estimates of sub-groups. Such a property is called ‘additivity’, and it is worth mentioning that this property can be described in different degrees. The complete additivity appears when the estimated distributions of the integrated sub-groups and the whole community are not significantly different (at which all the statistical properties of the two are significantly equal), while the partial additivity appears when only some of the statistical properties of them are significantly equal. Therefore, in this section, we are trying to answer the third question in our criteria list: *Can we describe our model as being additive (either completely or partially additive)? How much difference is there between applying the model on the whole dataset, and applying the model on subsets of data then aggregating?*

Due to the complexity and high dimensionality of our proposed model, it is difficult to theoretically verify the additive property. Therefore, we restrict ourselves to looking at this issue empirically through simulation. However, there is no clear cut methodology in doing so, besides that, such as in species estimation, sub-groups used for estimation are somewhat arbitrary and can be done at different levels of the taxonomic hierarchy (genus, family, phylum *etc.*), and it is well known that conclusions drawn from data can change according to how that data is aggregated [90]. Hence, we followed two schemes of creating subgroups from the already available design of the three artificial worlds in order to provide a comprehensive analysis of the situation. One scheme is based on generating one whole group then splitting it into two subgroups, while the other scheme is based on generating two subgroups then merging them into a whole one. As we are in the species context, we consider each given artificial world as a whole category that includes multiple taxa on different levels. In the current study, in the splitting scheme we decided to split each category (artificial world) into two taxa (subgroups), while in the merging scheme we independently create two taxa (two different sub-artificial worlds but same design as the whole category) to be merged

into one whole category.

The main idea of testing the model additivity revolves around checking whether the posterior estimation of the whole category is equivalent to the aggregation of the posterior estimates of the two taxa. Therefore, we investigated main summary statistics such as the posterior mean, posterior mode, 95% credible interval, and the deviation measure which is reflected by the CDFs curves and box-plots. An additional statistical tool, Kullback-Leibler divergence metric (KLD)¹ [66], is also computed to measure the difference between the probability distributions of the whole category and the combined taxa. We computed this measure in two directions; one is when we consider the distribution of the whole category as a true distribution and the distribution of the combined taxa as an approximation of it, while in the other direction we consider the distribution of the combined taxa as a true distribution and the distribution of the whole category as an approximation of it. Our insight of applying the two directions is that the minimum difference between the two directions indicates the higher closeness between the two distributions.

The adopted two schemes are performed as follows:

Splitting Scheme:

1. Given a generated artificial world, randomly split it into two taxa - namely $S1$ and $S2$. In this process, we split the number of authors and the number of discoveries, as well as splitting the number of species into C_{S1} and C_{S2} , where $C = C_{S1} + C_{S2}$.
2. Fit each taxon separately, then derive the posterior distribution of C_{S1} and C_{S2} and compute their posterior of estimation C_{S1}^* and C_{S2}^* , respectively. However, the fitting stage of the subgroups is a challenging one due to the change in their distributional structure resulting from the splitting procedure. Although the

¹KLD evaluates the difference between two distributions, say $P(x)$ and $P(y)$, by measuring the information loss when using one distribution instead of the other. Note that the notation $\text{KLD}(P(x)|P(y))$ is used to express the amount of the information lost when using $P(y)$ instead of $P(x)$. It is mathematically defined as the expectation of the logarithmic difference between $P(x)$ and $P(y)$, calculated with respect to $P(x)$.

random splitting of the number of discoveries can be considered as a random sampling from the species population (*i.e.* the parameters values of the species abundances are kept the same in the taxa as in the whole category), the random splitting of the number of authors does not imply that the latent effort is the same in the taxa as the whole category. Therefore, we need to consider splitting the effort mean as well. As indicated in Section 4.1, the latent effort $\{l_j\}_{j=1:T}$ is applied on a log-scale, $L_j = \log(l_j)$, so that L_j acts as a discrete Gaussian Markov process and l_j as a discrete Log-Gaussian Markov process, where the effort mean = $\exp(\mu_{L_1} + \sigma_L^2/2)$. Note that the log-effort process is an Intrinsic Gaussian Markov process where its mean L_{j-1} is determined by the mean of the first variable, μ_{L_1} . Hence, when we do the fitting for each taxon, we choose values of their log-effort parameters (*i.e.* μ_{L_1} and σ_L^2 of each taxon) such that the sum of their effort means is equal to the effort mean of the whole category *i.e.* $\exp(\mu_{L_1} + \sigma_L^2/2)_{whole} \simeq \exp(\mu_{L_1} + \sigma_L^2/2)_{S_1} + \exp(\mu_{L_1} + \sigma_L^2/2)_{S_2}$. However, since the effort mean is just one equation in two variables, the number of solutions are infinite. Thus, we considered an additional constraint that maintains the percentage of discoveries in the data; more explanatory details will be illustrated within the artificial worlds examples in the next subsections. Note that such practices are not going to be applied in the reality, because in reality if we already have the estimation of the whole category, there is no point of doing the splitting scheme. These suggested scenarios are explained here just to investigate the additivity property which is a desirable property to save our time and effort in the case where there are already estimations of sub-groups and we would like to aggregate them. It is worth mentioning that fitting a sub-group is subject to same principles of fitting whole group in that we should refer to the species literature, expert opinion, and statistical assumptions, along with pilot experimentations for specifying the distributional assumptions and parameters' values.

3. Combine the resultant samples of the fitting, and check how much the difference would be between C^* (out of the usual fitting of the whole category) and $C_{S_1+S_2}^*$ (out of the combined fitting of the two splitted taxa).

Merging Scheme:

1. At a given design (such as the design of World-A), generate two other artificial worlds (as subgroups or taxa) independently - namely M1 and M2 - according to the same design of the given world but with different number of species denoted by C_{M1} and C_{M2} for each taxon respectively, where $C = C_{M1} + C_{M2}$. However, the generating stage of the subgroups is a challenging stage in this scheme due to the ambiguity in how to create a subgroup of the same design of the whole category. Therefore, in addition to using two different values of the number of species (C_{M1} and C_{M2}) in generating the subgroups, three attempts are made to approach what we call it “same design”:

- Treatment (1): For generating each taxon, keep the same seed and same parametrization of the given whole category, except for the parameters of the latent log-effort process. For the parameters of this process, choose values such that the sum of the efforts means of the two taxa equals the effort mean of the whole category *i.e.* $\exp(\mu_{L1} + \sigma_L^2/2)_{whole} = \exp(\mu_{L1} + \sigma_L^2/2)_{M1} + \exp(\mu_{L1} + \sigma_L^2/2)_{M2}$. We borrowed this treatment from the splitting scheme, which is motivated by the idea of splitting the effort mean of the whole category over the two taxa so that, in the aggregating stage, the parametrization of the whole category will be kept in the combined taxa.
- Treatment (2): For generating each taxon, keep the same parametrization of the given whole category, but use different seeds. This treatment is motivated by the idea of using the parametrization of the given whole category as a reference point in characterizing the subgroups to be able to perform the fitting stage for these subgroups.
- Treatment (3): For generating each taxon, keep the same seed and same parametrization of the given whole category. This treatment is motivated by the idea of achieving $P(C_{M1} | \text{all data})$ and $P(C_{M2} | \text{all data})$ in order to check the additivity property through checking whether that $P(C | \text{all data}) = P(C_{M1} + C_{M2} | \text{all data})$. In generating each taxon, fixing every thing (the seed and the parametrization) to be the same as the ones of the whole category should, in theory, provide us with “all data”. However, satisfying this is not completely achievable since the number of discoveries is directly

affected by the number of species which is different in the two taxa from the whole category *i.e.* $\{\text{all data}\}_{whole} > \{\text{all data}\}_{Mi}$ since $\{\text{all data}\}_{whole} = \{\mathbf{t}, \mathbf{x}\}$, while $\{\text{all data}\}_{Mi} = \{\mathbf{t}_i, \mathbf{x}\}$, where $i = 1, 2$.

2. Fit each taxon separately, then derive the posterior distribution of C_{M1} and C_{M2} and compute their posterior of estimation C_{M1}^* and C_{M2}^* , respectively.
3. Merge the two taxa by combining their resultant samples of the fitting, and check how much the difference between C^* (out of the usual fitting of the whole category) and C_{M1+M2}^* (out of the merged fitting of the two taxa).

The above two schemes are exactly followed in World-A and World-B as will be detailed in the following Section 5.3.1 and Section 5.3.2, respectively. In World-C, we faced some challenges due to having skewness in the estimated posterior distribution, and adapting the (50%, 50%) division in creating two subgroups. Thus, we followed a special treatment for World-C to measure the additivity property as will be explained in Section 5.3.3. However, this part motivated some extended analysis for World-A and World-B as we will see in the last Section 5.3.4. Note that holistic-view plots and results are available in Appendix B, Section B.4.

5.3.1 Additivity in World-A Design:

In the Splitting Scheme, World-A is randomly split into taxon S1 and taxon S2 which represent 70% and 30%, respectively, of the whole category (World-A), so that $C_{S1} = 7000$ and $C_{S2} = 3000$. Figure 5.11 shows the number of discoveries

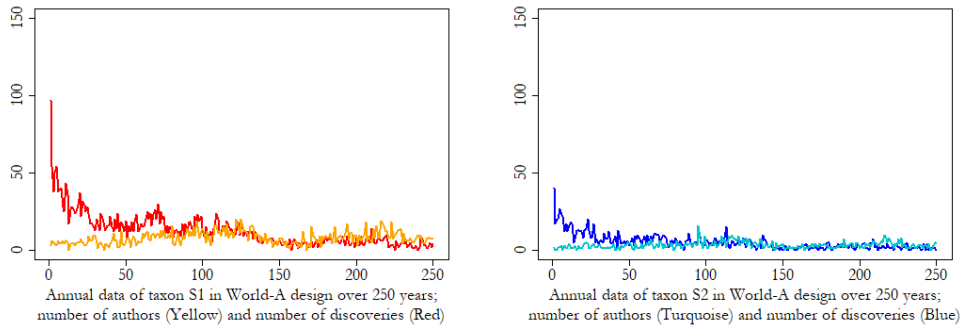


Figure 5.11: The resulting data of the Splitting Scheme of World-A.

and the number of authors of the two taxa after the splitting procedure. As we mentioned in the introduction of the splitting scheme, we need to choose different parameters for the effort process in the two taxa such that the effort mean after the aggregation stage would be the same as the effort mean of the whole category. The effort mean of World-A $\simeq \exp(5 + (0.1)^2/2)_{whole} \simeq 149.2$, thus we need to specify values for μ_{L_1} and σ_L in taxon S1 such that its effort mean = $\exp(\mu_{L_1} + \sigma_L^2/2)_{S1} \simeq 70\% \times \exp(5 + (0.1)^2/2)_{whole} \simeq 104.4$, and values for μ_{L_1} and σ_L in taxon S2 such that its effort mean = $\exp(\mu_{L_1} + \sigma_L^2/2)_{S2} = 30\% \times \exp(5 + (0.1)^2/2)_{whole} \simeq 44.7$. However, since the effort mean is just one equation in two variables, the number of solutions are infinite. Thus, we needed to consider additional constraints which is “approximately maintaining the percentage of the discoveries”. In World-A, the percentage of the discoveries $\simeq 44.7\%$, and after the splitting the percentage of the discoveries $\simeq 45.4\%$ and 42.9% in taxon S1 and S2, respectively. Therefore, we need to specify values for μ_{L_1} and σ_L in taxon S1 and taxon S2 such that we can closely maintain the above percentages, individually at each taxon and totally after aggregating the two taxa. After several trials (that include proposing a set of parameters’ values, re-generating S1 and S2 under the candidate values, and verifying the effort means and the percentage of the discoveries), we found that the following parametrization (Table 5.4) seems to be the best choice that approximately maintain the targeted effort mean and the percentage of the discoveries.

Table 5.4: Parametrization of whole category vs. two taxa of World-A.

Hyper-Parameters	World-A	Taxon S1	Taxon S2
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	{4, 2.5}	{4, 2.5}
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	{4.64, 0.094}	{3.82, 0.089}
$\theta_x = \alpha$	$\log(30)$	$\log(30)$	$\log(30)$

In taxon S1, the effort mean $\simeq \exp(4.64 + (0.094)^2/2)_{S1} \simeq 104$ and the discovery percentage $\simeq 46.2\%$, while in taxon S2, the effort mean $\simeq \exp(3.82 + (0.089)^2/2)_{S2} \simeq 45.8$ and the discovery percentage $\simeq 40.4\%$. Hence, in this selected parametrization, after aggregating the two taxa, the total effort mean $\simeq \exp(4.64 + (0.094)^2/2)_{S1} + \exp(3.82 + (0.089)^2/2)_{S2} \simeq 149.8$ with total discovery percentage $\simeq (46.2\% + 40.4\%)/2 \simeq 43.3\%$ which are very close to the given ones (149.2 and 44.7%, respectively) in the whole category of World-A. Figure

5.12 shows the fitting results (according to the above selected parametrization) of each taxon, which seem to be successful estimations at which the posterior mean and mode are close to the true values in both taxa, more details are available in Group B.8, Appendix B.

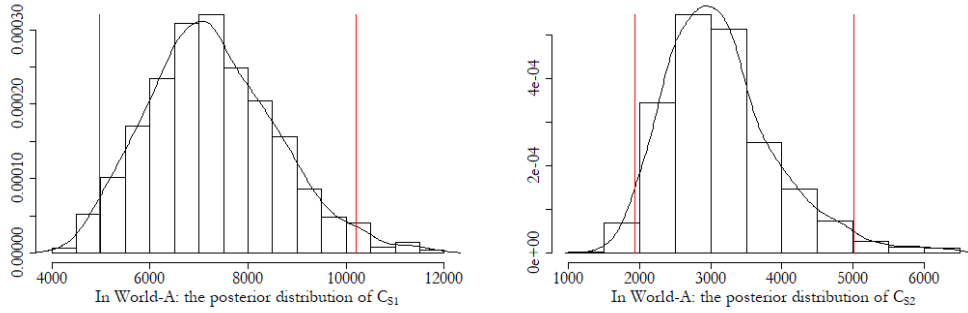


Figure 5.12: The fitting stage in the Splitting Scheme of the two taxa in World-A.

Going to the final stage of the splitting scheme, we got very close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.13. It seems that the central tendency measures of the whole category and the aggregated fitted taxa are very close especially the posterior means at

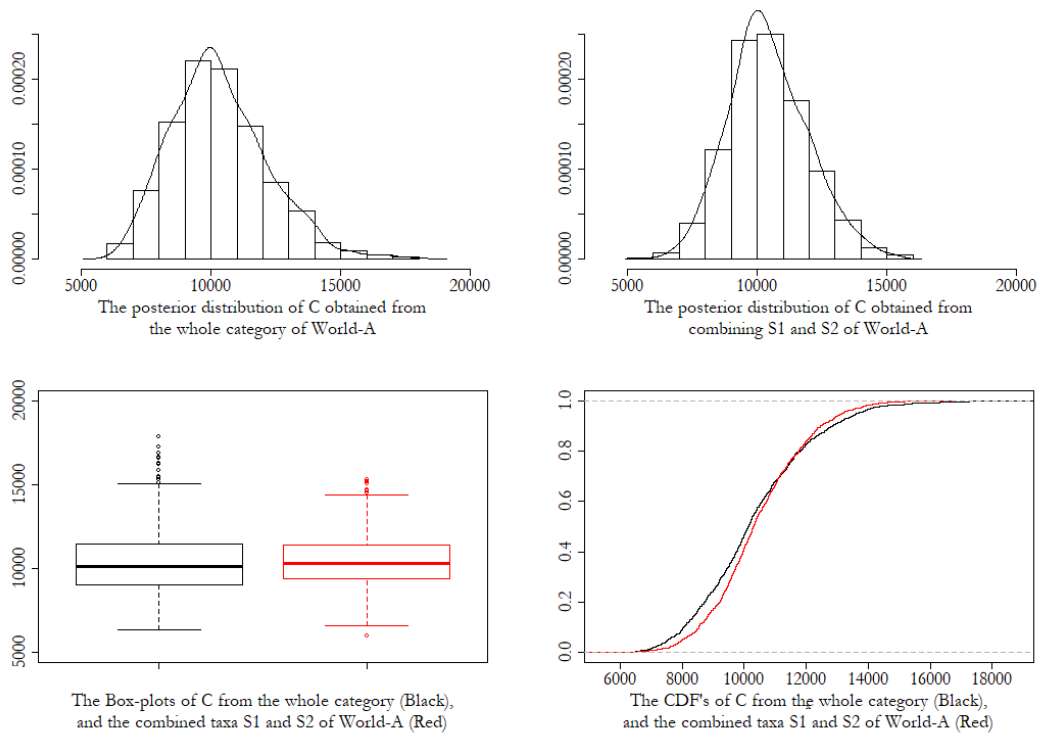


Figure 5.13: The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.

which $C^* = 10,320$ for the whole category, while $C_{S_1+S_2}^* = 10,436$. Although the aggregation treatment includes more uncertainty due to fitting two distributions that are associated with two Monte Carlo errors instead of one, it seems that the successful fitting of both resulted in less variational estimation (about 2.4 million) compared to the whole category fitting (about 3.4 million). This difference in the variation justifies the little difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.13. With respect to the KLD metric, is computed in both directions and we found that $\text{KLD}(P(C|\text{rest})|P(C_{S_1+S_2}|\text{rest})) = 0.003035$ and $\text{KLD}(P(C_{S_1+S_2}|\text{rest})|P(C|\text{rest})) = 0.003163$ *i.e.* the amount of the information lost when using the whole category distribution instead of the combined taxa distribution is approximately similar to vice versa, which indicates the very closeness of these two distributions.

In the Merging Scheme - Treatment (1), we adopted the same idea of splitting the effort mean, as explained in the introduction of the current section. Two sub-artificial worlds, taxon M1_T(1) and taxon M2_T(1), are generated “independently” according to the same parametrization of S1 and S2, respectively, that are used in Table 5.4, where $C_{M1_T(1)} = 7000$ and $C_{M2_T(1)} = 3000$, as shown in Figure 5.14.

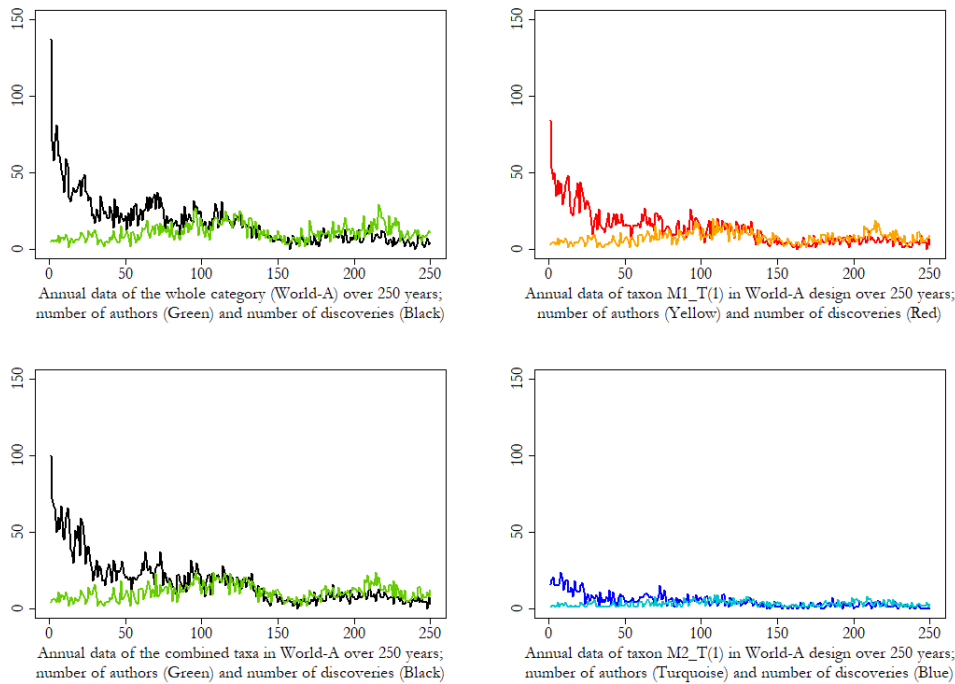


Figure 5.14: The resulting data of the Merging Scheme - Treatment (1) of World-A design.

It is worth mentioning that M1_T(1) and M2_T(1) are generated with the same seed that is used in generating the whole category (World-A). By comparing the top plot with the bottom plot of the left panel in this figure, we can see that there are still some differences between the observed data resulting from aggregating taxon M1_T(1) and taxon M2_T(1), and the data of the original world (World-A), where the current number of discoveries of the original world is $D = 4,468$, while in the combined world is $D_{M1_T(1)+M2_T(1)} = 4,448$. However, they are similar in the general pattern and they both share the same seed, same parameters of the sampling-discovery process, and almost the same effort mean, where $\exp(4.64 + (0.094)^2/2)_{M1_T(1)} + \exp(3.82 + (0.089)^2/2)_{M2_T(1)} \simeq 149.8$ and $\exp(5 + (0.1)^2/2)_{whole} \simeq 149.2$. Figure 5.15 shows the fitting results of each taxon which seem to be successful estimations, but with little over-estimation in taxon M1_T(1) and little under-estimation in M2_T(1), see Group B.8, Appendix B.

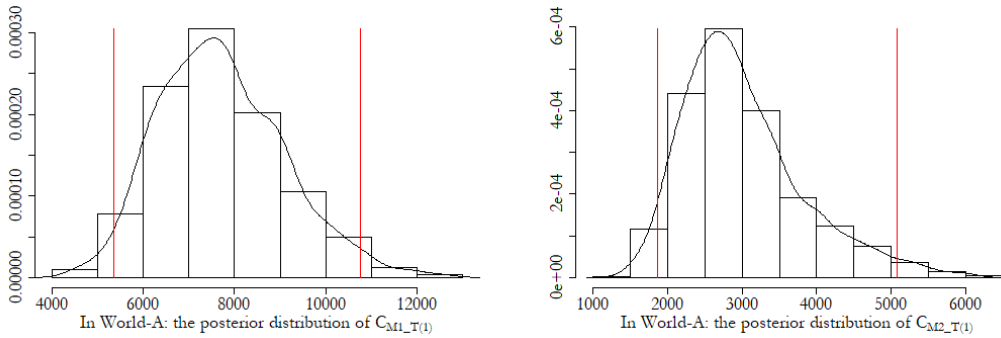


Figure 5.15: The fitting stage in the Merging Scheme - Treatment (1) of the two taxa in World-A.

Going to the final stage of the merging scheme - treatment (1), we also got close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.16, but with little over-estimation at which the posterior means $C^* = 10,320$ for the whole category and $C^*_{M1_T(1)+M2_T(1)} = 10,738$ for the combined taxa. Again, it seems that the aggregated fitting resulted in less variational estimation (about 2.6 million) compared to the fitting of the whole category (about 3.4 million), which justifies the difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.16. We also computed the KLD metric in both directions and we found that $\text{KLD}(P(C|\text{rest})|P(C_{M1_T(1)+M2_T(1)}|\text{rest})) = 0.001066$ and $\text{KLD}(P(C_{M1_T(1)+M2_T(1)}|\text{rest}) | P(C|\text{rest})) = 0.001061$ which means that the amount of the information

lost when using the whole category distribution instead of the combined taxa distribution is approximately close to vice versa, therefore, indicating the closeness of these two distributions.

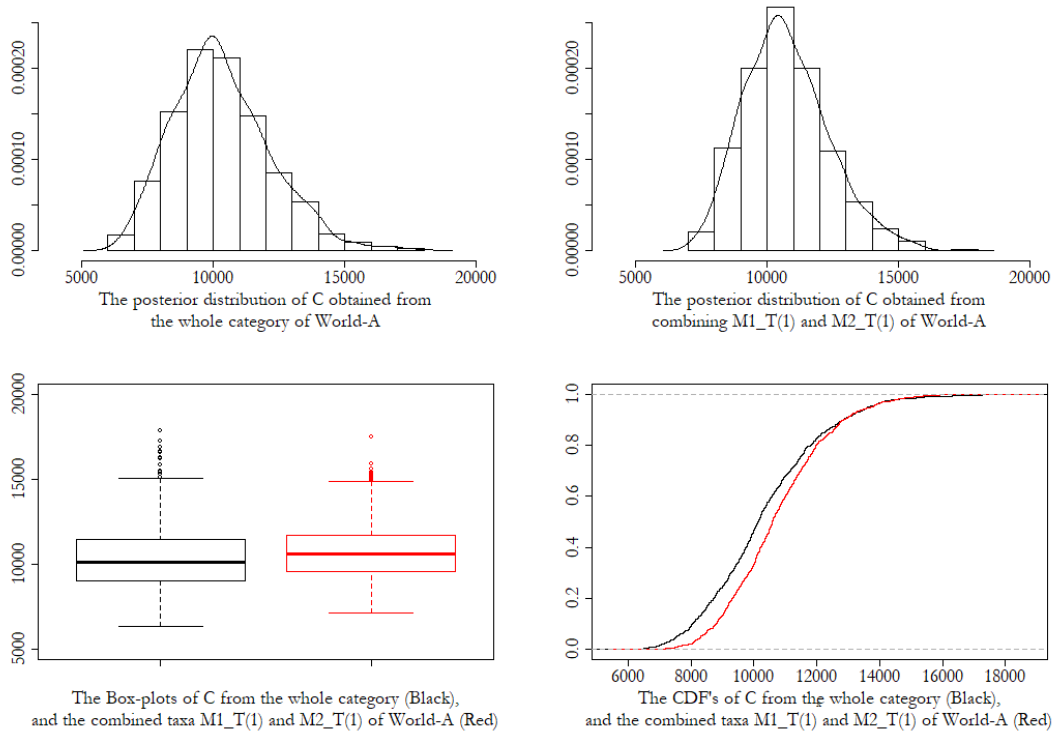


Figure 5.16: The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.

In the Merging Scheme - Treatment (2), two sub-artificial worlds, taxon M1_T(2) and taxon M2_T(2), of the same design as World-A are generated independently such that $C_{M1.T(2)} = 7000$ and $C_{M2.T(2)} = 3000$. They have the exact parametrization of World-A, but different seeds. After multiple trials with a variety of seeds, we managed to create two worlds that are close enough to represent subgroups of the whole category (World-A), taxon M1_T(2) with seed=94 and taxon M2_T(2) with seed= 76, while the whole category is of seed=1, as shown in Figure 5.17. By comparing the top plot with the bottom plot of the left panel in this figure, we can see that the observed data resulting from aggregating taxon M1_T(2) and taxon M2_T(2) is different in details from the data of the original world (World-A), where the current number of discoveries of the original world is $D = 4,468$, while in the combined world is $D_{M1.T(2)+M2.T(2)} = 4,924$. However, they are similar in the general pattern and they are considered samples from the

same underlying process.

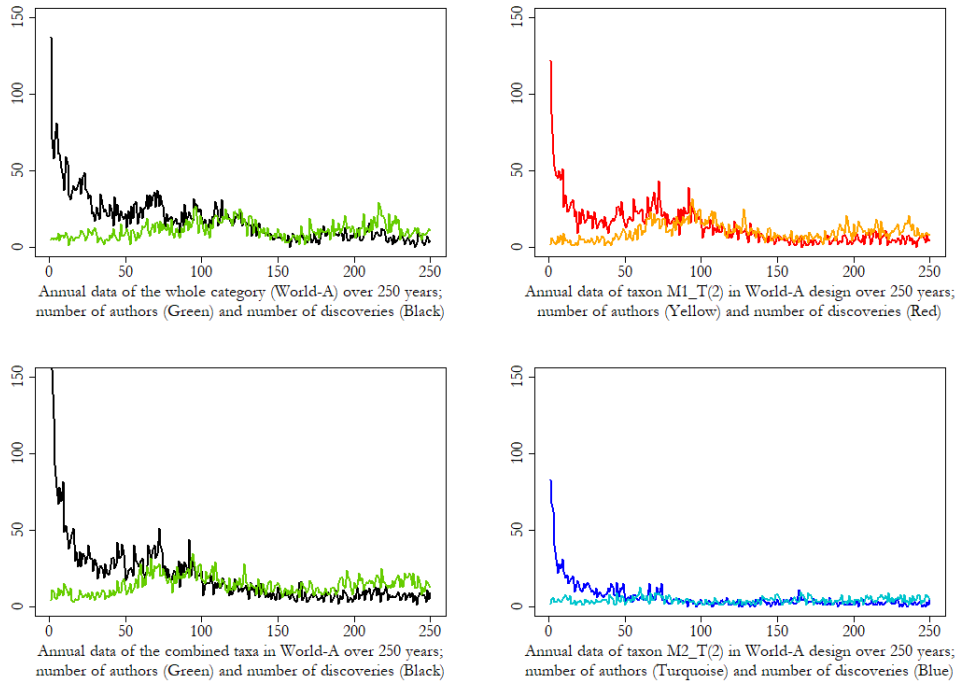


Figure 5.17: The resulting data of the Merging Scheme - Treatment (2) of World-A design.

Figure 5.18 shows the fitting results of each taxon which seem to be successful estimations at which the posterior mean and mode are close to the true values in both taxa, more details are available in Group B.9, Appendix B.

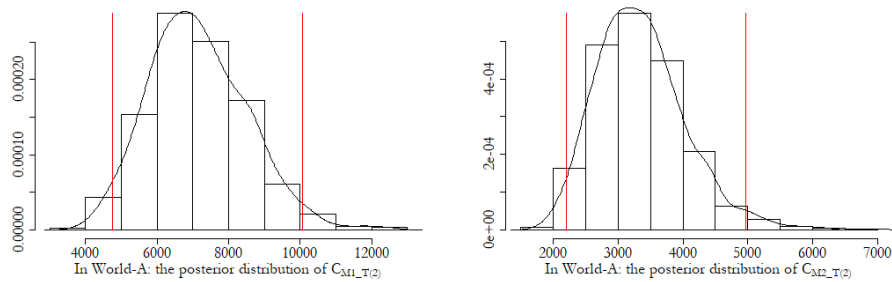


Figure 5.18: The fitting stage in the Merging Scheme - Treatment (2) of the two taxa in World-A.

Going to the final stage of the merging scheme - treatment (2), we also got close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.19. It seems that the central tendency measures are very close with posterior means $C^* = 10,320$ for the whole category, and $C^*_{M1.T(2)+M2.T(2)} = 10,517$ for the combined taxa. Again, it seems that the aggregated fitting resulted in less variational estimation (about 2.4 million) compared

to the fitting of the whole category (about 3.4 million), which justifies the little difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.19. We also computed the KLD metric in both directions and we found that $\text{KLD}(P(C|\text{rest}) | P(C_{M1.T(2)+M2.T(2)}|\text{rest})) = 0.003933$ and $\text{KLD}(P(C_{M1.T(2)+M2.T(2)} | \text{rest}) | P(C|\text{rest})) = 0.003968$ which means that the amount of the information lost when using the whole category distribution instead of the combined taxa distribution is approximately similar to vice versa, therefore, indicating the closeness of these two distributions.

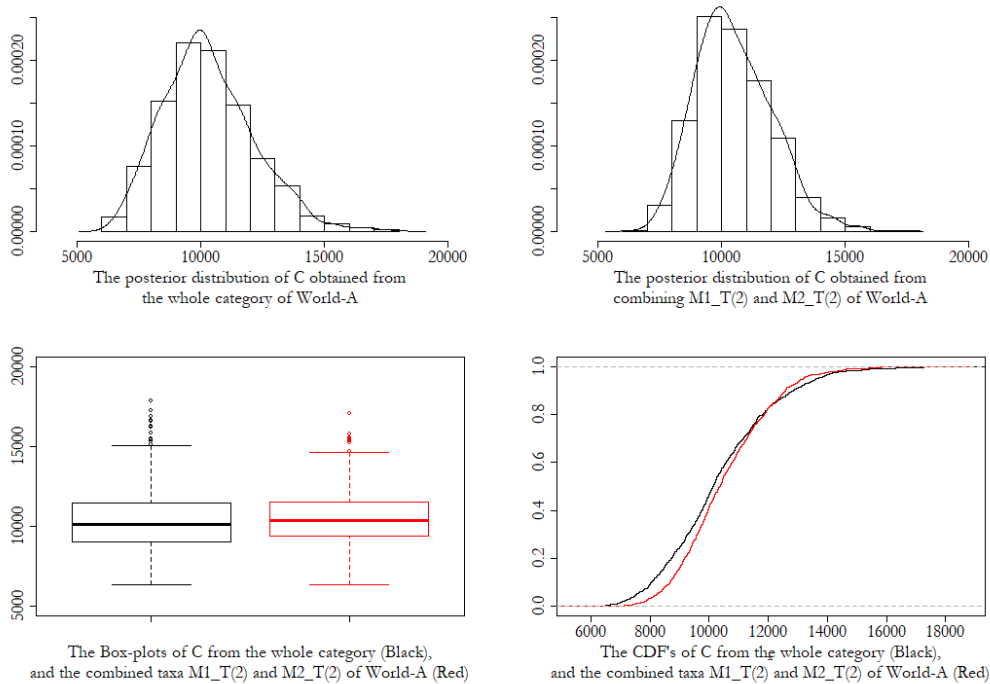


Figure 5.19: The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.

In the Merging Scheme - Treatment (3), again two sub-artificial worlds, taxon M1_T(3) and taxon M2_T(3), of the same design as World-A are generated independently such that $C_{M1.T(3)} = 7000$ and $C_{M2.T(3)} = 3000$. They have the exact parametrization of World-A with the same seed=1. Figure 5.20 shows the generated number of discoveries and the number of authors of the two taxa, as well as the fitting results of each taxon which seem to be successful estimations, more details are available in Group B.9, Appendix B. In this treatment, the current number of discoveries of the original world (World-A) is $D = 4,468$, while in the combined world is $D_{M1.T(3)+M2.T(3)} = 5,166$. In addition, we can see in

the top panel of Figure 5.20 that the number of authors is the same in both taxa, which is a results of conditioning on the whole data as detailed previously in the explanation of Treatment (3). However, in the aggregating stage, we are interested in combining the fitting results *i.e.* combining the posterior distributions of the two taxa not combining the given datasets.

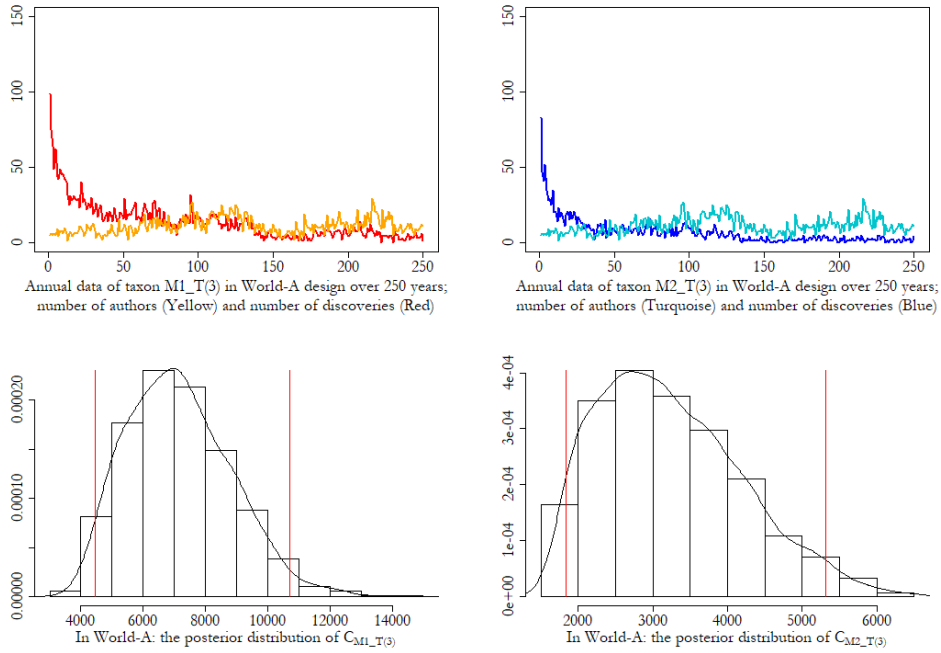


Figure 5.20: The generated data and the results of the fitting stage of the Merging Scheme - Treatment (3) of World-A design. The left panel belongs to taxon M1.T(3), while the right panel belongs to taxon M2.T(3).

Going to the final stage of the merging scheme - treatment (3), we also got very close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.19. It seems that the central tendency measures are very close with posterior means $C^* = 10,320$ for the whole category, and $C_{M1.T(3)+M2.T(3)}^* = 10,397$ for the combined taxa. Unlike the previous treatments, it seems that the aggregated fitting resulted in estimated variation that is equivalent to the one of fitting the whole category, where the former is about 3.6 million, while the latter is about 3.4 million. This reflects having the almost identical CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.21. We also computed the KLD metric in both directions and we found that $\text{KLD}(P(C|\text{rest}) | P(C_{M1.T(3)+M2.T(3)} | \text{rest})) = 0.002981$ and $\text{KLD}(P(C_{M1.T(3)+M2.T(3)} | \text{rest}) | P(C|\text{rest})) = 0.002918$ which means that the amount of the information lost when using the whole category distribution in-

stead of the combined taxa distribution is almost as the same as vice versa, and therefore, indicates the high closeness of these two distributions.

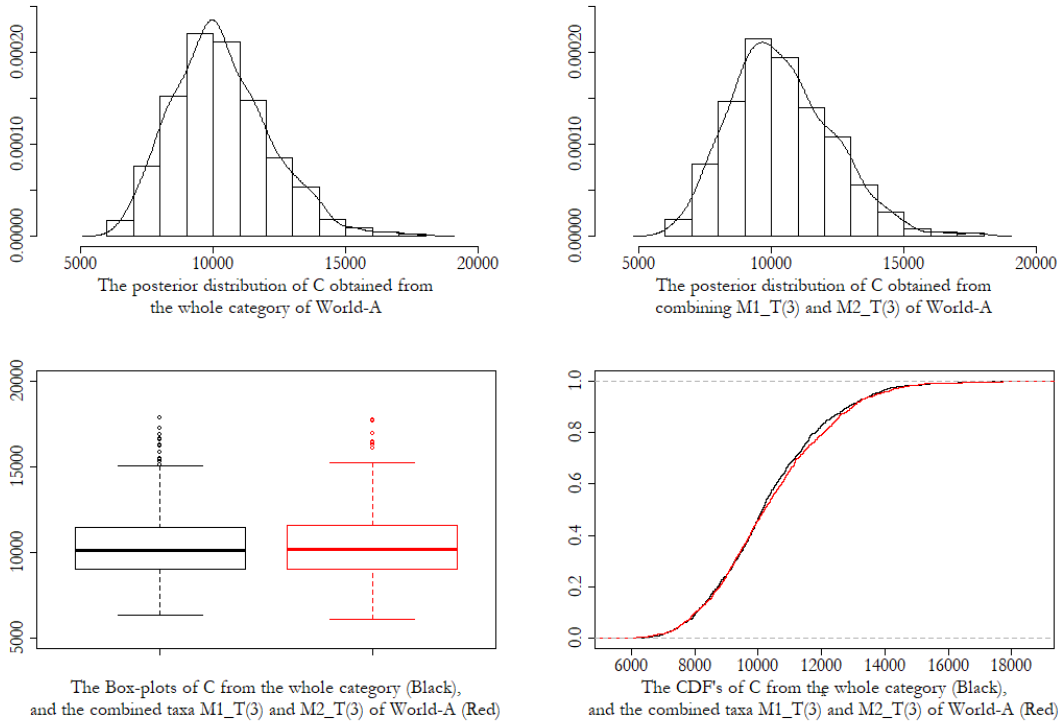


Figure 5.21: The aggregating stage of the Merging Scheme - Treatment (3) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-A.

5.3.2 Additivity in World-B Design:

In the Splitting Scheme, World-B is randomly split into taxon S1 and taxon S2 which represent 60% and 40%, respectively, of the whole category (World-B), so that $C_{S1} = 60,000$ and $C_{S2} = 40,000$. Figure 5.22 shows the number of discoveries and the number of authors of the two taxa after the splitting procedure. As we did in World-A, we need to choose different parameters for the effort process in the two taxa such that the effort mean after the aggregation stage would be the same as the effort mean of the whole category. The effort mean of World-B $\simeq \exp(7 + (0.15)^2/2)_{whole} \simeq 1109$, thus we need to specify values for μ_{L1} and σ_L in taxon S1 such that its effort mean = $\exp(\mu_{L1} + \sigma_L^2/2)_{S1} \simeq 60\% \times \exp(7 + (0.15)^2/2)_{whole} \simeq 665.4$, and values for μ_{L1} and σ_L in taxon S2 such that its effort mean = $\exp(\mu_{L1} + \sigma_L^2/2)_{S2} = 40\% \times \exp(7 + (0.15)^2/2)_{whole} \simeq 443.6$. However, since the effort mean is just one equation in two variables, the number of solutions are infinite. Thus,

we needed to consider additional constraints which is “approximately maintaining the percentage of the discoveries”. In World-B, the percentage of the discoveries $\simeq 31.9\%$, and after splitting the percentage of the discoveries $\simeq 31.9\%$ and 31.8% in taxon S1 and S2, respectively. Therefore, we need to specify values for μ_{L_1} and σ_L in taxon S1 and taxon S2 such that we can closely maintain the above percentages, individually at each taxon and totally after aggregating the two taxa.

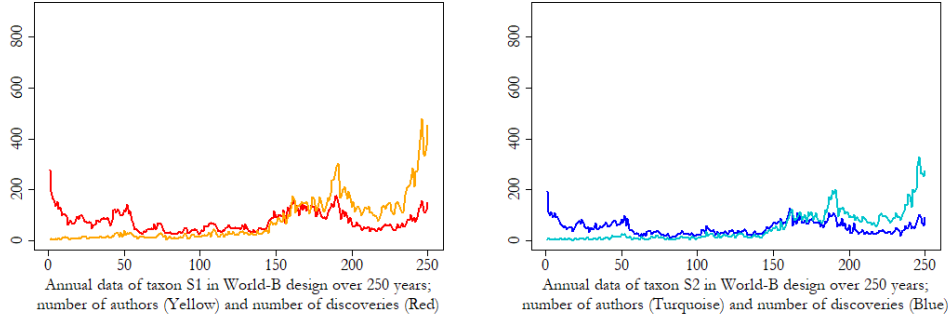


Figure 5.22: The resulting data of the Splitting Scheme of World-B.

After several trials (that include proposing a set of parameters’ values, re-generating S1 and S2 under the candidate values, and verifying the effort means and the percentage of the discoveries), we found that the parametrization shown in Table 5.5 seems to be the best choice that approximately maintain the targeted effort mean and the percentage of the discoveries. In taxon S1, the effort mean $\simeq \exp(6.49 + (0.146)^2/2)_{S_1} \simeq 665.6$ and the discovery percentage $\simeq 29.3\%$, while

Table 5.5: Parametrization of whole category vs. two taxa of World-B.

Hyper-Parameters	World-B	Taxon S1	Taxon S2
$\theta_N = \{\mu, \sigma\}$	{10,3.5}	{10,3.5}	{10,3.5}
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7,0.15}	{6.49,0.146}	{6.08,0.142}
$\theta_x = \alpha$	$\log(100)$	$\log(100)$	$\log(100)$

in taxon S2, the effort mean $\simeq \exp(6.08 + (0.142)^2/2)_{S_2} \simeq 441.5$ and the discovery percentage $\simeq 31.8\%$. Hence, in this selected parameters, after aggregating the two taxa, the total effort mean $\simeq \exp(6.49 + (0.146)^2/2)_{S_1} + \exp(6.08 + (0.142)^2/2)_{S_2} \simeq 1107$ with total discovery percentage $\simeq (29.3\% + 31.8\%)/2 \simeq 30.6\%$ which are very close to the given ones (1109 and 31.9%, respectively) in the whole category. Figure 5.23 shows the fitting results (according to the above selected

parametrization) of each taxon, which seem to be successful estimations at which the posterior mean and mode are close to the true values in both taxa, more details are available in Group B.11, Appendix B

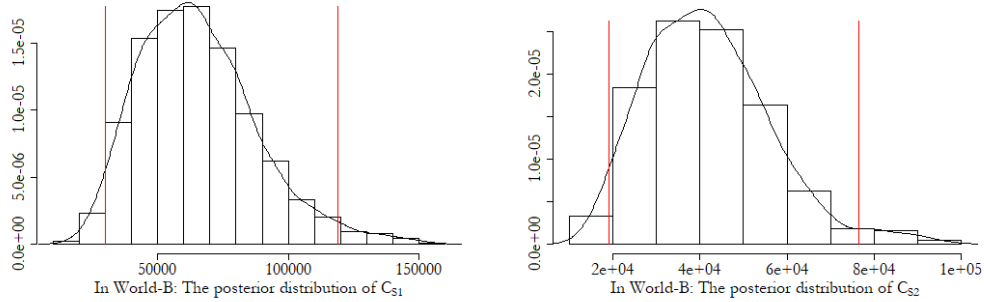


Figure 5.23: The fitting stage in the Splitting Scheme of the two taxa in World-B.

Going to the final stage of the splitting scheme, we got close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.24. It seems that the central tendency measures of the whole category and the aggregated fitted taxa are close where, for example, the posterior means $C^* = 109,793$ for the whole category and $C_{S1+S2}^* = 107,554$ for the combined fitting. In addition, the successful fitting of both resulted in less variational estimation

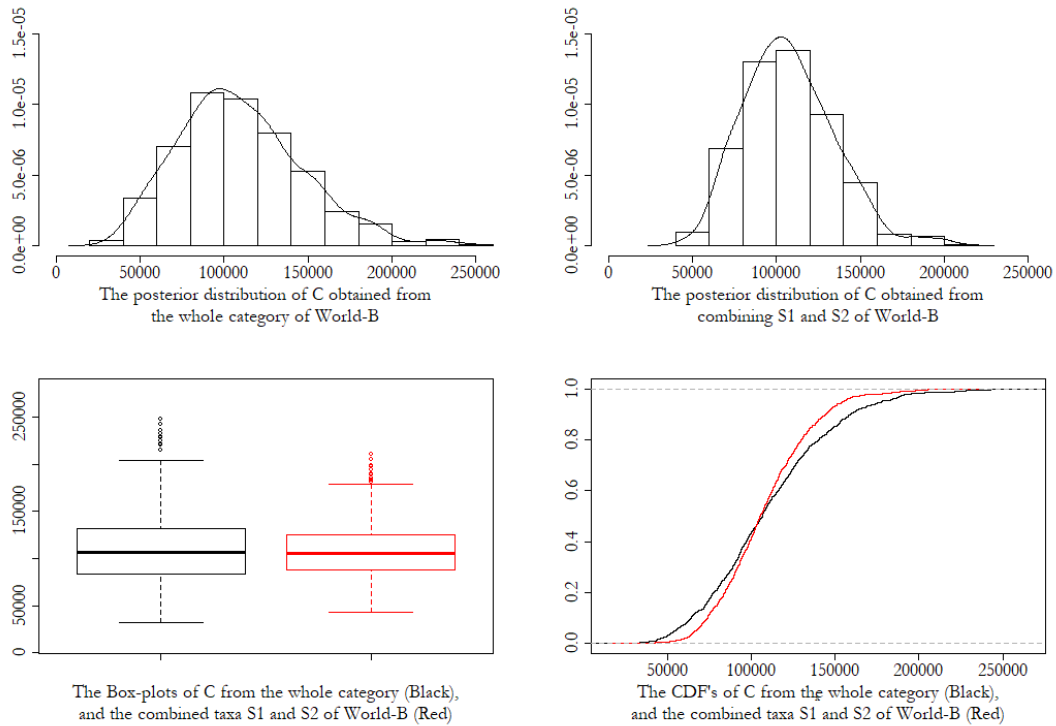


Figure 5.24: The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.

(about 0.75 billion) compared to the whole category fitting (about 1.4 billion). The difference in the variation justifies the difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.24. With respect to the KLD metric, it is computed in both directions and found that $\text{KLD}(P(C|\text{rest})|P(C_{S_1+S_2}|\text{rest})) = 0.001677$, while $\text{KLD}(P(C_{S_1+S_2}|\text{rest})|P(C|\text{rest})) = 0.001777$ *i.e.* the amount of the information lost when using the whole category distribution instead of the combined one is close to vice versa, which indicates the closeness of these two distributions.

In the Merging Scheme - Treatment (1), we adopted the same idea of splitting the effort mean. Two sub-artificial worlds, taxon M1_T(1) and taxon M2_T(1), are generated “independently” according to the same parametrization of S1 and S2, respectively, that are used in Table 5.5, where $C_{M1.T(1)} = 60,000$ and $C_{M2.T(1)} = 40,000$. M1_T(1) and M2_T(1) are generated with the same seed that is used in generating the whole category (World-B), see Figure 5.25.

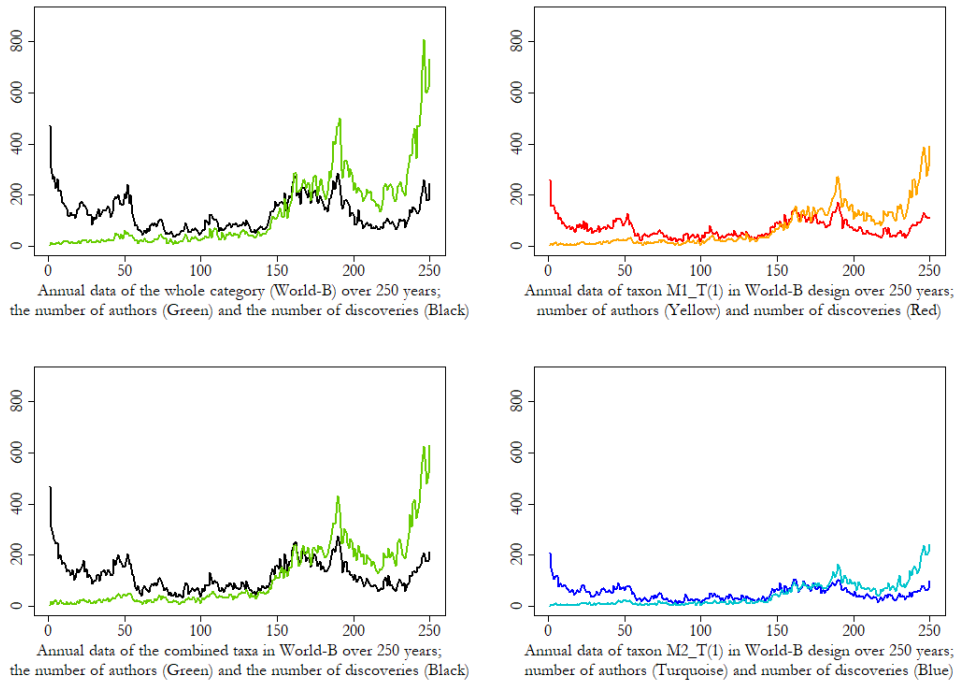


Figure 5.25: The resulting data of the Merging Scheme - Treatment (1) of World-B design.

By comparing the top plot with the bottom plot of the left panel in this figure, we can see that there are still some differences between the observed data resulting from aggregating taxon M1_T(1) and taxon M2_T(1) and the data of the

original world (World-B), where the current number of discoveries of the original world is $D = 31,865$, while in the combined world is $D_{M1.T(1)+M2.T(1)} = 30,269$. However, they are similar in the general pattern and they both share the same seed, same parameters of the sampling-discovery process, and almost the same effort mean, where $\exp(6.49 + (0.146)^2/2)_{M1.T(1)} + \exp(6.08 + (0.142)^2/2)_{M2.T(1)} \simeq 1107$ and $\exp(7 + (0.15)^2/2)_{whole} \simeq 1109$. Figure 5.26 shows the fitting results of each taxon which seem to be successful estimations, see Group B.11, Appendix B for more details.

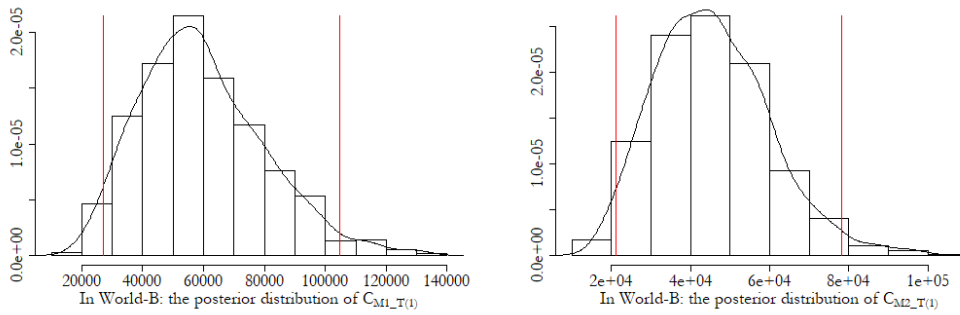


Figure 5.26: The fitting stage in the Merging Scheme - Treatment (1) of the two taxa in World-B.

Going to the final stage of the merging scheme - treatment (1), we also got approximately close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.27, at which the posterior means $C^* = 109,793$ for the whole category and $C^*_{M1.T(1)+M2.T(1)} = 105,199$ for the combined taxa. Again, it seems that the aggregated fitting resulted in less variational estimation (about 0.64 billion) compared to the fitting of the whole category (about 1.4 billion), which justifies the difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.27. We also computed the KLD metric in both directions and we found that in one direction the $\text{KLD}(P(C|\text{rest}) | P(C_{M1.T(1)+M2.T(1)}|\text{rest})) = 0.001512$, while in the other direction the $\text{KLD}(P(C_{M1.T(1)+M2.T(1)}|\text{rest}) | P(C|\text{rest})) = 0.001554$ which means that the amount of the information lost when using the whole category distribution instead of the combined taxa distribution is approximately close to vice versa, and therefore, indicating the closeness of these two distributions.

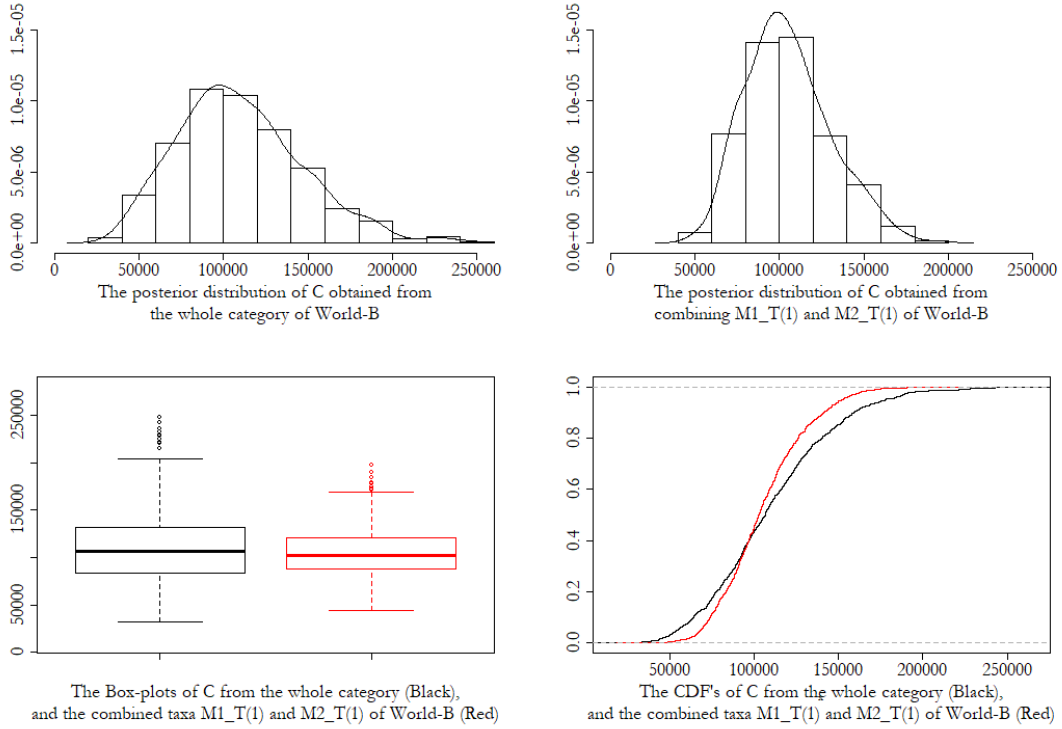


Figure 5.27: The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.

In the Merging Scheme - Treatment (2), two sub-artificial worlds, taxon M1_T(2) and taxon M2_T(2), of the same design as World-B are generated independently such that $C_{M1.T(2)} = 60,000$ and $C_{M2.T(2)} = 40,000$. They have the exact parametrization of World-B, but different seeds. After multiple trials with a variety of seeds, we managed to create two worlds that are close enough to represent subgroups of the whole category (World-B), taxon M1_T(2) with seed= -1365400275 and taxon M2_T(2) with seed= -1365400324 , while the whole category is of seed= -1365400193 , as shown in Figure 5.28. By comparing the top plot with the bottom plot of the left panel in this figure, we can see that the observed data resulting from aggregating taxon M1_T(2) and taxon M2_T(2) are different in details from the data of the original world (World-B), where the current number of discoveries of the original world is $D = 31,865$, while in the combined world is $D_{M1.T(2)+M2.T(2)} = 30,963$. However, they are similar in the general pattern and they both are considered as samples from the same underlying process. Figure 5.29 shows the fitting results of each taxon which seem to be

successful estimations, but with little under-estimation in the posterior mode of both taxon $M1_T(2)$ and taxon $M2_T(2)$, see Group B.12, Appendix B.

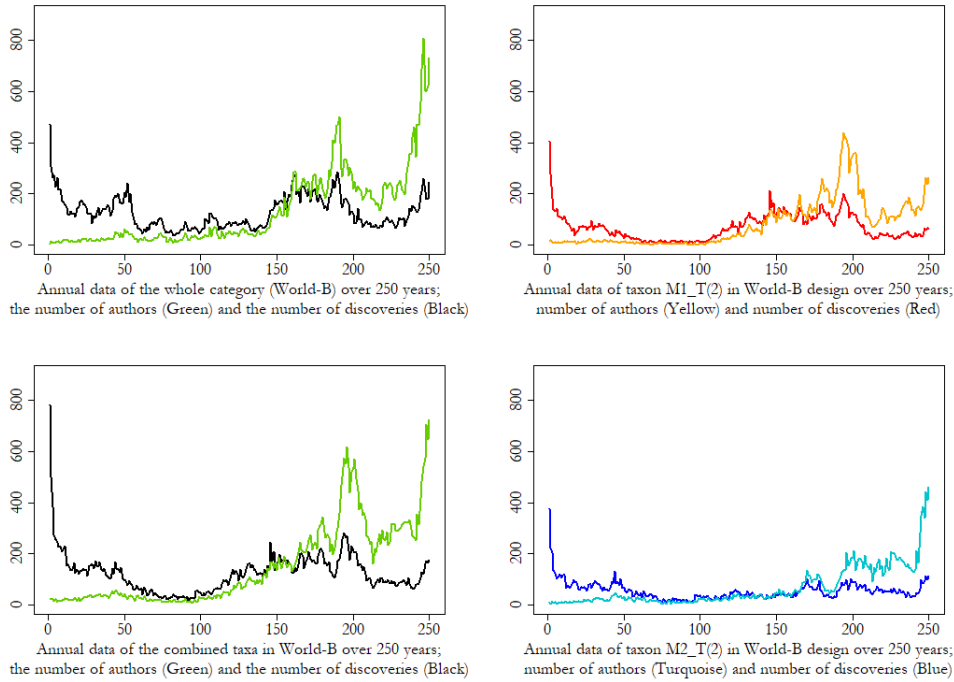


Figure 5.28: The resulting data of the Merging Scheme - Treatment (2) of World-B design.

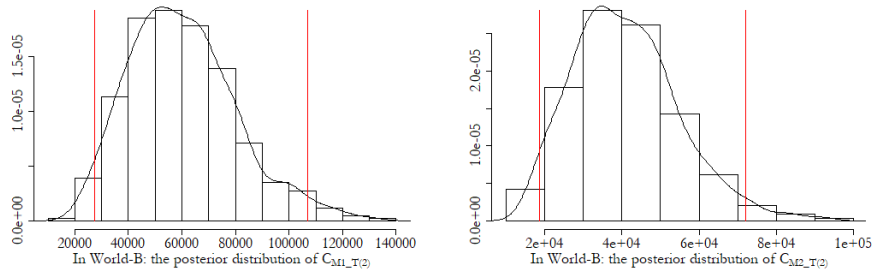


Figure 5.29: The fitting stage in the Merging Scheme - Treatment (2) of the two taxa in World-B.

Going to the final stage of the merging scheme - treatment (2), we also got approximately close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.30. It seems that the central tendency measures are very close with posterior means $C^* = 109,793$ for the whole category, and $C^*_{M1_T(2)+M2_T(2)} = 101,513$ for the combined taxa. Again, it seems that the aggregated fitting resulted in a less variational estimation (about 0.62 billion) compared to the fitting of the whole category (about 1.4 billion), which justifies the difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.30. We also computed the KLD metric in

both directions and we found that $\text{KLD}(P(C|\text{rest})|P(C_{M1.T(2)+M2.T(2)}|\text{rest})) = 0.001933$ and $\text{KLD}(P(C_{M1.T(2)+M2.T(2)}|\text{rest})|P(C|\text{rest})) = 0.002066$ which means that the amount of the information lost when using the whole category distribution instead of the combined taxa distribution is not that far to vice versa situation, and therefore, indicating a sort of closeness of these two distributions.

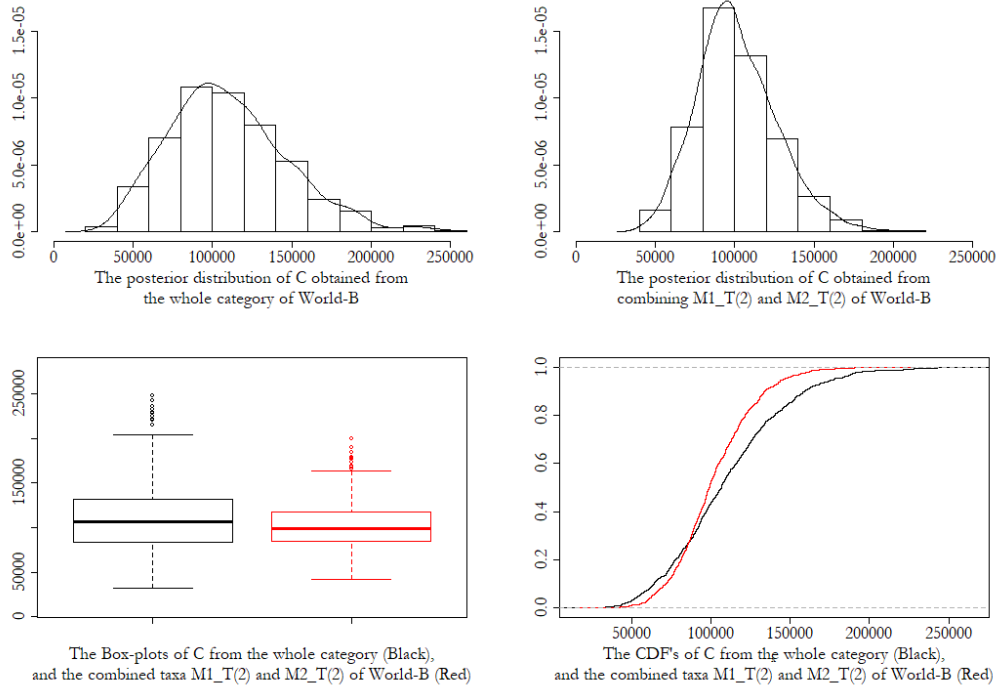


Figure 5.30: The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.

In the Merging Scheme - Treatment (3), again two sub-artificial worlds, taxon M1_T(3) and taxon M2_T(3), of the same design as World-B are generated independently such that $C_{M1.T(3)} = 60,000$ and $C_{M2.T(3)} = 40,000$. They have the exact parametrization of World-B with the same seed=-1365400193. Figure 5.31 shows the generated number of discoveries and the number of authors of the two taxa, as well as the fitting results of each taxon which seem to be successful estimations, more details are available in Group B.12, Appendix B. In this treatment, the current number of discoveries of the original world (World-B) is $D = 31,865$, while in the combined world is $D_{M1.T(3)+M2.T(3)} = 39,412$. In addition, we can see in the top panel of Figure 5.31 that the number of authors is the same in both taxa, which is the result of conditioning on the whole data as detailed previously in the explanation of Treatment (3). However, in the aggre-

gating stage, we are interested in combining the fitting results *i.e.* combining the posterior distributions of the two taxa not combining the given datasets.

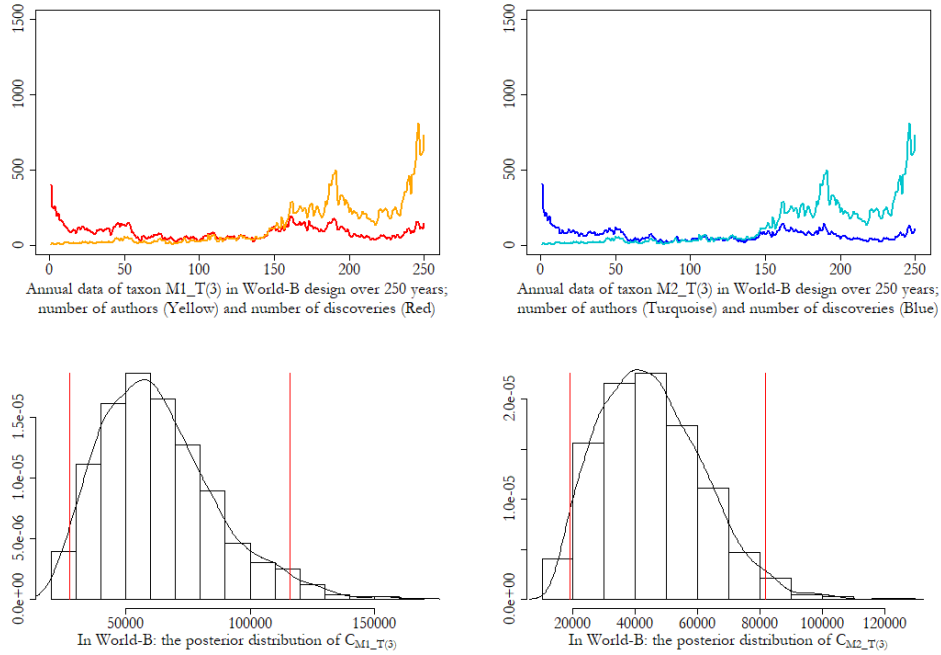


Figure 5.31: The generated data and the results of the fitting stage of the Merging Scheme - Treatment (3) of World-B design. The left panel belongs to taxon M1.T(3), while the right panel belongs to taxon M2.T(3).

Going to the final stage of the merging scheme - treatment (3), we also got very close results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.32. It seems that the central tendency measures are very close with posterior means $C^* = 109,793$ for the whole category, and $C_{M1.T(3)+M2.T(3)}^* = 108,551$ for the combined taxa. Compared with the previous treatments, it seems that the aggregated fitting resulted in estimated variation that is closer to the one of fitting the whole category, where the former is about 1.4 billion, while the latter is about 0.85 billion. This reflects the high closeness in the CDFs curves of the whole category and the aggregated taxa fitting (closer than the ones of the previous treatments in World-B), in the right bottom of Figure 5.32. We also computed the KLD metric in both directions and we found that $\text{KLD}(P(C|\text{rest}) | P(C_{M1.T(3)+M2.T(3)}|\text{rest})) = 0.002033$ and $\text{KLD}(P(C_{M1.T(3)+M2.T(3)}|\text{rest}) | P(C|\text{rest})) = 0.002003$ which means that the amount of the information lost when using the whole category distribution instead of the combined taxa distribution is approximately similar to the vice versa, and therefore, indicating the very closeness of these two distributions.

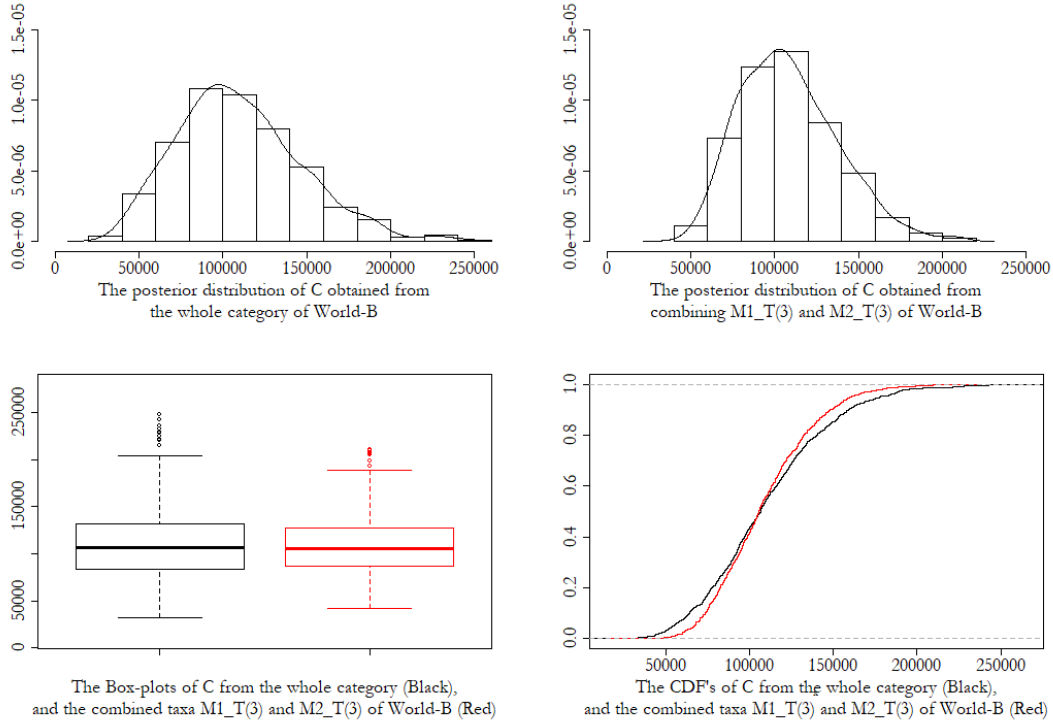


Figure 5.32: The aggregating stage of the Merging Scheme - Treatment (3) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-B.

5.3.3 Additivity in World-C Design:

In the Splitting Scheme, World-C is randomly split into taxon S1 and taxon S2, at which each taxon represents 50% of the whole category (World-C), so that $C_{S1} = C_{S2} = 25,000$. Figure 5.33 shows the number of discoveries and the number of authors of the two taxa after the splitting procedure. Again, we need to choose different parameters for the effort process in the two taxa such that the effort mean after the aggregation stage would be the same as the effort mean of the whole category. The effort mean of World-C $\simeq \exp(7 + (0.08)^2/2)_{whole} \simeq 1100$, and since we are splitting into equal portions taxa (50% & 50%), we need to specify values for μ_{L1} and σ_L only once for both taxon S1 and taxon S2 at which the effort mean = $\exp(\mu_{L1} + \sigma_L^2/2)_{S1} = \exp(\mu_{L1} + \sigma_L^2/2)_{S2} \simeq 50\% \times \exp(7 + (0.08)^2/2)_{whole} \simeq 550$. Again, the effort mean is just one equation in two variables, so the number of solutions are infinite. Thus, we need to consider additional constraints

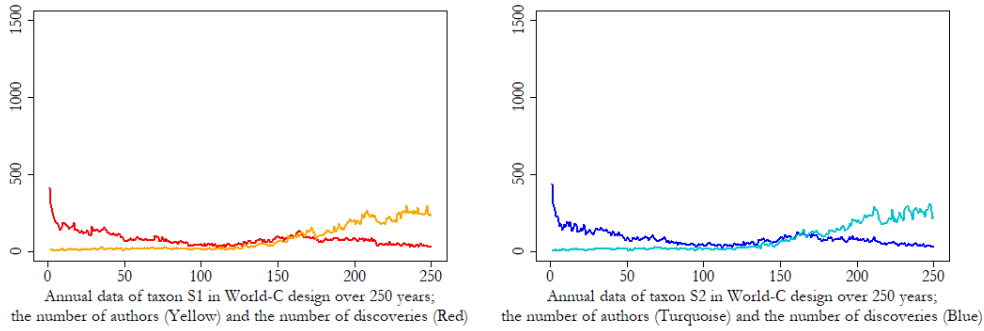


Figure 5.33: The resulting data of the Splitting Scheme of World-C.

which “approximately maintains the percentage of the discoveries”. In World-C, the percentage of the discoveries $\simeq 83\%$, and after the splitting the percentage of the discoveries $\simeq 82.6\%$ and 83.9% in taxon S1 and S2, respectively. Therefore, we need to specify values for μ_{L_1} and σ_L in both taxa such that we can closely maintain the above percentages, individually at each taxon and totally after aggregating the two taxa.

After several trials (that include proposing a set of parameters’ values, regenerating S1 and S2 under the candidate values, and verifying the effort means and the percentage of the discoveries), we found that the parametrization shown in Table 5.6 seems to be the best choice that approximately maintain the targeted effort mean and the percentage of the discoveries. In both taxa, the effort mean $\simeq \exp(6.31 + (0.078)^2/2) \simeq 551.7$ and the discovery percentage $\simeq 83.3\%$. Hence, in this selected parametrization, after aggregating the two taxa, the total effort mean $\simeq \exp(6.31 + (0.078)^2/2)_{S1} + \exp(6.31 + (0.078)^2/2)_{S2} \simeq 1103$ with total discovery percentage $\simeq (83.3\% + 83.3\%)/2 \simeq 83.3\%$ which are very close to the given ones (1100 and 83% , respectively) in the whole category.

Table 5.6: Parametrization of whole category vs. two taxa of World-C.

Hyper-Parameters	World-C	Taxon S1	Taxon S2
$\theta_N = \{\mu, \sigma\}$	{6,2}	{6,2}	{6,2}
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7,0.08}	{6.31,0.078}	{6.31,0.078}
$\theta_x = \alpha$	$\log(50)$	$\log(50)$	$\log(50)$

Figure 5.34 shows the fitting results (according to the above selected parametrization) of each taxon, which seem to be successful estimations at which the posterior

mean and mode are close to the true values in both taxa, more details are available in Group B.14, Appendix B.

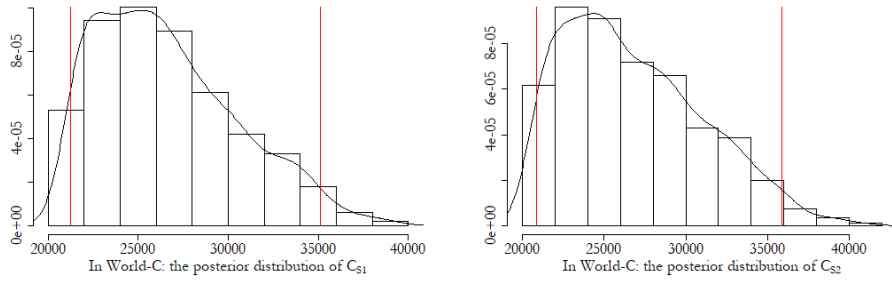


Figure 5.34: The fitting stage in the Splitting Scheme of the two taxa in World-C.

Going to the final stage of the splitting scheme, we got acceptable results, as shown in Figure 5.35. Although, we could not maintain the skewness in the aggregated fitting, the central tendency measures of the whole category and the aggregated fitted taxa are very close, at which the posterior means are $C_{whole}^*(mean) = 53,161$ and $C_{S1+S2}^*(mean) = 53,442$, while the posterior modes are

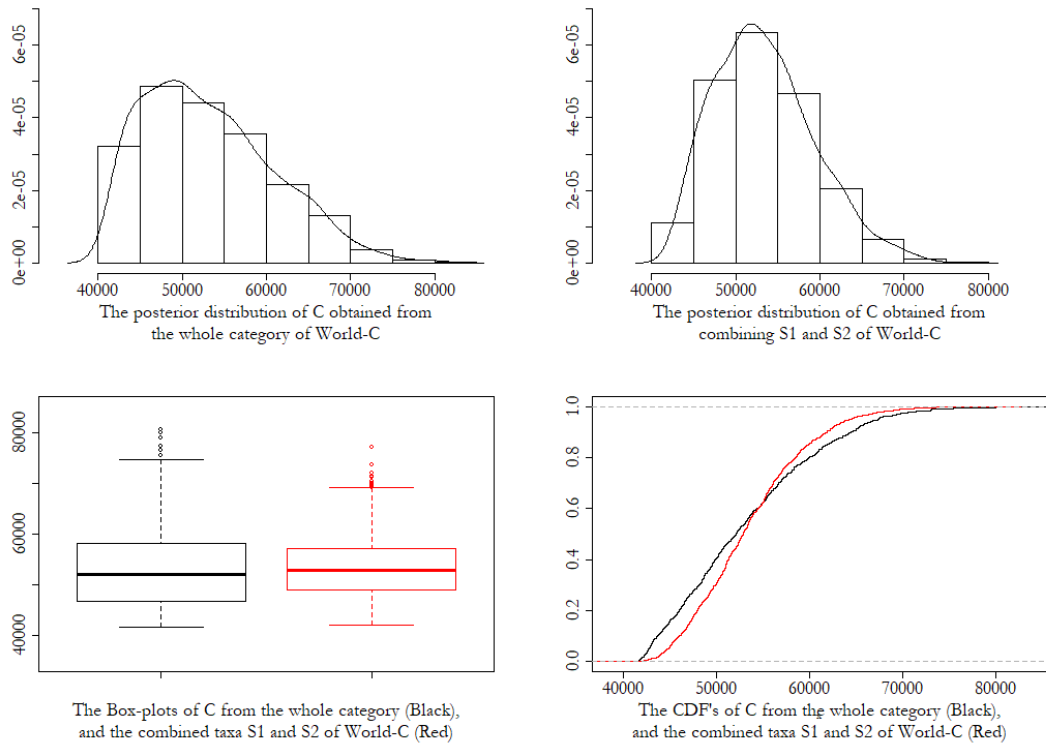


Figure 5.35: The aggregating stage of the Splitting Scheme showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.

$C_{whole}^*(mode) = 48,969$ and $C_{S1+S2}^*(mode) = 51,834$ for the whole category and the aggregated fitting, respectively. Again, the aggregation treatment includes more uncertainty due to fitting two distributions that are associated with two Monte Carlo errors instead of one, but it seems that the successful fitting of both resulted in less variational estimation (about 36 million) compared to the whole category fitting (about 59 million). The difference in the variation justifies the difference in the CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.35. With respect to the KLD metric, it is computed in both directions and we found that in one direction the $KLD(P(C|rest) | P(C_{S1+S2}|rest)) = 0.010169$, while in the other direction the $KLD(P(C_{S1+S2}|rest) | P(C|rest)) = 0.011392$ *i.e.* the amount of the information loss when using the whole category distribution instead of the combined taxa distribution is not that far to vice versa situation, which indicates a sort of closeness of these two distributions.

In the Merging Scheme - Treatment (1), we adopted the same idea of splitting the effort mean. Since we are adopting equal portions of taxa (50% & 50%), where the number of species in the two taxa are equal, their effort means are equal, and all the parametrization sets are equal, the generated two taxa will be identical. Therefore, we only need to generate one sub-artificial world to represent both taxon M1_T(1) and taxon M2_T(1). This taxon is generated according to the same parametrization of both S1 and S2 that is used in Table 5.6, where $C_{M1_T(1)} = C_{M2_T(1)} = 25,000$, and with the same seed that is used in generating the whole category (World-C), as shown in Figure 5.36. Note that since $M1_T(1) = M2_T(1)$, we will use one notation M_T(1) for both of them. By comparing the top plot with the bottom plot of the left panel in this figure, we can see that the differences are very small between the observed data resulting from aggregating two identical taxa *i.e.* $2 \times M_T(1)$ and the data of the original world (World-C), where the current number of discoveries of the original world is $D = 41,620$, while in the combined world is $D_{M_T(1)+M_T(1)} = 41,630$. In addition, they are very similar in the general pattern and they both share the same seed, same parameters of the sampling-discovery process, and almost the same effort mean, where $2 \times \exp(6.31 + (0.078)^2/2)_{M_T(1)} \simeq 1103$ and $\exp(7 + (0.08)^2/2)_{whole} \simeq 1100$.

The right bottom plot of Figure 5.36 shows the fitting results of $M.T(1)$ which seems to be successful estimation, see details in Group B.14, Appendix B.

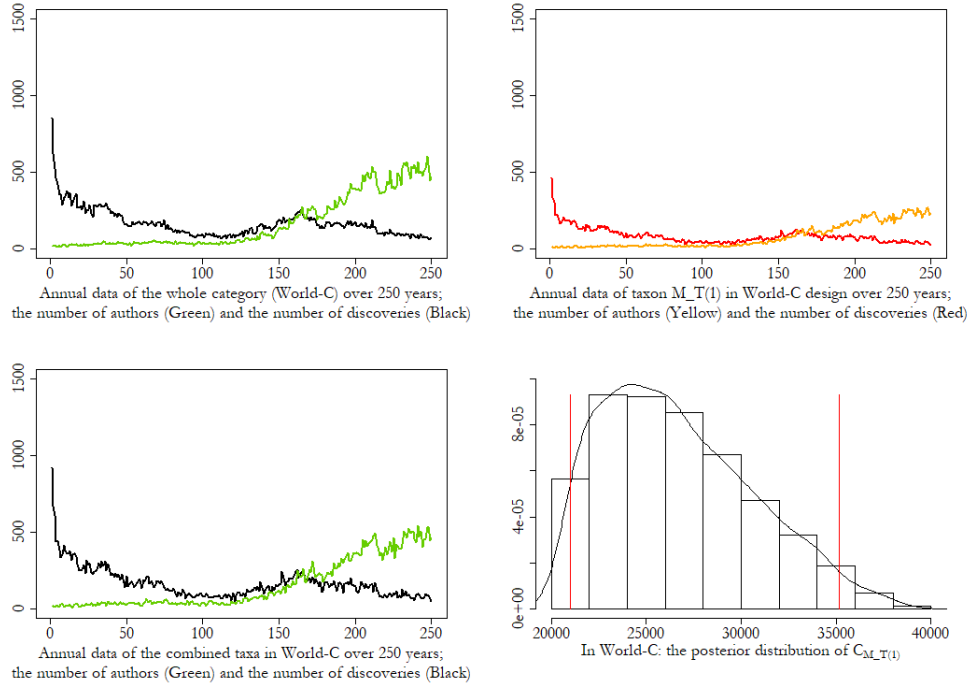


Figure 5.36: The resulting data of the Merging Scheme - Treatment (1) of World-C design.

Going to the final stage of the merging scheme - treatment (1), we got almost identical results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.37, at which the posterior means $C^* = 53,161$ for the whole category and $C_{M.T(1)+M.T(1)}^* = 53,376$ for the combined taxa. Unlike World-A and World-B, when we used two identical taxa, the aggregated fitting resulted in equivalent variation to the one of the whole category, which is a reflection of having equivalent CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.37. With respect to the KLD metric, it also reflects this high closeness of these two distributions, at which $\text{KLD}(P(C|\text{rest}) | P(C_{M.T(1)+M.T(1)}|\text{rest})) = 0.000758$ and $\text{KLD}(P(C_{M.T(1)+M.T(1)}|\text{rest}) | P(C|\text{rest})) = 0.000720$.

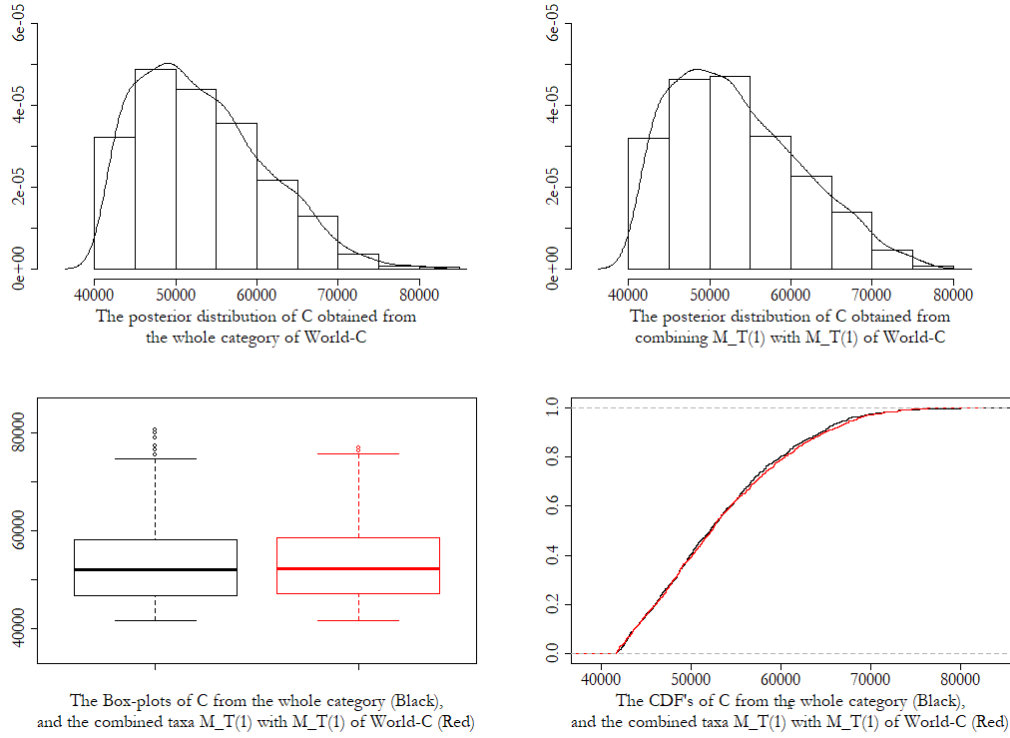


Figure 5.37: The aggregating stage of the Merging Scheme - Treatment (1) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.

In the Merging Scheme - Treatment (2), we also adopted the same idea of generating two identical sub-artificial worlds to represent taxon M1_T(2) and taxon M2_T(2). For simplicity, we will use one notation, M_T(2), to express both of them since they are identical. This sub-artificial world is independently generated from the same design as World-C such that $C_{M_T(2)} = 25,000$. It has the exact parametrization of World-C, but with different seed. After multiple trials with a variety of seeds, we managed to create a world that is close enough to represent a subgroup that occupies 50% of the whole category (World-C). Taxon M_T(2) is of seed=209523980, while the whole category is of seed=209522386, as shown in Figure 5.38. By comparing the top plot with the bottom plot of the left panel in this figure, we can see that the observed data resulting from aggregating the two identical taxa is different in details from the data of the original world (World-C), where the current number of discoveries of the original world is $D = 41,620$, while in the combined world is $D_{M_T(2)+M_T(2)} = 41,904$. However, they are similar in the general pattern and they both are considered as samples from the same underlying process. The right bottom plot of Figure 5.38 shows

the fitting results of $M.T(2)$ which seems to be successful estimation, more details are available in Group B.14, Appendix B.

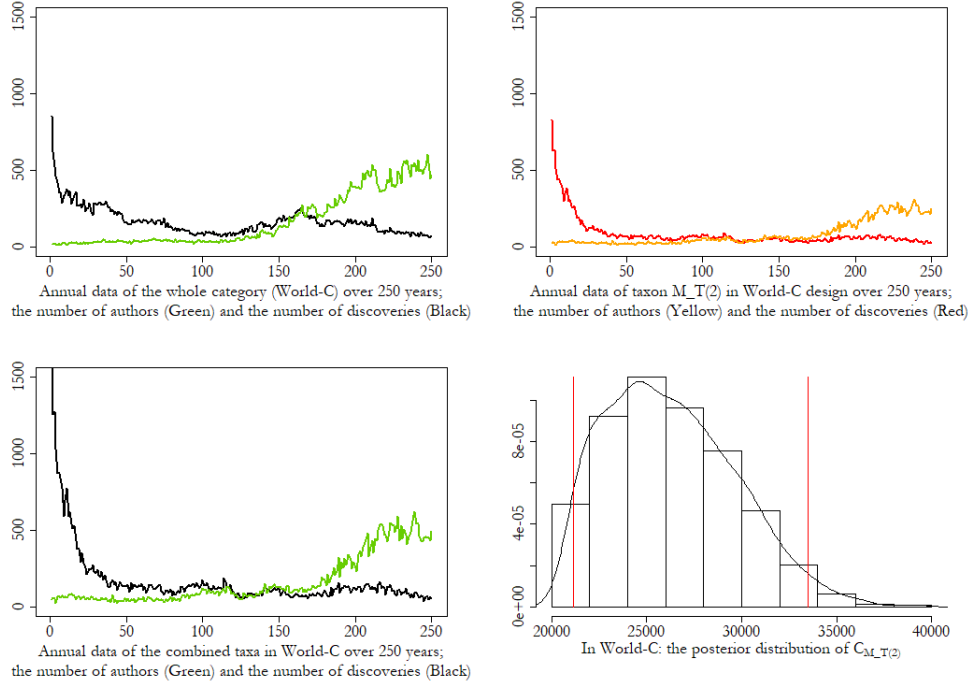


Figure 5.38: The resulting data of the Merging Scheme - Treatment (2) of World-C design.

Going to the final stage of the merging scheme - treatment (2), we also got almost identical results for the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.39, at which the posterior means $C^* = 53,161$ for the whole category and $C_{M.T(2)+M.T(2)}^* = 52,594$ for the combined taxa. Again, when we used two identical taxa, the aggregated fitting resulted in equivalent variation to the one of the whole category, which reflects having equivalent CDFs curves of the whole category and the aggregated taxa fitting, in the right bottom of Figure 5.39. With respect to the KLD metric, it also reflects this high closeness of these two distributions, at which $\text{KLD}(P(C|\text{rest}) | P(C_{M.T(2)+M.T(2)}|\text{rest})) = 0.002563$ and $\text{KLD}(P(C_{M.T(2)+M.T(2)}|\text{rest}) | P(C|\text{rest})) = 0.002525$.

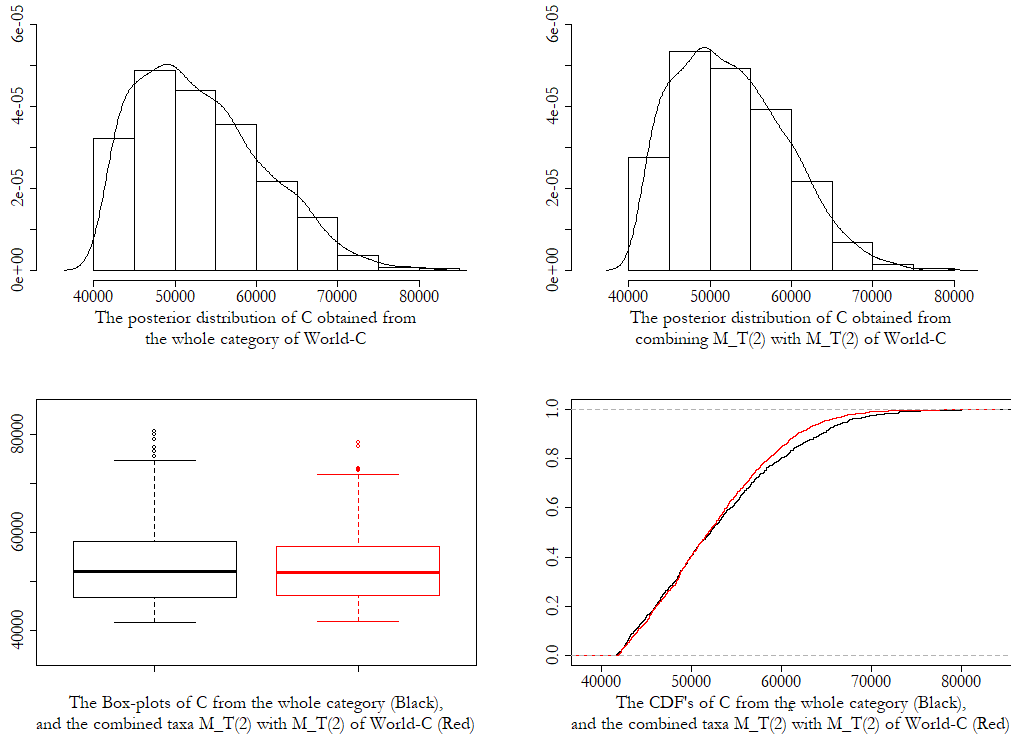


Figure 5.39: The aggregating stage of the Merging Scheme - Treatment (2) showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.

In the Merging Scheme - Treatment (3), two identical sub-artificial worlds are created to represent taxon $M1_{T(3)}$ and taxon $M2_{T(3)}$, and also we used a single notation, $M_{T(3)}$, to express both. The taxon $M_{T(3)}$ is independently generated from the same design as World-C and has the exact parametrization of World-C with the same seed=209522386, such that $C_{M_{T(3)}} = 25,000$. Figure 5.40 shows the generated number of discoveries and the number of authors of the

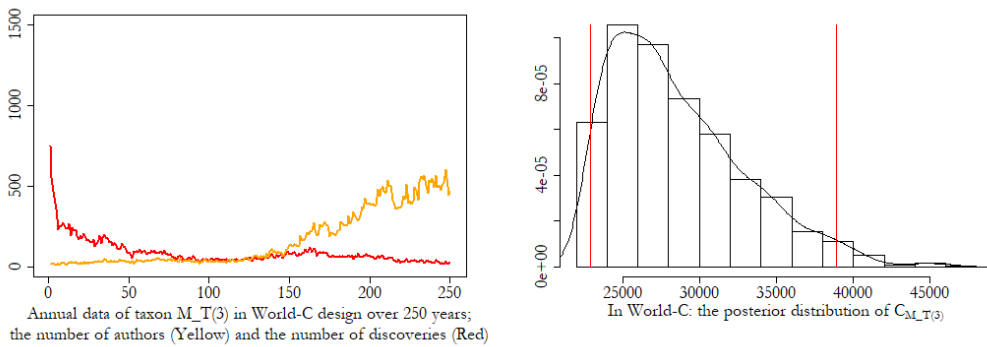


Figure 5.40: The generated data and the fitting results of the Merging Scheme - Treatment (3) of World-C.

taxon $M_T(3)$, as well as the fitting results which seems to be successful estimation, see more details in Group B.14, Appendix B. As we mentioned before, in the current treatment we are interested in combining the fitting results *i.e.* combining the posterior distributions of the two taxa not combining the given datasets.

Going to the final stage of the merging scheme - treatment (3), we got close (but a sort of truncated) results between the fitted whole category and the aggregated fitted taxa, as shown in Figure 5.41. The interpretation of this truncated pattern is that while the original artificial world (World-C) involves a high percentage of discoveries 83.24% at which $D = 41,620$ out of number of species $C = 50,000$, generating 50% subgroup from this world (under treatment (3)) resulted in a much higher number of discoveries 90.5% at which $D = 22,617$ out of a number of species $C = 25,000$. Therefore, the skewness in the subgroup is much higher than it is in the original world, where the concentration around the actual value of the number of species is very close to the left-end of the distribution. Consequently, in the aggregation stage, the combined number of discoveries $D_{M_T(3)+M_T(3)} = 45,234$ is much higher than the one of the original

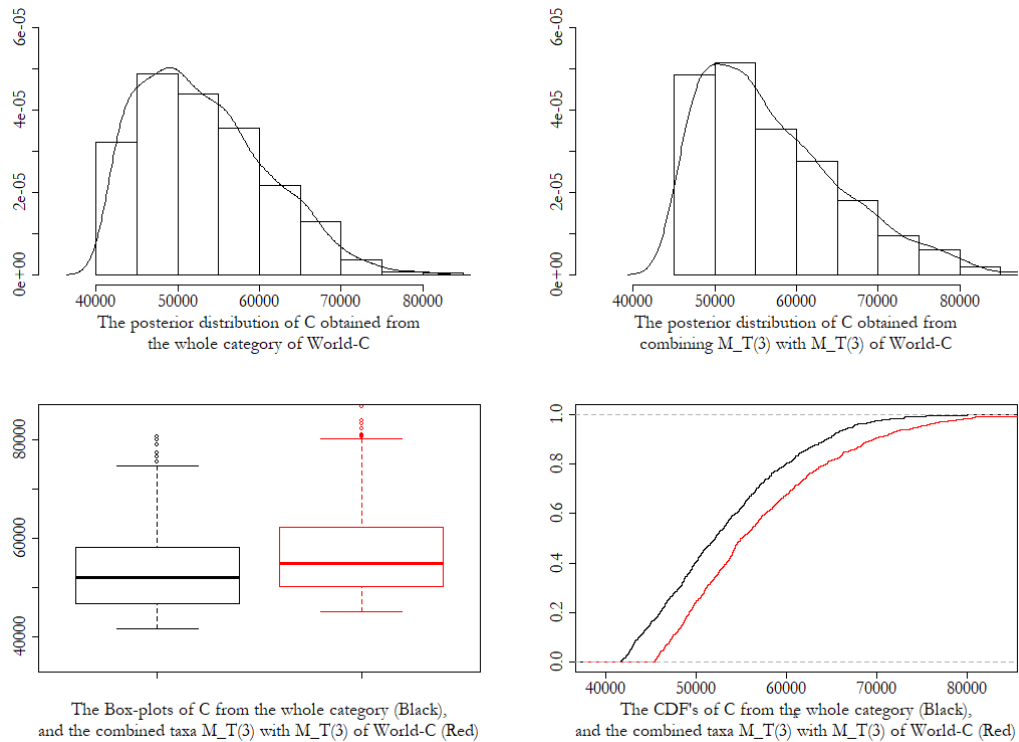


Figure 5.41: The aggregating stage of the Merging Scheme - Treatment (3), first trial, showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.

world $D = 41,620$, and is very close to the true value $C = 50,000$, which justifies the missing of the first bar in the left-end of the combined distribution. In other words, this bar covers the range $[40,000 - 45,000]$, while the support of combined distribution starts beyond this range from $D_{M.T(3)+M.T(3)} = 45,234$. This reflects getting almost identical (but shifted) CDFs curves and box-plots of the whole category and the aggregated taxa fitting, in the bottom panel of Figure 5.41.

Despite of the truncated and shifted patterns, it seems that the central tendency measures are still correctly informative, indicating very close results for the whole category (*e.g.* posterior means $C^* = 53,161$) and the combined taxa (*e.g.* posterior means $C_{M.T(3)+M.T(3)}^* = 57,159$). We also computed the KLD metric in both directions and we found that $\text{KLD}(P(C|\text{rest}) | P(C_{M.T(3)+M.T(3)}|\text{rest})) = 0.005493$ and $\text{KLD}(P(C_{M.T(3)+M.T(3)}|\text{rest}) | P(C|\text{rest})) = 0.005747$ which means that the amount of the information loss when using the whole category distribution instead of the combined taxa distribution is approximately close to vice versa, and therefore, indicates the closeness of these two distributions, where the shape of the CDFs are very close except that the aggregated one is shifted due the support issue that explained above.

To remedy the issue due to skewness, an attempt is made by re-fitting taxon M.T(3) but with wider range starting from 21,000 instead of 22,617 so that in the aggregation stage, we get a distribution that has a support starting from 42,000 instead of 45,234 to be able to include the missing bar in the above distribution, as shown in Figure 5.42. Although, we got tidier shape of the combined distribution, enlarging the range of the fitting resulted in having higher variation (about 89 million) in the combined distribution than the one in the original distribution (about 59 million). This reflects having the differences in CDFs curves and box-plots between the whole category and the aggregated taxa fitting, in the bottom panel of Figure 5.42. However, it seems that the central tendency measures of the combined fitting are still very close to the whole category with posterior means $C_{M1.T(3)+M2.T(3)}^* = 55,385$, and the KLD metric still reflects a closeness in these two distributions, where $\text{KLD}(P(C|\text{rest}) | P(C_{M1.T(3)+M2.T(3)}|\text{rest})) = 0.002093$ and $\text{KLD}(P(C_{M1.T(3)+M2.T(3)}|\text{rest}) | P(C|\text{rest})) = 0.002150$.

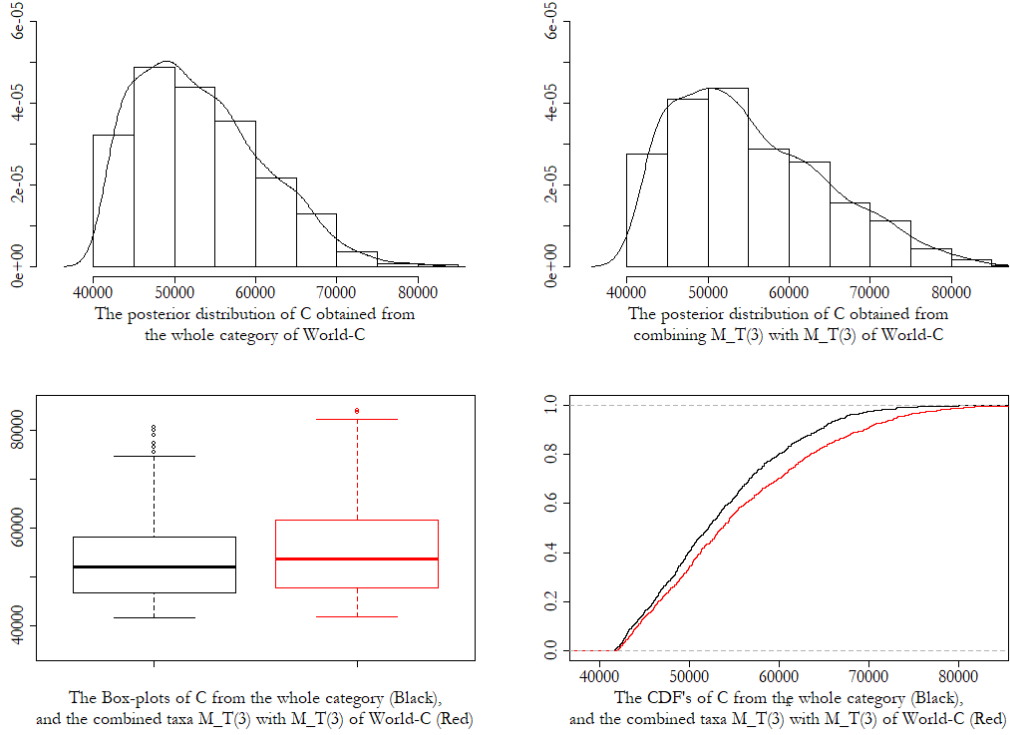


Figure 5.42: The aggregating stage of the Merging Scheme - Treatment (3), second trial, showing the final results of fitting the whole category and the combined fitting of the two taxa in World-C.

5.3.4 Additional Attempts in World-A & World-B:

Since measuring the model additivity is a challenging task in light of the complexity and high dimensionality of our proposed model, we tried to seek multiple ways to accomplish this task comprehensively. Throughout Section 5.3, we already adopted two main schemes with three treatments in one of them. However, we saw a noticeable performance in World-C when we adopted the idea of creating 50% & 50% two identical sup-groups in the merging scheme - treatment (1), which made us curious about using the same procedure in World-A and World-B.

Therefore, in World-A, we generated two identical taxa at which $C_{50\%A} = 5000$ for each, using the same seed and the same parametrization set of the whole category except for the effort mean where we used only the half *i.e.* $\{\mu_{L_1} = 4.31, \sigma_L = 0.092\}$. In the fitting stage, we also got almost identical results between the aggregated posterior distribution and the whole category posterior distribution, as shown in Figure 5.43. Again, the central tendency measures of

the whole category and the aggregated fitted taxa are very close, at which the posterior means are $C_{whole}^*(mean) = 10,320$ and $C_{50\%A}^*(mean) = 10,269$, while the posterior modes are $C_{whole}^*(mode) = 9,970$ and $C_{50\%A}^*(mode) = 9,994$ for the whole category and the aggregated fitting, respectively. With respect to the KLD metric, it also reflects this high similarity of these two distributions, at which $KLD(P(C|rest)|P(C_{50\%A+50\%A}|rest)) = 0.0011372$ and $KLD(P(C_{50\%A+50\%A}|rest)|P(C|rest)) = 0.0011375$.

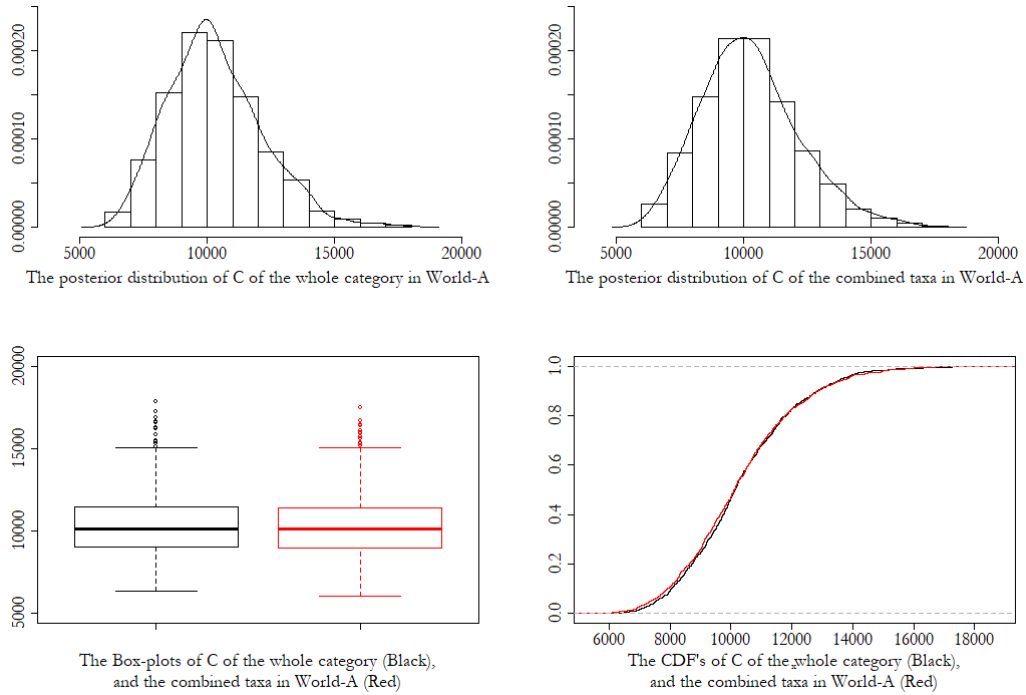


Figure 5.43: The aggregating stage of the Merging Scheme - Treatment (1) with 50% & 50% taxa of World-A, showing the final results of fitting the whole category and the combined fitting of the two taxa.

In World-B, we also generated two identical taxa at which $C_{50\%B} = 50,000$ for each, using the same seed and the same parametrization set of the whole category except for the effort mean where we used only the half *i.e.* $\{\mu_{L_1} = 6.31, \sigma_L = 0.143\}$. In the fitting stage, we got almost identical results between the aggregated posterior distribution and the whole category posterior distribution, as shown in Figure 5.44. The central tendency measures of the whole category and the aggregated fitted taxa are very close, at which the posterior means are $C_{whole}^*(mean) = 109,793$ and $C_{50\%B}^*(mean) = 110,030$, while the posterior modes are $C_{whole}^*(mode) = 97,160$ and $C_{50\%B}^*(mode) = 101,192$ for the

whole category and the aggregated fitting, respectively. With respect to the KLD metric, it also reflects the very closeness of these two distributions, at which $\text{KLD}(P(C|\text{rest}) | P(C_{50\%B+50\%B}|\text{rest})) = 0.002610$ and $\text{KLD}(P(C_{50\%B+50\%B}|\text{rest}) | P(C|\text{rest})) = 0.002677$.

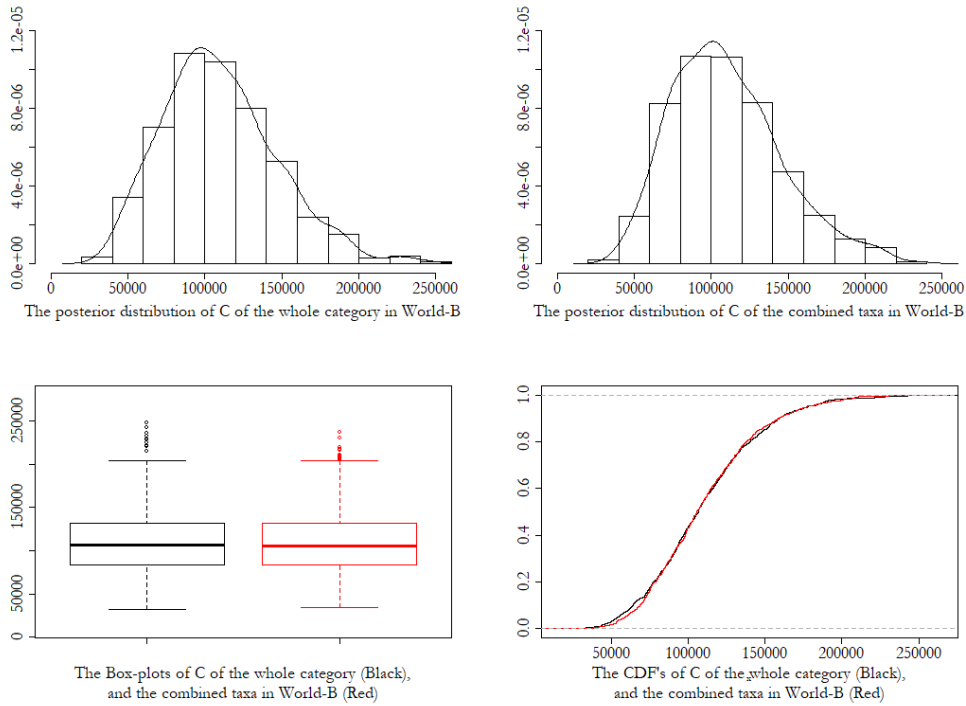


Figure 5.44: The aggregating stage of the Merging Scheme - Treatment (1) with 50% & 50% taxa of World-B, showing the final results of fitting the whole category and the combined fitting of the two taxa.

5.4 Chapter Summary:

This chapter involves three main measurements (accuracy, sensitivity, and additivity) in order to describe and evaluate our proposed model. Three artificial worlds (World-A, World-B, and World-C) are used as subjects of the measuring experiments to capture a wide range of scenarios in the model performance. In most of these experiments, the behaviour of the model is generally similar over the three artificial worlds. However, the similarity is higher between modelling World-A and World-B, while modelling World-C is slightly distinguished in the performance due to its skewness that is associated with high proportion of discoveries.

In the accuracy measure, we explored two aspects of the MC error. The first one is related to the selected final sample size used for estimating the number of species, while the second aspect is related to the number of iterations used in the fitting stage of modelling the number of species from which we obtain our estimation. Through the experiments, we found that all the three artificial worlds share a similar pattern in which that the estimation seems to be more consistent with increasing the sample size. However, a slightly biased estimation (overestimation) occurred with a sample size greater than 1000. It also appears that no significant enhancement can be made on this estimation beyond the number of iterations =1,000,000. Therefore, choosing the number of iterations to be 1,000,000 in the fitting stage with selecting a final sample of size $s = 1000$ seems to be a balanced trade-off to achieve a relatively less MC error with a relatively less biased estimation. See Appendix B, Section B.2 (Groups B.2 to Groups B.5) for more details and a comprehensive view.

In the sensitivity measure, we explored the behaviour of the estimation against the mis-parametrization, where we specified 20% over/under the actual values of the given five parameters, and tested each one at a time with fixing the rest. Through the experiments, we found that World-C is less sensitive against the mis-parametrization overall, while World-A and World-B share a similar pattern of being highly sensitive to the mis-parametrization of two parameters. These two parameters are the variation of the species abundances, and the expected number of specimens to be collected per author per year (on a log-scale). Therefore, we should be careful in choosing the values of these two in such a situation that is similar to World-A and World-B. However, although the sensitivity is generally less in World-C, the existence of the skewness caused relative deviations in the sensitivity of the log-scale parameters of the latent effort to be slightly higher than their case in World-A and World-B. Thus, caution is also needed in selecting the values of these parameters when having skewness in the distribution. See Appendix B, Section B.3 (Groups B.6 and Groups B.7) for more details and a comprehensive view.

In the additivity measure, we adopted two main schemes (splitting and merging), and the merging itself is branches into three treatments. We tried three different proportion-divisions in creating sub-groups out of the design of the given

three artificial worlds. Through the analysis we focused on the results of three main statistical tools (the summary statistics especially the central tendency ones, the CDFs curves and box-plots which also reflect the deviation measure, and the KLD metric). Focusing on the central tendency statistics, we found that in all the three artificial worlds, all the posterior means and modes values that are obtained from the aggregation step are very close to the true values of the number of species and to the estimated ones out of the whole category fitting. The derived values of KLD metric also provided similar implications; a closeness between the distributions out of the aggregation and the whole category distributions. With respect to the CDFs curves and box-plots, we found a variation in the results due to the variation in the standard deviations between the aggregated distributions and whole category distributions. However, by checking the 95% credible intervals of all these attempts (as shown in Group B.10, Group B.13, and Group B.15 in Appendix B, Section B.4), we found that the limits of the credible intervals resulting from the aggregation stage are either close or within the credible limits of the intervals resulting from the whole category fitting.

World-A and World-B are similar in response to the splitting and merging schemes, while World-C is distinguished in response due to its skewness. Therefore, in light of the complexity and high dimensionality of our proposed model and based on the current attempts, it might not be an easy statement to say that our model is additive, but our approach definitely provides alternative aggregated estimation that is equivalent to the whole category estimation and within 95% credible interval of the whole category estimation. However, this is not guaranteed without the need for some manual tuning to achieve this result especially in a real data scenarios (where there might be other hidden factors influencing the aggregation or estimation). See Appendix B, Section B.4 (Groups B.8 to Groups B.15) for more details and a comprehensive view.

Chapter 6

Model Application on Real Data

Heretofore, our proposed model has been analysed, characterized, implemented, and validated through artificial data in Chapter 4 and Chapter 5. In the current chapter, we will continue this process but through real data. In particular, datasets that are retrieved from two well known inventory systems of biodiversity: the Catalogue of Life (CoL) [85] and the World Register of Marine Species (WoRMS) [86]. The WoRMS contains contributions from 119 overlapped databases, while the CoL contains contributions from 164 distinct taxonomic databases, some of them are included in the WoRMS. Each of these inventory systems involves several attributes to describe species, however, the main ones are: classification, species name, authority, publication year, and habitat. The number of species descriptions has been growing since 1735 (the beginning of the Linnaean taxonomy), in WoRMS, it reached about 460,722 marine species names, including synonyms from which 235,608 are accepted as unique species. In CoL, it reached about 1,829,341 living and 38,145 extinct species, including common and synonyms¹.

However, two refined non-overlapping subsets from each WoRMS and CoL are currently available to be a subject of experimentation for our proposed model. It comprises 426,815 unique (no synonyms) terrestrial species from CoL databases including freshwater species, and 173,401 unique marine species from WoRMS databases. These species were discovered and for first-time described over 240 years, in the period between 1761 and 2000. The annual numbers of authors who were involved in these discoveries are also included in these subsets, where the total number of authors is 505,372 in the CoL, while it is 62,285 in the WoRMS, as shown in Figure 6.1. In the following sections, and based on these

¹Data websites are accessed on 2 Sep. 2020.

given datasets, our proposed model is applied to estimate the number of terrestrial species (including freshwater) and the number of marine species, in sections 6.1 and 6.4, respectively. Then, in Section 6.5, we seek for a global estimation by aggregating the above two estimates. This is followed by making comparisons with some previous work in Section 6.6. However, the CoL dataset is taken in particular as an example for a chronological analysis and validation of our model in Section 6.2 and Section 6.3, respectively.

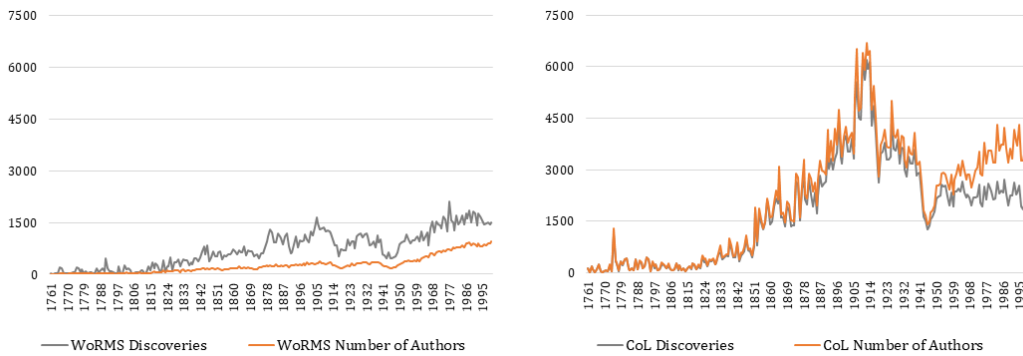


Figure 6.1: Two refined non-overlapping subsets from WoRMS (left) and CoL (right) inventory systems, over 240 years from 1761 to 2000, showing a big difference in the discovery and effort scales between the two subsets.

6.1 Model Fitting for CoL Data

In this section, we apply our model on the CoL dataset, at which we followed the same distributional specifications that we have detailed in Section 4.1 and shown in Figure 4.2. Although these specifications are justified by the species literature, expert opinion, and statistical assumptions, we faced some challenges in determining some of them such as the size of the domain (determined by the upper bound of the support, while the lower bound is always determined by the current number of discoveries D) of the prior uniform distribution for the number of species and the values of the model parameters. We also faced challenges in operating the model fitting itself due to computing-wise issues associated with the parallelization scheme. These two types of challenges are addressed in the following subsections.

6.1.1 Challenges in Model specifications

Since we adopted a simple case of dealing with fixed-value parametrization in the current implementation of our model, we need to elicit five values for the model parameters, and a value for the size of the prior support. In most cases, it is not a guarantee to have certain and complete information about these distributional specifications, thus we started with initial values and operated several pilot experiments for changing or tuning the selected values. As we learned from the sensitivity analysis in Section 5.2.1, our model is more sensitive to the variation of the species abundances σ_N and the expected number of specimens to be collected per author per year $\exp(\alpha)$. Therefore, we tried to carefully specify those two parameters by referring to the species literature and the field expert guidance².

However, it is tricky to summarize $\exp(\alpha)$ in just one fixed value in order to capture the annual individual taxonomic productivity over 240 years for a large inventory system such as CoL. Although [1] optimistically stated that a full-time author might examine a few thousand species in a year, this is not representative in our situation if we consider at least two main factors: first, the differing circumstances affecting the number of authors and their taxonomic productivity over a very long period such as 240 years (some of them are prolific and full-time, while others are not during such a period); the second factor is the broad variety of species that are included in the CoL (ranging from minuscule insects to gigantic animals or plants). Therefore, in the light of these two broad factors, we could say that $\exp(\alpha)$ might be ranging from few hundred to few thousand species, as a start we choose $\exp(\alpha) = 100$.

With respect to the species abundance variation, [87] found that on a log scale (base 2), species abundances followed a normal distribution with a standard deviation of approximately 3.5³. On a natural logarithm scale which we are currently using, this value transforms to approximately 2.45 (details about the transformation is available in Appendix A, Section A.3). Although [87] is an old

²via personal communication with M.J. Costello.

³In [87], a slightly different terminology was used to represent the variation and it is referred to by a parameter called ‘a’ that typically observed to be of value close to 0.2.

literature and the environmental circumstances probably developed over time, we relied on its suggestion as a start and choose $\sigma_N = 2.45$.

Regarding the other three parameters of our model, we relied on our intuition in noticing some similarity aspects between the given data of CoL and the ones of the artificial World-B and World-C. Thus we elicited values for those parameters that are close to the ones of the artificial World-B and World-C, so that $\mu_N = 10$, $\mu_{L_1} = 7$, and $\sigma_L = 0.2$. We started with prior support at which $C \in [D, 30D]$ and applied a small number of iterations ($iter = 10,000$). During the pilot experiments, we were monitoring the general pattern and scale of the change in the ABC generated datasets along with the given ones for changing/tuning the elicited parameter-values. However, after many trials between failed and close estimation (some samples of these results are available in Appendix B, Group B.17 - Group B.19), we changed some of our first parameter choices and end up with the best elicitation of parametrization for modelling the CoL dataset, as shown later in Section 6.1.3.

6.1.2 Challenges in Model Fitting

In the model fitting, we followed the same ABC algorithm that is detailed in Section 3.2.3, and implemented throughout Chapter 4 and Chapter 5. However, due to the large scale of occurrence in the given data of the CoL (*i.e.* large dataset in which the values of the number of discoveries and number of authors are large values, which are located on a large scale of the y-axis on Figure 6.1), we faced serious memory-consuming problem in the parallel computation, which caused task termination even in small number of iterations such as 10,000. A Windows machine with a six-core processor is robust against such issue, but we faced limitations in finding Windows-computing resources. Most of available resources are in Linux system which tends to terminate the coding task encountering the memory consumption. Developing the code to include error detection and track correction for such issues was not enough, because one of the errors cannot be technically tolerated in the Linux environment. This error mainly associated with one source, the generated linkage node (the sizes of the samples $\{n_j\}_{j=1:T}$), in particular, associated with their sum which became very large in the current application so that

when it was used later to generate $\{\tau_j\}_{j=1:T}$, it consumed a huge size of memory.

Therefore, to remedy this situation, we made two alterations to our adopted ABC algorithm: first, since the rate parameter became large in the current application, we used the Normal approximation of the Poisson process for generating n_j 's instead of just Poisson. Second, we included additional threshold to filter the sum of the generated n_j 's in the main experiment, based on extracted information from the pilot experiment. We followed the exact ABC algorithm of the pilot experiment (Section 3.2.3), but with additional 8th step in which we employed the given data of CoL to create an acceptable interval for $\sum_{j=1}^T n_j$ which involves values that are small enough to consumes less computational memory, but large enough to capture all the variety of patterns that represent the given data and so that guarantees the same accuracy of the final estimation. Thus, after obtaining the final acceptable 1000 samples, we compute $\sum_{j=1}^T x_j$ for each of the 1000 generated number of authors. Then, we compute the standard deviation of these sums *i.e.* $sd(\{\sum_{j=1}^T x_j\}_{1:1000})$. Finally, the acceptable interval of $\sum_{j=1}^T n_j$ of the given CoL dataset can be estimated as follows:

$$\alpha \left(\sum_{j=1}^T x_j \right)_{\text{CoL}} \pm 3\alpha \text{sd}(\{ \sum_{j=1}^T x_j \}_{1:1000}) \quad (6.1)$$

where, on estimation $(\sum_{j=1}^T n_j)_{\text{CoL}} = \alpha (\sum_{j=1}^T x_j)_{\text{CoL}}$, and $(\sum_{j=1}^T x_j)_{\text{CoL}}$ is the given sum of the number of authors in the CoL dataset, while $\{\sum_{j=1}^T x_j\}_{1:1000}$ is a set of sums in which each sum belongs to each generated series of the number of authors (in the final acceptable 1000 samples). The limits of the above interval represent our new threshold that is employed in the 7th step of the ABC algorithm of the main experiment (Section 3.2.3). We apply the same step but with different order, as here, we first start with generating \mathbf{n} from \mathbf{l} , then accept $sum(\mathbf{n})$ and proceed to the next step of generating \mathbf{N} , \mathbf{d} , and $\boldsymbol{\tau}^*$, if the $sum(\mathbf{n})$ is within the above limits in (formula 6.1). Reject otherwise, and go back for new generation, see Algorithm 2 of Section C.2 in Appendix C. It is worth mentioning that this new threshold is implemented on the three artificial worlds (World-A, World-B, and World-C from Chapter 4), and shown similar results to ones in Chapter 4, see Group B.16 in Appendix B, Section B.5.

6.1.3 Fitting Results of CoL

Table 6.1 shows best elicited parameters of modelling the CoL dataset, as well as the fitting results. After operating the pilot experiment and deriving the thresholds, we implemented intermediate main experiment on just 100,000 iterations to make sure that we are on the right track of our distributional specifications, see the fitting details and plots of the pilot and the intermediate experiments in Appendix B, Group B.20. This is followed by the final main experiment of 1,000,000 iterations, from which we estimated the number of terrestrial species (including freshwater).

Table 6.1: Parametrization and resulting statistics of modelling CoL dataset.

Model Parameterization of CoL		Fitting Results of CoL	
$C \in [D, 200D]$	$C \in [426,815 - 85,363,000]$	Posterior Mean	24,775,806
$\theta_N = \{\mu, \sigma\}$	{10, 4.7}	Posterior Mode	19,832,205
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{10.6, 0.15}	95% Credible Interval	[7,554,329 - 63,884,882]
$\theta_x = \alpha$	$\log(100)$	Posterior SD	13,358,331

Figure 6.2, right top plot, shows that the total number of terrestrial species (including freshwater) follows a skewed distribution, with a 95% probability interval ranging from about 7.6 to 64 million or 5 to 52 million species, according to the central (red lines) or 95% HPD interval (blue lines), respectively. The highest concentration of posterior mass is between 10 and 30 million species. If we adopted the posterior mode as point estimation for the total number of terrestrial species, we can say that about 19,405,390 species are yet to be discovered, which is about 45 times the current discoveries. In the bottom panel of Figure 6.2, we can see that the acceptable dark grey area (the generated number of authors and the number of the discoveries) is a good representative of the given CoL dataset. It captures the annual fluctuating behaviour of the given data at which a peak occurs around the 150th year (1910), and a trough around the 185th year (1945), followed by a recovery pattern.

If we track the fitting performance over the three number of iterations ($iter = 10,000$, $iter = 100,000$, and $iter = 1,000,000$) by comparing Figure 6.2 with the figures in Group B.20 of Section B.5 in Appendix B, we can see the

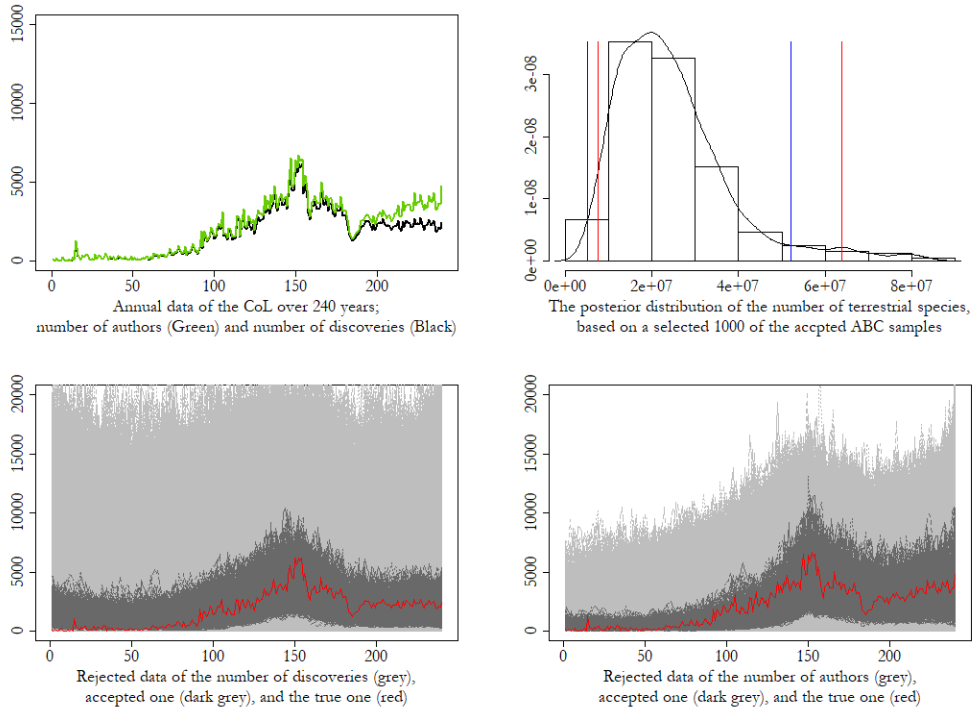


Figure 6.2: Fitting results of CoL dataset.

enhancement of the fitting associated with increasing the number of iterations; the shape of the posterior distribution becomes more tied with lighter tails. In addition, by observing the plots of the bottom panels, we can see that the generated data in the acceptable dark grey area are getting closer with increasing the number of iterations and better capturing the behaviour of the given CoL data.

6.2 Model Chronological Analysis

Since the type of our given data is a time series, and as time brings us two aspects of the information (accumulated knowledge and varying pattern), it is a subject of interest to measure to what extent our model is influenced by the time factor. In this section, we track our model performance (on CoL dataset) over time to check how our model is affected by the accumulated information over time and its varying pattern. As we noticed in the CoL dataset, there are two obvious turns in its pattern over 240 years, one is a peak around the 150th year (1910), and the other is a trough around the 185th year (1945). Therefore, we truncated the data twice at these turns as the following:

- At the first turn, from the 240-year data points, we obtained the first 150-year data points, and re-fitted this part of data pretending that we do not know the last 90-year data points.
- At the second turn, from the 240-year data points, we obtained the first 185-year data points, and re-fitted this part of data pretending that we do not know the last 55-year data points.

After re-fitting the truncated data, we confronted their fitting results with the fitting result of the whole data points over 240 years, as shown in Table 6.2 and Figure 6.3. In general, looking at the top right plot of Figure 6.3, we can see that in all the three time periods, the estimated posterior distributions of the number of terrestrial species are positively skewed, with peaks concentrated in the interval between 10 million and 30 million. However, considering the detailed behaviour of the estimation in the three time periods, there are some specific patterns addressed as follows:

Table 6.2: Statistics of fitting results of CoL dataset over the first 150 years, the first 185 years, and the whole period which is 240 years.

Fitting Results of CoL	Over first 150 years	Over first 185 years	Over whole 240 years
$C \in [D, \text{same upper limit}]$	[180,032 – 85,363,000]	[303,939 – 85,363,000]	[426,815 – 85,363,000]
Posterior Mean	17,828,147	23,066,547	24,775,806
Posterior Mode	13,139,901	17,195,969	19,832,205
95% Credible Interval	[4,973,074 – 57,632,151]	[7,265,953 – 60,105,763]	[7,554,329 – 63,884,882]
Posterior SD	12,848,302	13,066,394	13,358,331

1. With including more data over time, the pattern of the estimated distributions are moderately shifting to the right side over the number of species domain, as shown in the top right plot of Figure 6.3. This is justified by the fact that the prior domain (support) of the distribution starts from the current number of discoveries, D , which is different over these three time periods, where $D_{150} < D_{185} < D_{240}$, as shown in Table 6.2.
2. Considering the point estimation C^* , the ratio of the number of the undiscovered species to the number of the already discovered ones, $R = (C^* - D)/D$, is getting smaller over time and decreasing with almost the same rate, where

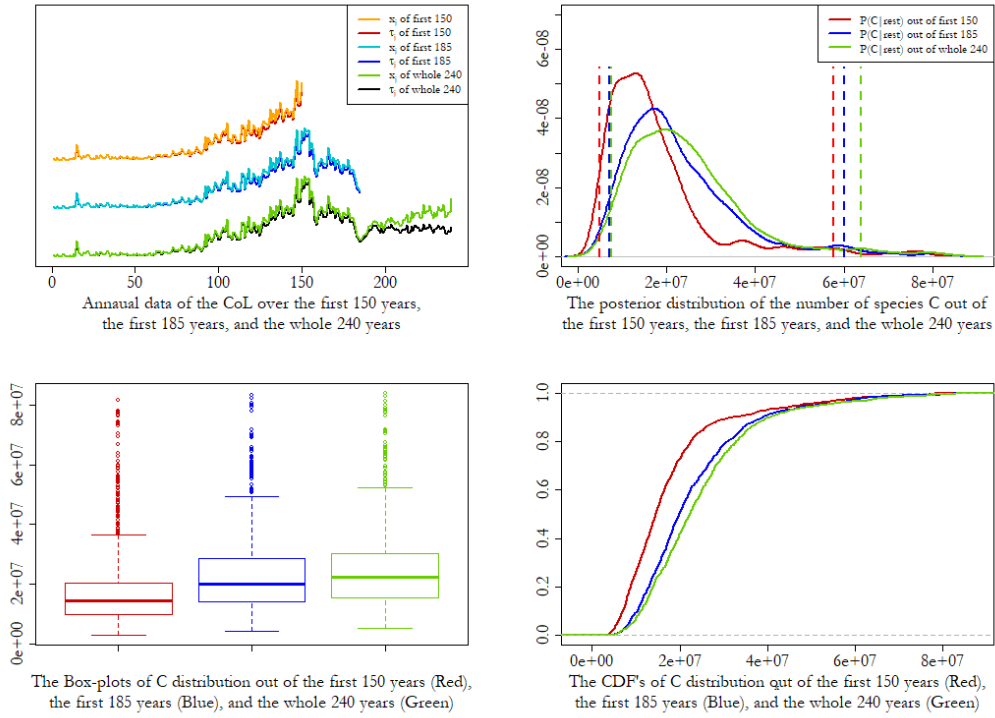


Figure 6.3: Plots of fitting results of CoL dataset over the first 150 years (Red), over the first 185 years (Blue), and over the whole period which is 240 years (Green).

$R_{150} = 98$, $R_{185} = 75$, and $R_{240} = 57$ if we considered the posterior mean as our target estimation, while $R_{150} = 72$, $R_{185} = 56$, and $R_{240} = 45$ if we considered the posterior mode as our target estimation.

3. The estimation of the 240-year period seems closer to the 185-year estimation, than the 150-year estimation, as shown in the bottom panel plots of Figure 6.3, and according to the summary statistics in Table 6.2. This is justified, since the knowledge difference in the former estimations is accounted for 55 years, while in the latter is accounted for 90 years.
4. Although in the 150-year period it seems that the number of discoveries is reaching a peak with increasing the number of authors (which may indicate that there are more species left to be discovered), the total estimation of the number of species out of this period, C_{150}^* , does not exceed the actual recent estimation C_{240}^* . On the other hand, having decreasing curves at the last pattern of the 185-year period did not conclude a less estimation for the number of species than C_{150}^* , instead with including more information over time, C_{185}^* is getting closer to C_{240}^* regardless of the trough occurrence. This might be an indication

that fitting our model using the ABC algorithm is getting more refined with the accumulated knowledge, but it is still robust against the big fluctuations in the varying pattern of the given data.

To sum up, in spite of the above specifications associated with time, the peak estimations of the 150-year and 185-year periods are still within the 95% credible interval of the 240-year time period. In addition, the prediction of these truncated periods, C_{150}^* and C_{185}^* , exceed the recent discoveries D_{240} , which meet our expectations that the prediction of a smaller previous period should at least reach the recent number of discoveries. All this indicate the consistency of our estimation over time.

6.3 Model Validation

In many cases, evaluating the model performance in the real-data context is not a straightforward task since we do not know the ground-truth. However, in our case it is beyond this due to the complexity and the high dimensionality of our model. Determining a ground-truth involves a validation itself which we will try to achieve in the next subsection. In addition, the conventional methods used to validate a model are not possible to be applied in our case. Therefore, we had to innovate a special way to validate our model, which is similar in principle to the residual diagnostic method of the regression analysis. However, we do not actually have residuals in the exact meaning, since a residual is a difference computed between a single-point observation and a single-point estimation, while in our case we are dealing with time series, where one time series is considered as one observation. Therefore, the Euclidean distance that we already used in our implementation of the ABC algorithm, which is computed between the observed time series and the estimated time series, is defined as a residual in our context to be employed for validating our model. In the current section we approach two point of views in validating our model, explained as follows:

6.3.1 Cross-sectional Validation

In this method, we focus our attention on the cross-sectional performance of our model, which is based on accumulated dataset up to date (in our case, it is the 240-year dataset of the CoL). We already obtained the fitting results of this dataset in Section 6.1.3, as shown in Table 6.1 and Figure 6.2. Since the posterior distribution of the number of terrestrial species is positively skewed, we believe that the posterior mode represents a better point estimation for the number of species. Therefore, we will validate our model through validating its point estimation (the posterior mode). In this process, we replaced the prior distribution of the number of species with a fixed value which is the estimated posterior mode C_{240}^* , and re-fitted the given dataset according to this new update in the model parametrization, as shown in Figure 6.4. The purpose of this fitting is to check the following two aspects:

- whether the point-estimate fitting is going to generate a dataset that is narrower or at least similar to the one generated by the original fitting of Section 6.1.3, *i.e.* check if

$\{\text{Accepted}(\boldsymbol{\tau}^*, \boldsymbol{x}^*)_{1000} | C_{240}^*\} \approx \{\text{Accepted}(\boldsymbol{\tau}^*, \boldsymbol{x}^*)_{1000} | P(C)\}$, where $P(C)$ is the prior distribution of C , and the number 1000 is the size of the accepted sample that contributes in estimating the posterior distribution of C .

- whether the distribution of the accepted total Euclidean distances out of the point-estimate fitting is similar or close to the one out of the original fitting, *i.e.* check if:

$P(\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000} | C_{240}^*) \approx P(\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000} | P(C))$, where $\{\mathcal{D}(\boldsymbol{\tau}^*, \boldsymbol{\tau}) + \mathcal{D}(\boldsymbol{x}^*, \boldsymbol{x})\}_{1000}$ is the accepted total Euclidean distances, explained previously in the ABC algorithm in Section 3.2.3.

If we verified the above two aspect, we can say that the point estimate of our model is valid which indicates the validation of our model in a cross-sectional point of view.

Figure 6.4 shows four plots of the accepted dark grey areas (for the number of discoveries and the number of authors of two ways of fitting) which represent

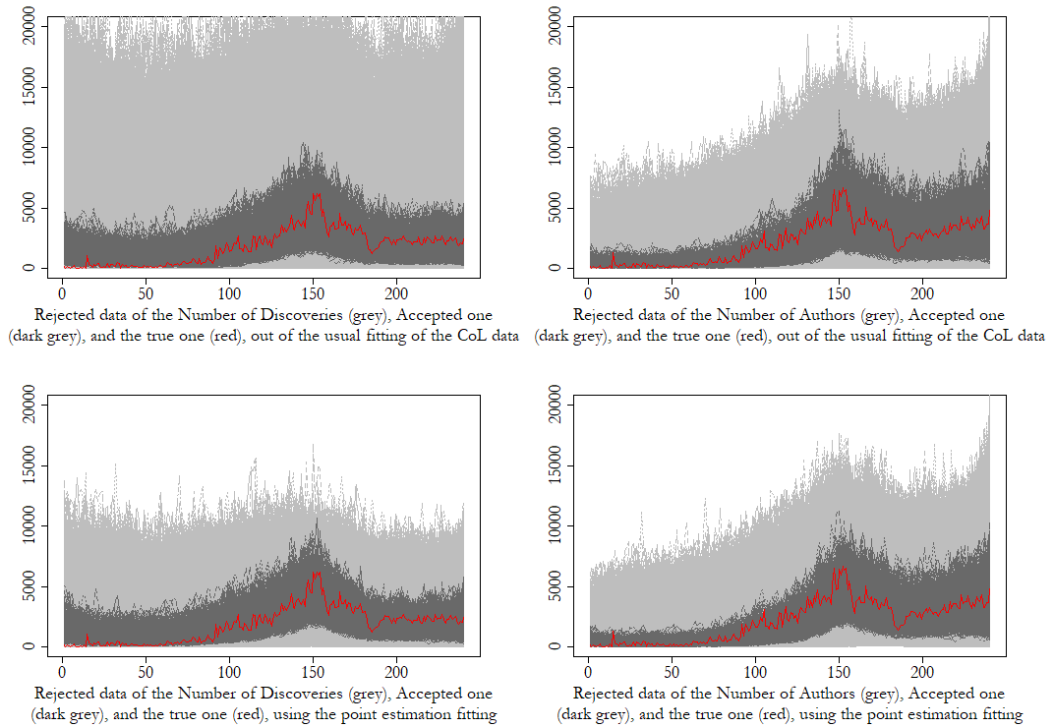


Figure 6.4: Fitting results of CoL dataset, the top panel belongs to the original fitting, while the bottom panel belongs to the point-estimate fitting.

the data that contribute in estimating the posterior distribution of the number of species. By comparing the accepted dark grey areas of the bottom panel plots with their confronted plots of the top panel, we can see that the results of the point-estimate fitting are as much as the original fitting results in capturing the general annual pattern of the given dataset, but narrower in enveloping this given dataset. This is an indication of valid estimation, as we expect the point estimate to generate data-series that are closer to the given one.

In addition, when we check the Euclidean distances, Figure 6.5 shows that the distribution of the Euclidean distances (measured in a standardized scale) of the point-estimate fitting is identical to the distribution of the ones of the original fitting. Note that before the standardization, the Euclidean distances of the point-estimate fitting is varying in a smaller scale than the one of the original fitting, since the point-estimate fitting results in generating data-series that are closer to the given CoL dataset than the ones generated by the original fitting. Therefore, we used the standardized scale to be able to make a fair comparison.

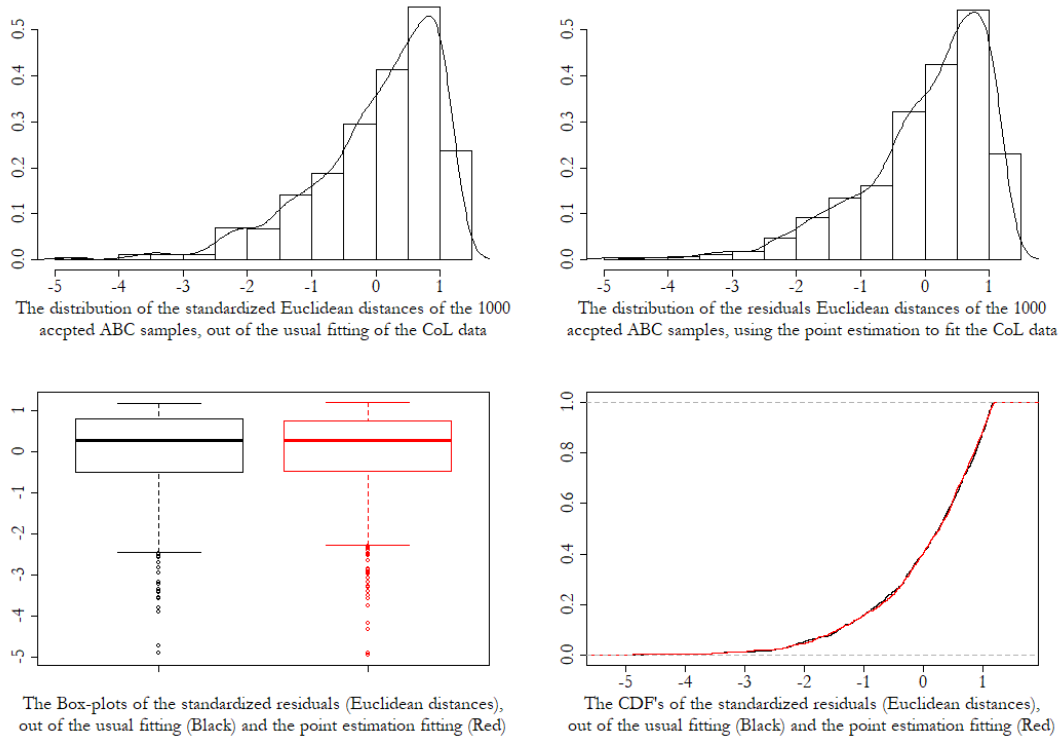


Figure 6.5: Distributions properties of the Euclidean distances (as a residual diagnostic) of the original fitting and the point-estimate fitting of the CoL dataset.

6.3.2 Longitudinal Validation

In the current method, we focus our attention on the longitudinal performance of our model, which is based on the model estimations over specific periods of time. Therefore, we will validate our model through validating its point estimates over the 150-year time period and the 185-year time period. Since the point estimation of 240-year time period is validated in the above method (Section 6.3.1), it will be taken as a ground-truth in the longitudinal validation. In each validation process, we replaced the prior distribution of the number of species with a fixed value which is the estimated posterior mode (C_{150}^* for the 150-year time period and C_{185}^* for the 185-year time period), and re-fitted the whole given dataset (over 240 years) according to this new update in the model parametrization. Such as the cross-sectional validation fitting, the purpose of the current fitting is also to check two aspects, one is related to the pattern of the accepted generated data and the other is related to the distribution of Euclidean distances that decided the

accepted data. However, here we applied two types of treatments in computing the Euclidean distances of the truncated past period, and each treatment branches into two parts, explained as follows:

First Treatment (FT):

- (FT.1): Computing the total Euclidean distances based on the whole given dataset (over 240 years), *i.e.* $\{\mathcal{D}(\boldsymbol{\tau}_{1:240}^*, \boldsymbol{\tau}_{1:240}) + \mathcal{D}(\boldsymbol{x}_{1:240}^*, \boldsymbol{x}_{1:240})\}_{1000} | C_T^*$, then select the accepted ones which decide the accepted dataset, where $T = 150, 185$.
- (FT.2): From the above accepted dataset, re-compute the total Euclidean distances based on just the last part of the data that represent the future to the truncated periods, *i.e.*

$$\{\mathcal{D}(\boldsymbol{\tau}_{T:240}^*, \boldsymbol{\tau}_{T:240}) + \mathcal{D}(\boldsymbol{x}_{T:240}^*, \boldsymbol{x}_{T:240})\}_{1000} | \{\text{Accepted}(\boldsymbol{\tau}_{1:240}^*, \boldsymbol{x}_{1:240}^*)_{1000} | C_T^*\}$$

Second Treatment (ST):

- (ST.1): Computing the total Euclidean distances based on the truncated dataset, *i.e.* $\{\mathcal{D}(\boldsymbol{\tau}_{1:T}^*, \boldsymbol{\tau}_{1:T}) + \mathcal{D}(\boldsymbol{x}_{1:T}^*, \boldsymbol{x}_{1:T})\}_{1000} | C_T^*$, then select the accepted ones which decide the accepted dataset (note that the accepted dataset is over 240 years but its Euclidean distances are computed on truncated part of the dataset).
- (ST.2): From the above accepted dataset, re-compute the total Euclidean distances based on just the last part of the data that represent the future to the truncated periods, *i.e.*

$$\{\mathcal{D}(\boldsymbol{\tau}_{T:240}^*, \boldsymbol{\tau}_{T:240}) + \mathcal{D}(\boldsymbol{x}_{T:240}^*, \boldsymbol{x}_{T:240})\}_{1000} | \{\text{Accepted}(\boldsymbol{\tau}_{1:T}^*, \boldsymbol{x}_{1:T}^*)_{1000} | C_T^*\}$$

If we verified the two targeted aspects related to the generated data pattern and the Euclidean distances distribution in the above two treatments, we can say that chronologically the point estimates of our model are valid which indicate the validation of our model on a longitudinal point of view. In the following, we will explore these aspects and treatments for both $T = 150$ and $T = 185$, in which we replaced the prior distribution of C by a fixed value C_{150}^* and C_{185}^* , respectively. For simplicity, we will check two treatments at a time, first the treatments (FT.1) and (ST.1) together, then the treatments (FT.2) and (ST.2).

In the 150-year time period, Figure 6.6 shows the generated datasets of the treatments (FT.1) and (ST.1) in the middle and bottom panels, respectively, against the generated dataset of the ground-truth fitting (given C_{240}^*) in the top panel. By comparing the accepted dark grey areas (for the number of discoveries and the number of authors) of the middle panel plots with their confronted plots of the top panel, we can see that the results of the (FT.1) fitting are very close to the C_{240}^* fitting results in capturing the general annual pattern and enveloping the given dataset. On the other hand, when we based the Euclidean distance computing only on the truncated dataset *i.e.* the first 150 data points (ST.1) as shown in the bottom panel, we can see some diversions in predicting the future

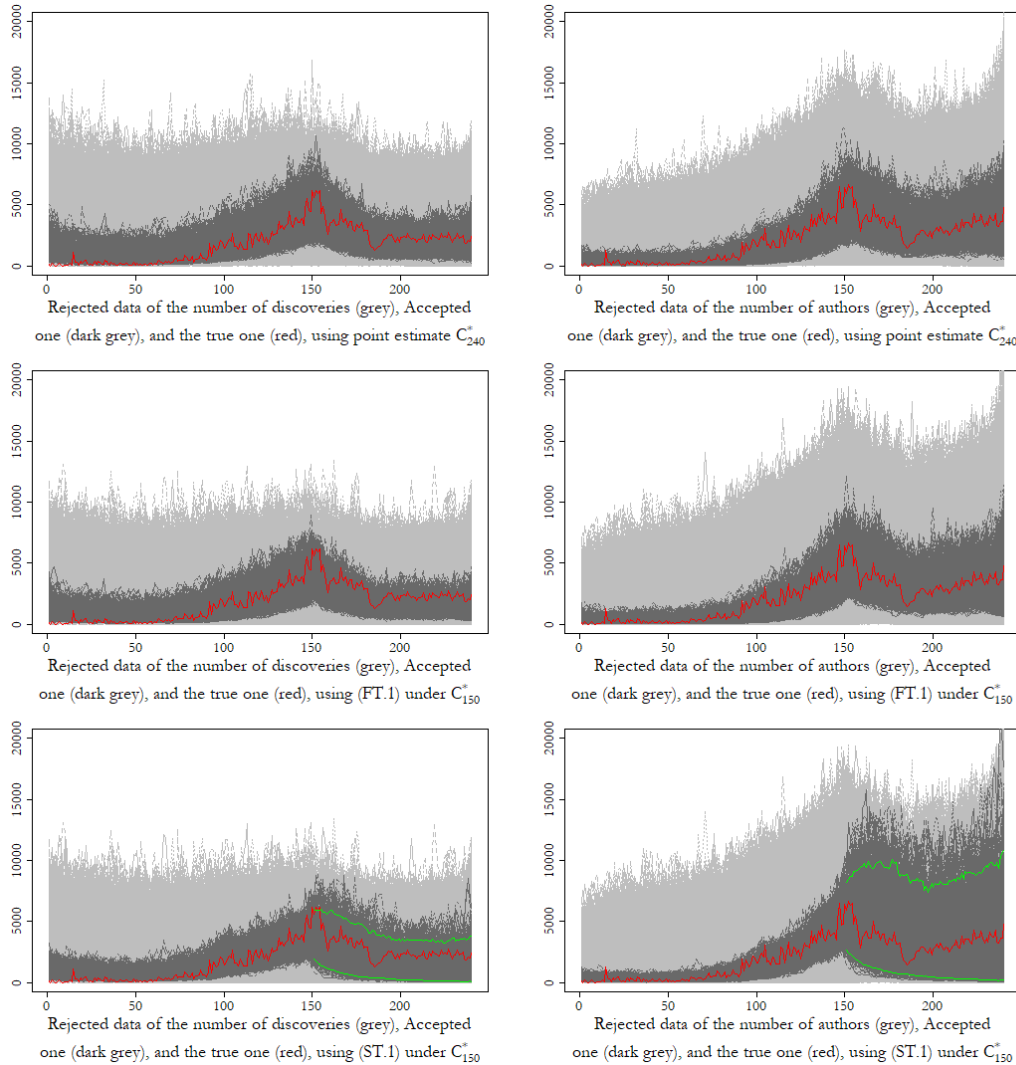


Figure 6.6: Fitting results of CoL, the top panel belongs to the ground-truth given C_{240}^* , while the middle and bottom panel belongs to the (FT.1) and (ST.1) given C_{150}^* , respectively. Note that the green lines in the bottom panel represent the 95% credible interval.

dataset (last 90 data points) especially in the number of authors in the right bottom plot. That is justified by the way that we applied the ABC algorithm, which is relied on computing the Euclidean distances on the given raw data, not on a sufficient statistic. However, when we check the Euclidean distances, Figure 6.7 shows that the distribution of the Euclidean distances (measured in a standardized scale) of the (FT.1) fitting and the (ST.1) fitting given C_{150}^* are both identical to the distribution of the ground-truth fitting given C_{240}^* , which is a reflection of a valid estimation.

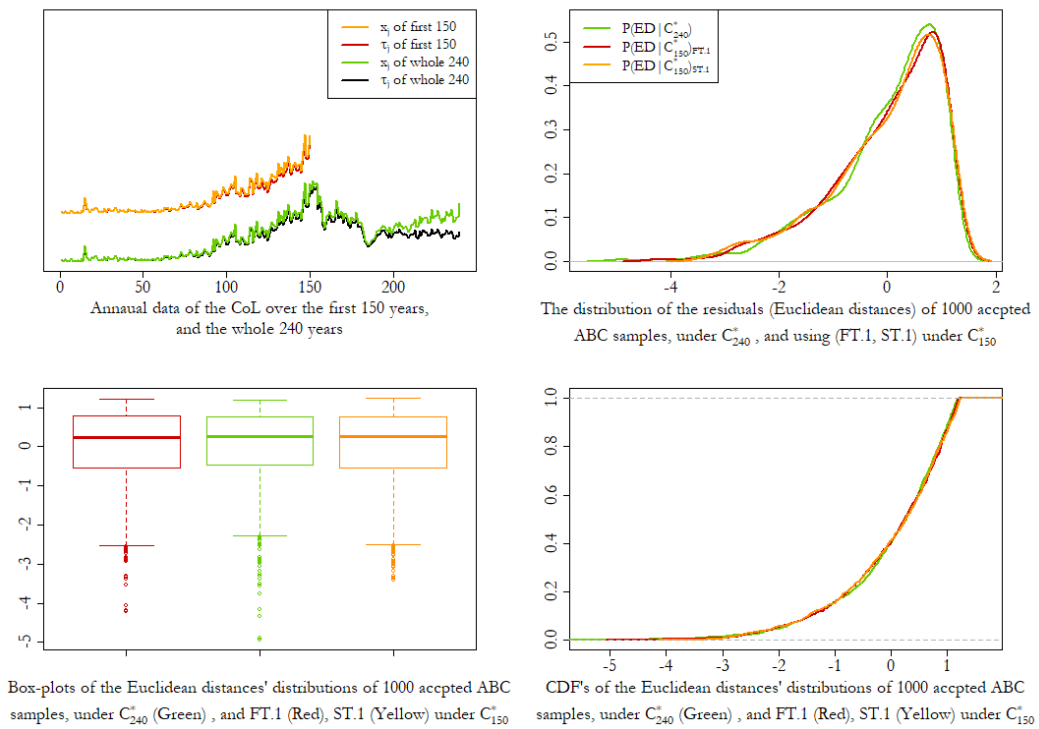


Figure 6.7: Distributions properties of the Euclidean distances (as a residual diagnostic) of the ground-truth fitting given C_{240}^* , and the (FT.1) and (ST.1) fittings given C_{150}^* .

With respect to the (FT.2) and (ST.2) treatments, after cutting-out the first 150 data points from the generated data (dark grey areas) in Figure 6.6, and re-computing the Euclidean distances only on the last 90 data points, Figure 6.8 shows that the distribution of the Euclidean distances (measured in a standardized scale) of the ground-truth fitting given C_{240}^* is identical to the distribution of the (FT.2) fitting given C_{150}^* . Moreover, it is very close (almost similar) to the distribution of the (ST.2) fitting given C_{150}^* , in spite of the higher variation in predicting the future dataset.

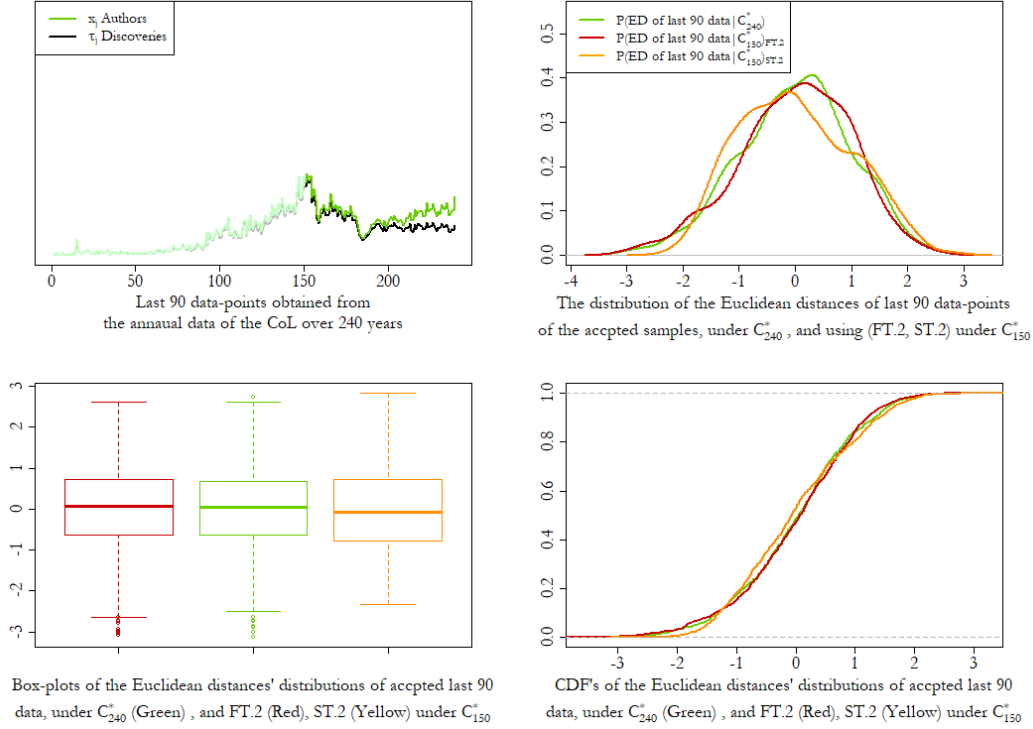


Figure 6.8: Distributions properties of the Euclidean distances (re-computed only on the last 90 data points of the generated datasets) of the ground-truth fitting given C_{240}^* , and the (FT.2) and (ST.2) fittings given C_{150}^* .

The 185-year time period, Figure 6.9 shows the generated datasets of the treatments (FT.1) and (ST.1) in the middle and bottom panels, respectively, against the generated dataset of the ground-truth fitting (given C_{240}^*) in the top panel. By comparing the accepted dark grey areas (for the number of discoveries and the number of authors) of the middle panel plots with their confronted plots of the top panel, we can see that the results of the (FT.1) fitting are almost as the same as the C_{240}^* fitting results in capturing the general annual pattern and enveloping the given dataset. On the other hand, when we based the Euclidean distance computing only on the truncated dataset *i.e.* the first 185 data points (ST.1) as shown in the bottom panel, we can see some diversions in predicting the future dataset (last 55 data points) especially in the number of authors in the right bottom plot, which is justified by the way that we applied the ABC algorithm, which is relied on computing the Euclidean distances on the given raw data, not on a sufficient statistic. However, when we check the Euclidean distances, Figure 6.10 shows that the distribution of the Euclidean distances (measured in a standardized scale) of the (FT.1) fitting and the (ST.1) fitting given C_{185}^* are both

identical to the distribution of the ground-truth fitting given C_{240}^* , which is a reflection of a valid estimation.

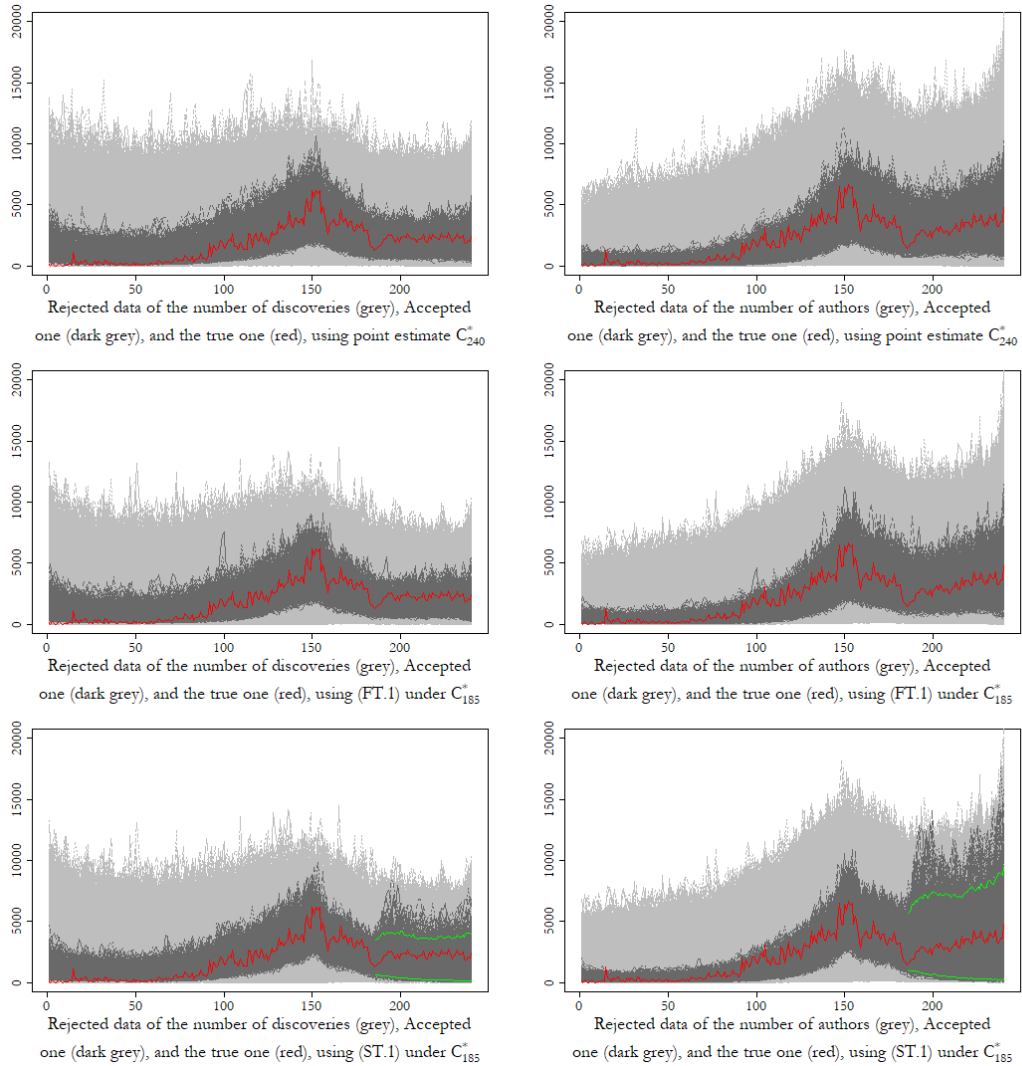


Figure 6.9: Fitting results of CoL, the top panel belongs to the ground-truth given C_{240}^* , while the middle and bottom panel belongs to the (FT.1) and (ST.1) given C_{185}^* , respectively. Note that the green lines in the bottom panel represent the 95% credible interval.

With respect to the (FT.2) and (ST.2) treatments, after cutting-out the first 185 data points from the generated data (dark grey areas) in Figure 6.9, and re-computing the Euclidean distances only on the last 55 data points, Figure 6.11 shows that the distribution of the Euclidean distances (measured in a standardized scale) of the ground-truth fitting given C_{240}^* is identical to the distributions of both the (FT.2) and (ST.2) fittings given C_{185}^* , in spite of the higher variation in generating (predicting) the future dataset in (ST.2).

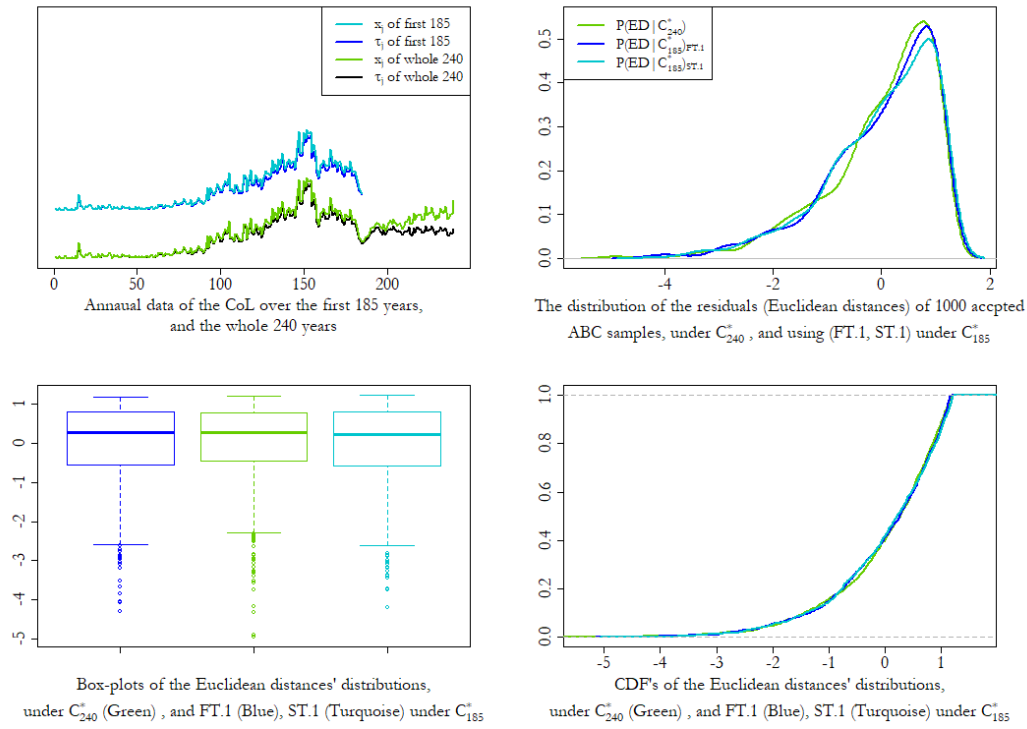


Figure 6.10: Distributions properties of the Euclidean distances (as a residual diagnostic) of the ground-truth fitting given C_{240}^* , and the (FT.1) and (ST.1) fittings given C_{185}^* .

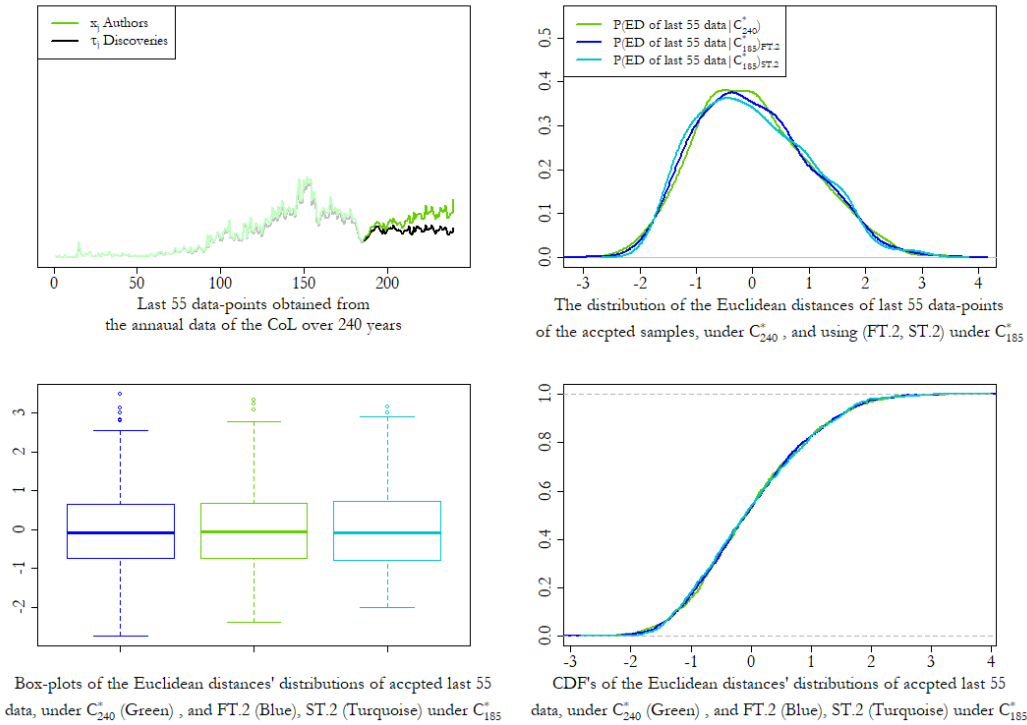


Figure 6.11: Distributions properties of the Euclidean distances (re-computed only on the last 55 data points of the generated datasets) of the ground-truth fitting given C_{240}^* , and the (FT.2) and (ST.2) fittings given C_{185}^* .

To sum up the longitudinal validation, it is worth emphasizing that the purpose of our model is to predict the number of species (given the annual observed dataset and considering all the other variables in the context of species discovery), not to predict the dataset itself in future time, and not to predict other variables in the context. Therefore, in our aim to validate our model in this respect, we tried to basically achieve this through implementing the first treatment (FT) in particular rather than the second treatment (ST). This is justified by the way that we applied the ABC algorithm, which relied on computing the Euclidean distances on the given raw dataset (the dataset in its exact nature as time series), not on a sufficient statistic. This way of implementation forces us to use the whole dataset (over 240 years) to compute the Euclidean distances even in the situations that we are using truncated period estimates (*e.g.* C_{150}^*). In these situations, our concern is to check the model performance with using the ABC algorithm applied on raw datasets, not to pretend that we do not know the data of the cut-off period (*e.g.* the last 90 data-points in the case of 150-time period) and try to predict this dataset (the last 90 data-points represents the future of the 150-time period dataset, which we may wish to predict). However, out of curiosity, we approached the second treatment (ST) that deals with predicting future dataset, which appeared to provide acceptable results although it is not really a benchmark to judge our model performance.

6.4 Model Fitting for WoRMS Data

In this section, we apply our model on the WoRMS dataset, at which we also followed the same distributional specifications that we have applied on CoL in Section 6.1, and previously detailed in Section 4.1 and shown in Figure 4.2. Since we noticed a good performance of the altered ABC algorithm (explained in Section 6.1.2, and Algorithm 2 of Section C.2 in Appendix C) on the CoL and the artificial worlds, we adopted the same algorithm for the WoRMS dataset. With respect to the distributional parametrizations, we faced some struggles (see Group B.21 of Section B.5 in Appendix B) until we end up with rational assumptions, explained in Section 6.4.1.

6.4.1 Model Parametrizations

For the parameters values, we based our choices on assuming that the community of the marine species (recorded in WoRMS databases) is just a sample from the same population of the terrestrial and freshwater species (recorded in CoL database). Therefore, the WoRMS dataset is assumed to be governed by the same distributional parametrizations of the species abundances of the CoL dataset, but they differ in the latent effort. Referring to Figure 6.1, in spite of the differences in the taxonomic efforts (represented in the number of authors) between the WoRMS and the CoL datasets, the “same population” assumption is supported by the similarity between them in the pattern of the number of the discoveries, if we neglected the occurrence scale, we can see the similarity pattern in Figure 6.12. Moreover, [13] (in which, one of the authors (M.J. Costello) is a biodiversity specialist) indicated this similarity by stating that “*As the marine environment includes all but one of the phyla on Earth, it provides a unique perspective on biodiversity*”. This may be interpreted as that we can consider the marine (recorded in WoRMS) and the terrestrial species (recorded in CoL) as two equivalent communities or samples coming from the same population.

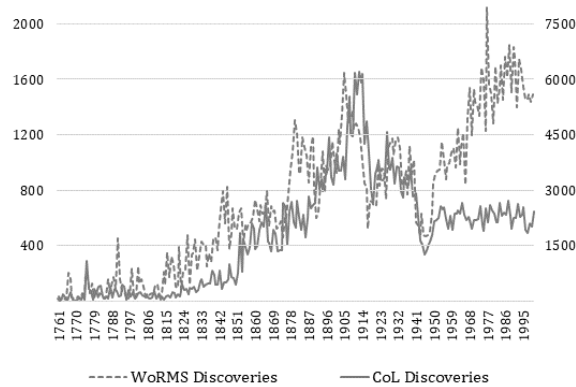


Figure 6.12: Similarity between the WoRMS and the CoL datasets in the discovery curves.

Therefore, based on the above assumption and on noticing the ratio of the difference between the WoRMS and the CoL datasets in the number of authors curves, we found that the total number of authors involved in discovering the WoRMS species represents about 12.325% of the one of the CoL. This implies that on a log-scale, the effort mean in the WoRMS data equals 12.325% of the

effort mean in the CoL data, *i.e.*

$$\exp(\mu_{L_1} + \sigma_L^2/2)_{\text{WoRMS}} = 12.325\% \times \exp(10.6 + 0.15^2/2)_{\text{CoL}} \simeq 5002$$

Hence, we need to specify values for μ_{L_1} and σ_L in the WoRMS dataset such that its effort mean $\simeq 5002$. However, since the effort mean is just one equation in two variables, we had to make few multiple trials for tuning the selected values until we end up with the best elicitation of parametrization as shown next.

6.4.2 Fitting Results of WoRMS

Table 6.3 shows the best elicited parameters of modelling the WoRMS dataset, as well as the fitting results. After operating the pilot experiment and deriving the thresholds, we implemented intermediate main experiment on just 100,000 iterations to make sure that we are on the right track of our distributional specifications, see the fitting details and plots of the pilot and the intermediate experiments in Group B.23 of Section B.5, Appendix B. This is followed by the final main experiment of 1,000,000 iterations, from which we estimated the number of marine species. It is worth mentioning that, before reaching our final results, we made an attempt at which $C \in [D, 300D]$. However, it resulted in having a posterior distribution that looked to be a heavy-tailed on the right-side *i.e.* the distribution appears to be truncated (suddenly goes down to zero) at the upper limit $300D$, see Group B.22 in Appendix B. This motivated us to repeat the whole experiments on a wider prior support, $C \in [D, 500D]$, which is the one that we finally adopted in Figure 6.13 in the current section and Group B.23 - Appendix B.

Table 6.3: Parametrization and resulting statistics of modelling WoRMS dataset.

Model Parameterization of WoRMS		Fitting Results of WoRM	
$C \in [D, 500D]$	$C \in [173,401- 86,700,500]$	Posterior Mean	31,859,192
$\theta_N = \{\mu, \sigma\}$	{10, 4.7}	Posterior Mode	17,651,757
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{8.5, 0.2}	95% Credible Interval	[9,396,661 - 76,398,957]
$\theta_x = \alpha$	log(100)	Posterior SD	18,253,295

Figure 6.13, right top plot, shows that the total number of marine species follows a skewed distribution, concentrated around 20 million species, ranging

from 9.4 to 76.4 million species in 95% central credible interval (red lines), and ranging from 6.3 to 71.9 million in a tighter 95% HPD interval (blue lines) with a difference of 1.4 million species between the two intervals. If we adopted the mode as point estimation for the total number of marine species, we can say that about 17,478,356 species are yet to be discovered, which is about 100 times the current discoveries. In the bottom panel of Figure 6.13, we can see that the acceptable dark grey area (the generated number of authors and the number of the discoveries) is a good representative of the given WoRMS dataset. It captures the annual fluctuating behaviour of the given data at which a first local maximum peak occurs around the 150th year (1910), followed by a deeper trough around the 185th year (1945), and end up with an increasing pattern.

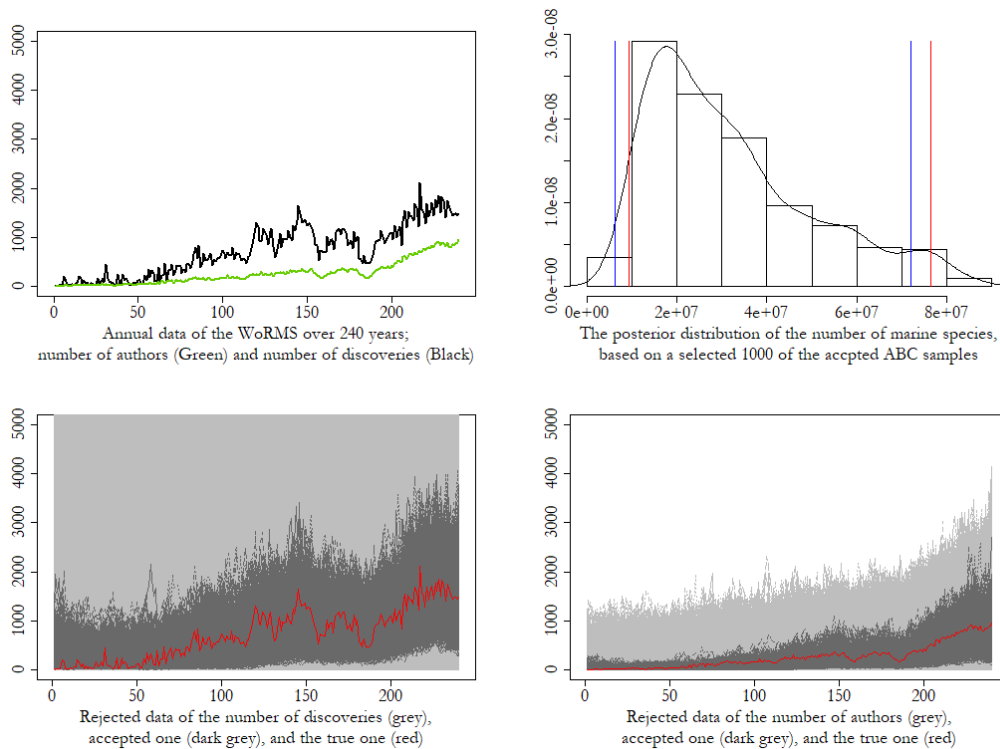


Figure 6.13: Fitting results of the CoL dataset.

If we track the fitting performance over the three number of iterations ($iter = 10,000$, $iter = 100,000$, and $iter = 1,000,000$) by comparing Figure 6.13 with the figures in Group B.23 - Appendix B, we can see the enhancement of the fitting associated with increasing the number of iterations; the shape of the posterior distribution becomes more tied with lighter tails. In addition, by

observing the plots of the bottom panels, we can see that the generated data in the acceptable dark grey area are getting closer with increasing the number of iterations and better capturing the behaviour of the given WoRMS data.

6.5 Global Estimation

In this section we attempt to estimate the number of species on a global scale which includes the number of terrestrial, freshwater, and marines species. We already estimated the number of terrestrial and freshwater species based on the CoL dataset (Section 6.1.3), and estimated the number of marines species based on WoRMS dataset (Section 6.4.2). In this instance, we shall recall the analysis we accomplished in Section 5.3, and concluded in Section 5.4 (pages 138-139), which stated that although we can not ultimately prove that our proposed model is additive, our approach definitely provides alternative aggregated estimation (through aggregating multiple resulted estimates after performing multiple-group fittings) that is equivalent to the whole category estimation (when we combine all datasets together first, then perform one whole fitting).

The experiments based on the artificial worlds (Section 5.3) showed that in all the three artificial worlds, all the central tendency statistics that are obtained from the aggregating step are very close to the true values of the number of species and to the estimated ones out of the whole category fitting. The derived values of KLD metric also provided similar implications; a closeness between the distributions out of the aggregation and the whole category distributions. Therefore, we will estimate the global number of species through aggregating the estimations results of CoL and WoRMS *i.e.* aggregating their selected final samples, then obtain the posterior distribution of total number of species C_{Global} .

Figure 6.14 shows the subgroups estimations (of CoL and WoRMS datasets) along with the global estimation of the number of species resulting from their aggregating. From the bottom plots, we can see that the resulting posterior distribution of C_{Global} is also positively skewed, with 95% central credible interval (red lines) ranging from 25.3 to 111.7 million species, a slightly tighter 95% HPD

interval (blue lines) ranging from 19.9 to 102.9 million species where the difference between the HPD and the central intervals is 3.4 million species, and inter-quartile interval ranging from 39 to 70 million species. The point estimation states that the total number of species on the global is $C_{\text{Global}}^* = 56,634,999$ if we adopted the posterior mean, $C_{\text{Global}}^* = 51,580,110$ if we adopted the posterior median, and $C_{\text{Global}}^* = 45,784,117$ if we adopted the posterior mode.

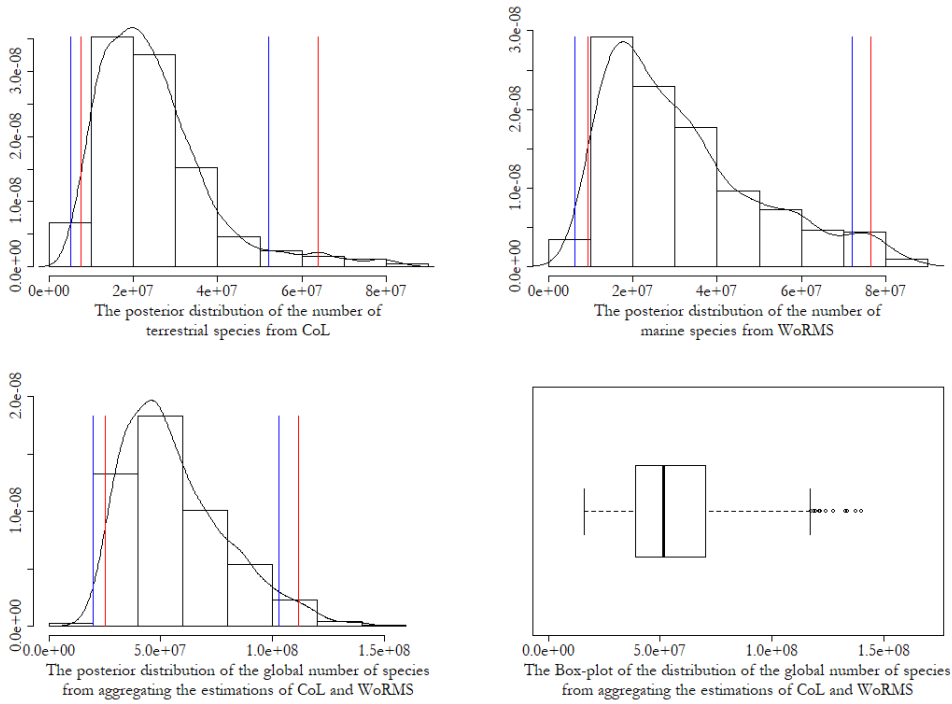


Figure 6.14: Global estimation results out of aggregating the estimation results of CoL and WoRMS datasets.

It is interesting to notice that the estimated global mean is exactly equal to the sum of the estimated means of CoL and WoRMS which is (56,634,999) total species, and the estimated global standard deviation is very close to the sum of the estimated standard deviations of CoL and WoRMS as both of them around 22 million ($\sigma_{\text{Global}} = 22,998,412$, while $\sigma_{\text{CoL}} + \sigma_{\text{WoRMS}} = 22,619,187$). However, the estimated global mode is a slightly higher where $C_{\text{CoL}}^*(\text{mode}) + C_{\text{WoRMS}}^*(\text{mode}) = 37,483,961$ total species, and the estimated global 95% central credible and HPD intervals are slightly wider with $[16,950,990 - 140,283,839]_{\text{CoL+WoRMS}}$ and $[11,413,877 - 124,025,173]_{\text{CoL+WoRMS}}$, respectively. These differences are expected due to the skewness (asymmetry) of the estimated distributions.

6.6 Comparing with Previous Work

As we explored in Section 2.1, there are two main approaches in the literature of estimating the number of classes/species: the sampling-theoretic and the data-analytic, where the former deals with cross-sectional datasets, while the latter mostly deals with longitudinal. On the other hand, our proposed approach integrates these two approaches in one comprehensive framework. However, the type of the observed datasets that we are dealing with is longitudinal, which makes it difficult to compare our model with models of the sampling-theoretic approach. Therefore, we will focus our attention to compare our model with ones of the data-analytic approach, in particular the related recent work suggested by [64] [21] [13].

In these references, a model of non-homogeneous renewal process (NHRP) was first suggested by [64] and applied on a selected dataset of European marine species. Then, was contrasted with other related approaches, with application on a wider collection of dataset from the European Register of Marine Species inventory system (ERMS) [21]. Finally, this model was extended on a global scale to include marine and terrestrial (with freshwater) species from WoRMS and CoL, respectively, and its results were contrasted (by observing patterns, not rigorous approach) with the taxonomic effort curves [13].

6.6.1 Brief Description of NHRP Model

In this model, the discovery process is described by a non-homogeneous renewal process (NHRP), a type of point process, in which t represents the time location (discovery dates) at which the points (discovery events) occur, and $N(t)$ represents the cumulative number of points (discoveries) at a given time. This process assumed to behave as “S-shape” patten fluctuating around a mean that is formulated as logistic function,

$$E\{N(t)\} = \frac{N}{(1 + \exp(-\psi_1(t - \psi_2)))} \quad (6.2)$$

In this function, there are two parameters ψ_1 and ψ_2 that control the rate of the logistic growth which is in our case the rate of species discovery, where ψ_2 is the

time of greatest rate of discovery, and ψ_1 is a scaling factor that determines how quickly the curve reaches its maximum value. N is the maximum value of the curve which is in our case the expected total number of species *i.e.* $\lim_{t \rightarrow \infty} E\{N(t)\} = N$. In this process, the times between discovery events are assumed to be independent and identically distributed Gamma with parameters θ_1 and θ_2 which are linked to ψ_1 and ψ_2 through the linkage between NHRP and Gamma distribution. As we mentioned in Section 2.2.2 (pages 24-25), having the two Gamma parameter allow to capture the mean and the under-dispersion/over-dispersion behaviour in the process, unlike the exponential distribution that is used in the non-homogeneous Poisson process which assumes that the dispersion is equal to the mean.

The Bayesian inference is applied in the NHRP modelling, in which there is a joint posterior distribution that is used to estimate the model's parameters and simulate the inter-discovery times in order to predict the total number of species. Through MCMC algorithms, a multiple of samples of $\{\psi_1^{(i)}, \psi_2^{(i)}, \theta_1^{(i)}, \theta_2^{(i)}\}$; $i = 1, 2, \dots, m$ are generated, and at each sample i , these generated parameters are used to simulate successive future inter-discovery times $s_1^{(i)}, s_2^{(i)}, \dots$ until we face a very large simulated value such as $s_n^{(i)} = 10,000$ years. At this point, the simulation stops under the assumption that this time interval (10,000 years) between discoveries is beyond human life expectancy, and consequently the eventual time point is $t_\infty^{(i)} = t_{n-1}^{(i)} + s_n^{(i)}$. Finally, $N(t)^{(i)}$ is computed, and its distribution (over the index i) is taken as an approximate of the predictive posterior distribution of the eventual future number of discoveries that is assumed to be the total number of species, $P\{N(\infty) | t_1^{(i)}, t_2^{(i)}, \dots, t_\infty^{(i)}\}$. From this distribution, summary statistics (mean, mode, median, 95% credible interval) are derived as estimation of the total number of species.

6.6.2 Our Model Results vs NHRP Model Results

Before going through the estimation results of the two models, it is worth mentioning that we aimed to use the same datasets that are used in the latest reference that applied the NHRP model [13], in order to make fair comparisons. However, due to noticing some differences between datasets of that reference and the ones that we are currently using (note that there is on-going update for these datasets

in their inventory systems, and we have the recent updated version), we asked one of the authors (S.P. Wilson) to re-fit the current datasets according to the NHRP model and derive the required estimates. The way he implemented the fitting process is explained in his words (via a personal communication) in the following:

“The code to implement the logistic function approach of [64] is written in C. This is a pure curve-fitting approach and only uses the discovery times as its input. It was applied to the CoL data. This data set was too large to be handled by the code, which was solved by dividing the yearly discovery counts by 40 and rounding to an integer, then fitting this smaller data set, obtaining an estimate of the number of species remaining to be discovered and multiplying that back by 40. The WoRMS data had a similar problem and that was solved similarly, but by dividing the number of discoveries per year by 10.”, S.P. Wilson.

With respect to the global estimation, since the logistic function is not additive, an alternative way that involves combining the CoL and the WoRMS datasets first then fitting them according to the NHRP model is attempted. However, the model failed in doing so, and consequently, another solution is provided by deriving the global estimations from the summary statistics of fitting CoL and WoRMS. Therefore, the global mean can be obtained from just adding the estimated means of the CoL and WoRMS datasets, and since the estimated posterior distribution of each dataset seems symmetric (according to the NHRP model), the global mode and median can also be obtained in the same way; by adding the estimated modes and estimated medians of the CoL and WoRMS datasets. With respect to the 95% credible interval, it is derived by adding $\pm 2\sqrt{sd(\text{CoL})^2 + sd(\text{WoRMS})^2}$ to the obtained global median, under the assumption that these variances can be added, where the standard deviations $sd(\text{CoL})$ and $sd(\text{WoRMS})$ each can be roughly computed as quarter the width of the 95% credible interval of CoL and WoRMS datasets, respectively .

Table 6.4 shows the NHRP model fitting results of the CoL, WoRMS datasets, and their global estimation confronted to the fitting results of our model, where we can see huge differences between the estimation of the two approaches. Our proposed model suggests a positively skewed distribution for the given datasets, while the NHRP model suggests a symmetric distribution for the same data. The

results of our model are ranging on ten-million scale, while the NHRP results are ranging on a hundred-thousands scale. The 95% credible interval provided by our model is reasonably wide enough to describe a probability distribution, while 95% credible interval provided by the NHRP model is very narrow, as shown in Figure 6.15. Moreover, the method used to derive the global credible interval in the NHRP model is based on the assumption of adding the variances of the CoL and WoRMS datasets, which can be considered only if there is no correlation between the given datasets, and this not necessary to be always true. In fact, there is a correlation in our case between the CoL and WoRMS datasets, as shown in Group B.24 of Section B.5, Appendix B (top panel).

Table 6.4: Estimation results of our proposed model vs. NHRP model.

Summary Statistics	Estimation Results of our Model			Estimation Results of NHRP Model		
	CoL	WoRMS	Global	CoL	WoRMS	Global
Mode	19,832,205	17,651,757	45,784,117	481,615	280,000	761,615
Median	22,293,127	26,980,422	51,580,110	480,615	280,000	760,615
Mean	24,775,806	31,859,192	56,634,999	480,665	280,000	760,665
95% CI	[7,554,329 - 63,884,882]	[9,396,661 - 76,398,957]	[25,292,129 - 111,737,642]	[475,415 - 486,115]	[274,000 - 287,000]	[752,196 - 769,034]

This substantially less estimation of NHRP model can be justified by the fact that the main property of the Logistic function is that it is a function of “S-shape” which means it should level up quickly at the end, representing a quick asymptotic estimation. However, the cumulative curves of the given datasets especially the WoRMS (in red color in the plots of the right panel of Figure 6.15) do not seem to infer an anticipation of levelling up pattern by the year 2100, and it might be not a representative assumption to enforce the Logistic function as a mean. Moreover, the NHRP model estimated the total number of marine species to be roughly half of the estimated total number of terrestrial species, which might be inconsistent if we consider the fact that the range of biodiversity in marine species is almost equivalent to the one of the terrestrial species [13], and the fact the number of marine discoveries is still obviously less and what is left is expected to be much more (see Figure 6.1). All this might invoke some doubts in the estimation results of the NHRP model on the current datasets.

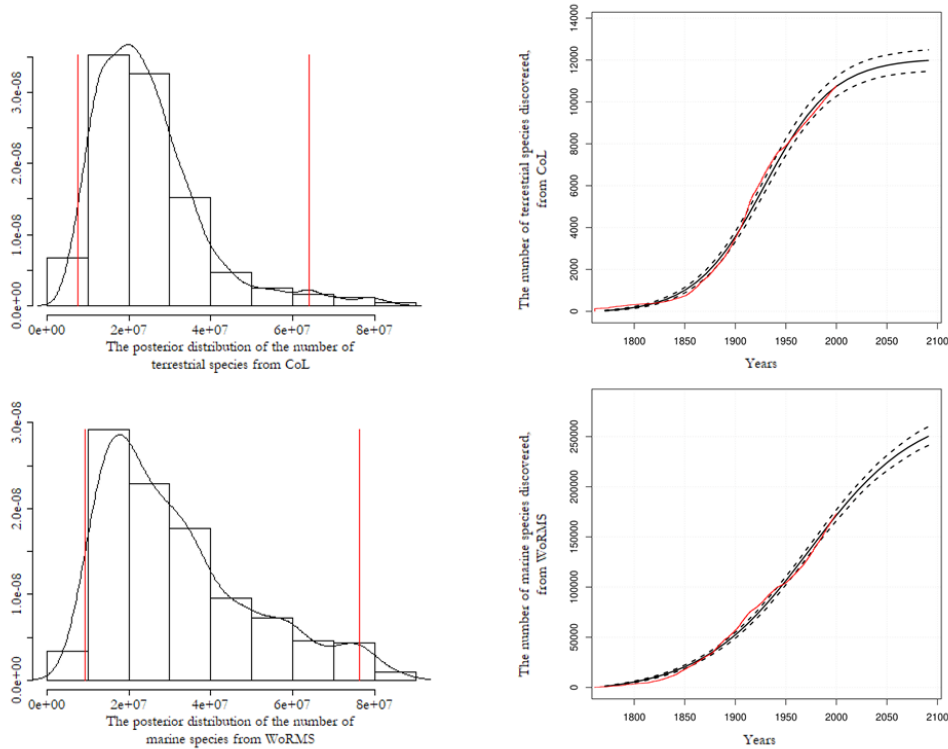


Figure 6.15: Our Model results (left panel) vs the NHRP Model results (right panel) of fitting the CoL and WoRMS datasets. The left panel shows the estimated posterior distributions, along with their 95% credible intervals (red lines). In the right panel, the red lines represent the annual cumulative number of discovery curves, the black solid lines represent the median fitting curves, and the black dashed lines represent the estimated 95% credible intervals.

Comments on Similarities & Differences

- Both models adopt Bayesian inference and approximate a posterior distribution for the target variable (total number of species), from which summary statistics are derived as estimations. However, the size of the two models is completely different, the NHRP model involves two variables and four parameters which are considered random variables, while our model involves six variables and five parameters which are considered fixed in the current implementation, though they can be treated as random variables if we would like to capture more variability.
- Both models adopt Bayesian computation techniques. However, the MCMC was the adopted algorithm in the NHRP model, while this algorithm was inefficient to be applied in our case and instead we adopted the ABC algorithm.
- The NHRP model predicts the number of species through predicting the even-

tual number of discoveries *i.e.* the estimated number of discoveries in future time (that assumed to be infinite) represents the estimated number of species. In contrast, our model use the number of discoveries as an input and employs it (along with a co-variate discovery efforts) to predict the number of species. In doing this, our model assumes that the number of species has a fixed distribution. This assumption is based on the fact that the emergence of a new species takes millions of years of bio-interactions which is beyond the human life expectancy so that, in our witness the number of species is time-fixed random variable not time-varying random variable.

- As a consequent result of the above explanation, the estimations of the NHRP model must be directly associated with time, while the estimations of our model are to some extent released from the time factor (we saw evidence from Section 6.2 when C_{150}^* and C_{185}^* are verified to be within the 95% credible interval of C_{240}^*).
- Although the NHRP model assumed to capture under/over-dispersions, it deals with cumulative curves of the given datasets and thereby ignores the annual variation in the data-patterns, while our model deals with the original nature of the given datasets and captures all the time variational patterns.
- The current way of applying the NHRP model assumes symmetric pattern of the posterior distribution, while it is not necessary a true reflection of the given datasets, as we saw in the estimations of our model, the distribution of the CoL and WoRMS datasets are positively skewed.
- The NHRP model is limited in dealing with large datasets, which forced ad hoc treatments to overcome such obstacles (*e.g.* dividing the annual data points by 40 in the CoL data), and also failed to perform a global fitting for the combined datasets. In contrast, our model is robust in this respect and we only needed to involve additional threshold to deal with the large dataset (as explained in Section 6.1.2).
- The NHRP model is obviously not additive model since it is based on the Logistic function which is not closed under addition. On the other hand, our

model is not proved to be not additive. Hence, in cases where we have to employ the additivity property such as the global estimation, the NHRP model can only provides crude or compelled solutions, while our model can provides us with promising alternative solutions (as explained in Section 5.3) though our model is still capable of fitting a global combined datasets.

From the above description, we can see that there is a significant difference in the way of implementing the two models, difference in their objectives, and their fitting results. Due to these differences, one of them cannot be considered a benchmark to evaluate other models against and one can only consider the positive and negative aspects of each. However, if we consider the species global estimates over the species literature that collectively ranged from 1.5 million to 100 million species on Earth [13], we can see that our global estimate (45 million species on Earth) lies within this range, but in a more reasonable scale, see Table 6.5 for miscellaneous estimates and Table 6.6 for global estimates.

Table 6.5: Examples of miscellaneous species estimates (in millions) of different taxa levels.

Reference	Estimate (millions)
Based on description rates using a subset of WoRMS 2009 [13].	0.3 of marine
Based on NHRP model of [64] and using a subset of WoRMS 2012 [91].	0.32 – 0.76 of marine
Based on expert opinion of proportions of undescribed species in regions of the world [92].	> 1 of marine
By extrapolation from proportion of Brachyura in Europe [93].	1.5 of marine
By assuming a global 6:1 ratio of fungi to vascular plants and that there are up to 270,000 species of vascular plants [94].	1.6 of fungi
Based on description rate using subset of CoL 2009 [13].	1.6 – 1.7 of terrestrial
By extrapolation from Sulawesi rain forest Hemiptera [95].	1.84 – 2.57 of global terrestrial insect
By extrapolation from rate of discovery of higher taxa using classification of valid species from WoRMS 2011 [2].	2.2 of marine
By extrapolation from a review of knowledge of insect families which found that there is about 350,000 described species of beetles [96].	1 – 3 of terrestrial insect
By extrapolation from benthos samples off Australia [97].	1 – 5 of marine
By extrapolation from host specificity of tropical floras and species richness of tropical insects to global patterns of biodiversity [98].	4 – 6 of global arthropods
By extrapolation from known fauna and regions [99].	4.9 – 6.6 of global arthropods

Reference	Estimate (millions)
By extrapolation from ratios of insect to plant species [100].	2.75 – 8.75 of global arthropods
Reviewed by Sabrosky [101].	2.5 – 10 of global insect
By extrapolation based on relationship between body size and abundance [102].	10 of global arthropods
By extrapolation from Extensive quantitative sampling of deep-sea macrofaunal communities [103].	> 1– > 10 of marine
Based on two models at which most parameters are represented by probability distributions, and Latin hypercube sampling was applied to each distribution [104] [105].	3.9 to 13.7 of terrestrial tropical arthropods
Based on a refined probabilistic models using Monte Carlo sampling and a technique known as probability bounds analysis [106].	2.4–20 & median 6.1 of terrestrial tropical arthropods
Estimation based on computational tool, called Prophinder, for predicting prophage in prokaryotic genomes. This tool is based on algorithm that uses the coding sequences CDS (a portion of a gene’s DNA or RNA that codes for protein) [107].	cover 100 of phage species
Estimation based on an analysis, at which all the phage-encoded ORFs (a program that searches for open reading frames in the DNA sequence) in Gen-Bank were compared against every other and grouped together using a BLASTE (a program that searches protein databases using a protein query) [108].	up to 100 of phage species

Table 6.6: Examples of global species estimates (in millions).

Reference	Estimate (millions)
Reviewed and objected by Erwin [109].	1.5
Estimation through aggregating description rates of two subsets of CoL and WoRMS 2009 [13].	1.8 – 2
Estimation by extrapolation from the better sampled temperate regions to the tropics. Raven noted that for the well documented species of large organisms there are roughly twice as many tropical as temperate species, and if the same ratio holds true for other organisms (with what is already described and two-thirds of these being temperate), then the true global total would be 3 -5 million or more [110].	3 – 5

Reference	Estimate (millions)
Estimation by extrapolation from rate of discovery of higher taxa using classification of valid species from CoL 2010 [2].	8.7
Estimation by extrapolating from counts of beetle species in a Panamanian tree species through estimates of total arthropod diversity per tree species to the percentages of species limited to each tree species to the total of tree species in tropical rain forests. Erwin found that the beetle diversity suggest far higher levels of insect and other arthropod diversity in tropical rain forests than had previously been estimated for the entire world fauna and flora. He hypothesized that instead of the current estimate of 1.5 million species on Earth, there are 30 million species of insects alone so that the estimate would be globally higher [109].	> 30
Estimation by extrapolation from a review of knowledge of several factors: the structure of food webs, the relative abundance of species, the number of species and of individuals in different categories of body size, along with other determinants of the commonness and rarity of organisms [48].	10 – 50
Estimation based on re-analysis of Erwin’s calculation [109] (in essence agreed with it), with using additional data from Indonesian forests [111].	10 – 80
Estimation by extrapolation from a review of knowledge of several factors: estimated current levels of biodiversity, the fact that the life of the planet remains mostly unexplored at the species and infra-species levels, hyper-diversity in some of taxa that are mostly targeted by study, the relatively little effort that should be expanded on some highly diverse taxa, and extinction rate [112].	100

6.7 Chapter Summary

In this chapter our model is applied on well known real datasets (CoL and WoRMS), at which we estimated two main categories of species, the terrestrial and marine (including freshwater). We faced two challenges in specifying the model parametrization, and operating the adapted ABC algorithm due to the huge size of the given datasets. However, we overcame the former obstacle by referring to some resources in the field and performing several pilot experiments to determine and refine our parameter-choices. With respect to the latter obstacle, we had to alter the ABC algorithm by including additional threshold. By combining the

fitting results of the given two datasets (CoL and WoRMS), we also performed a global estimation for the number of species.

However, based on the CoL only, we performed some extra analysis on our model (in addition to the one that we made in Chapter 5). First, we tracked the model performance over truncated periods of time that are expected to make critical influence on shaping future prediction of our model. The analysis concluded that our model is still capable of making desirable results. Second, we attempted to validate our model in several ways: one by checking if the point estimation would generate a range of data that is similar to the one which is originally used to derive the point estimation itself, and the other by exploiting the Euclidean distances as residuals, and checking their distributions. We ended this chapter by making a comparison between our model and a previous work which is the NHRP model. We believe that this comparison showed powerful points in our model as explained in the last section.

Chapter 7

Discussion and Conclusions

The number of kinds/classes problem has been defined and extensively analysed in the literature (Chapter 2), and our proposed framework adds to it (Chapter 3). Our model is a synthesis in response to the analysis of this problem, which went through stages of an evaluation process (from Chapter 4, Chapter 5, to Chapter 6). This involves analysing, characterizing, validating, and implementing our proposed model through artificial and real datasets. The following sections highlight the main points of the current study:

7.1 Summary of the Study Analysis & Synthesis

The number of kinds/classes literature involved two perspectives in tackling this problem either through sampling actions and employing combinatorics rules or through class-discovery curves and employing extrapolation methods. These perspectives are provided by two main approaches: ‘*Sampling-Theoretic Approach*’ and ‘*Data Analytic Approach*’, respectively. The methods of the former approach deals with cross-sectional data points, while the methods of the latter approach extends the scope of the data type to also include longitudinal data points.

On the other hand, our proposal involves a comprehensive perspective in tackling this problem by integrating the above two approaches in one framework that comprises our target (*Sampling-Discovery process*) and a co-variate (*Latent Effort process*). Under the umbrella of both processes we have four factors: the class-abundances N_k which gives us a sampling probability for each class, sampling action (covered by d_i , t_i , and n_j , where d_i is the number of samples until facing i th new class, t_i is the time point at which the i th new class is discovered,

and n_j is the sample size collected per time unit j), discovery curve τ_j , and the employed effort in the discovery (covered by the latent effort l_j and its proxy x_j). Therefore, our model involves six variables with a set of parameters, in which (N_k, d_i, t_i) represent the cross-sectional variables, while (n_j, l_j, x_j) represent the longitudinal variables. Recall that τ_j is just another form of t_i , but on a longitudinal scale over the time units j .

The strength of our novel proposal is that it introduces an integrated trans-disciplinary approach for estimating the number of classes, at which Bayesian inference is used. However, this makes it subject to modelling and parameter specifications which might be a difficult task in some situations where our ignorance about the target problem exceeds our knowledge. Another trade-off of having a comprehensive approach, is dealing with a complex and high dimensional model. However, while the well-known techniques such as Bayes factor and MCMC were too difficult to use for our model, we found that the approximate Bayesian computation (ABC) can cope with such complexity and high dimensionality, and it is efficiently used to run our model. It is worth mentioning that throughout developing our ABC algorithm, we concretely found how it is beneficial to have a co-variate, not only as a main influence on our target variable, but to be practically employed in creating threshold values in order to build an efficient computation code for operating our model.

In the current study, we implemented our model in the ecology field where our target is to estimate the number of species C on a local scale (in certain communities such as marine species) or on a global scale (all species that exist on Earth). In this context, we adopted the following distributional assumptions that are guided by statistical theories, species literature, and computational convenience:

- Prior knowledge of $C \sim \text{Uniform}[D, mD]$,
where D is the current number of discoveries, and $(m > 1)$ is a scalar to specify the width of the prior support (domain).
- Species abundances $\{N_k\}_{k=1:C} \sim \text{i.i.d Lognormal}(\mu_N, \sigma_N)$.
- Inter-discovery samples $\{d_i\}_{i=1:D} \sim \text{joint Geometric}(1 - \sum_{k=1}^{i-1} N_k/N)$.

- Latent effort $\{L_j\}_{j=1:T} \sim$ Intrinsic Gaussian Markov process (L_{j-1}, σ_L) , where $L_j = \log(l_j)$, and L_1 is chosen to follow \sim Normal (μ_{L_1}, σ_L) .
- Sample sizes $\{n_j\}_{j=1:T} \sim$ Non-Homogeneous Poisson process($\exp(L_j)$), where T is the time length in years.
- Proxy of latent effort: Number of authors $\{x_j\}_{j=1:T} \sim$ Non-Homogeneous Poisson process($\exp(L_j)/\exp(\alpha)$), where $\exp(\alpha)$ is the expected number of specimens to be collected per author per year j , while $\exp(L_j)$ is the effort rate.
- Number of discoveries $\{\tau_j\}_{j=1:T}$ are derived as $\tau_j = \lfloor \sum_{\forall i} \mathbb{1}(j-1 < t_i \leq j) \rfloor_{\forall j}$, where $t_i \sim$ Degenerate distribution($\mathbb{1}(\delta_i \leq \sum_{r=1}^j n_r)$), and $\{\delta_i - \delta_{i-1}\} = d_i$. Recall that the parameter of the Degenerate distribution is just an indicator of satisfying the model constraints *i.e.* under holding the model constraints $(\delta_i \leq \sum_{r=1}^j n_r) \equiv (d_i \leq \sum_{r=1}^j n_r - \delta_{i-1})$, the distribution of $\{t_i\}_{i=1:D}$ will be $P(t_{1:D}|d_{1:D}, n_{1:T}) = 1$.

From the above six variables, $\{\tau_j\}_{j=1:T}$ and $\{x_j\}_{j=1:T}$ are given information represented in longitudinal datasets. In the current implementation of our model, we chose the above five parameters $\{\mu_N, \sigma_N, \mu_{L_1}, \sigma_L, \alpha\}$ to be fixed values. These values are already known in the artificial worlds, while in the real worlds they are determined by referring to species literature and specialist guidance along with several pilot experiments for tuning their selected values.

7.2 Main Findings of the Study Evaluation

In this section we list the main points concluded from evaluating the performance of our model while it was implemented on both real and artificial datasets. The artificial datasets are represented by three artificial worlds (World-A, World-B, and World-C) that allow to capture a wide range of scenarios for the model performance. The real datasets are represented by terrestrial, freshwater, and marine species datasets that are retrieved from two well known inventory systems of biodiversity (CoL and WoRMS). Through many experiments we made in the current study, we conclude the following:

7.2.1 Main findings from the artificial data experiments

- Our model showed accurate performance in estimating the number of species. However, ABC, as a simulation-based method, is subject to Monte Carlo error and two aspects of this were investigated: one aspect is related to the number of iterations used in the fitting stage of modelling the number of species from which we obtain the target estimations, while the other aspect is associated with the selected final sample size used for approximating the posterior distribution of the number of species and deriving the point and interval estimations. We found that our model provides consistent estimations with relatively small MC error and relatively small biased estimation when we choose a number of iterations ($iter = 1,000,000$) and a final sample size ($s = 1000$). With achieving this degree of consistency and accuracy, we believe that our model is reliable.
- Our model showed a moderate degree of robustness in the light of model misspecification. In the current study, the source of sensitivity mainly appeared against two parameters: the variation of the class-abundances (σ_N), and the expected number of items to be collected per author per time-unit on a log-scale ($\exp(\alpha)$). In the former parameter, the model provides a relatively over-estimation of the number of species associated with over-parametrization, and under-estimation associated with under-parametrization. It is an opposite case in the latter parameter, where the model provides a relatively over-estimation of the number of species associated with under-parametrization, and under-estimation associated with over-parametrization. However, there is a special case (in World-C) where the degree of sensitivity became smaller when there is a skewness associated with high proportion of discoveries. In general, caution is needed in selecting the values of model parameters especially for (σ_N) and ($\exp(\alpha)$). Caution can be interpreted through referring to target literature, specialist judgement, and pilot experiments.
- Our model showed a degree of additivity property. We can describe the additivity property on different degrees, where the complete additivity appears when the distributions of the integrated sub-estimations and the whole estimation are significantly equal (at which all the statistical properties of the two

are significantly equal), while the partial additivity appears when only some of the statistical properties of the integrated sub-estimations and the whole estimation are significantly equal. In the current study, we measured the model additivity through investigating its performance over two schemes (splitting and merging) of creating subgroups from the already available design of the three artificial worlds (where these worlds are used to represent the whole categories).

We found that the central tendency statistics that are obtained from the aggregation step are very close to the true values of the number of species (that are specified by the underlying artificial world) and to the estimated ones out of the whole category fitting. There are some differences in their standard deviations which justifies the differences in their CDFs curves and box-plots. These differences might be due to the correlation that exists between the given datasets (the number of discoveries and the number of authors) which cannot be avoided. However, when we used the KLD metric in two directions (one direction is when we considered the whole category distribution as the true model and the aggregated distribution as an approximation of it, while the other direction is when we considered the aggregated distribution as the true model and the whole category distribution as an approximation of it), we found a closeness between the results of the two directions of the KLD metric, indicating existence of additivity.

Although our model is not completely additive, it definitely provides alternative aggregated estimations that are equivalent to the whole category estimations, in which the point estimators are very close and the estimated 95% credible intervals of the aggregated distributions are either close or within the estimated 95% credible intervals of the whole category distributions. Therefore, we believe that it is a minimum risk if we adopted the aggregated estimation rather than the whole estimation, especially in cases when the sub-groups estimations are already made, and the whole estimation is actually demanding (in time, effort, and computation), as we witnessed in modelling the real datasets.

7.2.2 Main findings from the real data experiments

- Our model showed a consistent performance over time. It is worth noticing that time brings us two aspects of the information (accumulated knowledge and varying pattern). However, as our main assumption is that the number of species is a fixed-time variable, therefore, we expect two point: first, the fluctuating pattern of the number of discoveries and the number of authors should not have a great influence on estimating the number of species. Second, the accumulated knowledge should contribute to refining the estimation of the number of species. This seems to be what we achieved through the chronological analysis of our model.

The first point appeared in the fact that the total estimation of the number of species out of the 150-year period, C_{150}^* , does not exceed the actual recent estimation C_{240}^* , although in this period it seems that the number of discoveries is reaching a peak with increasing the number of authors, which may indicate that there are a lot more species left to be discovered. In addition, having decreasing curves at the last pattern of the 185-year period did not conclude a less estimation for the number of species than C_{150}^* , instead with including more information over time, C_{185}^* is getting closer to C_{240}^* regardless of the trough occurrence. This might be an indication that fitting our model using the ABC algorithm is getting more refined with the accumulated knowledge, but it is still robust against the big fluctuations in the varying pattern of the given data.

The second point appeared in the fact that the prediction of the truncated periods, C_{150}^* and C_{185}^* , do exceed the recent discoveries D_{240} , which meet our expectations that the prediction of a smaller previous period should at least reach the recent number of discoveries (definitely not less). In addition, the peak estimations of the 150-year and 185-year periods are still within the 95% credible interval of the 240-year time period, in spite of the specific differences associated with time.

- Our model showed a valid performance in the real data context. In the current study, we verified our model validation from two point of view: cross-sectional

validation and longitudinal validation. Since the conventional methods used to validate a model can not be applied in our case, we had to innovate a special way to validate our model, which is similar in principle to the residual diagnostic method of regression analysis. In the cross-sectional validation, we compared the usual method of our model fitting (in which we have a prior uniform distribution for C) with the situation when we replaced the prior distribution of C with the posterior point estimation, where the purpose is to check the validation of the point estimation of our model. We found that the distribution of the Euclidean distances (measured in a standardized scale) of the point-estimate fitting is identical to the distribution of the ones of the original fitting (the usual method). Moreover, the generated datasets out of the point-estimate fitting are as much as the ones out of the original fitting in capturing the general annual pattern of the given datasets, but narrower in enveloping this given data, which is in line with our expectation that the point estimate should generate data-series that are closer to the given one.

In the longitudinal validation, we replaced the prior distribution of C by the posterior point estimation of the truncated dataset (once by C_{150}^* and once by C_{185}^*). Then, we compared the point-estimate fitting using C_{240}^* with the point-estimate fitting using C_{150}^* , and with the point-estimate fitting using C_{185}^* . The comparison was based on two ways of computing the Euclidean distances: one was based on the whole-period data points, and the other was based on the last predicted part of the datasets (the last 90 data-points in the C_{150}^* case, and the last 55 data-points in the C_{185}^* case). We followed two treatments denoted by (FT) and (ST), explained in details in Section 6.3.2. In all of the treatments, we found that the distribution of the Euclidean distances (measured in a standardized scale) of the point-estimate fitting given C_{240}^* is the similar to the one given C_{150}^* , and also similar to the one given C_{185}^* . This is a reflection of a valid estimation. It is worth emphasizing that the purpose of our model is to predict the number of species (given the annual observed dataset, and considering all the other variables in the context of species discovery), not to predict the dataset itself in future time, and not to predict other variables in the context. We basically achieved this through implementing (FT) way, while the (ST) way is implemented just out of curiosity to predict the future

datasets which appeared to be acceptable results although it is not really a benchmark to judge our model performance.

- Our model over-performs the NHRP model in multiple aspects. Through implementing our model and the NHRP model on the same datasets (CoL and WoRMS), we found that our model is more reflective in describing the target problem including its context, more comprehensive in capturing the varying pattern of the given datasets, and more flexible and robust in dealing with large datasets. With respect to the global estimation, the NHRP model failed in performing a global estimation on the combined (CoL+WoRMS) dataset, while our model is absolutely capable to perform such estimation (this is proofed through some pilot experiments). However, due to the time limitation to continue this attempt and go through additional fitting experiments for the combined (CoL+WoRMS) dataset, we contented ourselves with making global estimation based on the extensive aggregation analysis that we accomplished on the artificial world setting. In conclusion, if we consider species estimates over the species literature that collectively ranges from 1.5 million to over 100 million species on Earth [13], we can see that our estimations lie within this range, but in a more reasonable scale.

7.3 Discussion & Suggestions for Future Work

Since our proposal is a comprehensive framework that covers multiple aspects (statistical/computation-wise and application-wise), it offers multiple research opportunities. To elaborate this statement, there are three aspects that shape our framework, and a fourth aspect that represents the context of applying our framework. The three aspects are: the structure of the model, the method of operating the model, and a decision/monitorization rule, see Figure 7.1. The structural aspect revolves around the number of variables affecting our target, the distributional assumptions of the adopted variables, the nature of the distributions' parameters, and the prior support of our target. With respect to the operational aspect, we adopted the ABC method which is subject to a distance metric, threshold, and also affected by the chosen prior support of our target.

The third aspect is the decision/monitorization rule which is related to two types of decisions (static and dynamic), and applied on the above two aspects. Focusing on the structure, the static decisions are concerned with specifying the number of variables and their distributional assumptions, also determining the sufficient nature of the adopted parameters (*e.g.* determining whether a parameter is sufficient to be fixed, or in form of a function, or it should be varying as a hyper-parameter). Such decisions need a collaboration between statisticians and specialists, in addition to pilot experiments. The pilot experiments can be made through simulation (artificial data), or through complete real databases that can be used as a learning data. Focusing on operating our model, the static decisions are concerned with specifying the distance metric and determining the nature and the number of the thresholds. This kind of decisions depends on the complexity of the structure of the adopted model, and also needs aid from pilot experimentation.

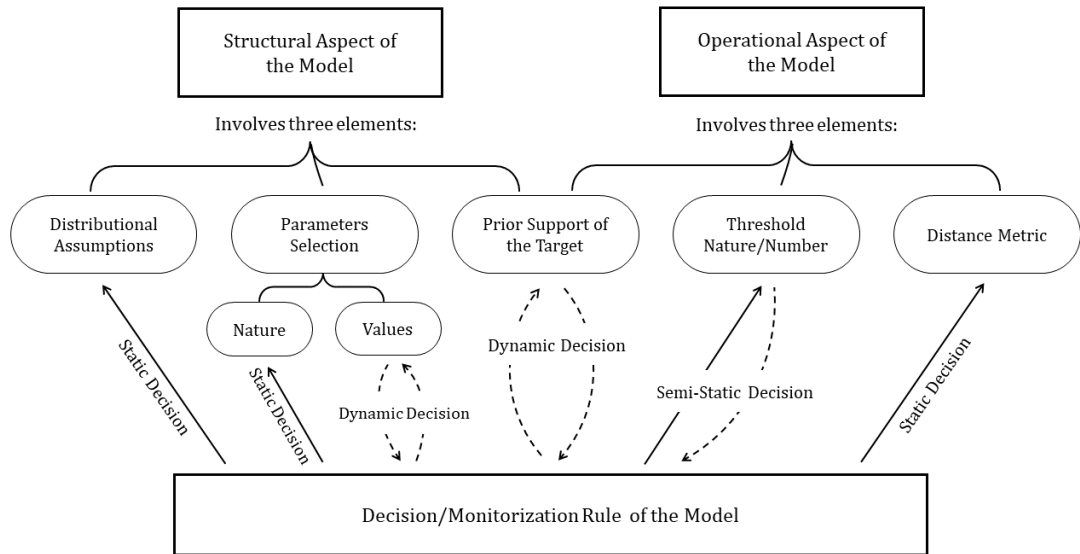


Figure 7.1: The three aspects of our proposed model. The solid arrows represent the static decisions that should be taken at the beginning of the modelling process, while the dashed arrows represent the dynamic decisions that are taken for tuning the model and enhancing its performance. Note that the semi-static decisions appear in a situation where we already specified the nature and number of the thresholds but for some specific occasions we need an additional filtering step.

Once the static decisions are made, the role of the dynamic decisions comes to monitor and tune the model performance. The dynamic decisions are concerned with determining the sufficient prior support of our target, and determining the

values of the adopted parameters. In some situations, some static decisions turn to be dynamic such as determining whether to keep a fixed parameter or change it to a hyper-parameter after multiple trials on a specific dataset, or determining the number of the thresholds (as we faced in the WoRMS and CoL data) where it is necessary to include additional threshold (filtering step) to overcome the memory overflow problem (explained in details in Section 6.1.2)

In the following we are listing some examples of the research opportunities under each aspect:

1. In the structural aspect:

- In the adopted version of our proposed model there are six variables; however, it might be more accurate to include additional variables or use some of them to replace existing ones. For example, in the species richness problem, the time factor might not be the only dominant influence on our target variable, and the geographic location factor might have a high contribution in this problem. The number of discovered species might be highly influenced by type-locality of the discovery locations. This variable can be considered as an additional influence on the discovery events along with the species abundances, or it can be considered as an influence factor on the species-abundance itself. It would be interesting to investigate this aspect.
- As mentioned earlier our model is subject to a variety of distributional assumptions that can be selected for different reasons. Hence, it would be interesting to explore other distributional specifications in the current subject (species richness). For example, for simplicity and computational convenience we chose the Uniform distribution as a prior knowledge of our target variable C , and the Geometric distribution for the inter-discovery samples d_i under the assumption of sampling with replacement so that the samples are equally likely to be drawn. It would be interesting to try different distributions such as the truncated Poisson as a prior distribution of C where the parameter rate $\lambda \geq D$, and Hyper-Geometric distribution for the discovery events where we can sample without replacement. In addition, the Lognormal distribution that we chose for the species abundances N_k is

one of the suggestions in the species literature, and exploring other distributions such as Poisson Lognormal might be interesting. With respect to the latent effort proxy (the number of authors) x_j and the sample sizes collected annually n_j , we chose them to be non-homogeneous Poisson processes, but to include more variability especially in the number of authors, it might be interesting to try the non-homogeneous renewal processes where the process mean evolves differently from its variance.

- Since our model is represented as a high dimensional Bayesian network, this allows for employing hyper-parametrization where the model parameters are treated as random variables rather than fixed values. The current application of our model includes five parameters, in which two of them (μ_N, σ_N) have a static and timeless influence *i.e.* these parameters ultimately govern the species-abundance distribution over all the time period, while the latent effort parameters (μ_{L_1}, σ_L) have a dynamic longitudinal influence through the Markov property where the mean μ_{L_1} affects the random variable L_1 and the value taken for L_1 will be used as a mean for the next L_2 and so on. On the other hand, the 5th parameter $\exp(\alpha)$ has a longitudinal influence on the number of authors but static at each time point. The inter-discovery samples and the sample sizes are dynamically parametrized by the species abundance and the latent effort variables, respectively, through the Bayesian network. However, we can increase the degree of dynamics in our model by including hyper-parameters which means making some of these parameters variables, especially σ_N and $\exp(\alpha)$ at which our model seems to be highly sensitive. We can choose σ_N to be a random variable from Uniform or Gamma distribution. With respect to $\exp(\alpha)$, we can capture more variability in the process governed by this parameter either by treating the parameter as an exponential function or as a random variable from Uniform or Exponential distributions.

However, when we adopt any of the suggested hyper-parametrization, different possible scenarios might occur. It could be possible to apply the same ABC algorithm that is adopted in the current study, in which we start to generate values for σ_N and $\exp(\alpha)$ from their specified distributions and feed them into the network so that each generated ABC sample has

different values of σ_N and $\exp(\alpha)$. This scenario might work well with a similar accepting rate of the generated datasets, but presumably there will be more variation in the estimated posterior of C due to adding more variation into the system through the hyper-parametrization. Another scenario that might happen is that the generated values of σ_N and $\exp(\alpha)$ are not consistent with the prior values of C so that the rejection rate becomes higher. In this case, we might end up with significantly increasing the number of iterations to be able to reach certain number of sufficiently accepted samples that contribute in estimating the posterior distribution of our target C , but this will probably lead to inefficient computation time. Or, we might end up with changing the way of applying the ABC algorithm either by including another filtering steps or using a different type of ABC such as ABC-PMC.

2. In the operational aspect:

- The adopted version of the ABC algorithm is based on the *Euclidean* distance metric. In the light of the complexity and high dimensionality of our model, we found difficulty to derive a sufficient statistic to be the base of the distance measure. Hence, we computed this metric based on the exact nature of the given annual datasets. It is worth mentioning that using the data in this way helped us in selecting and tuning our parameters' choices through noticing how the annual patterns of the generated datasets can capture the pattern of the given datasets (through the accepting/rejecting plots). However, using such distance metric had to be associated with two filtering steps: one is a threshold on the generated authors data, and the second is a specified percentage to select the accepted final total data. In addition, when applying our model on very large datasets such as the CoL and WoRMS, we had to include a third filtering step to overcome the memory overflow problem. Thus, it would be interesting to develop our ABC algorithm by investigating the possibility of deriving a sufficient statistic or building a special formula for our distance metric or threshold and, therefore, create fewer number of filtering steps.

- Based on a pilot experiment of the current study (see Group B.25 of Section B.6, Appendix B), we find that there are no significant differences between our adopted distance metric (*Euclidean*) and using other distance metrics such as *Manhattan* and *Minkowski*. However, as a part of developing our ABC algorithm, it might be interesting to investigate this aspect more with other distributional assumptions or on other research subjects to check its influence on the performance of our model.
- The adopted version of the ABC algorithm provided us with the parallelization process which is crucial in our high dimensional complex model. In the case of small datasets (as the ones that we used in the artificial worlds), the parallelization process helped us to make our model significantly efficient, while in case of large datasets (as the ones that we used in the real world), the parallelization process is in fact essential to make our model operate in the first place. However, as part of developing the operating method of our model, we would be interested in applying other types of the ABC algorithm such as ABC-PMC and ABC-SMC.

3. In the decisional aspect:

- In the current study, our dynamic decisions to decide whether our choices of the parameters' values are representative, are made through observing generated (accepting/rejecting) plots. Our rule is to check the closeness of the generated datasets from the given ones. The closeness is defined by that the accepted area of the generated data should cover the given data (so that the given data is roughly placed in the middle of the accepted area), and should capture the annual pattern of the given data. Obvious examples of violating this rule are faced in the trials of selecting the parameters' values of the WoRMS dataset, see right-side plots (a, b, c, and d) of Group B.17 in Appendix B. The action to take when violating the rule is to re-consider the chosen values or in the worst case changing the nature of the parameter (*e.g.* change it from being fixed to be hyper-parameter). However, our current decisions rule is still subjective, and it would be interesting to develop

a sort of automated formula or test to assess the parameter specification of our model. It is worth mentioning that even if we adopted the hyper-parametrization, we would face similar challenges or might be worse due to increasing the uncertainty level. Therefore, investigating the possibility of developing an objective decision rule regarding the parameters would be of great help in the case of hyper-parametrization.

- Another subjective decision rule is determining sufficient prior support of our target variable C . As we mentioned at the end of Section 2.3.4, we can run multiple pilot simulations to suggest an upper bound of C , and our current decision rule (to decide whether our suggestion is sufficient) is made through observing the plot of the estimated posterior distribution of C . The right-side tail of the distribution should be light (not heavy tail) such that it is obviously no point for exploring larger values than the suggested upper bound. However, it would be more convenient to derive an automated way such as adding some constraints to make this task easier. Nevertheless, we should emphasize that our proposed framework is meant to solve problems in critical research-areas. Therefore, in spite of our ambitions to operate our model in automated way, the judgement of a statistician and a specialist is still of a high contribution.

4. In the application aspect:

- With respect to the artificial worlds, in the current study, our model is applied on three different scenarios: one is when the discovery rate and the number of authors are following each other in a stationary pattern as appeared in World-A, and as an increasing pattern as shown in World-B. The third scenario is when the discovery rate is increasing while the number of authors is decreasing (shown in World-C). However, a fourth interesting scenario to be explored is when the discovery rate is increasing while the number of authors stays constant or stationary. However, this fourth scenario is not possible to be well modelled in the light of our current parameters specifications, because we adopted a fixed value for $\exp(\alpha)$ which

implies that the increase of the discovery rate is highly correlated with the increased effort that is interpreted by increasing the number of author (if $\exp(\alpha)$ is fixed), *i.e.* with this high correlation, it seems impossible to see a pattern in which the discovery rate increases and the authors numbers are stationary. Thus, making $\exp(\alpha)$ as an exponent function or as a hyperparameter will ensure seeing this fourth scenario which will imply that the increasing discovery rate is due to increasing the sampling action made by an active author (not due to increasing the number of authors in the field).

- With respect to the real world, in the current study, our model is applied on one of the important research-areas (species richness) that witnessed intensive and extensive research efforts. Although there are still many alternative ways in operating our model in the species richness field, as indicated above, it would be interesting to apply our model on other fields where we can deal with other types of data, such as data of software-bugs discovery in software engineering, where the latent effort proxy could be the number of users or licence holders registered with the software over the time that faults are being reported. Other interesting type of data is the oil-well discovery data in the petroleum industry, where oil-wells can be classified into distinguished types (classes) according to some major distinct characteristics, and the latent effort proxy could be the amount of funding spent on the discovery process *etc.*

Bibliography

- [1] M.J. Costello, R.M. May, and N.E. Stork. Can We Name Earth’s Species Before They Go Extinct? *science*, 339(6118):413–416, 2013.
- [2] C. Mora, D.P. Tittensor, Sina Adl, A.G.B. Simpson, and B. Worm. How Many Species Are There on Earth and in the Ocean? *PLoS Biol*, 9(8):e1001127, 2011.
- [3] S.P. Wilson and F.J. Samaniego. Nonparametric Analysis of the Order-Statistic Model in Software Reliability. *IEEE Transactions on Software Engineering*, 33(3):198–208, 2007.
- [4] X. Yu, J. Liu, Z. Yang, X. Jia, Q. Ling, and S. Ye. Learning from Imbalanced Data for Predicting the Number of Software Defects. In *2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)*, pages 78–89, 2017.
- [5] N.D. Singpurwalla and S.P. Wilson. *Statistical Methods in Software Engineering: Reliability and Risk*. Springer Science & Business Media, 2012.
- [6] B. Efron and R. Tibshirani. Estimating the Number of Unseen Species: How Many Words Did Shakespeare Know? *Biometrika*, 63(3):435–447, 1976.
- [7] D.R. McNeil. Estimating an Author’s Vocabulary. *Journal of the American Statistical Association*, 68(341):92–96, 1973.
- [8] J. Bunge and M. Fitzpatrick. Estimating the Number of Species: A Review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.
- [9] L.A. Goodman. On the Estimation of the Number of Classes in a Population. *The Annals of Mathematical Statistics*, 20(4):572–579, 1949.
- [10] D.P. Bebbler, F.H.C. Marriott, K.J. Gaston, S.A. Harris, and R.W. Scotland. Predicting Unknown Species Numbers Using Discovery Curves. *Proceedings*

- of the Royal Society of London B: Biological Sciences*, 274(1618):1651–1658, 2007.
- [11] M.J. Costello, C.S. Emblow, and B.E. Picton. Long Term Trends in the Discovery of Marine Species New to Science Which Occur in Britain And Ireland. *Journal of the Marine Biological Association of the United Kingdom*, 76(1):255–257, 1996.
- [12] K.J. Gaston. Body Size and Probability of Description: the Beetle Fauna of Britain. *Ecological Entomology*, 16(4):505–508, 1991.
- [13] M.J. Costello, S.P. Wilson, and B. Houlding. Predicting Total Global Species Richness Using Rates of Species Description And Estimates of Taxonomic Effort. *Systematic Biology*, 61(5):871–883, 2012.
- [14] S.P. Wilson and S.Ó. Ríordáin. Optimal Software Testing across Version Releases. In R.S. Kenett, F. Ruggeri, and F.W. Faltin, editors, *Analytic Methods in Systems and Software Testing*, page 65. Wiley Online Library, 2018.
- [15] D.J. Martin and K.S. Loomis. *Building Teachers: A Constructivist Approach to Introducing Education*. Cengage Learning, 2013.
- [16] A.H. Maslow. A Theory of Human Motivation. *Psychological Review*, 50(4):370, 1943.
- [17] H. Elderfield. *The Oceans and Marine Geochemistry*. Elsevier, 2006.
- [18] S.K. Dash. Marine Microbial Biodiversity: Current Approaches. In C.S.K. Mishra and A.A. Juwarkar, editors, *Environmental Biotechnology*, pages 443–457. APH Publishing, 2007.
- [19] M.J. Costello, M. Lane, S.P. Wilson, and B Houlding. Factors Influencing When Species are First Named and Estimating Global Species Richness. *Global Ecology and Conservation*, 4:243–254, 2015.
- [20] M.J. Costello. Developing Species Information Systems: The European Register of Marine Species (ERMS). *Oceanography*, 13(3):48–55, 2000.

- [21] M.J. Costello and S.P. Wilson. Predicting the Number of Known and Unknown Species in European Seas Using Rates of Description. *Global Ecology and Biogeography*, 20(2):319–330, 2011.
- [22] O. Frank. Estimation of the Number of Connected Components in a Graph by Using a Sampled Subgraph. *Scandinavian Journal of Statistics*, 5(4):177–188, 1978.
- [23] M. Knott. Models for Cataloguing Problems. *The Annals of Mathematical Statistics*, 38(4):1255–1260, 1967.
- [24] W.C. Hou and G. Ozsoyoglu. Statistical Estimators for Aggregate Relational Algebra Queries. *ACM Transactions on Database Systems (TODS)*, 16(4):600–654, 1991.
- [25] W.C. Hou, G. Ozsoyoglu, and B.K. Taneja. Statistical Estimators for Relational Algebra Expressions. In SIGACT, SIGMOD, and SIGART, editors, *Proceedings of the Seventh ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 276–287. ACM Press, Austin, Texas, 1988.
- [26] A. Shlosser. On Estimation of the Size of the Dictionary of a Long Text on the Basis of a Sample. *Engineering Cybernetics*, 19(1):97–102, 1981.
- [27] M. Harwit. Cosmic Discovery: The Search, Scope, and Heritage of Astronomy. *Brighton: Harvester Press, 1981*, 1981.
- [28] W.W. Esty. Estimation of the Number of Classes in a Population and the Coverage of a Sample. *Mathematical Scientist*, 10:41–50, 1985.
- [29] V.M. Kalinin. Functionals Related to the Poisson Distribution and Statistical Structure of a Text. In Ju.V. Linnik, editor, *Articles on Mathematical Statistics and the Theory of Probability*, pages 202–220, 1965.
- [30] H.S. Sichel. The GIGP Distribution Model with Applications to Physics Literature. *Czechoslovak Journal of Physics*, 36(1):133–137, 1986.
- [31] H.S. Sichel. Parameter Estimation for a Word Frequency Distribution Based on Occupancy Theory. *Communications in Statistics-Theory and Methods*, 15(3):935–949, 1986.

- [32] H.S. Sichel. Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist*, 11(1):45–72, 1986.
- [33] A. Chao. Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11(4):265–270, 1984.
- [34] M.T. Chao. From Animal Trapping to Type-Token. *Statistica Sinica*, 2:189–201, 1992.
- [35] A. Chao and S.M. Lee. Estimating the Number of Classes via Sample Coverage. *Journal of the American statistical Association*, 87(417):210–217, 1992.
- [36] R.A. Fisher, A.S. Corbet, and C.B. Williams. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *The Journal of Animal Ecology*, 12(1):42–58, 1943.
- [37] J.K. Ord and G.A. Whitmore. The Poisson-Inverse Gaussian Distribution as a Model for Species Abundance. *Communications in Statistics-Theory and Methods*, 15(3):853–871, 1986.
- [38] D. Zelterman. Robust Estimation in Truncated Discrete Distributions with Application to Capture-Recapture Experiments. *Journal of Statistical Planning and Inference*, 18(2):225–237, 1988.
- [39] K.P. Burnham and W.S. Overton. Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals. *Biometrika*, 65(3):625–633, 1978.
- [40] K.P. Burnham and W.S. Overton. Robust Estimation of Population Size When Capture Probabilities Vary Among Animals. *Ecology*, 60(5):927–936, 1979.
- [41] A. Chao. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, 43(4):783–791, 1987.
- [42] A.D. Carothers. Capture-Recapture Methods Applied to A Population with Known Parameters. *The Journal of Animal Ecology*, pages 125–146, 1973.

- [43] B. Brainerd. On the Relation Between Types and Tokens in Literary Text. *Journal of Applied Probability*, 9(3):507–518, 1972.
- [44] P. de Caprariis, R.H. Lindemann, and C.M. Collins. A Method for Determining Optimum Sample Size in Species Diversity Studies. *Journal of the International Association for Mathematical Geology*, 8(5):575–581, 1976.
- [45] B.M. Hill. Posterior Moments of the Number of Species in a Finite Population and the Posterior Probability of Finding a New Species. *Journal of the American Statistical Association*, 74(367):668–673, 1979.
- [46] B.M. Hill. Invariance and Robustness of the Posterior Distribution of Characteristics of a Finite Population, with Reference to Contingency Tables and the Sampling of Species. *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 383–395, 1980.
- [47] W.A. Lewins and D.N. Joanes. Bayesian Estimation of the Number of Species. *Biometrics*, pages 323–328, 1984.
- [48] R.M. May. How Many Species Are There on Earth? *Science*, 241(4872):1441–1449, 1988.
- [49] M.J. Gibbons, A.J. Richardson, M.V. Angel, E. Buecher, G. Esnal, M.A. Fernandez Alamo, R. Gibson, H. Itoh, P. Pugh, R. Boettger-Schnack, and E. Thuesen. What Determines the Likelihood of Species Discovery in Marine Holozooplankton: Is Size, Range or Depth Important? *Oikos*, 109(3):567–576, 2005.
- [50] F.A. Zapata and D. Ross-Robertson. How Many Species of Shore Fishes are There in the Tropical Eastern Pacific? *Journal of Biogeography*, 34(1):38–51, 2007.
- [51] T.M. Blackburn and K.J. Gaston. A Critical Assessment of the Form of the Interspecific Relationship Between Abundance and Body Size in Animals. *Journal of Animal Ecology*, 66(2):233–249, 1997.
- [52] R.A. Cooper, P.A. Maxwell, J.S. Crampton, A.G. Beu, C.M. Jones, and B.A. Marshall. Completeness of the Fossil Record: Estimating Losses Due to Small Body Size. *Geology*, 34(4):241–244, 2006.

- [53] B.J. Finlay, J.O. Corliss, G. Esteban, and T. Fenchel. Biodiversity at the Microbial Level: The Number of Free-Living Ciliates in the Biosphere. *The Quarterly Review of Biology*, 71(2):221–237, 1996.
- [54] S.J. Hall and S.P. Greenstreet. Global Diversity and Body Size. *Nature*, 383(6596):133–133, 1996.
- [55] C.D.L. Orme, D.L.J. Quicke, J.M. Cook, and A. Purvis. Body Size Does Not Predict Species Richness Among the Metazoan Phyla. *Journal of Evolutionary Biology*, 15(2):235–247, 2002.
- [56] P.M. Hammond. Species Inventory. In WCMC and B. Groombridge, editors, *Global Biodiversity: Status of the Earth's Living Resources*, pages 17–39. Chapman & Hall, 1992.
- [57] M.J. Costello, M. Coll, R. Danovaro, P. Halpin, H. Ojaveer, and P. Miloslavich. A Census of Marine Biodiversity Knowledge, Resources, and Future Challenges. *PloS one*, 5(8):e12110, 2010.
- [58] J.J. Sepkoski. Alpha, Beta, or Gamma: Where Does All the Diversity Go? *Paleobiology*, 14(3):221–234, 1988.
- [59] P.J.D. Lamshead and G. Boucher. Marine Nematode Deep-Sea Biodiversity-Hyperdiverse or Hype? *Journal of Biogeography*, 30(4):475–485, 2003.
- [60] N.E. Stork. Biodiversity: World of Insects. *Nature*, 448(7154):657, 2007.
- [61] K. Dolphin and D.L.J. Quicke. Estimating the Global Species Richness of an Incompletely Described Taxon: An Example Using Parasitoid Wasps (Hymenoptera: Braconidae). *Biological Journal of the Linnean Society*, 73(3):279–286, 2001.
- [62] P.M. Hammond. Practical Approaches to the Estimation of the Extent of Biodiversity in Speciose Groups. *Phil. Trans. R. Soc. Lond. B*, 345(1311):119–136, 1994.
- [63] A.R. Solow and W.K. Smith. On Estimating the Number of Species From the Discovery Record. *Proceedings of the Royal Society of London B: Biological Sciences*, 272(1560):285–287, 2005.

- [64] S.P. Wilson and M.J. Costello. Predicting Future Discoveries of European Marine Species by Using a Non-Homogeneous Renewal Process. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(5):897–918, 2005.
- [65] M.J. Costello. *European Register of Marine Species: A Check-List of the Marine Species in Europe and a Bibliography of Guides to Their Identification*. Paris: Museum National D’histoire Naturelle, 2001.
- [66] A. Gelman, H.S. Stern, J.B. Carlin, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.
- [67] T.A. Stephenson. An introduction to bayesian network theory and usage. Technical report, IDIAP Research Report, Dalle Molle Institute for Perceptual Artificial Intelligence, Switzerland, 2000.
- [68] A.F.M. Smith. Bayesian Computational Methods. *Philosophical Transactions of the Royal Society of London. Series A: Physical and Engineering Sciences*, 337(1647):369–386, 1991.
- [69] P. Kythe and P. Puri. *Computational Methods for Linear Integral Equations*. Springer Science & Business Media, 2011.
- [70] Yu.V. Prokhorov and V. Statulevicius. *Limit Theorems of Probability Theory*, volume 6. Springer Science & Business Media, 2000.
- [71] M. Chiachio, J.L. Beck, J. Chiachio, and G. Rus. Approximate Bayesian Computation by Subset Simulation. *SIAM Journal on Scientific Computing*, 36(3):A1339–A1358, 2014.
- [72] I. Hazra, M.D. Pandey, and N. Manzana. Approximate Bayesian Computation (ABC) Method for Estimating Parameters of the Gamma Process using Noisy Data. *Reliability Engineering & System Safety*, 198:106780, 2020.
- [73] B.M. Turner and T. Van-Zandt. A Tutorial on Approximate Bayesian Computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.
- [74] M.A. Beaumont. Approximate bayesian computation. *Annual review of statistics and its application*, 6:379–403, 2019.

- [75] B. Everett. *An Introduction to Latent Variable Models*. Springer Science & Business Media, 2013.
- [76] I.J. Busch-Vishniac. *Electromechanical Sensors and Actuators*. Springer Science & Business Media, 1998.
- [77] J.A. Mee, J.R. Post, H. Ward, K.L. Wilson, E. Newton, and A. Cantin. Interaction of Ecological and Angler Processes: Experimental Stocking in An Open Access, Spatially Structured Fishery. *Ecological Applications*, 26(6):1693–1707, 2016.
- [78] W.S. Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Pearson Addison Wesley, 2006.
- [79] E. Baldrige, D.J. Harris, X. Xiao, and E.P. White. An Extensive Comparison of Species-Abundance Distribution Models. *PeerJ*, 4:e2823, 2016.
- [80] S. Thurner, R. Hanel, B. Liu, and B. Corominas-Murtra. Understanding Zipf’s Law of Word Frequencies Through Sample-Space Collapse in Sentence Formation. *Journal of the Royal Society Interface*, 12(108):20150330, 2015.
- [81] K. Gao and T.M. Khoshgoftaar. A Comprehensive Empirical Study of Count Models for Software Fault Prediction. *IEEE Transactions on Reliability*, 56(2):223–236, 2007.
- [82] P.J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711–732, 1995.
- [83] R.E. Kass and A.E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.
- [84] C.S. Bos. A Comparison of Marginal Likelihood Computation Methods. In *Compstat*, pages 111–116. Springer, 2002.
- [85] CoL Editorial Board (2020). *Catalogue of Life*. Available from <http://www.catalogueoflife.org>, Accessed 2020-04-01.
- [86] WoRMS Editorial Board (2020). *World Register of Marine Species*. Available from <http://www.marinespecies.org>, Accessed 2020-04-01.

- [87] F.W. Preston. The Commonness and Rarity of Species. *Ecology*, 29(3):254–283, 1948.
- [88] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC press, 2005.
- [89] S.M. Edie, P.D. Smits, and D. Jablonski. Probabilistic Models of Species Discovery and Biodiversity Comparisons. *Proceedings of the National Academy of Sciences*, 114(14):3666–3671, 2017.
- [90] C.E. Gehlke and K. Biehl. Certain Effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association*, 29(185A):169–170, 1934.
- [91] W. Appeltans et al. The Magnitude of Global Marine Species Diversity. *Current Biology : CB*, 22(23):2189–2202, 2012.
- [92] J.E. Winston. Systematics and Marine Conservation. *Systematics, Ecology, and the Biodiversity Crisis*. Columbia University Press. New York, pages 144–168, 1992.
- [93] P. Bouchet. The Magnitude of Marine Biodiversity. *The Exploration of Marine Biodiversity: Scientific and Technological Challenges*, pages 31–62, 2006.
- [94] D.L. Hawksworth. The Fungal Dimension of Biodiversity: Magnitude, Significance, and Conservation. *Mycological Research*, 95(6):641–655, 1991.
- [95] I.D. Hodkinson and D. Casson. A Lesser Predilection for Bugs: Hemiptera (Insecta) Diversity in Tropical Rain Forests. *Biological Journal of the Linnean Society*, 43(2):101–109, 1991.
- [96] K.J. Gaston. The Magnitude of Global Insect Species Richness. *Conservation biology*, 5(3):283–296, 1991.
- [97] G.C.B. Poore and G.D.F. Wilson. Marine Species Richness. *Nature*, 361(6413):597–598, 1993.

- [98] V. Novotny, Y. Basset, S.E. Miller, G.D. Weiblen, B. Bremer, L. Cizek, and P. Drozd. Low Host Specificity of Herbivorous Insects in A Tropical Forest. *Nature*, 416(6883):841–844, 2002.
- [99] N. Stork and K. Gaston. Counting Species One by One. *New Scientist*, 127(1729):43–47, 1990.
- [100] K.J. Gaston. Regional Numbers of Insect and Plant Species. *Functional Ecology*, pages 243–247, 1992.
- [101] C.W. Sabrosky. How Many Insects Are There? *Systematic Zoology*, 2(1):31–36, 1953.
- [102] R.M. May. How Many Species? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 330(1257):293–304, 1990.
- [103] J.F. Grassle and N.J. Maciolek. Deep-Sea Species Richness: Regional and Local Diversity Estimates from Quantitative Bottom Samples. *The American Naturalist*, 139(2):313–341, 1992.
- [104] A.J. Hamilton, Y. Basset, K.K. Benke, P.S. Grimbacher, S.E. Miller, V. Novotny, G. A. Samuelson, N.E. Stork, G.D. Weiblen, and J.D.L. Yen. Quantifying Uncertainty in Estimation of Tropical Arthropod Species Richness. *The American Naturalist*, 176(1):90–95, 2010.
- [105] A.J. Hamilton, Y. Basset, K.K. Benke, P.S. Grimbacher, S.E. Miller, V. Novotny, G. A. Samuelson, N.E. Stork, G.D. Weiblen, and J.D.L. Yen. Erratum: Quantifying Uncertainty in Estimation of Tropical Arthropod Species Richness (American Naturalist 176 (90-95)). *American Naturalist*, 177(4):544–545, 2011.
- [106] A.J. Hamilton, V. Novotny, E.K. Waters, Y. Basset, K.K. Benke, P.S. Grimbacher, S.E. Miller, G.A. Samuelson, G.D. Weiblen, J.D.L. Yen, et al. Estimating Global Arthropod Species Richness: Refining Probabilistic Models Using Probability Bounds Analysis. *Oecologia*, 171(2):357–365, 2013.
- [107] G. Lima-Mendez, J. Van Helden, A. Toussaint, and R. Leplae. Prophinder: A Computational Tool for Prophage Prediction in Prokaryotic Genomes. *Bioinformatics*, 24(6):863–865, 2008.

- [108] F. Rohwer. Global Phage Diversity. *Cell*, 113(2):141, 2003.
- [109] T.L. Erwin. Tropical Forests: Their Richness in Coleoptera and Other Arthropod Species. *The Coleopterists Bulletin*, 1982.
- [110] P.H. Raven. Disappearing Species, A Global Tragedy. *Futurist*, 19(5):8–14, 1985.
- [111] N.E. Stork. Insect Diversity: Facts, Fiction and Speculation. *Biological journal of the Linnean Society*, 35(4):321–337, 1988.
- [112] P.R. Ehrlich et al. Biodiversity Studies: Science and Policy. *Science*, 253(5021):758–762, 1991.

Appendix A

Mathematical Equations

The current appendix includes three sections. In the first section, we introduce some equations related to Chapter 2, Section 2.2.1. In the second section, we introduce the equations of the proposed model of Chapter 4, Section 4.1. The last section includes some equations related to Chapter 6, Section 6.1.1. These equations explain the derivation of the standard deviation value of the species abundances, as suggested by [87].

A.1 Some Equations from the Literature Review

Some of the estimators are explained in Chapter 2, particularly in Section 2.1.1 but without explicitly showing their formulas. However, they are listed below in case they are subject of interest. The general setting of these equations is already explained at the beginning of Section 2.1.1. Here are notations that are commonly used for all these equations: n is the size of the sample drawn from a population of size N and partitioned into C classes with proportions p_i . Theoretically speaking, this sample should represent the population such that $n = \sum_{i=1}^C n_i$, where n_i is the number of the sampled items from the i th class. The total number of classes in the sample is $c = \sum_{j=1}^n c_j$, where c_j is the number of the classes involving j items in the drawn sample.

$$\hat{C}_{Goodman1} = c + \sum_{j=1}^n (-1)^{j+1} \left[\frac{(N - n + j - 1)!(n - j)!}{(N - n - 1)!n!} \right] c_j \quad (\text{A.1})$$

$$\hat{C}_{Shlosser} = c + c_1 \left[\sum_{i=1}^n i q (1 - q)^{i-1} c_i \right]^{-1} \sum_{j=1}^n (1 - q)^j c_j \quad (\text{A.2})$$

where q is the limit for the sampling fraction: $n/N \rightarrow q \in (0, 1)$.

$$\hat{C}_{Goodman2} = c + \sum_{j=1}^n (-1)^{j+1} \left[\frac{1-p}{p} \right]^j c_j \quad (\text{A.3})$$

where p is the probability of being in sample n , and it is same for all classes.

$$\hat{C}_{Sichel} = \frac{2}{\hat{\theta}_1 \hat{\theta}_2} \quad (\text{A.4})$$

where $\hat{\theta}_1$ and $\hat{\theta}_2$ are the parameters of the Inverse Gaussian proposed as an approximate PDF for p_i 's.

$$\hat{C}_{Chao1} = c + \frac{c_1^2}{2c_2} \quad (\text{A.5})$$

$$\hat{C}_{Chao2} = \frac{c}{\hat{u}_{Good}} + \left[\frac{n(1 - \hat{u}_{Good})}{\hat{u}_{Good}} \right] \hat{\gamma}^2 ; \quad \hat{u}_{Good} = 1 - \frac{c_1}{n} \quad (\text{A.6})$$

where $\hat{\gamma}$ is the estimated coefficient of variance, and \hat{u}_{Good} is the Good's estimate for coverage.

$$\hat{C}_{Poisson} = \frac{c}{1 - \widehat{P_0(F)}} \quad (\text{A.7})$$

where $\widehat{P_0(F)}$ is the probability that an arbitrary class will occur zero times in the drawn sample, which follows Poisson distribution (governed by a rate parameter that is described by a distribution F).

A.2 Equation Explanation of the Proposed Model

Here is explanation of how we moved from equation (4.6) to equation (4.7) in Chapter 4, Section 4.1:

The final mathematical formulation of the full joint probability distribution of our model within the species context is given by,

$$P(C, \mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}, \mathbf{x}, \mathbf{t}) =$$

$$P(C) \times P(\mathbf{N} | C) \times P(\mathbf{d} | \mathbf{N}) \times P(\mathbf{L}) \times P(\mathbf{n} | \mathbf{L}) \times P(\mathbf{x} | \mathbf{L}) \times P(\mathbf{t} | \mathbf{d}, \mathbf{n}) =$$

$$\begin{aligned}
& \left\{ \frac{1}{mD - D} \right\} \times \left\{ \prod_{k=1}^C \frac{1}{N_k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[\frac{-1}{2} \left(\frac{\log(N_k) - \mu}{\sigma} \right)^2 \right] \right\} \times \\
& \left\{ \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \right\} \times \left\{ \prod_{j=2}^T \frac{1}{\sqrt{2\pi\sigma_L^2}} \exp \left[\frac{-1}{2} \left(\frac{L_j - L_{j-1}}{\sigma_L} \right)^2 \right] \right\} \times \\
& \left\{ \prod_{j=1}^T [\exp(L_j)]^{n_j} \frac{\exp[-\exp(L_j)]}{n_j!} \right\} \times \left\{ \prod_{j=1}^T \left[\frac{\exp(L_j)}{\exp(\alpha)} \right]^{x_j} \exp \left[-\frac{\exp(L_j)}{\exp(\alpha)} \right] / x_j! \right\} \times \{ 1 \}
\end{aligned} \tag{A.8}$$

By excluding the fixed values from the above expression, the full joint distribution can be proportionally expressed as follows,

$$\begin{aligned}
P(C, \mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}, \mathbf{x}, \mathbf{t}) \propto & \left\{ \prod_{k=1}^C \frac{1}{N_k} \exp \left[\frac{-1}{2} \left(\frac{\log(N_k) - \mu}{\sigma} \right)^2 \right] \right\} \times \left\{ \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \right\} \times \\
& \left\{ \prod_{j=2}^T \exp \left[\frac{-1}{2} \left(\frac{L_j - L_{j-1}}{\sigma_L} \right)^2 \right] \right\} \times \left\{ \prod_{j=1}^T [\exp(L_j)]^{n_j} \frac{\exp[-\exp(L_j)]}{n_j!} \right\} \times \\
& \left\{ \prod_{j=1}^T \left[\frac{\exp(L_j)}{\exp(\alpha)} \right]^{x_j} \exp \left[-\frac{\exp(L_j)}{\exp(\alpha)} \right] / x_j! \right\}
\end{aligned} \tag{A.9}$$

■ Simplifying the 1st term from equation (A.9),

$$\prod_{k=1}^C \frac{1}{N_k} \exp \left[\frac{-1}{2} \left(\frac{\log(N_k) - \mu}{\sigma} \right)^2 \right] = \left(\prod_{k=1}^C \frac{1}{N_k} \right) \exp \left[\frac{-1}{2\sigma^2} \sum_{k=1}^C (\log(N_k) - \mu)^2 \right] \tag{A.10}$$

■ Simplifying the 3rd term from equation (A.9),

$$\prod_{j=2}^T \exp \left[\frac{-1}{2} \left(\frac{L_j - L_{j-1}}{\sigma_L} \right)^2 \right] = \exp \left[\frac{-1}{2\sigma_L^2} \sum_{j=2}^T (L_j - L_{j-1})^2 \right] \tag{A.11}$$

■ Simplifying the 4th term from equation (A.9),

$$\begin{aligned}
\prod_{j=1}^T [\exp(L_j)]^{n_j} \frac{\exp[-\exp(L_j)]}{n_j!} &= \prod_{j=1}^T \frac{1}{n_j!} \exp(n_j L_j) \exp[-\exp(L_j)] = \\
& \left(\prod_{j=1}^T \frac{1}{n_j!} \right) \exp \left[\sum_{j=1}^T n_j L_j - \sum_{j=1}^T \exp(L_j) \right]
\end{aligned} \tag{A.12}$$

■ Simplifying the 5th term from equation (A.9),

$$\begin{aligned} \prod_{j=1}^T \left[\frac{\exp(L_j)}{\exp(\alpha)} \right]^{x_j} \exp \left[-\frac{\exp(L_j)}{\exp(\alpha)} \right] / x_j! &= \prod_{j=1}^T \frac{1}{x_j!} \exp [x_j(L_j - \alpha)] \exp [-\exp(L_j - \alpha)] = \\ \left(\prod_{j=1}^T \frac{1}{x_j!} \right) \exp \left[\sum_{j=1}^T x_j(L_j - \alpha) - \sum_{j=1}^T \exp(L_j - \alpha) \right] &\quad (A.13) \end{aligned}$$

By substituting the 1st, 3rd, 4th, and 5th terms of equation (A.9) with equations (A.10), (A.11), (A.12), and (A.13), respectively, and rearranging some terms, we reach the final proportional formula of the full joint distribution,

$$P(C, \mathbf{N}, \mathbf{d}, \mathbf{L}, \mathbf{n}, \mathbf{x}, \mathbf{t}) \propto$$

$$\begin{aligned} &\left(\prod_{k=1}^C \frac{1}{N_k} \right) \left(\prod_{j=1}^T \frac{1}{x_j!} \right) \left(\prod_{j=1}^T \frac{1}{n_j!} \right) \left\{ \prod_{i=2}^D \left(\frac{\sum_{k=1}^{i-1} N_k}{N} \right)^{d_i-1} \left(1 - \frac{\sum_{k=1}^{i-1} N_k}{N} \right) \right\} \\ &\exp \left\{ \frac{-1}{2\sigma^2} \sum_{k=1}^C (\log(N_k) - \mu)^2 + \frac{-1}{2\sigma_L^2} \sum_{j=2}^T (L_j - L_{j-1})^2 + \sum_{j=1}^T (n_j + x_j)L_j - \alpha x_j - (1 + e^{-\alpha})e^{L_j} \right\} \quad (A.14) \end{aligned}$$

A.3 Explanation of the Species Abundance Variation Provided by [87]

[87] found that on a log scale (base 2), species abundances followed a normal distribution, at which he described the species abundances “*commonness*” as a relative concept so that we can say one species is twice as common as another *etc.* He suggested that a natural series of groups representing commonness would run as a sequence of octaves of frequency. He justified his proposal as: “*It is the most natural grouping possible, but our ultimate justification for adopting it is not its naturalness, but the fact that it works. It is a geometric, or logarithmic series, and this paper produces evidence that commonness of species appears to be a simple*

Gaussian curve on a geometric base (i.e. “lognormal” curve)”. He formulated the octave Gaussian curve as follows,

$$n = n_0 e^{-(aR)^2} \quad (\text{A.15})$$

where ‘ n_0 ’ is the number of species in the octave, ‘ R ’ is the number of octaves to left or right of the mode of the curve, ‘ n ’ is the number in an octave distance R octaves from the mode, and ‘ a ’ is a constant to be calculated from the experimental evidence, which was noticed to be close to the value 0.2.

He explained that theoretically the curve extends to infinity so that the total number of species theoretically available for observation can be expressed as,

$$N = \int_{-\infty}^{+\infty} n dR = n_0 \frac{\sqrt{\pi}}{a} \quad (\text{A.16})$$

By substituting the term n in equation (A.16) by its value in equation (A.15), and re-write the result as an integral of the exponential part of the Gaussian distribution ($\mu = 0, \sigma$) as shown in equation (A.18), we can derive the formula of the standard deviation of species abundances:

$$\int_{-\infty}^{+\infty} e^{-(aR)^2} dR = \sqrt{\pi} \left(\frac{1}{a} \right) \quad (\text{A.17})$$

$$\int_{-\infty}^{+\infty} e^{-\left(\frac{x}{\sigma\sqrt{2}}\right)^2} dx = \sqrt{\pi} \left(\sigma\sqrt{2} \right) \quad (\text{A.18})$$

Thus, the standard deviation of species abundances is given by: $\sigma = 1/a\sqrt{2}$, and by using the approximate value $a = 0.2$, the estimated standard deviation on a log (base 2) scale is $\sigma = 3.54$.

By using the transformation formula $\{\log_e(x) = \log_2(x)/\log_2(e)\}$, we got our estimated value of the standard deviation of species abundances ($\sigma = 3.54/\log_2(e) = 2.45$) on a natural logarithmic scale.

Appendix B

Groups of Tables and Plots

The current appendix consists of six sections related to the artificial and real data experiments. It involves mixed groups of plots and tables, at which each group is presented in a manner that gives the reader a holistic view to allow for making comparisons and noticing the related numerical values. Therefore, the used caption in this appendix is titled by the name ‘Group’.

B.1 Simulation Results Related to Chapter 4

Group B.1: Summary information (design and results) of the three artificial worlds.

Artificial World-A with number of Species $C = 10,000$

World Parameterization		Fitting Results	
$C \in [D, 5D]$	$C \in [4,468 - 22,340]$	Posterior Mean	10,320
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	9,970
$\theta_L = \{L_1, \sigma_L\}$	{5, 0.1}	95% Credible Interval	[7,222 - 14,163]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	1,850

Artificial World-B with number of Species $C = 100,000$

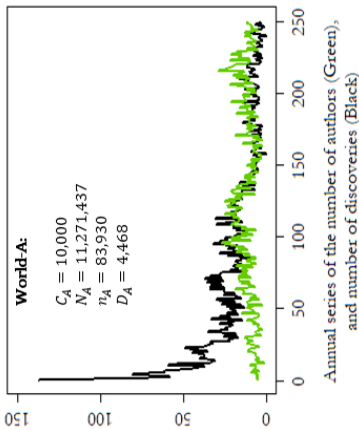
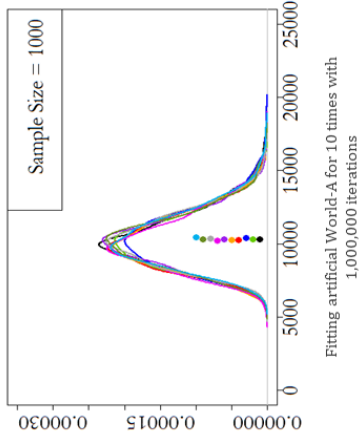
World Parameterization		Fitting Results	
$C \in [D, 8D]$	$C \in [31,865 - 254,920]$	Posterior Mean	109,793
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	97,160
$\theta_L = \{L_1, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[48,927 - 188,961]
$\theta_x = \alpha$	$\log(100)$	Standard Deviation	37,053

Artificial World-C with number of Species $C = 50,000$

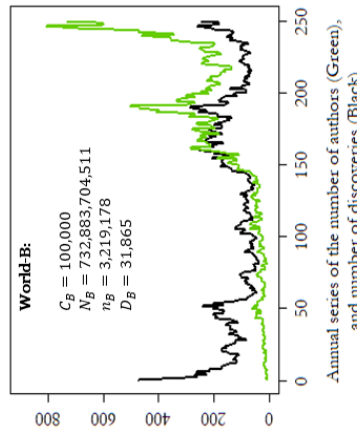
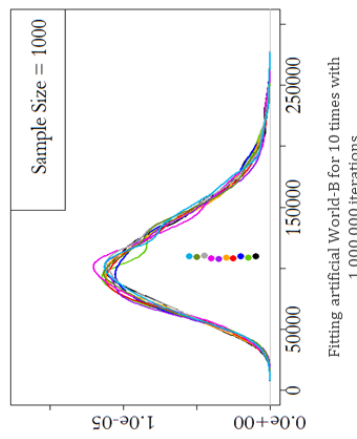
World Parameterization		Fitting Results	
$C \in [D, 3D]$	$C \in [41,620 - 124,860]$	Posterior Mean	53,161
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	48,969
$\theta_L = \{L_1, \sigma_L\}$	{7, 0.08}	95% Credible Interval	[42,280 - 69,765]
$\theta_x = \alpha$	$\log(50)$	Standard Deviation	7,703

B.2 Accuracy Results Related to Chapter 5

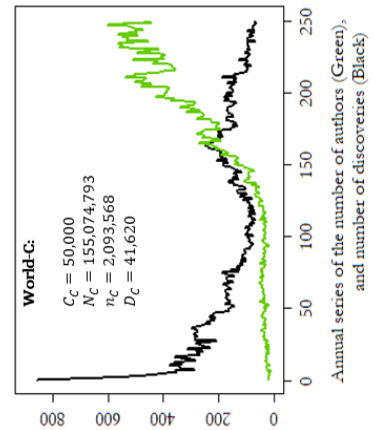
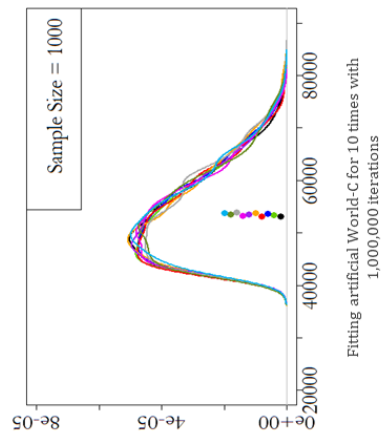
	Mean	Mode	95% CI	SD
Fitting 1	10,320	9,970	[7,222 – 14,163]	1,850
Fitting 2	10,294	10,235	[7,269 – 14,260]	1,801
Fitting 3	10,389	10,165	[7,237 – 14,605]	1,918
Fitting 4	10,265	9,779	[7,244 – 14,171]	1,758
Fitting 5	10,260	9,748	[7,084 – 13,996]	1,782
Fitting 6	10,311	10,267	[7,172 – 13,959]	1,774
Fitting 7	10,221	9,668	[7,147 – 14,300]	1,875
Fitting 8	10,347	10,119	[7,461 – 13,942]	1,729
Fitting 9	10,317	10,383	[7,233 – 14,258]	1,792
Fitting 10	10,433	9,751	[7,318 – 14,391]	1,828



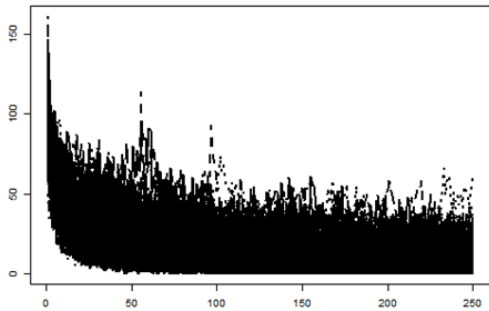
	Mean	Mode	95% CI	SD
Fitting 1	109,793	97,160	[48,927 – 188,961]	37,053
Fitting 2	108,716	92,217	[49,786 – 192,040]	36,506
Fitting 3	109,741	94,710	[51,320 – 192,761]	36,753
Fitting 4	108,176	91,059	[48,539 – 190,156]	35,909
Fitting 5	108,543	95,521	[48,355 – 189,696]	36,051
Fitting 6	107,473	100,919	[50,966 – 183,659]	35,183
Fitting 7	108,426	101,202	[49,664 – 185,502]	35,543
Fitting 8	110,483	104,203	[51,329 – 188,449]	35,885
Fitting 9	109,063	95,888	[52,964 – 187,719]	35,348
Fitting 10	109,824	101,613	[52,793 – 182,097]	35,736



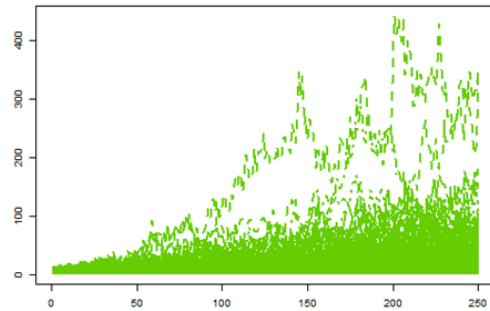
	Mean	Mode	95% CI	SD
Fitting 1	53,161	48,969	[42,280 – 69,765]	7,703
Fitting 2	53,324	46,036	[42,177 – 70,724]	7,897
Fitting 3	53,596	47,494	[42,182 – 70,567]	7,908
Fitting 4	53,129	46,225	[42,076 – 69,735]	7,755
Fitting 5	53,826	48,712	[42,422 – 71,807]	8,178
Fitting 6	53,474	47,759	[42,438 – 70,907]	7,800
Fitting 7	53,303	50,167	[42,176 – 70,770]	7,717
Fitting 8	53,949	47,631	[42,366 – 71,579]	8,187
Fitting 9	53,486	51,305	[42,028 – 72,093]	7,954
Fitting 10	53,771	49,736	[42,182 – 70,967]	7,867



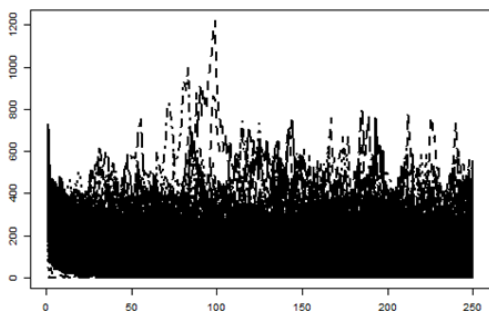
Group B.2: Summary of the accuracy measure (1st aspect of MC error) over the three artificial worlds.



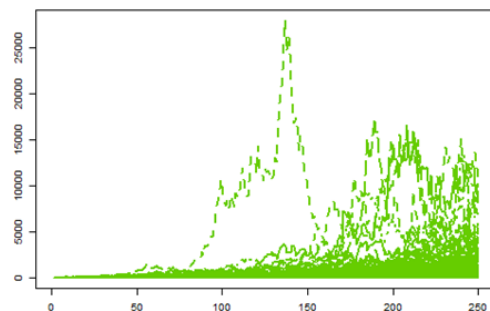
Annual data of the number of discoveries of 500 artificial worlds of design World-A with generating different seeds, over 250 years.



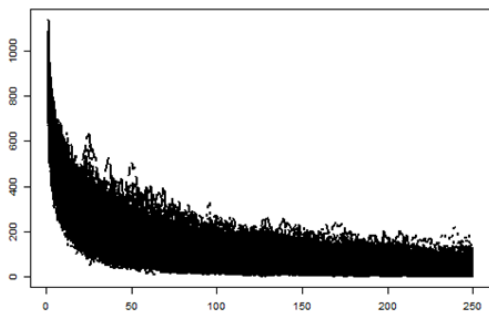
Annual data of the number of authors of 500 artificial worlds of design World-A with generating different seeds, over 250 years.



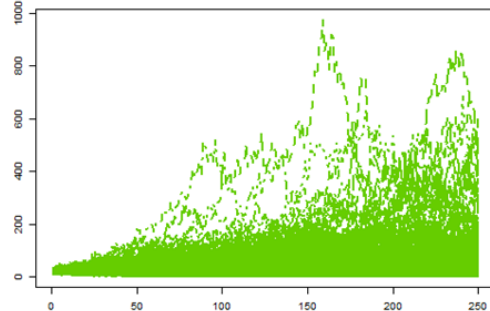
Annual data of the number of discoveries of 500 artificial worlds of design World-B with generating different seeds, over 250 years.



Annual data of the number of authors of 500 artificial worlds of design World-B with generating different seeds, over 250 years.



Annual data of the number of discoveries of 500 artificial worlds of design World-C with generating different seeds, over 250 years.



Annual data of the number of authors of 500 artificial worlds of design World-C with generating different seeds, over 250 years.

Group B.3: The 500 generated datasets of the three artificial worlds. In the current number of discoveries D , the coefficient of variation (CV) is higher in World-B at which $CV = 60\%$, while $CV = 30\%$ in World-A and $CV = 20\%$ in World-C.

Group B.4: Summary (statistics) of the accuracy measure (2nd aspect of MC error) over the three artificial worlds.

500 artificial worlds of World-A design with number of Species $C = 10,000$

Iterations NO.	Average of Post. Means	SD of Post. Means	95% CI of Post. Means	Average of Post. Modes	SD of Post. Modes	95% CI of Post. Modes
100,000	10,220	1,598	[6,625 – 12,738]	9,598	1,511	[6,652 – 12,412]
500,000	10,109	1,373	[6,868 – 12,257]	9,677	1,330	[7,114 – 12,151]
1,000,000	10,083	1,321	[6,944 – 12,176]	9,763	1,278	[7,204 – 12,246]
1,500,000	10,066	1,295	[7,040 – 12,138]	9,746	1,232	[7,233 – 12,122]

500 artificial worlds of World-B design with number of Species $C = 100,000$

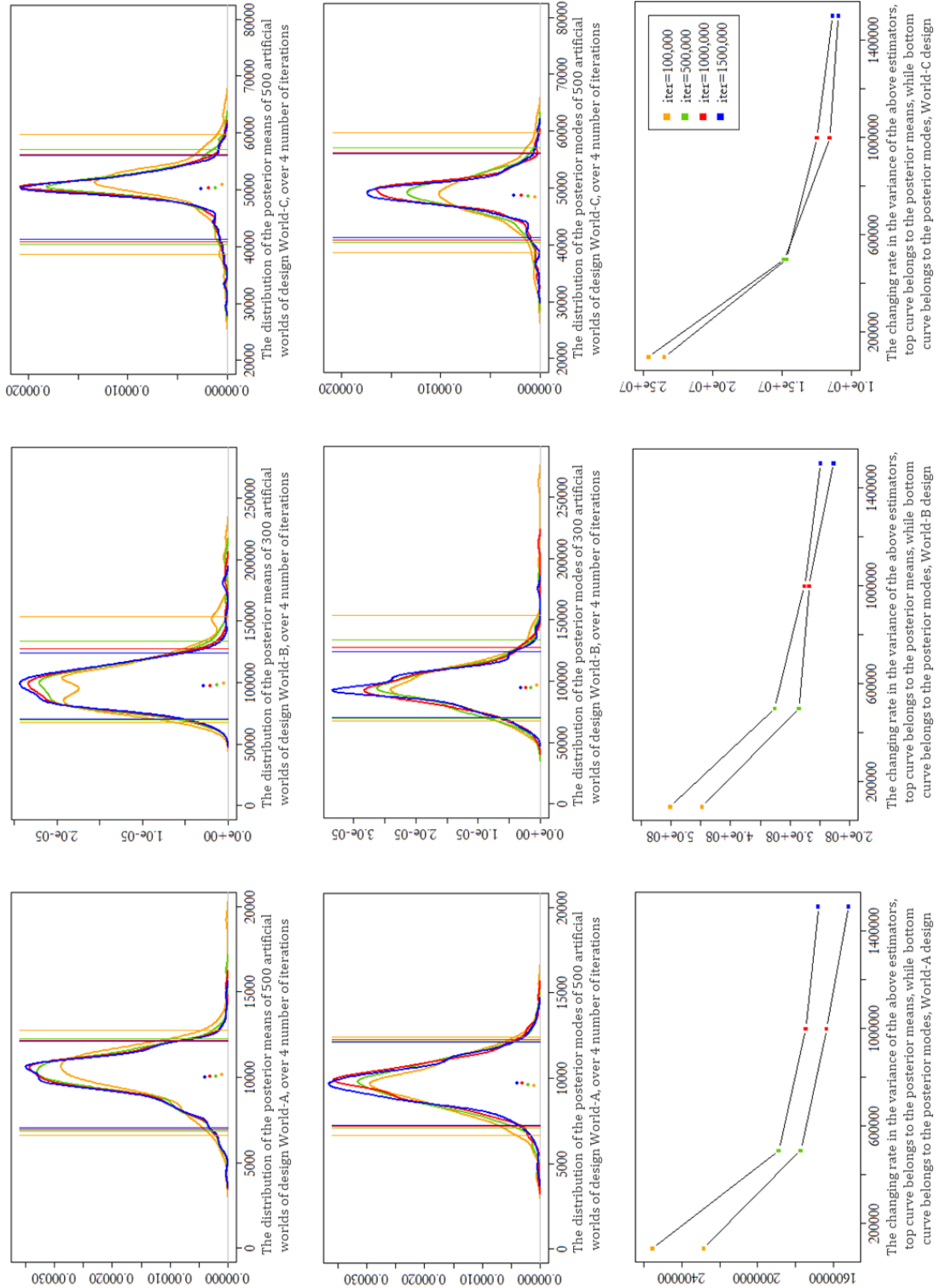
Iterations NO.	Average of Post. Means	SD of Post. Means	95% CI of Post. Means	Average of Post. Modes	SD of Post. Modes	95% CI of Post. Modes
100,000	80,474	30,708	[30,003 – 147,904]	81,461	27,363	[33,400 – 129,697]
500,000	80,470	27,712	[31,103 – 128,353]	80,777	24,256	[34,761 – 123,622]
1,000,000	80,525	26,794	[31,518 – 124,089]	81,174	24,109	[34,578 – 123,401]
1,500,000	80,509	26,258	[31,708 – 122,514]	81,229	23,562	[34,601 – 122,502]

300 artificial worlds of World-B design with number of Species $C = 100,000$

Iterations NO.	Average of Post. Means	SD of Post. Means	95% CI of Post. Means	Average of Post. Modes	SD of Post. Modes	95% CI of Post. Modes
100,000	99,820	22,389	[67,949 – 153,563]	97,154	21,165	[62,515 – 139,344]
500,000	98,490	18,037	[69,837 – 133,502]	95,082	16,863	[64,789 – 127,713]
1,000,000	98,125	16,601	[70,640 – 127,537]	95,492	16,366	[67,097 – 128,248]
1,500,000	97,864	15,754	[70,561 – 124,163]	95,479	15,048	[65,485 – 124,566]

500 artificial worlds of World-C design with number of Species $C = 50,000$

Iterations NO.	Average of Post. Means	SD of Post. Means	95% CI of Post. Means	Average of Post. Modes	SD of Post. Modes	95% CI of Post. Modes
100,000	50,978	4,847	[38,710 – 59,690]	48,516	4,962	[35,859 – 57,792]
500,000	50,494	3,831	[40,459 – 57,144]	48,679	3,854	[39,542 – 55,574]
1,000,000	50,383	3,531	[40,850 – 56,185]	48,776	3,397	[40,086 – 54,283]
1,500,000	50,336	3,368	[41,289 – 56,068]	48,853	3,305	[41,034 – 55,003]



Group B.5: Summary (plots) of the accuracy measure (2nd aspect of MC error) over the three artificial worlds.

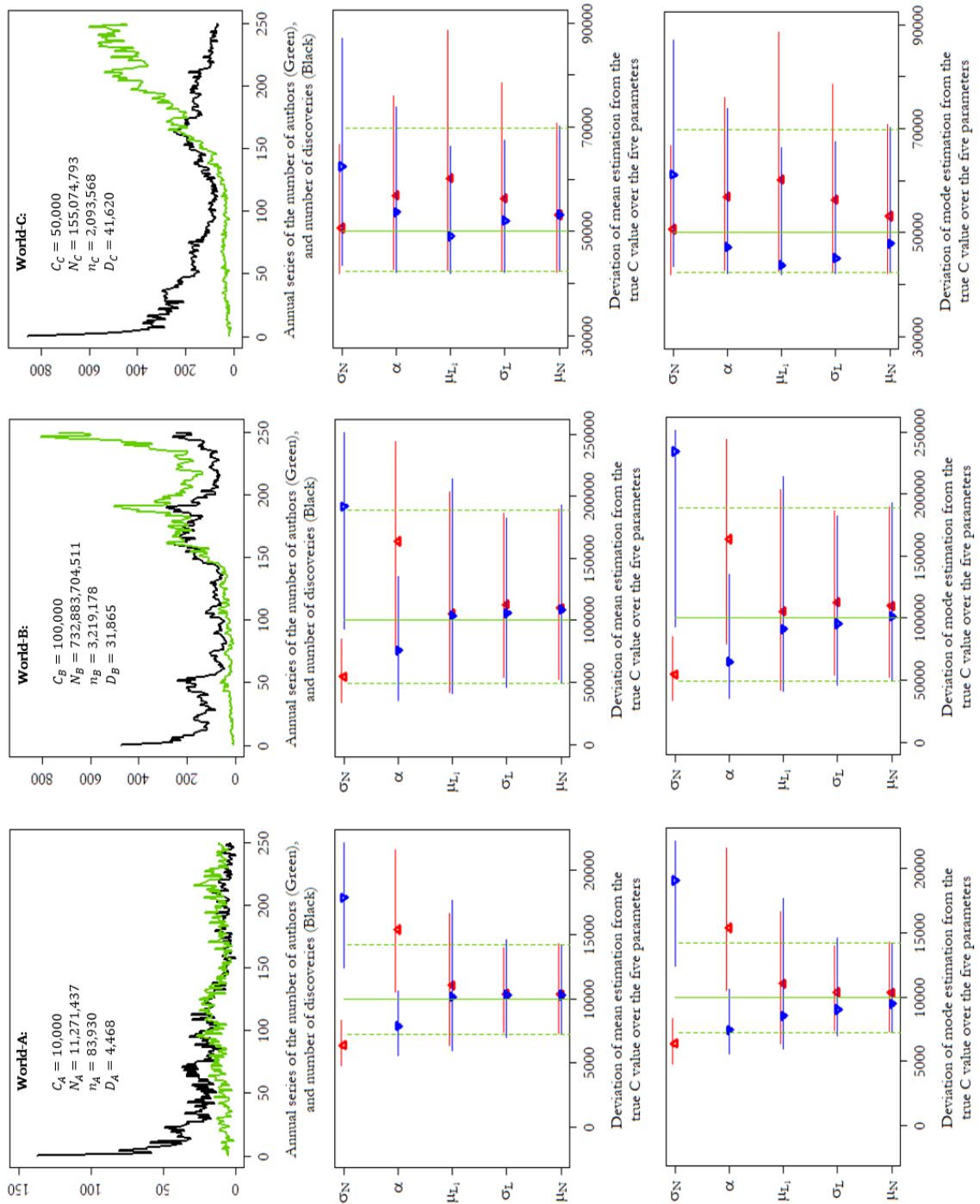
B.3 Sensitivity Results Related to Chapter 5

Group B.6: Summary (statistics) of the sensitivity measure over the three artificial worlds.

20% Under/Over Parameterization in World-A	Post. Means	Post. Mode	95% Credible Interval
$N_k \sim \text{LogNormal}(\mu = 3.2, \sigma = 2.5)$	10,322	10,182	[7,285 - 14,250]
$N_k \sim \text{LogNormal}(\mu = 4.8, \sigma = 2.5)$	10,273	9,467	[7,189 - 14,080]
$N_k \sim \text{LogNormal}(\mu = 4, \sigma = 2)$	6,397	6,179	[4,781 - 8,319]
$N_k \sim \text{LogNormal}(\mu = 4, \sigma = 3)$	17,835	19,038	[12,394 - 22,140]
$L_1 \sim \text{Normal}(\mu_{L_1} = 4, \sigma_L = 0.1)$	11,038	10,806	[6,355 - 16,648]
$L_1 \sim \text{Normal}(\mu_{L_1} = 6, \sigma_L = 0.1)$	10,149	8,515	[5,995 - 17,662]
$L_1 \sim \text{Normal}(\mu_{L_1} = 5, \sigma_L = 0.08)$	10,359	10,562	[7,394 - 13,992]
$L_1 \sim \text{Normal}(\mu_{L_1} = 5, \sigma_L = 0.12)$	10,297	9,024	[7,015 - 14,570]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(15.2)))$	15,375	14,759	[10,487 - 21,523]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(59.2)))$	7,869	7,428	[5,600 - 10,584]

20% Under/Over Parameterization in World-B	Post. Means	Post. Mode	95% Credible Interval
$N_k \sim \text{LogNormal}(\mu = 8, \sigma = 3.5)$	109,924	95,536	[52,048 - 190,148]
$N_k \sim \text{LogNormal}(\mu = 12, \sigma = 3.5)$	108,885	101,511	[50,168 - 193,199]
$N_k \sim \text{LogNormal}(\mu = 10, \sigma = 2.8)$	54,536	50,599	[33,633 - 85,082]
$N_k \sim \text{LogNormal}(\mu = 10, \sigma = 4.2)$	191,970	234,506	[93,469 - 251,989]
$L_1 \sim \text{Normal}(\mu_{L_1} = 5.6, \sigma_L = 0.15)$	105,516	96,451	[41,470 - 203,861]
$L_1 \sim \text{Normal}(\mu_{L_1} = 8.4, \sigma_L = 0.15)$	104,036	91,045	[41,299 - 214,037]
$L_1 \sim \text{Normal}(\mu_{L_1} = 7, \sigma_L = 0.12)$	112,726	99,149	[53,657 - 186,204]
$L_1 \sim \text{Normal}(\mu_{L_1} = 7, \sigma_L = 0.18)$	105,681	95,418	[46,104 - 182,774]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(39.8)))$	163,752	183,404	[79,266 - 243,988]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(251.2)))$	75,887	64,961	[35,675 - 135,361]

20% Under/Over Parameterization in World-C	Post. Means	Post. Mode	95% Credible Interval
$N_k \sim \text{LogNormal}(\mu = 4.8, \sigma = 2)$	53,139	46,247	[42,076 - 70,746]
$N_k \sim \text{LogNormal}(\mu = 7.2, \sigma = 2)$	53,192	47,809	[42,289 - 70,101]
$N_k \sim \text{LogNormal}(\mu = 6, \sigma = 1.6)$	50,625	44,073	[41,942 - 66,728]
$N_k \sim \text{LogNormal}(\mu = 6, \sigma = 2.4)$	62,427	61,127	[43,374 - 87,074]
$L_1 \sim \text{Normal}(\mu_{L_1} = 5.6, \sigma_L = 0.08)$	60,149	47,996	[42,432 - 88,570]
$L_1 \sim \text{Normal}(\mu_{L_1} = 8.4, \sigma_L = 0.08)$	49,051	43,673	[41,848 - 66,367]
$L_1 \sim \text{Normal}(\mu_{L_1} = 7, \sigma_L = 0.064)$	56,278	48,185	[42,279 - 78,493]
$L_1 \sim \text{Normal}(\mu_{L_1} = 7, \sigma_L = 0.096)$	52,012	45,072	[42,032 - 67,522]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(22.9)))$	56,898	58,679	[42,650 - 75,879]
$x_j \sim \text{Poisson}(\exp(L_j)/\exp(\log(109.3)))$	53,664	47,210	[42,117 - 73,791]



Group B.7: Summary (plots) of the sensitivity measure over the three artificial worlds. In the middle and bottom panels, the solid green line represents the true value of C , while the dashed green lines represent the 95% credible interval of C . The (blue) dots along with the blue horizontal lines represent the point estimations given the over-parametrization and their 95% credible intervals, respectively, while the (red) dots along with the red horizontal lines represent the point estimations given the under-parametrization and their 95% credible intervals, respectively.

B.4 Additivity Results Related to Chapter 5

Group B.8: Summary of the parametrizations and the fitting results of the subgroups (S1 and S2) out of the Splitting Scheme, and subgroups (M1_T(1) and M2_T(1)) out of the Merging Scheme - Treatment (1) of World-A.

Taxon S1 split from Artificial World-A, with number of Species $C_{S1} = 7,000$

World Parameterization		Fitting Results	
$C_{S1} \in [D, 5D]$	$C_{S1} \in [3,180 - 15,900]$	Posterior Mean	7,285
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	7,060
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{4.64, 0.094}	95% Credible Interval	[4,972 - 10,207]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	1,316

Taxon S2 split from Artificial World-A, with number of Species $C_{S2} = 3,000$

World Parameterization		Fitting Results	
$C_{S2} \in [D, 5D]$	$C_{S2} \in [1,288 - 6,440]$	Posterior Mean	3,151
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	2,937
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{3.82, 0.089}	95% Credible Interval	[1,938 - 5,001]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	789

Taxon M1_T(1) generated as World-A (Seed=1), with number of Species $C_{M1_T(1)} = 7,000$

World Parameterization, Seed=1		Fitting Results	
$C_{M1_T(1)} \in [D, 5D]$	$C_{M1_T(1)} \in [3,236 - 16,180]$	Posterior Mean	7,720
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	7,557
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{4.64, 0.094}	95% Credible Interval	[5,368 - 10,755]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	1,375

Taxon M2_T(1) generated as World-A (Seed=1), with number of Species $C_{M2_T(1)} = 3,000$

World Parameterization, Seed=1		Fitting Results	
$C_{M2_T(1)} \in [D, 5D]$	$C_{M2_T(1)} \in [1,212 - 6,060]$	Posterior Mean	3,019
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	2,679
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{3.82, 0.089}	95% Credible Interval	[1,865 - 5,078]
$\theta_x = \alpha$	$\log(30)$	Standard Deviation	811

Group B.9: Summary of the parametrizations and the fitting results of the subgroups (M1.T(2) and M2.T(2)) out of the Merging Scheme - Treatment (2), and subgroups (M1.T(3) and M2.T(3)) out of the Merging Scheme - Treatment (3) of World-A.

Taxon M1_T(2) generated as World-A (Seed=1), with number of Species $C_{M1_T(2)} = 7,000$

World Parameterization, Seed=94		Fitting Results	
$C_{M1_T(2)} \in [D, 5D]$	$C_{M1_T(2)} \in [3,407 - 17,035]$	Posterior Mean	7,157
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	6,790
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	95% Credible Interval	[4,766 - 10,075]
$\theta_x = \alpha$	log(30)	Standard Deviation	1,374

Taxon M2_T(2) generated as World-A (Seed=1), with number of Species $C_{M2_T(2)} = 3,000$

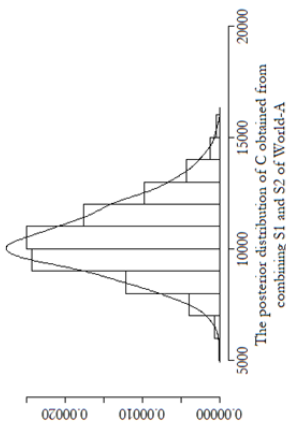
World Parameterization, Seed=76		Fitting Results	
$C_{M2_T(2)} \in [D, 5D]$	$C_{M2_T(2)} \in [1,517 - 7,585]$	Posterior Mean	3,360
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	3,168
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	95% Credible Interval	[2,203 - 4,964]
$\theta_x = \alpha$	log(30)	Standard Deviation	694

Taxon M1_T(3) generated as World-A (Seed=1), with number of Species $C_{M1_T(3)} = 7,000$

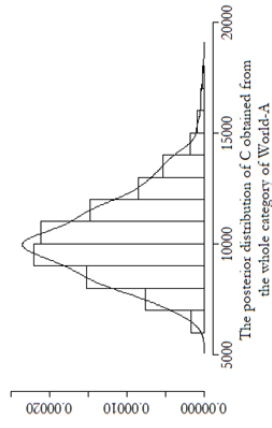
World Parameterization, Seed=1		Fitting Results	
$C_{M1_T(3)} \in [D, 5D]$	$C_{M1_T(3)} \in [3,429 - 17,145]$	Posterior Mean	7,162
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	6,958
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	95% Credible Interval	[4,460 - 10,709]
$\theta_x = \alpha$	log(30)	Standard Deviation	1,671

Taxon M2_T(3) generated as World-A (Seed=1), with number of Species $C_{M2_T(3)} = 3,000$

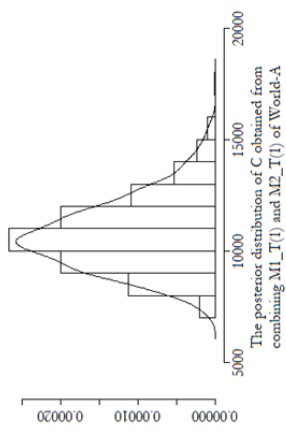
World Parameterization, Seed=1		Fitting Results	
$C_{M2_T(3)} \in [D, 5D]$	$C_{M2_T(3)} \in [1,737 - 8,685]$	Posterior Mean	3,234
$\theta_N = \{\mu, \sigma\}$	{4, 2.5}	Posterior Mode	2,715
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{5, 0.1}	95% Credible Interval	[1,841 - 5,314]
$\theta_x = \alpha$	log(30)	Standard Deviation	951



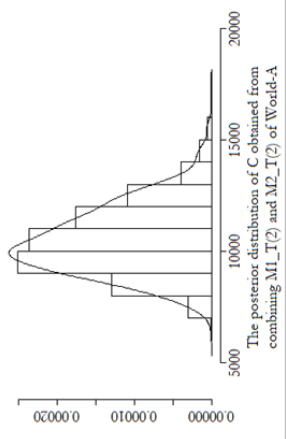
Aggregating Results of S1 and S2	
Current (D_{S1+S2})	4,468
Posterior Mean	10,436
Posterior Mode	10,020
95% Credible Interval	[7,682 – 13,779]
Standard Deviation	1,537
KLD($S1+S2$ Whole)	0.003163



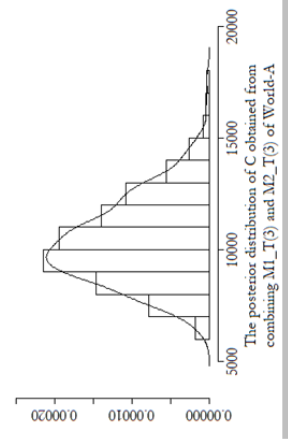
Fitting Results of Whole Category (World-A)	
Current (D)	4,468
Posterior Mean	10,320
Posterior Mode	9,970
95% Credible Interval	[7,222 – 14,163]
Standard Deviation	1,850
KLD(Whole S1+S2)	0.003035
KLD(Whole M1_T(1)+M2_T(1))	0.001066
KLD(Whole M1_T(2)+M2_T(2))	0.003933
KLD(Whole M1_T(3)+M2_T(3))	0.002981



Aggregating of M1_T(1) and M2_T(1)	
Current ($D_{M1_T(1)+M2_T(1)}$)	4,448
Posterior Mean	10,738
Posterior Mode	10,409
95% Credible Interval	[8,079 – 14,330]
Standard Deviation	1,606
KLD($M1_T(1)+M2_T(1)$ Whole)	0.001061



Aggregating of M1_T(2) and M2_T(2)	
Current ($D_{M1_T(2)+M2_T(2)}$)	4,924
Posterior Mean	10,517
Posterior Mode	9,911
95% Credible Interval	[7,883 – 13,913]
Standard Deviation	1,551
KLD($M1_T(2)+M2_T(2)$ Whole)	0.003968



Aggregating of M1_T(3) and M2_T(3)	
Current ($D_{M1_T(3)+M2_T(3)}$)	5,166
Posterior Mean	10,397
Posterior Mode	9,645
95% Credible Interval	[7,162 – 14,421]
Standard Deviation	1,919
KLD($M1_T(3)+M2_T(3)$ Whole)	0.002918

Group B.10: Summary of the fitting results of the whole category (World-A) against all combined taxa: ($S1+S2$), ($M1_T(1) + M2_T(1)$), ($M1_T(2) + M2_T(2)$), and ($M1_T(3) + M2_T(3)$).

Group B.11: Summary of the parametrizations and the fitting results of the subgroups (S1 and S2) out of the Splitting Scheme, and subgroups (M1.T(1) and M2.T(1)) out of the Merging Scheme - Treatment (1) of World-B.

Taxon S1 split from Artificial World-B, with number of Species $C_{S1} = 60,000$

World Parameterization		Fitting Results	
$C_{S1} \in [D, 5D]$	$C_{S1} \in [19,153 - 153,224]$	Posterior Mean	65,612
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	61,829
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.49, 0.146}	95% Credible Interval	[30,128 - 118,744]
$\theta_x = \alpha$	log(100)	Standard Deviation	22,557

Taxon S2 split from Artificial World-B, with number of Species $C_{S2} = 40,000$

World Parameterization		Fitting Results	
$C_{S2} \in [D, 5D]$	$C_{S2} \in [12,712 - 101,696]$	Posterior Mean	41,941
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	40,336
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.08, 0.142}	95% Credible Interval	[19,079 - 76,494]
$\theta_x = \alpha$	log(100)	Standard Deviation	14,378

Taxon M1_T(1) generated as World-B (Seed= -1365400193), with $C_{M1,T(1)} = 60,000$

Parameterization, Seed= -1365400193		Fitting Results	
$C_{M1,T(1)} \in [D, 5D]$	$C_{M1,T(1)} \in [17,550 - 140,400]$	Posterior Mean	59,595
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	55,360
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.49, 0.146}	95% Credible Interval	[27,031 - 104,676]
$\theta_x = \alpha$	log(100)	Standard Deviation	20,408

Taxon M2_T(1) generated as World-B (Seed= -1365400193), with $C_{M2,T(1)} = 40,000$

Parameterization, Seed= -1365400193		Fitting Results	
$C_{M2,T(1)} \in [D, 5D]$	$C_{M2,T(1)} \in [12,719 - 101,752]$	Posterior Mean	45,605
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	44,010
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.08, 0.142}	95% Credible Interval	[21,067 - 78,124]
$\theta_x = \alpha$	log(100)	Standard Deviation	14,492

Group B.12: Summary of the parametrizations and the fitting results of the subgroups (M1.T(2) and M2.T(2)) out of the Merging Scheme - Treatment (2), and subgroups (M1.T(3) and M2.T(3)) out of the Merging Scheme - Treatment (3) of World-B.

Taxon M1_T(2) generated as World-B (Seed= -1365400193), with $C_{M1_T(2)} = 60,000$

Parameterization, Seed= -1365400275		Fitting Results	
$C_{M1_T(2)} \in [D, 5D]$	$C_{M1_T(2)} \in [17,235 - 137,880]$	Posterior Mean	60,439
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	52,499
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[27,415 - 106,912]
$\theta_x = \alpha$	log(100)	Standard Deviation	20,134

Taxon M2_T(2) generated as World-B (Seed= -1365400193), with $C_{M2_T(2)} = 40,000$

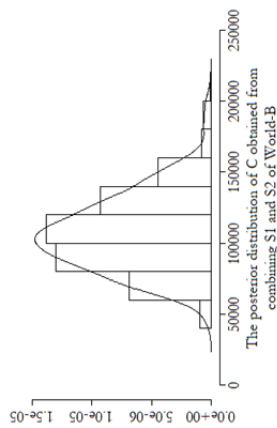
Parameterization, Seed= -1365400324		Fitting Results	
$C_{M2_T(2)} \in [D, 5D]$	$C_{M2_T(2)} \in [13,728 - 109,824]$	Posterior Mean	41,074
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	34,800
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[18,733 - 71,989]
$\theta_x = \alpha$	log(100)	Standard Deviation	13,785

Taxon M1_T(3) generated as World-B (Seed= -1365400193), with $C_{M1_T(3)} = 60,000$

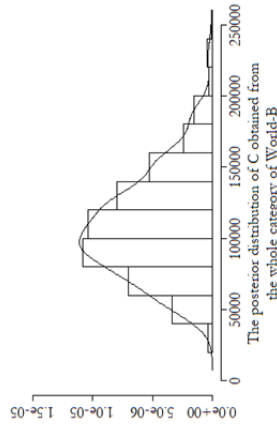
Parameterization, Seed= -1365400193		Fitting Results	
$C_{M1_T(3)} \in [D, 5D]$	$C_{M1_T(3)} \in [22,297 - 178,376]$	Posterior Mean	63,190
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	57,543
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[27,477 - 115,991]
$\theta_x = \alpha$	log(100)	Standard Deviation	23,287

Taxon M2_T(3) generated as World-B (Seed= -1365400193), with $C_{M2_T(3)} = 40,000$

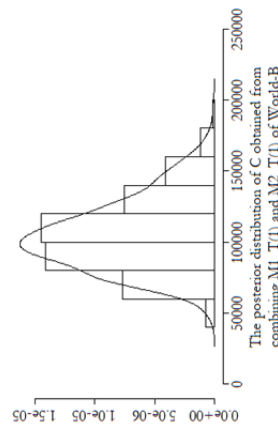
Parameterization, Seed= -1365400193		Fitting Results	
$C_{M2_T(3)} \in [D, 5D]$	$C_{M2_T(3)} \in [17,115 - 136,920]$	Posterior Mean	45,361
$\theta_N = \{\mu, \sigma\}$	{10, 3.5}	Posterior Mode	40,630
$\theta_L = \{\mu_{L_1}, \sigma_L\}$	{7, 0.15}	95% Credible Interval	[18,953 - 81, 805]
$\theta_x = \alpha$	log(100)	Standard Deviation	16,476



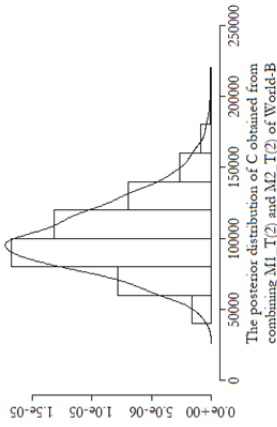
Aggregating Results of S1 and S2	
Current (D_{S1+S2})	31,865
Posterior Mean	107,554
Posterior Mode	102,746
95% Credible Interval	[62,671 - 164,625]
Standard Deviation	27,304
KLD($S1+S2$ Whole)	0.001777



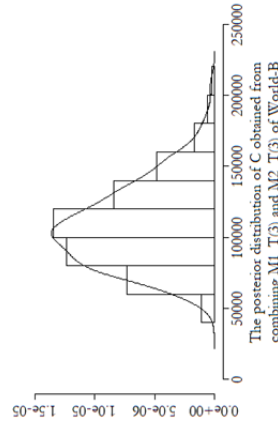
Fitting Results of Whole Category (World-B)	
Current (D)	31,865
Posterior Mean	109,793
Posterior Mode	97,160
95% Credible Interval	[48,927 - 188,961]
Standard Deviation	37,053
KLD(Whole S1+S2)	0.001677
KLD(Whole M1_T(1)+M2_T(1))	0.001512
KLD(Whole M1_T(2)+M2_T(2))	0.001933
KLD(Whole M1_T(3)+M2_T(3))	0.002033



Aggregating of M1_T(1) and M2_T(1)	
Current ($D_{M1_T(1)+M2_T(1)}$)	30,269
Posterior Mean	105,199
Posterior Mode	98,834
95% Credible Interval	[63,768 - 160,087]
Standard Deviation	25,304
KLD($M1_T(1)+M2_T(1)$ Whole)	0.001554



Aggregating of M1_T(2) and M2_T(2)	
Current ($D_{M1_T(2)+M2_T(2)}$)	30,963
Posterior Mean	101,513
Posterior Mode	95,221
95% Credible Interval	[58,706 - 156,508]
Standard Deviation	24,944
KLD($M1_T(2)+M2_T(2)$ Whole)	0.002066



Aggregating of M1_T(3) and M2_T(3)	
Current ($D_{M1_T(3)+M2_T(3)}$)	39,412
Posterior Mean	108,551
Posterior Mode	102,797
95% Credible Interval	[60,628 - 172,654]
Standard Deviation	29,127
KLD($M1_T(3)+M2_T(3)$ Whole)	0.002003

Group B.13: Summary of the fitting results of the whole category (World-B) against all combined taxa: ($S1+S2$), ($M1_T(1) + M2_T(1)$), ($M1_T(2) + M2_T(2)$), and ($M1_T(3) + M2_T(3)$).

Group B.14: Summary of the parametrizations and the fitting results of the subgroups (S1 and S2) out of the Splitting Scheme, subgroup (M.T(1) out of the Merging Scheme - Treatment (1), subgroup (M.T(2) out of the Merging Scheme - Treatment (2), and subgroup (M.T(3) out of the Merging Scheme - Treatment (3) of World-C.

Taxon S1 and S2 split from Artificial World-C, with number of Species $C_{S1} = C_{S2} = 25,000$

World Parameterization		Fitting Results of S1 and S2, respectively		
$C_S \in [D, 3D]$	$C_{S1} \in [20,647 - 61,941]$ $C_{S2} \in [20,973 - 62,919]$	Posterior Mean	26,629	26,813
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	25,200	24,352
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.31, 0.078}	95% CI	[21,251 - 35,131]	[20,902 - 35,870]
$\theta_x = \alpha$	log(50)	SD	3,911	4,241

Taxon M_T(1) generated as World-C (Seed = 209522386), with $C_{M,T(1)} = 25,000$

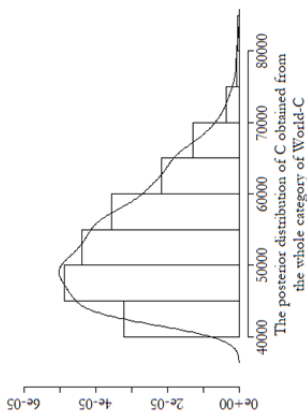
Parameterization, Seed= 209522386		Fitting Results	
$C_{M,T(1)} \in [D, 3D]$	$C_{M,T(1)} \in [20,815 - 62,445]$	Posterior Mean	26,688
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	24,222
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{6.31, 0.078}	95% Credible Interval	[21,007 - 35,156]
$\theta_x = \alpha$	log(50)	Standard Deviation	3,935

Taxon M_T(2) generated as World-C (Seed= 209522386), with $C_{M,T(2)} = 25,000$

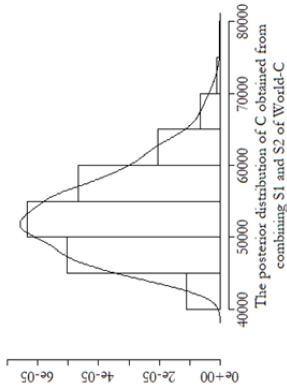
Parameterization, Seed= 209523980		Fitting Results	
$C_{M,T(2)} \in [D, 3D]$	$C_{M,T(2)} \in [20,952 - 62,856]$	Posterior Mean	26,297
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	24,639
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{7, 0.08}	95% Credible Interval	[21,168 - 33,495]
$\theta_x = \alpha$	log(50)	Standard Deviation	3,389

Taxon M_T(3) generated as World-C (Seed = 209522386), with $C_{M,T(3)} = 25,000$

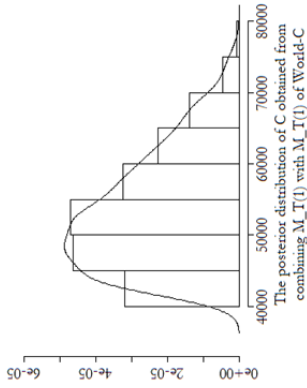
Parameterization, Seed= 209522386		Fitting Results of 1 st Trial and 2 nd Trial, respectively		
$C_{M,T(3)} \in [D, 3D]$	$C_{M,T(3)} \in [22,617 - 67,851]$	Posterior Mean	28,579	27,693
$\theta_N = \{\mu, \sigma\}$	{6, 2}	Posterior Mode	25,165	25,158
$\theta_L = \{\mu_{L1}, \sigma_L\}$	{7, 0.08}	95% CI	[22,916 - 38,917]	[21,274 - 38,166]
$\theta_x = \alpha$	log(50)	SD	4,386	4,711



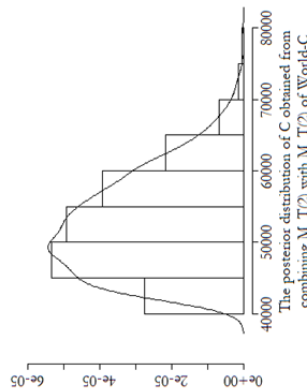
Fitting Results of Whole Category (World-C)	
Current (D)	41,620
Posterior Mean	53,161
Posterior Mode	48,969
95% Credible Interval	[42,280 – 69,765]
Standard Deviation	7,703



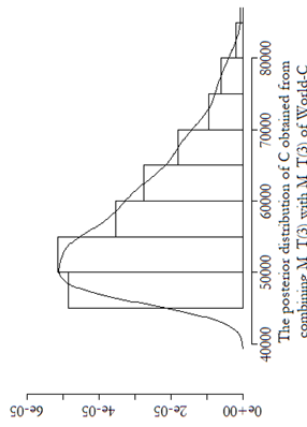
Aggregating Results of S1 and S2	
Current (D_{S1+S2})	41,620
Posterior Mean	53,442
Posterior Mode	51,834
95% Credible Interval	[43,984 – 66,827]
Standard Deviation	5,985
KLD($S1+S2$ Whole)	0.011392
KLD(Whole $S1+S2$)	0.010169



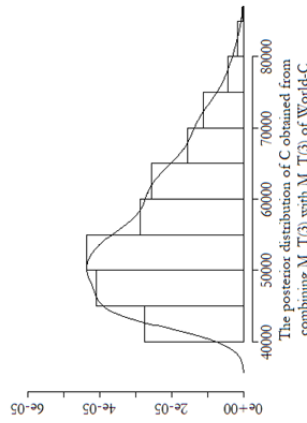
Aggregating of $M_1.T(1)$ and $M_1.T(1)$	
Current ($D_{M_1.T(1)+M_1.T(1)}$)	41,630
Posterior Mean	53,376
Posterior Mode	48,443
95% Credible Interval	[42,014 – 70,312]
Standard Deviation	7,870
KLD($M_1.T(1)+M_1.T(1)$ Whole)	0.000720
KLD(Whole $M_1.T(1)+M_1.T(1)$)	0.000758



Aggregating of $M_1.T(2)$ and $M_1.T(2)$	
Current ($D_{M_1.T(2)+M_1.T(2)}$)	41,904
Posterior Mean	52,594
Posterior Mode	49,279
95% Credible Interval	[42,335 – 66,989]
Standard Deviation	6,779
KLD($M_1.T(2)+M_1.T(2)$ Whole)	0.002525
KLD(Whole $M_1.T(2)+M_1.T(2)$)	0.002563



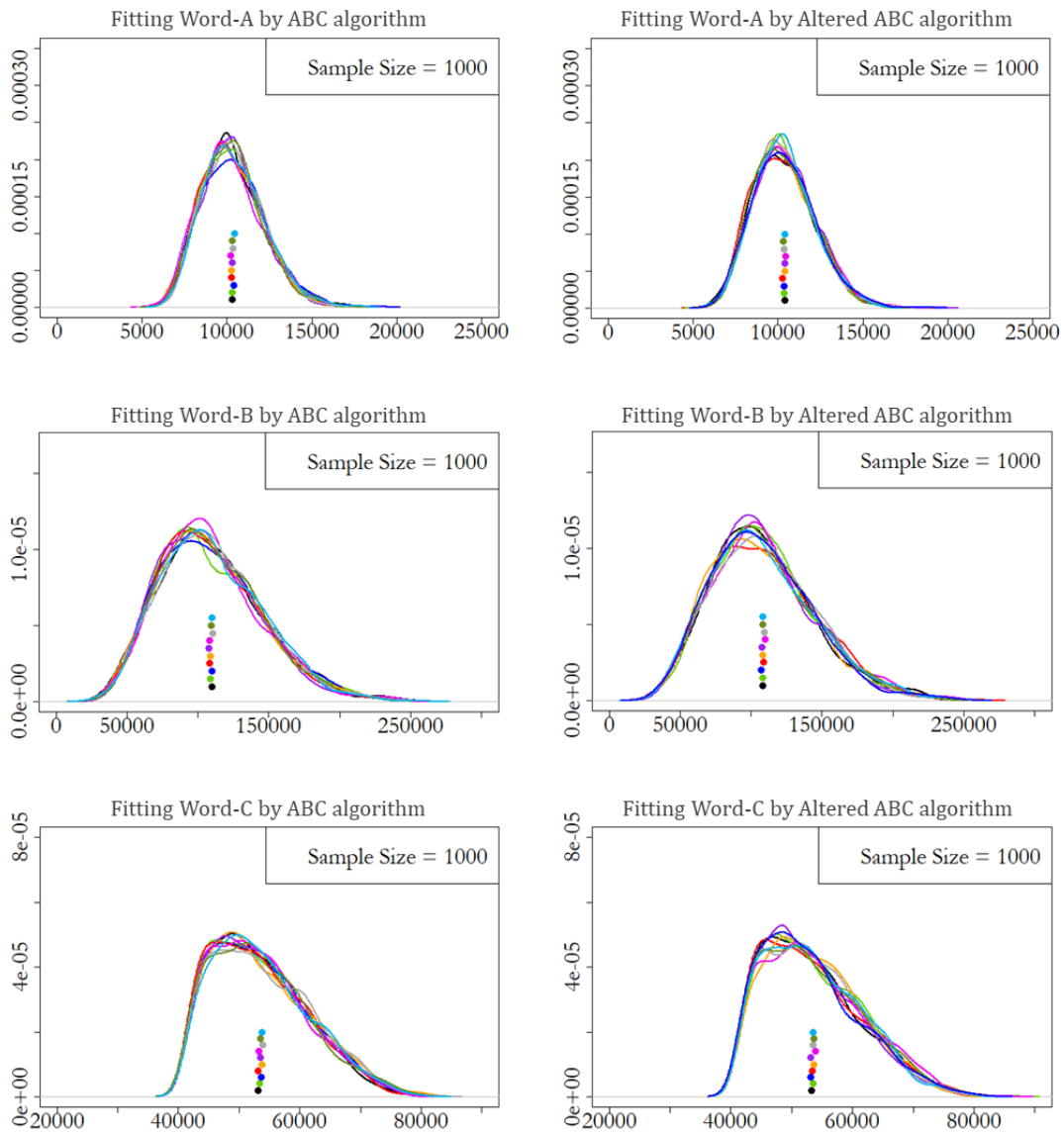
1 st Trial: Aggregating of $M_1.T(3)$ and $M_1.T(3)$	
Current ($D_{M_1.T(3)+M_1.T(3)}$)	45,234
Posterior Mean	57,159
Posterior Mode	50,330
95% Credible Interval	[45,832 – 77,834]
Standard Deviation	8,771
KLD($M_1.T(3)+M_1.T(3)$ Whole)	0.005747
KLD(Whole $M_1.T(3)+M_1.T(3)$)	0.005493



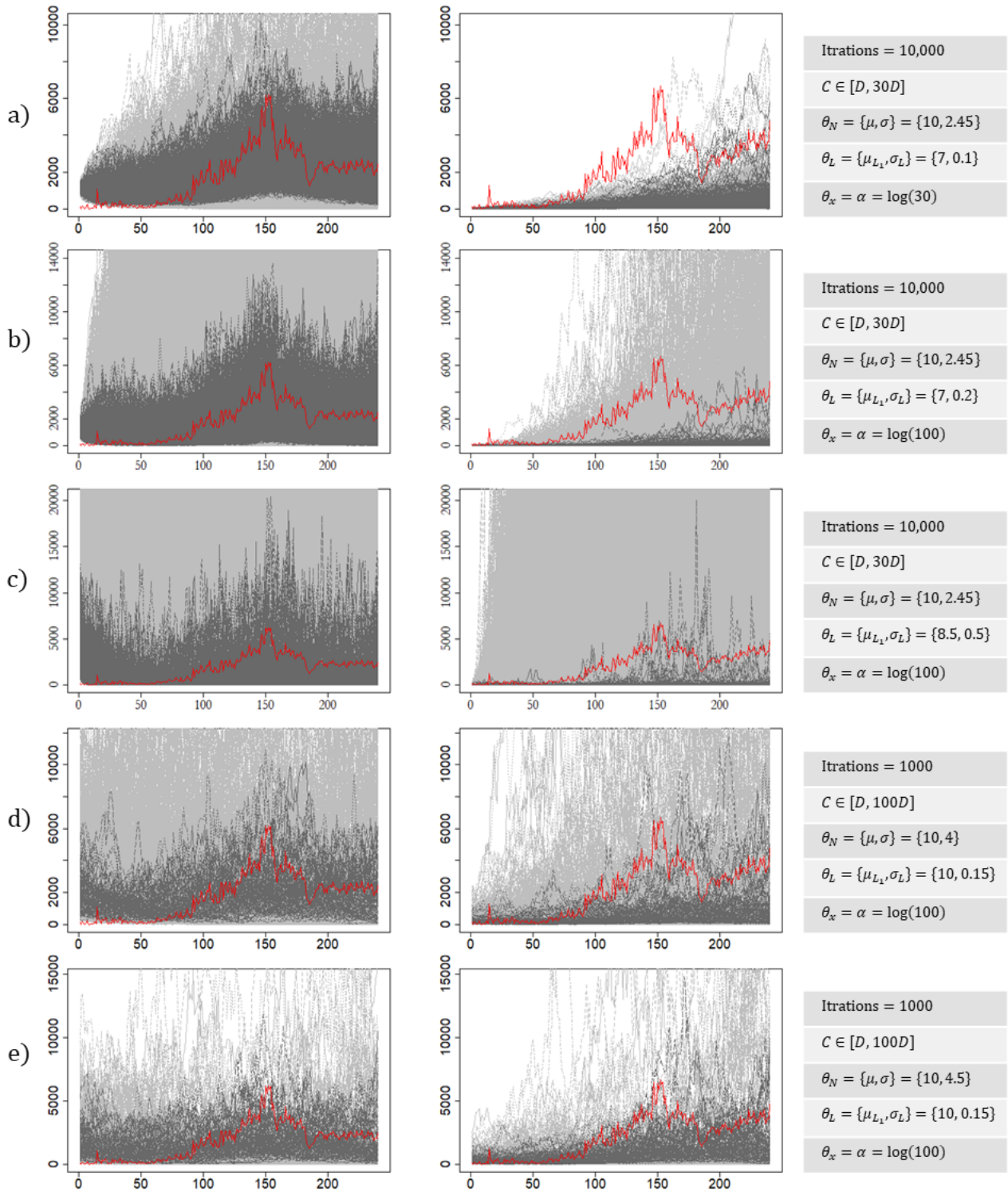
2 nd Trial: Aggregating of $M_1.T(3)$ and $M_1.T(3)$	
Setting ($D_{M_1.T(3)+M_1.T(3)}$)	42,000
Posterior Mean	55,385
Posterior Mode	50,316
95% Credible Interval	[42,548 – 76,333]
Standard Deviation	9,422
KLD($M_1.T(3)+M_1.T(3)$ Whole)	0.002150
KLD(Whole $M_1.T(3)+M_1.T(3)$)	0.002093

Group B.15: Summary of the fitting results of the whole category (World-C) against all combined taxa: ($S1+S2$), ($M1.T(1) + M2.T(1)$), ($M1.T(2) + M2.T(2)$), and ($M1.T(3) + M2.T(3)$).

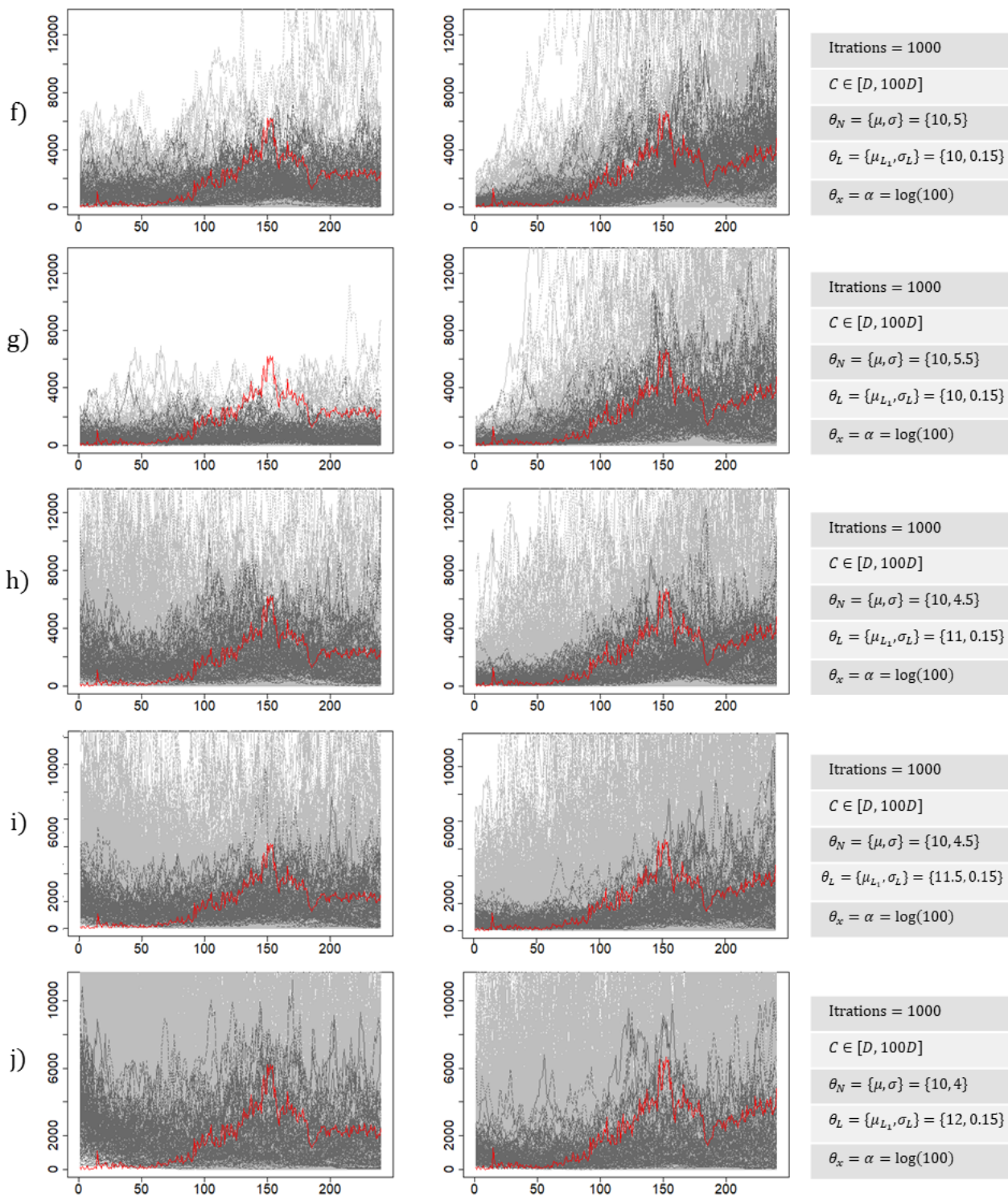
B.5 Results Related to Chapter 6



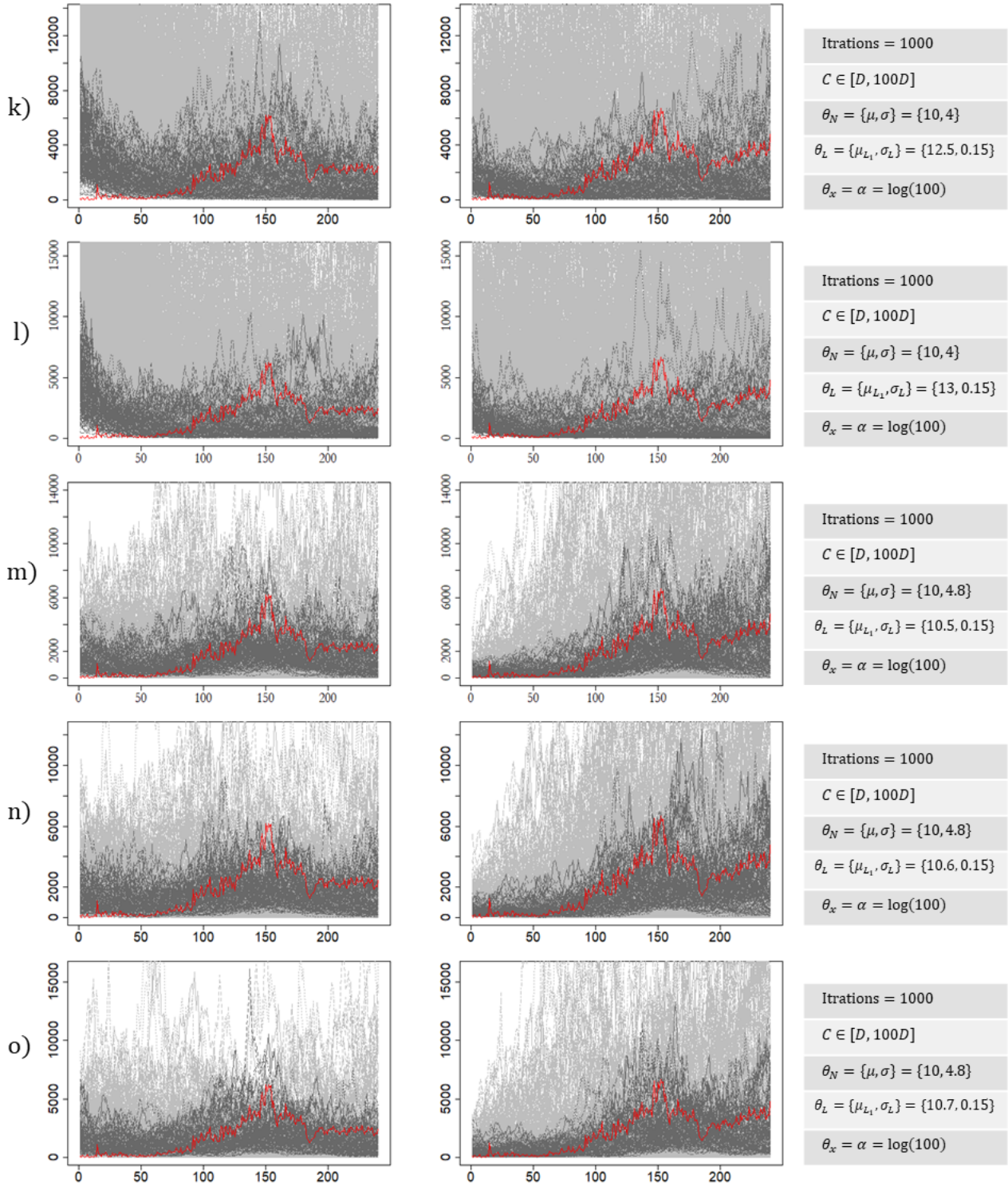
Group B.16: Ten times-fitting results (out of 1,000,000 iterations, and selected final sample $s=1000$) showing similar accuracy in the performance of both: the ABC algorithm that we adopted in all artificial experiments (left panel), and the altered ABC algorithm that includes an additional filtering step and applied on the real experiments (right panel).

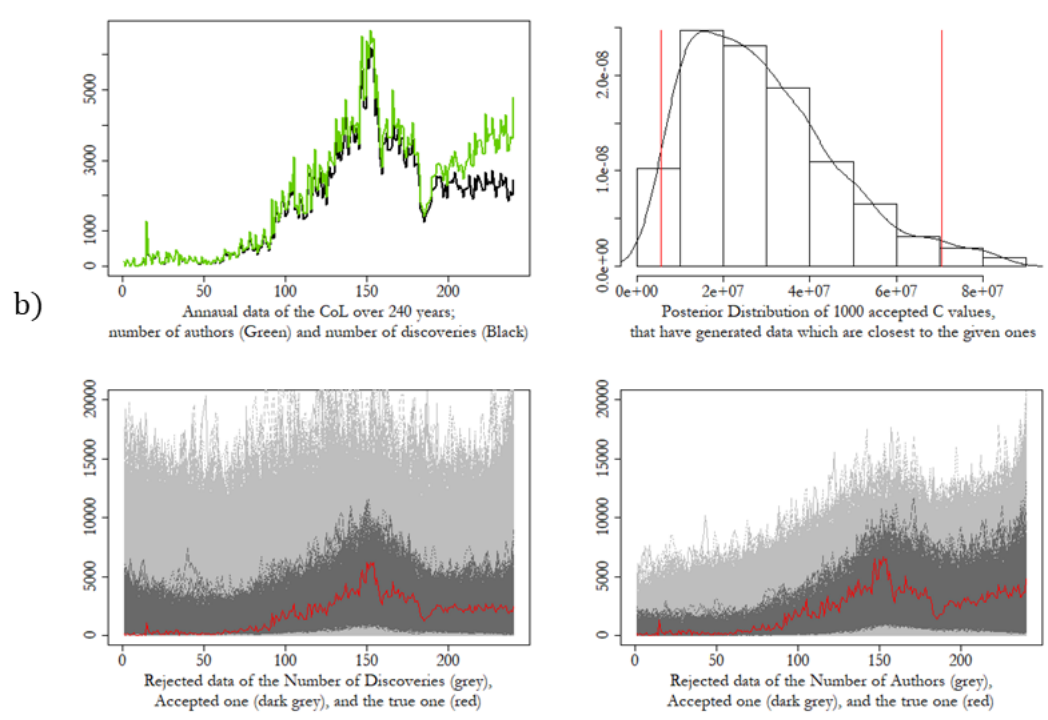
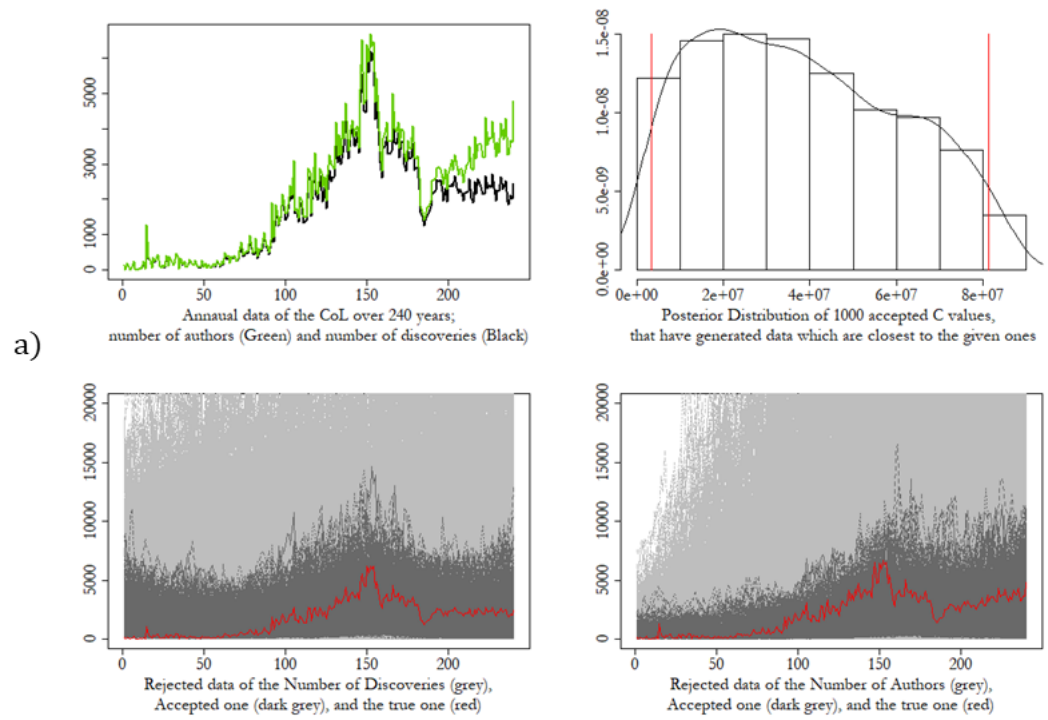


Group B.17: A sample of the CoL fitting trials (pilot experiments). The left plot from each figure belongs to the number of discoveries (rejected (grey), accepted (dark grey), and true curve (red)), and the right plots belong to the number of authors. The chosen parameters of each trial are shown next to each figure. Note that, we started with 10,000 iterations with final sample size =1000 in the figures (a), (b) and (c), but then we found that 1000 iterations with final sample size =100 was enough to make decisions on the rest of the trials.

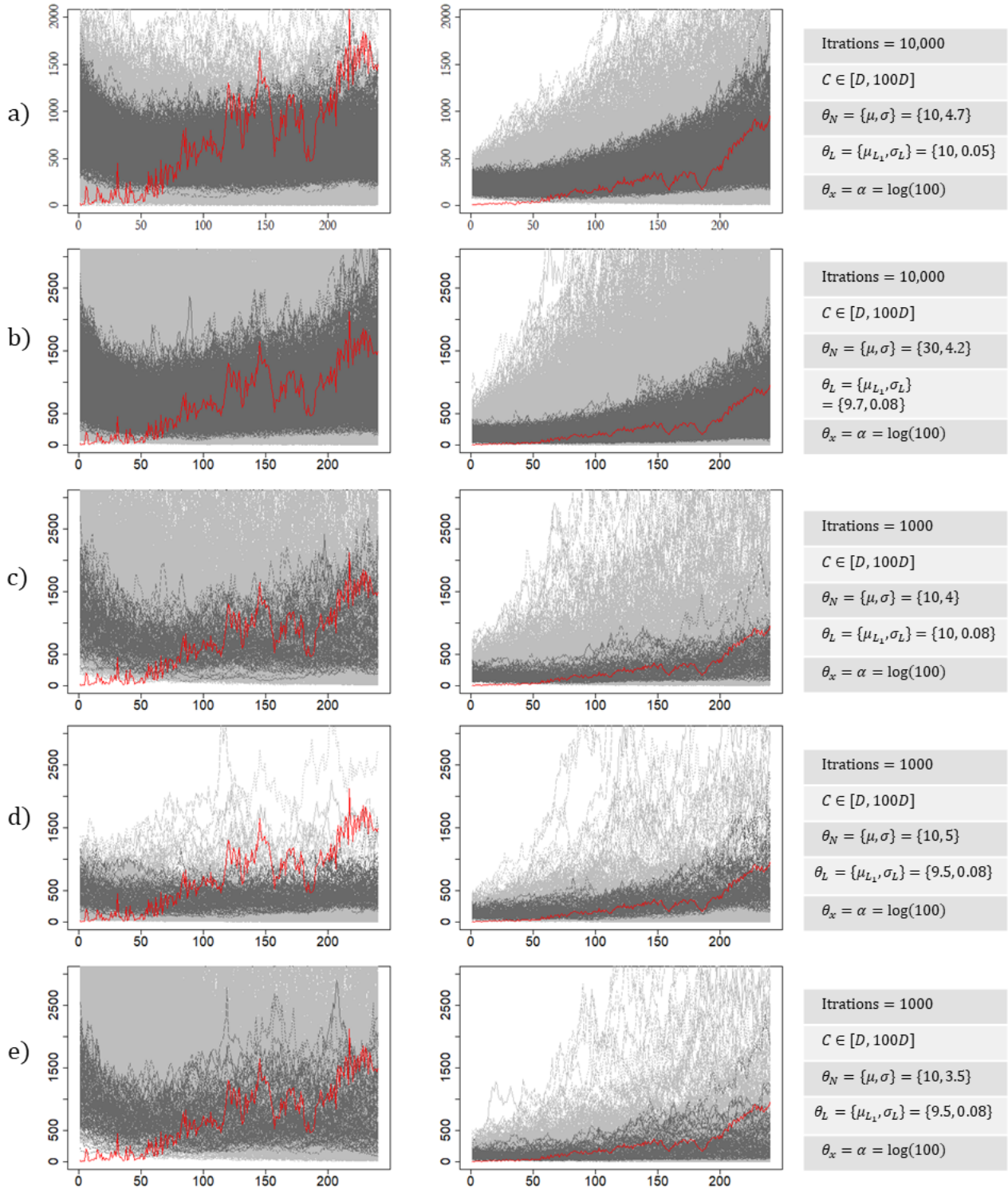


Group B.18: Continue, a sample of the CoL fitting trials (pilot experiments). The left plot from each figure belongs to the number of discoveries (rejected (grey), accepted (dark grey), and true curve (red)), and the right plots belong to the number of authors. The chosen parameters of each trial are shown next to each figure.

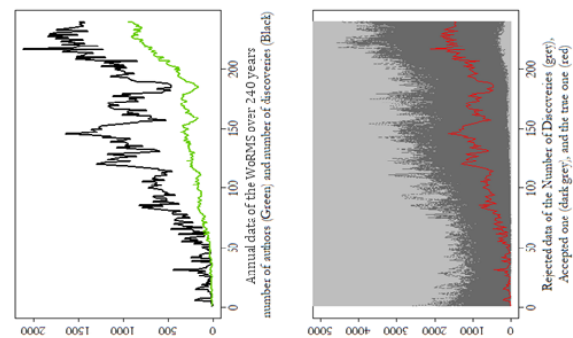
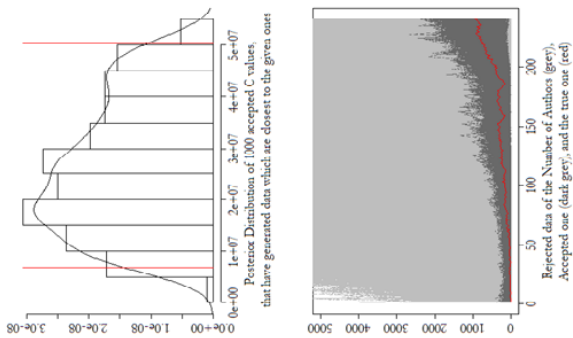




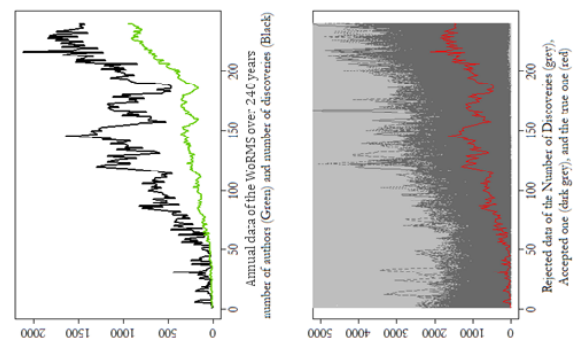
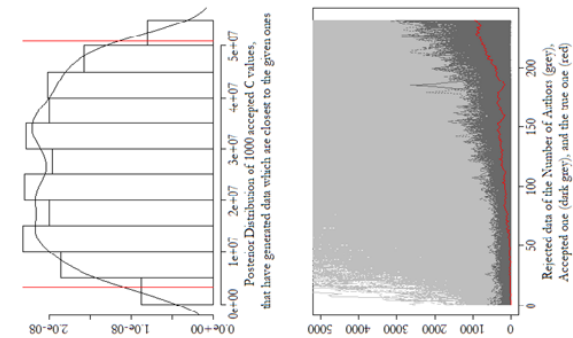
Group B.20: The adopted fitting results of the CoL dataset, Figure (a) belongs to the 10,000 iterations, while Figure (b) belongs to the 100,000 iterations.



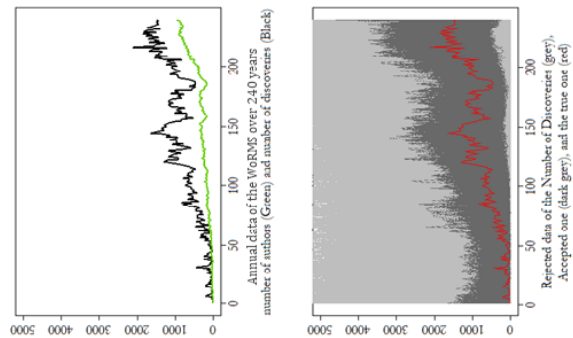
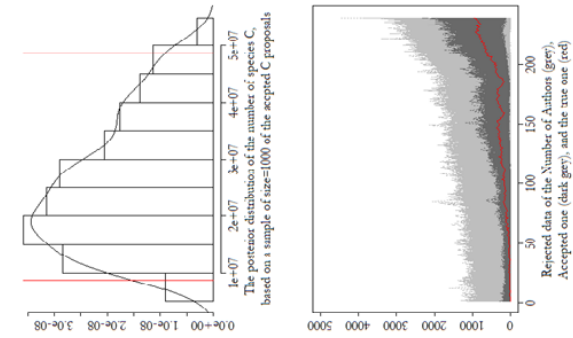
Group B.21: A sample of the WoRMS fitting trials (pilot experiments). The left plot from each figure belongs to the number of discoveries (rejected (grey), accepted (dark grey), and true curve (red)), and the right plots belong to the number of authors. The chosen parameters of each trial are shown next to each figure. Note that, we started with 10,000 iterations with final sample size =1000 in the figures (a) and (b), but then we found that 1000 iterations with final sample size =100 was enough to make decisions on the rest of the trials.



b)

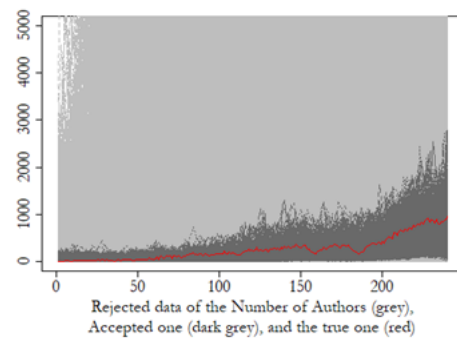
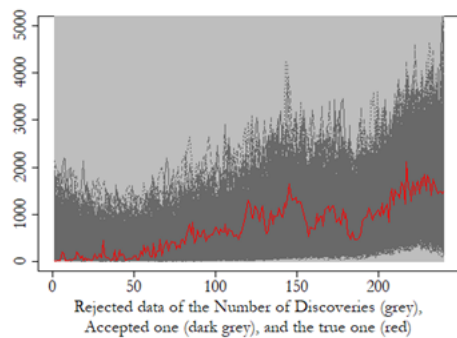
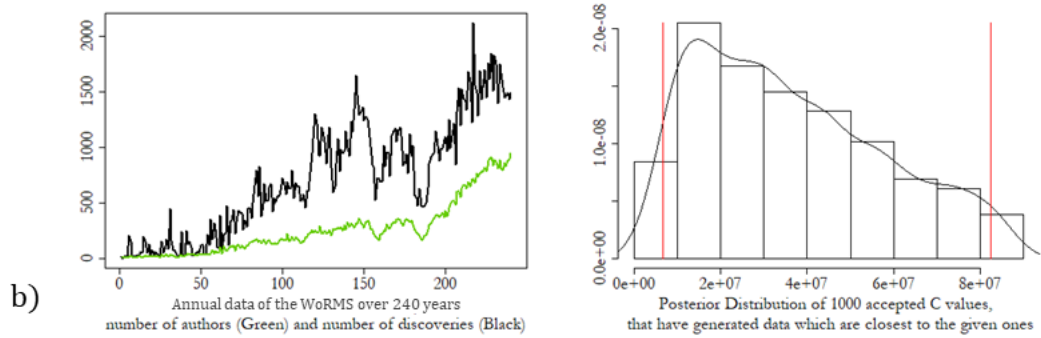
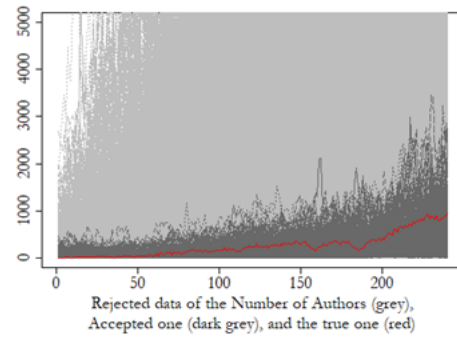
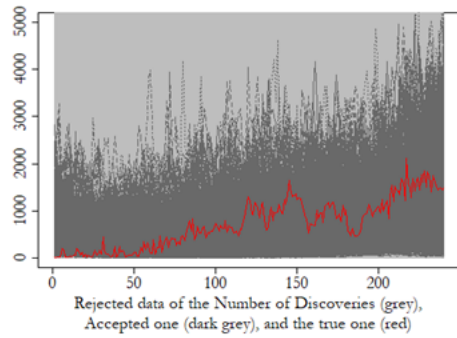
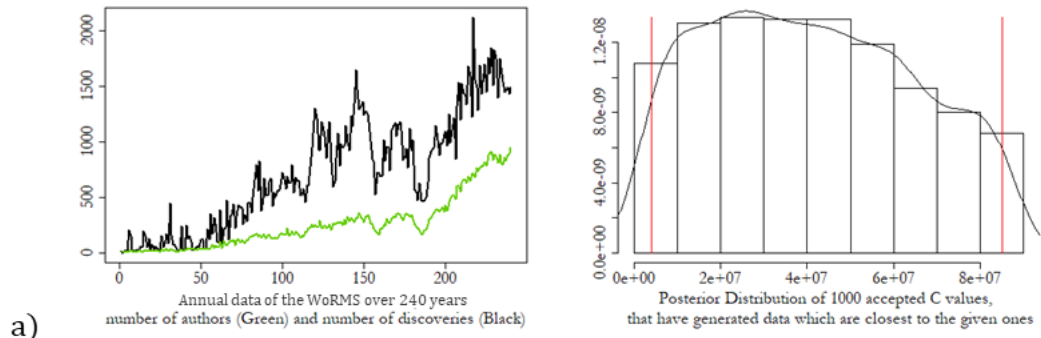


a)

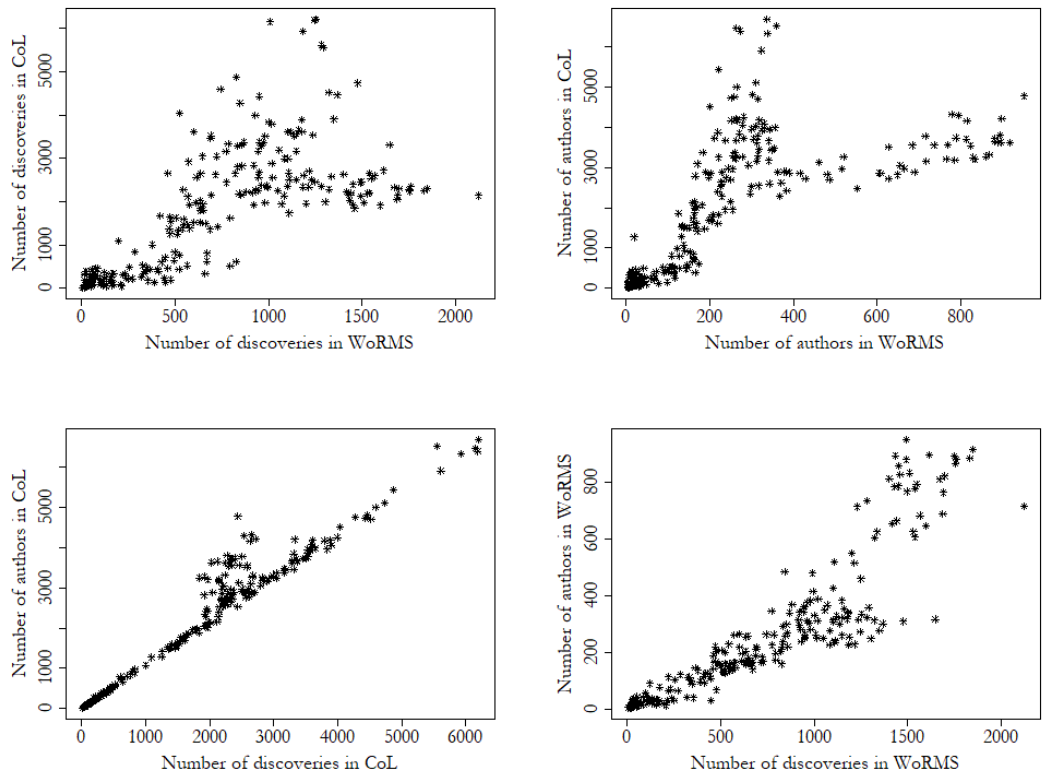


c)

Group B.22: The first attempt of the fitting results of the WoRMS dataset with prior range $[D, 300D]$, Figure (a) belongs to the 10,000 iterations, Figure (b) belongs to the 100,000 iterations, while Figure (c) belongs to the 1,000,000 iterations.

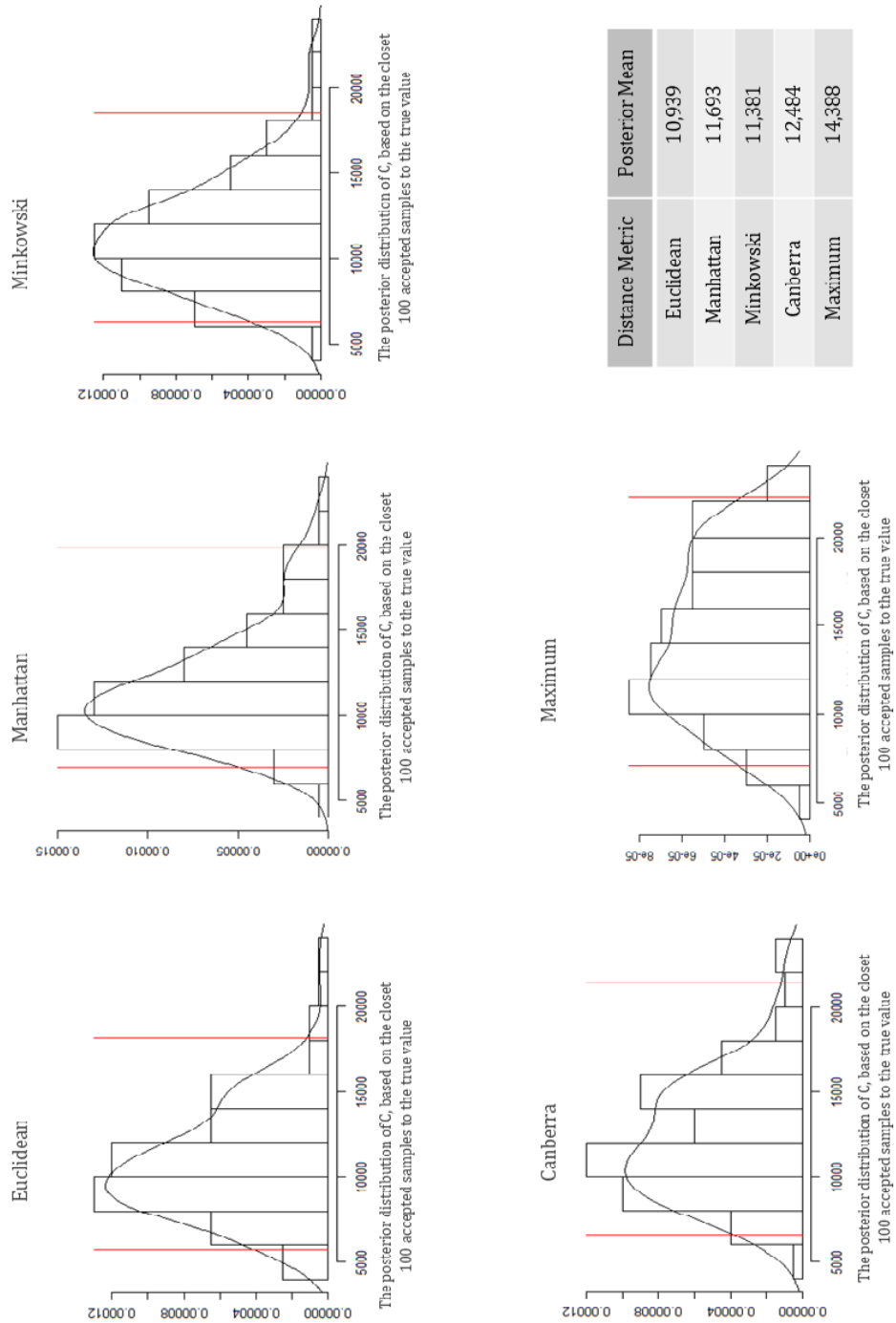


Group B.23: The adopted fitting results of the WoRMS dataset with prior range $[D, 500D]$, Figure (a) belongs to the 10,000 iterations, while Figure (b) belongs to the 100,000 iterations.



Group B.24: Correlations among CoL and WoRMS datasets. The top panel reflects the correlations between the CoL and WoRMS datasets, where the top left plot shows a significant correlation in the number of discoveries ($\rho_{\text{pearson}} = 0.71$ and $\rho_{\text{spearman}} = 0.79$), while the top right plot shows a significant correlation in the number of authors ($\rho_{\text{pearson}} = 0.67$ and $\rho_{\text{spearman}} = 0.86$). The bottom panel reflects the correlations between the number of discoveries and the number of authors within each dataset, where the bottom left plot belongs to the CoL with a significant correlation ($\rho_{\text{pearson}} = 0.97$ and $\rho_{\text{spearman}} = 0.97$), while the bottom right plot belongs to the WoRMS with a significant correlation ($\rho_{\text{pearson}} = 0.90$ and $\rho_{\text{spearman}} = 0.95$).

B.6 Distance Metric Results



Group B.25: An initial attempt for a small pilot experiments with just 10,000 iterations and a final sample size=100, to check whether there is another distance metric that over-performs the Euclidean distance metric. The implemented artificial world parametrized by $\mu_N = 10$, $\sigma_N = 3$, $\mu_{L_1} = 6$, $\sigma_L = 0.15$, and $\alpha = 5$, where $T = 250$, $D = 4,753$, $C = 10,000$. It seems that there is no significant difference between them except for the ‘Maximum’ metric. This is justified by the fact that they all (except for the ‘Maximum’ metric) are equivalently monotonic transformations.

Appendix C

Pseudo Codes

The current appendix includes two sections. In the first section, we introduce the pseudo code for the adopted ABC algorithm that is implemented on the artificial datasets throughout Chapter 4 and Chapter 5. In the second section, we introduce the pseudo code for the altered ABC algorithm that is implemented on the real datasets in Chapter 6.

C.1 Artificial-Data Algorithm

Algorithm 1 Pseudo code for the adopted ABC algorithm

- 1: Define: number of iterations ($iter$), and size of the accepted final sample (s).
 - 2: Define: values for parameters ($\theta_N, \theta_l, \theta_x$), and upper limit (mD) for uniform prior of C .
 - 3: Define: distance metric threshold (ϵ). ▷ // obtained via a pilot trial.
 - 4: Import: data $\tau_{1:T}$ and $x_{1:T}$.
 - 5: **for** (each iteration index $i = 1, \dots, iter$) **do**
 - 6: Simulate $C^{(i)}$ from $P(C)$.
 - 7: Simulate $l_{1:T}^{(i)}$ from $P(l_{1:T} | \theta_l)$.
 - 8: Simulate $x_{1:T}^{(i)}$ from $P(x_{1:T} | l_{1:T}^{(i)}, \theta_x)$.
 - 9: Compute distance metric $\mathcal{D}_x^{(i)}$ between given $x_{1:T}$ and simulated $x_{1:T}^{(i)}$.
 - 10: **if** ($\mathcal{D}_x^{(i)} > \epsilon$) **then**
 - 11: skip this step, and go back for a new iteration.
 - 12: **else** ▷ // continue simulating the rest.
 - 13: Simulate $N_{1:C^{(i)}}^{(i)}$ from $P(N_{1:C^{(i)}} | C^{(i)}, \theta_N)$.
 - 14: Simulate $n_{1:T}^{(i)}$ from $P(n_{1:T} | l_{1:T}^{(i)})$.
 - 15: Simulate $d_{1:D}^{(i)}$ from $P(d_{1:D}^{(i)} | N_{1:C^{(i)}}^{(i)})$.
 - 16: Calculate $\tau_{1:T}^{(i)}$ from $d_{1:D}^{(i)}$ and $n_{1:T}^{(i)}$ ▷ // using Equation 3.2.
 - 17: Compute distance metric $\mathcal{D}_\tau^{(i)}$ between given $\tau_{1:T}$ and simulated $\tau_{1:T}^{(i)}$.
 - 18: Calculate $\mathcal{D}_{total}^{(i)} = \mathcal{D}_\tau^{(i)} + \mathcal{D}_x^{(i)}$
 - 19: Store $\{C^*\} = C^{(i)}$, and $\{\mathcal{D}_{total}^*\} = \mathcal{D}_{total}^{(i)}$ ▷ // end of the for loop.
 - 20: Take the s accepted $\{C^*\}$ with the smallest $\{\mathcal{D}_{total}^*\}$ to be a final sample from $P(C | \tau_{1:T}, x_{1:T}, \theta_N, \theta_l, \theta_x)$.
 - 21: Derive summary statistics and plots from the above final sample of size s .
-

C.2 Real-Data Algorithm

Algorithm 2 Pseudo code for the (altered) ABC algorithm

- 1: Define: number of iterations ($iter$), and size of the accepted final sample (s).
 - 2: Define: values for parameters $(\theta_N, \theta_l, \theta_x)$, and upper limit (mD) for uniform prior of C .
 - 3: Define: distance metric threshold (ϵ_0). ▷ // obtained via a pilot trial.
 - 4: Define: two limit thresholds (ϵ_1) and (ϵ_2). ▷ // obtained via same pilot trial.
 - 5: Import: data $\tau_{1:T}$ and $x_{1:T}$.
 - 6: **for** (each iteration index $i = 1, \dots, iter$) **do**
 - 7: Simulate $C^{(i)}$ from $P(C)$.
 - 8: Simulate $l_{1:T}^{(i)}$ from $P(l_{1:T} | \theta_l)$.
 - 9: Simulate $x_{1:T}^{(i)}$ from $P(x_{1:T} | l_{1:T}^{(i)}, \theta_x)$.
 - 10: Compute distance metric $\mathcal{D}_x^{(i)}$ between given $x_{1:T}$ and simulated $x_{1:T}^{(i)}$.
 - 11: **if** ($\mathcal{D}_x^{(i)} > \epsilon_0$) **then**
 - 12: skip this step, and go back for a new iteration.
 - 13: **else**
 - 14: Simulate $n_{1:T}^{(i)}$ from $P(n_{1:T} | l_{1:T}^{(i)})$.
 - 15: **if** ($(\sum_{r=1}^{Tj} n_r^{(i)} < \epsilon_1)$ or $(\sum_{r=1}^{Tj} n_r^{(i)} > \epsilon_2)$) **then**
 - 16: skip this step, and go back for a new iteration.
 - 17: **else** ▷ // continue simulating the rest.
 - 18: Simulate $N_{1:C^{(i)}}^{(i)}$ from $P(N_{1:C^{(i)}} | C^{(i)}, \theta_N)$.
 - 19: Simulate $d_{1:D}^{(i)}$ from $P(d_{1:D}^{(i)} | N_{1:C^{(i)}}^{(i)})$.
 - 20: Calculate $\tau_{1:T}^{(i)}$ from $d_{1:D}^{(i)}$ and $n_{1:T}^{(i)}$ ▷ // using Equation 3.2.
 - 21: Compute distance metric $\mathcal{D}_\tau^{(i)}$ between given $\tau_{1:T}$ and $\tau_{1:T}^{(i)}$.
 - 22: Calculate $\mathcal{D}_{\text{total}}^{(i)} = \mathcal{D}_\tau^{(i)} + \mathcal{D}_x^{(i)}$
 - 23: Store $\{C^*\} = C^{(i)}$, and $\{\mathcal{D}_{\text{total}}^*\} = \mathcal{D}_{\text{total}}^{(i)}$ ▷ // end of the for loop.
 - 24: Take the s accepted $\{C^*\}$ with the smallest $\{\mathcal{D}_{\text{total}}^*\}$ to be a final sample from $P(C | \tau_{1:T}, x_{1:T}, \theta_N, \theta_l, \theta_x)$.
 - 25: Derive summary statistics and plots from the above final sample of size s .
-